IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 March 2024, accepted 12 April 2024, date of publication 18 April 2024, date of current version 26 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391057

RESEARCH ARTICLE

Facial Emotion Recognition Combining Auxiliary Classifiers and Multiscale CBAM Attention Mechanisms

YUJIE SHANG[1], FEI YAN[1,2], YUNQING LIU[1,2], QI LI[1], AND QIONG ZHANG[1]
[1]School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
[2]Jilin Provincial Science and Technology Innovation Center of Intelligent Perception and Information Processing, Changchun 130022, China

Corresponding author: Fei Yan (yanf@cust.edu.cn)

This work was supported by the Science and Technology Development Plan Project of Jilin Province, China, under Grant 20240304145SF.

ABSTRACT Facial emotions are the most intuitive way to react to changes in inner emotions. We propose a facial emotion recognition method that combines auxiliary classifiers (Acs) and multi-scale CBAM (MCBAM) by improving the Xception network model. And we design a lightweight network model AMDCNN. We introduce Acs in the middle layers of the model. The features extracted from the middle layer portion of the model are utilized to aid in emotion recognition. Finally, the recognition results of the Acs and the main classifier are adaptively weighted and fused to obtain better emotion recognition results. This enables better utilization of the feature information extracted from the intermediate layers and further reduces the feature loss caused by the downsampling process of the convolutional layer. The CBAM does not increase the number of parameters and computation too much, and it enables the model to better focus on the important areas of the face. We apply it to the proposed lightweight model and improve it further. The width is increased by introducing a multi-branch convolutional structure and utilizing convolutional layers with different kernel sizes. This allows for more adequate spatial and channel features during feature extraction, allowing the model to more accurately focus on important facial regions. Our proposed model was experimentally validated on datasets of FER2013, FERPlus, RAF-DB and CK+, with accuracies of 69.82%, 85.40%, 86.77% and 99.49%, respectively. The number of parameters of the proposed model is only 1.6M. Our model is a good competitive advantage compared with other lightweight models.

INDEX TERMS Facial emotion recognition, lightweight, Acs, MCBAM.

I. INTRODUCTION

In daily life socialization, facial expressions can convey rich emotional information and make socialization more vivid. With the development of society and the continuous innovation of face recognition technology, the research on facial emotion recognition is becoming increasingly popular. In 2012, Hinton and other researchers [1] applied Deep Convolutional Neural Networks (DCNN) to image recognition and achieved recognition results that far exceeded those of traditional algorithms. After that, large-scale DCNN models such as VGGNet [2], GooGleNet [3] and ResNet [4] have emerged. These large models are further enhanced by increasing the number of model convolution layers and increasing the depth of the network. However, these improvements in modeling effectiveness come at the expense of the amount of hardware storage and computation. After more intensive research, attention mechanisms such as SENet [5], ECANet [6], STN [7] and GENet [8] were also introduced into the model. Among them, CBAM proposed by Woo et al. [9] can better help the model to capture the important features in the image and is sought after by researchers. The CBAM incorporates channel attention and spatial attention and is more applicable to lightweight models. By introducing these channel attention mechanisms(CAM) and spatial attention mechanisms(SAM) to make the model more focused on the important regions of recognition. Zhang et al. [10] proposed a lightweight DCNN with a convolutional block attention module by combining DNN and CBAM. Liao et al. [11] proposed a RCL-Net network

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

model. The structure consists of two main branches, the ResNet-CBAM residual attention branch and the local binary feature (LBP) extraction branch. Nan et al. [12] proposed a lightweight A-MobileNet model. Introducing attention module in MobileNetV1 model to enhance local feature extraction for facial expressions.

However, these research methods directly ignore the feature information lost by the model during the downsampling process. This characterization information is not effectively utilized. Especially in some cases where the dataset is small, the number of images is limited, and the image quality is low, these lost feature information also has extremely important recognition ability. Therefore, we propose a facial emotion recognition method that combines auxiliary classifiers (Acs) and multi-branch CBAM (MCBAM). By introducing Acs in the middle layers of the model to further improve the utilization of features in the middle layers of the model and reduce the impact of feature loss during downsampling. By improving the CBAM and introducing a multi-branch structure, the model is able to better focus on important regions and obtain more effective feature information. We designed a novel lightweight model for facial emotion recognition based on the above proposed approach.

The main contributions of this paper can be summarized as follows:

1) We propose a facial emotion recognition method that combines Acs and MCBAM to improve the utilization of important features in the middle layer of DCNN. Then, we designed a novel facial emotion recognition lightweight model AMDCNN for facial emotion recognition.

2) Introducing a multi-branch convolutional layer in the CBAM for further improvement.Using different scales of convolution allows the module to better model the spatial relationships between features. This allows the model to weight and utilize important features more accurately and can focus more on important facial areas.

3) Improve the utilization of features in the middle layer of the model by introducing Acs in the model. Combining MCBAM can utilize the important intermediate level features more effectively. Finally, we validate the enhancement effect of the model's intermediate level features for facial emotion recognition.

