IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2023, accepted 9 June 2023, date of publication 13 June 2023, date of current version 28 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285781

RESEARCH ARTICLE

Expression Recognition Based on Multi-Regional Coordinate Attention Residuals

JIANGHAI LAN[1], XINGGUO JIANG[1,2], GUOJUN LIN[1,2], XU ZHOU[1], SONG YOU[1], ZHEN LIAO[1], AND YANG FAN[1]

[1]School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China
[2]Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China

Corresponding author: Xingguo Jiang (179828432@qq.com)

This work was supported in part by the Scientific Research Foundation of Sichuan University of Science and Engineering under Grant 2019RC12, in part by the Open Foundation of Artificial Intelligence Key Laboratory of Sichuan Province under Grant 2020RZJ03, and in part by the Postgraduate Innovation Fund of Sichuan University of Science and Engineering under Grant Y2022146.

ABSTRACT Facial expression recognition is an important research direction of emotion computing and has broad application prospects in human-computer interaction. However, noise such as illumination and occlusion in natural environment brings many challenges to facial expression recognition. In order to solve the problems such as low recognition rate of facial expression in natural environment, unable to highlight the characteristics of facial expression in global facial research, and misclassification caused by the similarity between negative expressions. In this paper, a multi-region coordinate attentional residual expression recognition model (MrCAR) is proposed. The model is mainly composed of the following three parts: 1) multi-region input: MTCNN is used for face detection and alignment processing, and the eyes and mouth parts are further cropped to obtain multi-region pictures. Through multi-region input, local details and global features are more easily obtained, which reduces the influence of complex environmental noise and highlights the facial features. 2) Feature extraction module: On the basis of residual element, CA-Net and multi-scale convolution were added to obtain coordinate residual attention module, through which the model's ability to distinguish subtle changes of expression and the utilization rate of key features were improved; 3) Classifier: Arcface Loss is used to enhance intra-class tightness and inter-class difference at the same time, thus reducing the wrong classification of negative expressions by the model. Finally, the accuracy rates of CK+, JAFFE, FER2013 and RAF-DB were 98.78%, 99.09%, 74.50% and 88.26%, respectively. The experimental results show that compared with many advanced models, the MrCAR model in this paper is more competent for the task of expression classification.

INDEX TERMS Deep learning, artificial intelligence, expression recognition, CA-Net, Arcface loss.

I. INTRODUCTION

People's facial expressions can directly show their inner feelings, and people can fully communicate with each other through the changes of facial expressions. Many ''unspoken'' situations refer to the subtle expressions of communication. The application of facial expression research is very wide, such as: driver's fatigue detection, criminal psychology, student classroom education, film and television special effects, etc.

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

Facial expression recognition is a key technology of computer vision, mainly in the accurate classification of different kinds of expressions. Facial expression recognition mainly includes three aspects: face detection, facial feature extraction and expression classification. Among them, facial feature extraction and expression classification are important steps and also the key to improve the accuracy of face expression recognition. Facial feature extraction mainly includes traditional methods and methods based on deep learning. The commonly used traditional methods are mainly geometric structure-based and representation-based feature extraction methods, one of which requires precise localization of the

face key, and the other extracts features mainly through texture information of the image. However, the traditional feature extraction methods are based on manual production and mainly rely on historical experience, which is prone to feature loss. At present, methods based on deep learning have become the mainstream method for facial expression recognition. Jiang et al. [1] proposed a method based on representation reinforcement network (RRN) combined with transfer learning for facial expression recognition. Mao et al. [2] proposed a deep convolutional neural network based on regions of interest for facial expressions.

However, in natural environments, expression recognition is often affected by many noises, such as lighting, occlusion, and the presence of many non-face regions, which brings many troubles to facial expression feature extraction. In addition, the within-class similarity between similar expressions and the unbalanced distribution of the datasets themselves can easily lead to wrong classification. Previous single deep learning methods often achieve unsatisfactory results when dealing with such situations. In order to solve these problems and better adapt to the complex environment of facial expression recognition, this paper proposes a new facial expression recognition algorithm, the main contributions are as follows:

- A multi-region input training method is proposed: firstly, face detection, cropping and alignment operations are performed by MTCNN (Multi-task Cascaded Convolutional Networks) [3] to remove the redundant non-face background to obtain the global input; and further cropping of eye and mouth related regions to obtain the local input; finally, the global input and the local input are fed into the network for training to extract features. Finally, the global input and local input are fed together into the network for training and feature extraction. This training method can extract both local detail features and global overall features, and can also effectively reduce the impact of noise in the images and improve the robustness of the network model.
- Improvements to the residual network: introduce CA-Net on the basis of residual units to improve the model's ability to extract features of key parts through direction-aware and position-sensitive information; and add multi-scale convolution to enhance the perceptual field of the network model through convolution kernels of different sizes so that the model can better capture the detailed features of expression changes; add a maximum pooling layer to optimize the Down-sampling method to reduce information loss.
- Arcface Loss is used instead of the traditional Softmax Loss [4]: expanding the inter-class distance and reducing the intra-class distance to reduce the misclassification rate of the model for similar negative expressions.
- The proposed multi-region coordinate attention residual-based face expression recognition model achieves better results on the CK