II. RELATED WORK

DCNN is currently the main research method for facial emotion recognition. With the development of research, various large-scale DCNN models have emerged. Considering the computationally intensive nature of large network models and the high number of parameters, researchers have begun to conduct multifaceted studies to reduce the memory occupancy and computational resource consumption of the models. Szegedy et al. [13] proposed the InceptionV3 model, which introduces the concept of Acs. François [14] proposed the Xception model, which greatly reduces the number of parameters in the model by introducing depthwise separable convolution. Depthwise separable convolution decomposes the traditional convolution operation into depthwise convolution and pointwise convolution. This design allows Xception to have fewer parameters while maintaining high performance. Minaee et al. [15] proposed a DeepEmotion attention mechanism. The accuracy of facial emotion recognition can be further improved by focusing on salient facial regions through convolutional attention networks. Daihong et al. [16] proposed a facial emotion recognition method based on the attention mechanism. The method incorporates a self-attention mechanism and a channel-attention mechanism to improve the model's ability to extract globally important features. Shen and Xu [17] proposed a facial expression recognition model based on multi-channel attention residual network. Their proposed model achieved 72.7%, 98.8% and 93.33% accuracies on FER2013, CK+ and Jaffe datasets respectively. He [18] proposed a multi-branch attention CNN based on multi-branch structure to recognize facial expressions. Extraction of facial expression features by multi-branch architecture and further feature fusion. Emotion recognition is then performed in conjunction with CBAM. The model recognition rates on the FER2013, FERPLUS and CK+ datasets were 69.49%, 84.63% and 99.39%, respectively. Burrows [19] designed a lightweight model with a parameter count of only 1.5M size by combining CNN and Generative Adversarial Networks (GAN). The seven classifications against the homemade dataset were able to achieve the accuracy of 58.71%. Joseph and Mathew [20] achieved 67.18% accuracy on the FER2013 dataset by designing a lightweight CNN. Putro et al. [21] proposed a multi-view real-time facial emotion detector based on a lightweight CNN. The model can distinguish between specific facial components. The computational effort of the convolution operation is reduced by using the convolution of cross-stage partial (CSP) method. Chang et al. [22] started their experiments by utilizing LibreFace, an open source toolkit for facial expression analysis. They designed a model with a parameter count of 43M and achieved 82.79% accuracy on the RAF-DB dataset. In this paper, we focus on designing a lightweight facial emotion recognition model by improving the Xception network model. The utilization of features in the middle layers of the model is improved by introducing Acs and MCBAM to achieve better experimental results.

III. PROPOSED METHODOLOGY

In facial emotion recognition research, many of the important feature information is lost when DCNN are used for downsampling feature extraction. In this paper, we propose a lightweight facial emotion recognition method to improve the utilization of important features in the intermediate level of DCNN. We then designed a network model AMDCNN combining Auxs and MCBAM by improving XCEPTION. MCBAM is a further improvement of the CBAM attention mechanism. It uses multi-branch convolutional layers and utilizes convolutional kernels of different scales for fea-

ture extraction. This allows for fuller characterization and enables the model to focus more accurately on important features. These intermediate level important features are then better applied to facial emotion recognition by introducing Acs at the intermediate layer of the DCNN for assisted classification.

A. STRUCTURE OF THE PROPOSED MODEL

The AMDCNN model designed in this paper is shown in Figure 1. It is mainly divided into five modules: Shallow Feature Extraction Module (SFEM), Deep Feature Extraction Module (DFEM), Auxiliary Classifier Module (ACM), Main Classifier Module (MCM), and Adaptive Weighted Fusion Module (AWFM). The model performs layer-by-layer feature extraction to extract the feature map of $C \times H \times W$ from the input facial image. C, H and W are the number of channels, height, and width, respectively. The model utilizes MCBAM to better focus on the extracted important features. Acs are then introduced after the DFEM to utilize these intermediate-level features for assisted classification. And the deep high-level features extracted by the model at the end are passed to the main classifier for classification. Finally, the outputs of the main classifier and all Acs are adaptively weighted and fused to output the final facial emotion recognition results. Table 1 shows the contents of the model and the corresponding parameters.

TABLE 1. Overall architecture of model.

Layer	Input Size	Kernel Size	Stride
InputLayer	(48,48,1)	-	-
Conv1	(48,48,1)	(3,3)	1
Conv2	(46,46,16)	(3,3)	1
DFEM1	(44,44,16)	(3,3)	1
ACM1	(22,22,32)	(3,3)	1
DFEM2	(22,22,32)	(3,3)	1
ACM2	(11,11,64)	(3,3)	1
DFEM3	(11,11,64)	(3,3)	1
ACM3	(6,6,128)	(3,3)	1
DFEM4	(6,6,128)	(3,3)	1
MCM	(3,3,256)	(3,3)	1
AWFM	7	-	-

B. DFEM

DFEM, unlike SFEM for shallow feature extraction, is not used to extract shallow edge features and texture features for facial expressions. DFEM is mainly composed of two parts: the standard convolution block and the residual block. It is mainly used to extract deep features that can be used for facial emotion recognition, and the specific structure is shown in Figure 1. Standard convolutional blocks are mainly used to extract deep features. And the residual block is used to establish a direct information transfer path to facilitate gradient back propagation and solve the gradient vanishing and gradient explosion problems. To ensure model's accuracy, the model does not use depth-separable convolution for feature extraction. Finally, a linear summation is used to fuse the standard convolutional block with the features extracted from the residual block.

The standard convolutional block consists mainly of two standard convolutional layers of 3×3 convolutional kernels, a BatchNormalization layer and a layer of Relu activation function and a Maxpooling layer as well as MCBAM.

MCBAM is mainly a further improvement to CBAM. In this work, we extract spatial features by employing multi-scale convolutional layers. The extracted spatial features are then further feature spliced and fused to more fully weight the important spatial features. MCBAM also consists of CAM and Multiscale Spatial Attention Module (MSAM) as shown in Figure 2. The MCBAM attention operation can be expressed by the following Equation (1) and Equation (2).

$$F_{CAM} = M_{CAM}(F) \otimes F \tag{1}$$

$$F_{MSAM} = M_{MSAM}(F_{CAM}) \otimes F_{CAM} \tag{2}$$

In Equation (1) and Equation (2), F denotes the input feature map of the MCBAM attention module. $\otimes$ denotes element-by-element multiplication. M_{CAM} denotes the channel attention extraction operation. And M_{MSAM} denotes the spatial dimension extraction operation.

1)The overall block diagram of CAM attention is shown in Figure 3, and the attention calculation formula is shown as Equation (3).

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)) \\ &= \sigma(W_1(W_0(F^c_{avg})) + \sigma(W_1(W_0(F^c_{max})) \end{aligned} \tag{3}$$

In Equation (3), F is the input feature map. σ is the sigmoid function. W_0 and W_1 are the weights.

The overall block diagram of MSAM attention is shown in Figure 4. In MSAM, we use three different scales of convolutional layers of 5×5, 3×3 and 1×1 for further extraction of spatial features. The introduction of a multi-scale convolution operation allows the module to better model the spatial relationships between features and improve the performance of the module. The MSAM attention formulas are shown as Equation (4) to Equation (9).

$$F' = M_C(F) \tag{4}$$

$$\begin{aligned} M'_{S_1}(F') &= f^{5\times5}([AvgPool(F); MaxPool(F)]) \\ &= f^{5\times5}([F^S_{avg}; F^S_{max}]) \end{aligned} \tag{5}$$

$$\begin{aligned} M'_{S_2}(F') &= f^{3\times3}([AvgPool(F); MaxPool(F)]) \\ &= f^{3\times3}([F^S_{avg}; F^S_{max}]) \end{aligned} \tag{6}$$

$$\begin{aligned} M'_{S_3}(F') &= f^{1\times1}([AvgPool(F); MaxPool(F)]) \\ &= f^{1\times1}([F^S_{avg}; F^S_{max}]) \end{aligned} \tag{7}$$

$$M'_S(F') = Concatenate(M'_{S_1}(F'), M'_{S_2}(F'), M'_{S_3}(F')) \tag{8}$$

$$M_S(F') = \sigma(M'_S(F')) \tag{9}$$

As Equation (4) to Equation (9) shown, MSAM learns the spatial information based on the output feature map of CAM. Firstly, the spatial attention module applies average pooling and maximum pooling operations and obtains

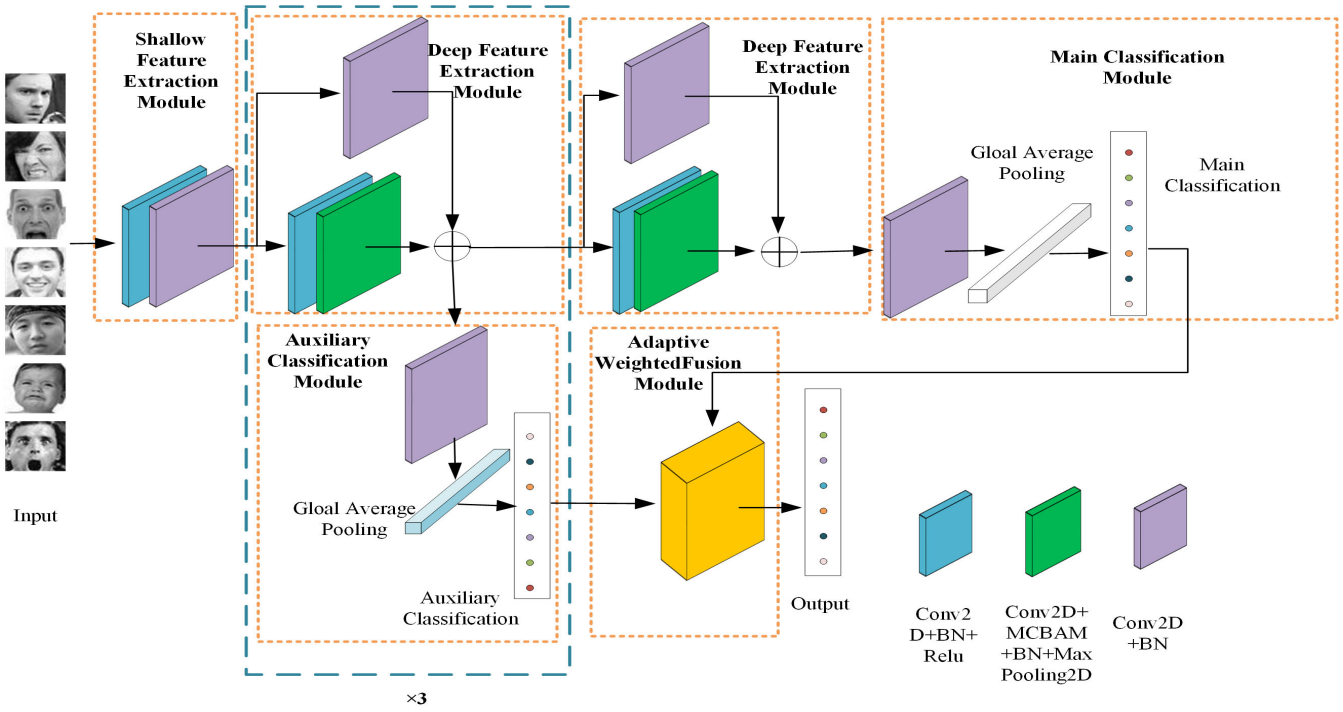

FIGURE 1. Overall framework diagram of the model.

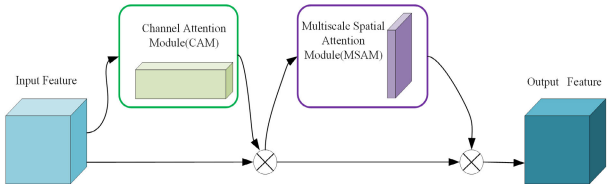

FIGURE 2. The overview of MCBAM.

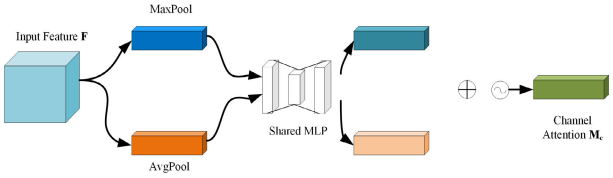

FIGURE 3. Diagram of CAM.

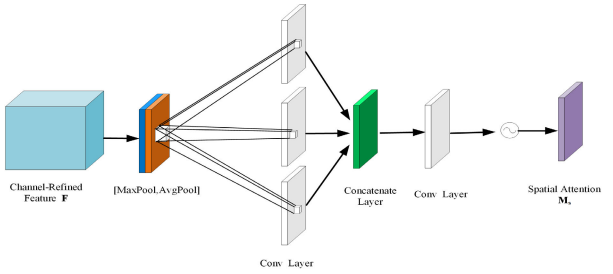

FIGURE 4. Diagram of MSAM.

two feature maps F^S_{avg} and F^S_{max}. And a new feature map is generated by joining the two feature maps. Then, multi-branch convolutional layers with three different convolutional layers 5×5, 3×3 and 1×1 are used to extract the multi-scale spatial features $M'_{S_1}(F')$, $M'_{S_2}(F')$ and $M'_{S_3}(F')$. Further features are fused into $M'_S(F')$. Finally a spatial attention map $M_S(F')$ is used with sigmoid functions to generate the spatial attention map.

MCBAM contains two main attention mechanisms, CAM and MSAM. CAM can learn the weights of different channels of the feature map. CAM does this by capturing the correlation between different feature channels and then automatically selecting the most relevant channel features. While MSAM mainly learns the feature maps of specific locations. Through the multi-scale feature extraction approach, MSAM allows the model to focus more accurately on specific regions in the image, which helps the model to better understand the spatial structure of the image. This combination of spatial and channel attention mechanisms allows the model to learn and utilize intermediate-level important features in a more targeted manner. The model's performance and generalization capabilities have also improved.

The residual block with convolutional layers, on the other hand, is capable of spatially localizing the input features. The residual block is able to extract higher-level features, further increasing the nonlinear and expressive capabilities of the model.

C. AWFM

The AWFM, proposed in this paper, focuses on the adaptive weighted fusion of the outputs of ACM and MCM. The final output is the output of the model facial emotion recognition.

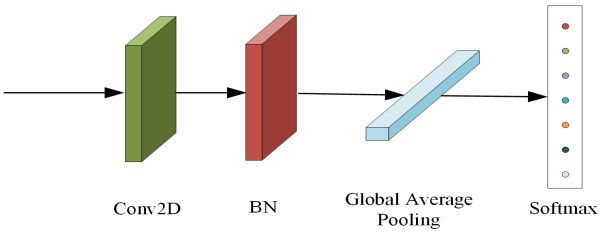

FIGURE 5. Diagram of classifier module.

The proposed auxiliary and primary classifier modules in this paper are shown in Figure 5. Both use the same structure and contain a standard convolutional layer, a BatchNormalization layer and an adaptive average pooling layer. ACM is mainly introduced after the DFEM to improve the utilization of important features at the intermediate level of network downsampling by introducing Acs.

In the final classification prediction, the classification results of the auxiliary and primary classifiers are fused using adaptive weighted fusion. Auxiliary and primary classifiers make predictions at different network hierarchies. Acs are usually predicted in the middle layer of the model, utilizing important features extracted from the middle layer. Main classifier, on the other hand, makes predictions at the deepest level, utilizing more abstract and semantically rich feature information. The multi-scale feature fusion can improve the efficiency of the model in utilizing important features at different levels. AWFM dynamically adjusts the fusion weights as the model training can be based on the performance and confidence of each classifier. Higher weights are assigned to better performing classifiers, while lower weights are assigned to poorer performing classifiers. This adaptive weighted fusion improves the recognition performance and generalization of the model.

In this paper, the Softmax activation function is used to accomplish the multi-classification task of emotion recognition. And the AM-Softmax loss function is used to calculate the model loss and optimize the probability distribution of each category independently. The AM-Softmax loss function was proposed by Wang et al. [23] to enable face features with larger class spacing and smaller intra-class distance. By limiting the angular differences between categories, overlap between categories can be reduced. It improves the model's ability to discriminate between categories. The formulation of the AM-Softmax Loss function can be expressed as equation (10).

$$L_{AMS} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i} e^{s\cdot cos\theta_j}} \quad (10)$$

where s is a learnable parameter to adjust the distance between categories. m is an angular interval parameter used to enhance angular differences between categories.

In the model of this paper, for any given category i, Softmax function can be expressed as Equation (11).

$$P_i = \frac{exp(x_i)}{\sum_{i=1}^{N} exp(x_j)} \quad (11)$$

Prior to classification, we used adaptive mean pooling. Adaptive mean pooling layer reduces the spatial dimensions of the feature map and retains more spatial information. The output feature is M_a. The final facial expression polarity can be obtained by inserting the feature map of the previous adaptive mean pooling layer into the softmax layer, as shown in Equation (12).

$$P = Softmax(w_c M_a + b_c) \quad (12)$$

where w_c is the weight matrix and b_c is the bias.

The final adaptive weighting formula is shown in Equation (13) and Equation (14).

$$P_{all} = P_1 \times \alpha_1 + P_2 \times \alpha_2 + P_3 \times \alpha_3 + P_4 \times \alpha_4 \quad (13)$$

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1 \quad (14)$$

where P_1, P_2, P_3 and P_4 are the output accuracies of the three Acs and the main classifier of the model. α_1, α_2, α_3 and α_4 are the weighting parameters. And P_{all} is the accuracy of the final weighted fused output.

IV. EXPERIMENTATION

A. DATASETS

We use four datasets, FER2013 [24], FERPlus [25], RAF-DB [26] and CK+ [27] for training and testing. And we compare the results with other experiments.

1) FER2013

These images in the FER2013 dataset were taken by matching sentiment keywords. There are 35887 images in this dataset and the image size within the dataset is 48 × 48. Out of 35887 images, 28709 images are used for training, 3589 images are used for validation of the model and 3589 images are used for testing the model.

2) FERPLUS

FERPlus is extended from FER2013. Some of the original images were relabeled. The dataset has a total of 35485 images and the image size within the dataset is 48 × 48. There are eight categories in this dataset. Here, we select the seven categories consistent with the FER2013 dataset for facial emotion recognition.

3) RAF-DB

RAF-DB is a large-scale real-world expression database, which contains 29,672 highly diverse facial images. Image size is 224 × 224. This experiment mainly uses the data of the basic expression part, which contains 12271 training samples and 3068 test samples.

4) CK+

The CK+ dataset is an extension of the CohnKanada dataset, published in 2010. It contains a total of 7 basic emotion categories. The image size within the dataset is 48 × 48 and the total number of images reaches 981.

B. EVALUATION METRICS

In this experiment, *Accuracy*, *Precision*, *Recall* and $F1-score$ for facial emotion recognition are used for multi evaluation. *Accuracy* indicates the ratio of the number of correctly predicted samples to the total number of samples. It measures the accuracy of the model in samples that are predicted to be positive examples. *Precision* is the proportion of all samples classified as positive cases that are indeed positive cases. *Recall* is the proportion of all samples that are truly positive examples that are correctly categorized as positive examples. $F1-score$ is the reconciled mean of *Precision* and *Recall*, which combines misclassification and underclassification of the classifier. The formula for these metrics are shown as Equation (15) to Equation (18):

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

where TP refers to the number of positive category samples that the model correctly predicts as positive. TN refers to the number of negative category samples that the model correctly predicts as negative. FP refers to the number of negative category samples that the model incorrectly predicts as positive. FN refers to the number of positive category samples that the model incorrectly predicts as negative.

C. EXPERIMENTAL SETUP

The experiments in this paper are based on the Tensorflow framework, Windows operating system and a GPU graphics card of NVIDIA GeForce RTX 3090 with 24GB of video memory. We use the machine learning system TensorFlow, and the cross-platform computer vision and machine learning software library OpenCV for model construction. Finally we optimize using the Adam optimizer [28]. In the preprocessing stage, data enhancement is performed by operations such as random horizontal flips, random vertical flips, perspective transformations, normalization processes, and grayscale map transformations [29]. To further balance the data during training, we set a balanced policy to equalize the weight of the input data during training by calling the compute_sample_weight function. The specific parameter settings are shown in the Table 2.

TABLE 2. The experimental parameter setting.

The Experimental Parameters	Value
epochs	300
batch_size	32 / 8
learning_rate	0.0001
patience	50
compute_sample_weight	balanced

V. RESULTS

A. EXPERIMENTAL RESULTS AND COMPARISON

We used several metrics to evaluate our model AMDCNN. In total, experiments were performed on four datasets and the computational effort of the model was calculated. The specific experimental results are shown in Table 3. For the three datasets, Fer2013, FERPLUS and RAF-DB, we used a batch size of 32 for training. Whereas the CK+ dataset has less data, we used a batch size of 8 for training. The model could achieve accuracies of 0.6982, 0.8540, 0.8677 and 0.9954 on the four datasets, respectively. The computational complexity of the model on the Fer2013,FERPLUS and RAF-DB datasets is only 138.98 MFlops. And the computational complexity of the model on the RAF-DB dataset is only 633.98M.

1) EXPERIMENTS WITH THE FER2013 DATASET

The AMDCNN model designed for this experiment was trained and tested on the FER2013 dataset with an accuracy of 0.6982 on the test set. The confusion matrix generated by the model was shown in Figure 6. As can be seen from the figure, the model recognizes happy with the accuracy of 0.89, sad and neutral with accuracies of more than 0.70. The remaining four emotions were recognized with low accuracy. Fear's accuracy was only 0.48. The main reason may be that Fear's facial expression changes are more complex. The model can easily misrecognize it as other expressions. Moreover, the number of training sets is relatively small, which leads to

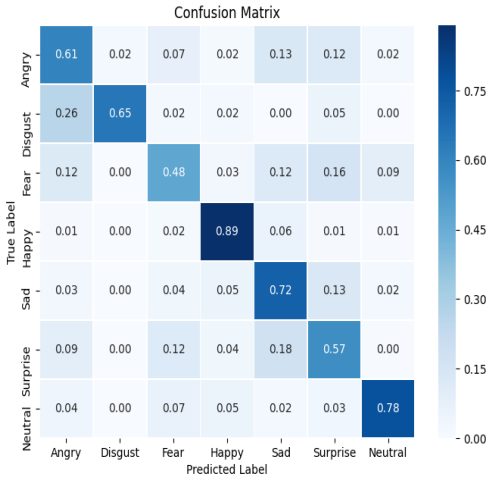

FIGURE 6. Confusion matrix on FER2013 dataset.

TABLE 3. Experimental results on four datasets.

Dataset	Batch_Size	Accuracy	Precision	Recall	F1-score	Flops(M)
Fer2013	32	0.6982	0.6985	0.6980	0.6980	138.98
FERPLUS	32	0.8540	0.8535	0.8540	0.8534	138.98
RAF-DB	32	0.8677	0.8708	0.8677	0.8686	633.31
CK+	8	0.9949	0.9954	0.9949	0.9950	138.98

poor results in feature extraction and training of the model. Table 4 shows the results of our model in comparison with other methods.

2) EXPERIMENTS WITH THE FERPLUS DATASET

The AMDCNN model proposed in this experiment was trained and tested on the FERPlus dataset and the accuracy of the test set was 0.8540. The confusion matrix generated by the model was shown in Figure 7. As can be seen from the figure, the model recognized around 0.90 of both happiness and sadness. And the accuracy of recognizing anger and neutral emotions was over 0.80. The remaining emotions of fear, disgust and surprise were also recognized with over 0.60 accuracies. Among other things, disgust is easily misidentified as anger. Whereas surprise is easily misidentified as sadness. The main reason may be that there are some similarities between these expressions, which would make the model more difficult to distinguish. The model's recognition on the FERPLUS dataset is much better than the experimental results on the FER2013 dataset, reflecting the higher accuracy of the recalibrated dataset. This also shows that the accuracy of the dataset calibration has a crucial impact on the results of the model training. Table 5 shows the results of our model in comparison with other methods.

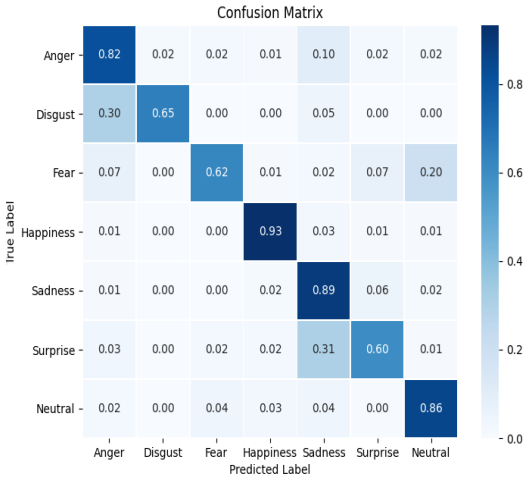

FIGURE 7. Confusion matrix on FERPLUS dataset.

3) EXPERIMENTS WITH THE RAF-DB DATASET

The AMDCNN model proposed in this experiment was trained and tested on the RAF-DB dataset and the accuracy of the test set was 0.8677. The confusion matrix generated by the model was shown in Figure 8. As can be seen from the figure,

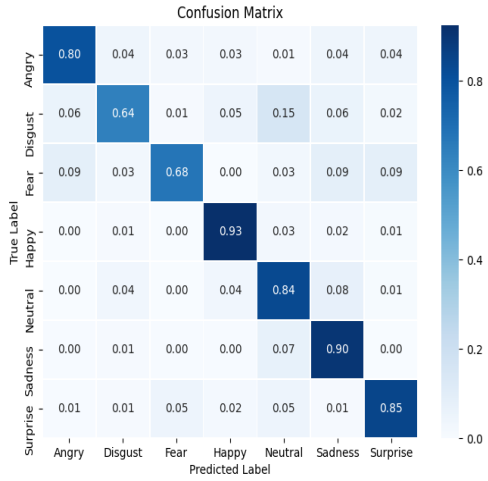

FIGURE 8. Confusion matrix on RAF-DB dataset.

the model recognized happy and sadness with more than 0.90 accuracies. And the recognition rates for angry, neutral and surprise were over 0.80. Each category of the dataset had a recognition accuracy higher than 0.60. The model is able to recognize individual expressions effectively. Table 6 shows the results of our model in comparison with other methods.

4) EXPERIMENTS WITH THE CK+ DATASET

The AMDCNN model proposed in this experiment was trained and tested on the CK+ dataset with 0.9949 accuracy on the test set. The confusion matrix generated by the model was shown in Figure 9. Due to the small size of the CK+ dataset and the unobstructed face, it is easier to perform facial emotion feature extraction and feature learning. As can be seen from the figure, the model has been trained to recognize every emotion except for surprise, which is slightly misrecognized. All other categories are able to recognize each emotion completely and correctly. Table 7 shows the results of our model in comparison with other methods.

B. ABLATION EXPERIMENTS

The ablation experiments of the method proposed in this paper are shown in the table below. Our baseline model is the DCNN model without the use of Acs and attention mechanisms. And the number of Acs increases as the model feature extraction progresses. The ablation experiment is shown in Table 8. As can be seen from the table, the utilization of the model for the intermediate level of important features is further improved with the use of MCBAM and Acs. There

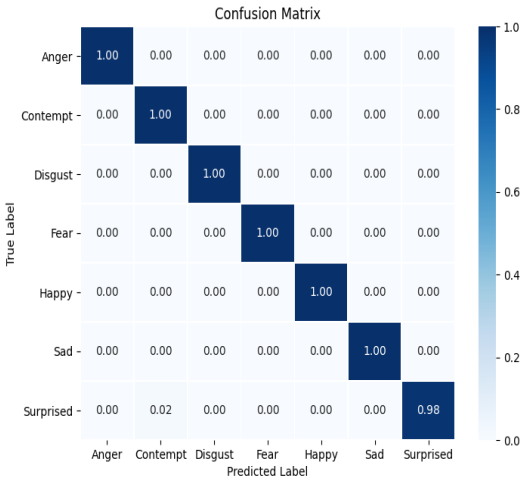

FIGURE 9. Confusion matrix on CK+ dataset.

TABLE 4. Comparison of different methods on the FER2013 dataset.

Methods	Parameters(M)	Accuracy
CNN [30]	0.93	0.6577
CNNCraft-net [31]	-	0.69
MTCNN [32]	0.05	0.67
BReG-Net-50 [33]	3.1	0.6921
Dar, et al. [34]	5.31	0.642
E-FCNN [35]	-	0.6657
MobileNetV3 [36]	4.2	0.639
Białek,et al. [37]	3.45	0.6763
He [18]	-	0.6949
Ours	**1.6**	**0.6982**

TABLE 5. Comparison of different methods on the FERPLUS dataset.

Methods	Parameters(M)	Accuracy
R. Saabni et al. [38]	-	0.8510
DenseNet [39]	0.17	0.8428
Rethink-Self-SSL [40]	-	0.7591
VGG13	9.41	0.8437
MobileNet3 [39]	3.88	0.8167
ShuffleNet2 [20]	1.26	0.8044
Ours	**1.6**	**0.8540**

TABLE 6. Comparison of different methods on the RAF-DB dataset.

Methods	Parameters(M)	Accuracy
Muhamad et al. [21]	2	0.8491
Ada-CM [41]	-	0.8532
A-MobileNet [12]	3.4	0.8449
Rethink-Self-SSL [40]	-	0.8554
Saurav,et al. [42]	4.81	0.84
WS-LGAN [43]	-	0.8507
Ours	**1.6**	**0.8677**

is also a better weighting effect for the important features of facial expressions. To further validate the role of the Acs, we performed a significance analysis of the model. The introduction of the Acs was finally verified to have a significant improvement in the recognition accuracy of the model by independent samples t-test ($p<0.001$).

TABLE 7. Comparison of different methods on the CK+ dataset.

Methods	Parameters(M)	Accuracy
FERNET [45]	44.7	0.9970
DTL-I-ResNet18 [46]	-	0.98
Daul CNNs [44]	1.08	0.9854
E-FCNN [35]	-	0.9495
Ours	**1.6**	**0.9949**

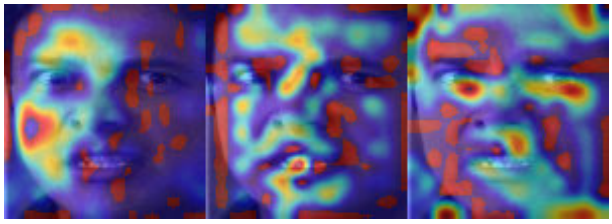

FIGURE 10. Grad-CAM heat map comparison diagram.

C. VISUALIZATION AND ANALYSIS

Although ablation experiments have been performed as described above, we further performed visualization analysis to observe the effect of MCBAM. Using the Grad-CAM visualization technique, we output heat maps under three different models: baseline, baseline+CBAM and baseline+MCBAM. The specific results are shown in Figure 10. As can be seen from the figure, with the introduction of the attention mechanism, the model is able to further focus on the middle region of the face. It is beneficial to further extract facial features related to emotion recognition. And by further enhancing the CBAM, the model is able to better focus on areas such as the eyes, nose and mouth. It can be seen that by introducing MCBAM, the model is able to extract more accurate key features, which is conducive to better emotion recognition by the model.

VI. DISCUSSION

In this paper, we propose a facial emotion recognition method to improve the utilization of important features in the intermediate level of DCNN and design an AMDCNN model. The model further improves the utilization of intermediate-level important features of the model by introducing the MCBAM and Acs. The MCBAM is able to more accurately weight important intermediate-level feature information through a multi-branch structure, allowing the model to better focus on important feature information. The classifier results are then fused with adaptive weighting to obtain the final recognition results of facial expressions. After experimental validation, the intermediate-level feature utilization of the AMDCNN model is greatly improved, and the recognition accuracy of the whole model is further improved. The comparison of the experimental results with other lightweight models can also show that our model has a more competitive advantage.

TABLE 8. Ablation experiment.

Methods	FER2013	FERPLUS	RAF-DB	CK+
baseline	0.6766	0.7922	0.7851	0.9798
baseline+CBAM	0.6840	0.8055	0.8162	0.9831
baseline+MCBAM	0.6888	0.8143	0.8305	0.9864
baseline+MCBAM+1ACs	0.6919	0.8255	0.8420	0.9898
baseline+MCBAM+2ACs	0.6933	0.8426	0.8591	0.9913
baseline+MCBAM+3ACs	**0.6982**	**0.8540**	**0.8677**	**0.9949**

In the four datasets, for the CK+ dataset, the post-training model test results were 0.9949. But this only applies to frontal and unobstructed facial expression images collected in the laboratory and lacks robustness and practicality. The facial emotion recognition results are not as accurate in the three real datasets FER2013, FERPLUS and RAF-DB. On the latter two datasets, a recognition accuracy of about 0.86 can be achieved. However, the recognition accuracy on the FER2013 dataset is only 0.6982. We can notice from Figure 6 and Figure 7 that the model can easily misidentify disgust as anger. Besides, sad and surprise are more easily misidentified with each other as the other. This may be due to the high similarity between the two types of expressions. Therefore, the model still needs to pay more attention to the important regions that can distinguish between different expressions.

VII. CONCLUSION AND FUTURE WORK

In this paper, in order to further improve the utilization of important features in the intermediate level of DCNN, an AMDCNN network model is designed for facial emotion recognition research. By conducting experiments on four datasets, FER2013, FERPlus, RAF-DB and CK+, it is confirmed that the proposed model has better facial emotion recognition. The AMDCNN model achieved accuracies of 0.6982, 0.8540, 0.8677 and 0.9949 on the four datasets. In addition, we further validated the validity of Acs through independent samples t-tests. For MCBAM, we also performed a visualization analysis. Compared to CBAM, the improved MCBAM has better feature weighting, allowing the model to place more emphasis on the more important facial emotional features. Compared to other lightweight models, our model parameter count is only 1.6M, but it is still highly competitive.

In the future, we hope to further improve our model. The cropped facial image and the original facial image are input into the model separately by pre-processing by cropping the important areas of the face. Finally, feature fusion is performed, which enables the model to better distinguish different emotions and achieve better classification results.

ACKNOWLEDGMENT

The authors would like to thank "Changchun Computing Center" and "Eco-Innovation Center" for providing inclusive computing power and technical support of MindSpore during the completion of this article.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[3] 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.

[4] 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.

[5] 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.

[6] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[7] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2017–2025.

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9423–9433.

[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[10] Y. Zhang, X. Zou, S. Yu, L. Huang, W. Wang, S. Zhao, and X. Wang, "DNN-CBAM: An enhanced DNN model for facial emotion recognition," *J. Intell. Fuzzy Syst.*, vol. 43, no. 5, pp. 5673–5683, Sep. 2022.

[11] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP," *Sensors*, vol. 23, no. 9, p. 4204, Apr. 2023.

[12] J. 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, 2022.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[15] 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.

[16] J. Daihong, H. Yuanzheng, D. Lei, and P. Jin, "Facial expression recognition based on attention mechanism," *Scientific Program.*, vol. 2021, pp. 1–10, Mar. 2021.

[17] T. Shen and H. Xu, "Facial expression recognition based on multi-channel attention residual network," *CMES-Comput. Model. Eng. Sci.*, vol. 135, no. 1, pp. 540–560, 2023.

[18] Y. He, "Facial expression recognition using multi-branch attention convolutional neural network," *IEEE Access*, vol. 11, pp. 1244–1253, 2023.

[19] H. Burrows, J. Zarrin, L. Babu-Saheer, and M. Maktab-Dar-Oghaz, "Realtime emotional reflective user interface based on deep convolutional neural networks and generative adversarial networks," *Electronics*, vol. 11, no. 1, p. 118, Dec. 2021.

[20] J. L. Joseph and S. P. Mathew, "Facial expression recognition for the blind using deep learning," in *Proc. IEEE 4th Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Sep. 2021, pp. 1–5.

[21] M. D. Putro, D.-L. Nguyen, A. Priadana, and K.-H. Jo, "An efficient multi-view facial expression classifier implementing on edge device," in *Proc. Asian Conf. Tntelligent Inf. Database Syst.* Singapore: Springer, 2022, pp. 517–529.

[22] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, "LibreFace: An open-source toolkit for deep facial expression analysis," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8205–8215.

[23] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[24] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, Daegu, (South) Korea. Berlin, Germany: Springer, 2013, pp. 117–124.

[25] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.

[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] H. Alshamsi, V. Kepuska, and H. Meng, "Real time automated facial expression recognition app development on smart phones," in *Proc. 8th IEEE Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2017, pp. 384–392.

[28] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

[29] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, 2nd ed. Berlin, Germany: Springer, 2012, pp. 421–436.

[30] A. Agrawal and N. Mittal, "Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy," *Vis. Comput.*, vol. 36, no. 2, pp. 405–412, Feb. 2020.

[31] A. Mostafa, H. University, H. El-Sayed, M. Belal, H. University, and H. University, "Facial expressions recognition via CNNCraft-net for static RGB images," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 4, pp. 410–421, Aug. 2021.

[32] N. Zhou, R. Liang, and W. Shi, "A lightweight convolutional neural network for real-time facial expression detection," *IEEE Access*, vol. 9, pp. 5573–5584, 2021.

[33] B. Hasani, P. S. Negi, and M. H. Mahoor, "BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1023–1036, Apr. 2022.

[34] T. Dar, A. Javed, S. Bourouis, H. S. Hussein, and H. Alshazly, "Efficient-SwishNet based system for facial emotion recognition," *IEEE Access*, vol. 10, pp. 71311–71328, 2022.

[35] J. Shao and Q. Cheng, "E-FCNN for tiny facial expression recognition," *Int. J. Speech Technol.*, vol. 51, no. 1, pp. 549–559, Jan. 2021.

[36] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[37] C. Bialek, A. Matiolanski, and M. Grega, "An efficient approach to face emotion recognition with convolutional neural networks," *Electronics*, vol. 12, no. 12, p. 2707, Jun. 2023.

[38] R. Saabni and A. Schclar, "Facial expression recognition using combined pre-trained ConvNets," *Comput. Sci. Inf. Technol.*, vol. 95, pp. 95–106, Jul. 2020.

[39] G. Zhao, H. Yang, and M. Yu, "Expression recognition method based on a lightweight convolutional neural network," *IEEE Access*, vol. 8, pp. 38528–38537, 2020.

[40] B. Fang, X. Li, G. Han, and J. He, "Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning," *IEEE Access*, vol. 11, pp. 45547–45558, 2023.

[41] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "Towards semi-supervised deep facial expression recognition with an adaptive confidence margin," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4166–4175.

[42] S. Saurav, R. Saini, and S. Singh, "EmNet: A deep integrated convolutional neural network for facial emotion recognition in the wild," *Int. J. Speech Technol.*, vol. 51, no. 8, pp. 5543–5570, Aug. 2021.

[43] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local–global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, 2020.

[44] S. Saurav, P. Gidde, R. Saini, and S. Singh, "Dual integrated convolutional neural network for real-time facial expression recognition in the wild," *Vis. Comput.*, vol. 38, no. 3, pp. 1083–1096, Mar. 2022.

[45] C. Gupta, M. Kumar, A. K. Yadav, and D. Yadav, "FERNET: An integrated hybrid DCNN model for driver stress monitoring via facial expressions," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 37, no. 3, Mar. 2023, Art. no. 2357002.

[46] R. Helaly, S. Messaoud, S. Bouaafia, M. A. Hajjaji, and A. Mtibaa, "DTL-I-ResNet18: Facial emotion recognition based on deep transfer learning and improved ResNet18," *Signal, Image Video Process.*, vol. 17, no. 6, pp. 2731–2744, Sep. 2023.

YUJIE SHANG was born in Henan, China, in 1998. He received the B.E. degree from Changchun University of Science and Technology (CUST), in 2021, where he is currently pursuing the M.E. degree. His research interests include deep learning, digital image processing, and emotion recognition.

FEI YAN was born in Shandong, China, in 1987. He received the B.E. degree from Changchun University, in 2009, the M.E. degree from Changchun University of Science and Technology (CUST), in 2012, and the Ph.D. degree from Jilin University, in 2016. He is currently a Lecturer with CUST. His research interests include pattern recognition and intelligent information processing.

YUNQING LIU was born in Henan, China, in 1970. He received the B.E., M.E., and Ph.D. degrees from Changchun University of Science and Technology (CUST), Changchun, in 1994, 1998, and 2009, respectively. He was a Professor and a Master Instructor. His research interests include radar signal processing, laser communication, and digital synchronization.

QI LI was born in Jilin, China, in 1996. She received the B.S. degree in automation from Jilin Engineering Normal University, in 2017. She is currently pursuing the Ph.D. degree in information and communication engineering with Changchun University of Science and Technology. Her research interests include intelligent information processing and EEG signal processing.

QIONG ZHANG received the M.S. and Ph.D. degrees from the College of Instrumentation and Electrical Engineering, Jilin University, Changchun, China, in 2015 and 2018, respectively. She is currently a Lecturer with the School of Electronics and Information Engineering, Changchun University of Science and Technology. Her research interests include data processing and inversion.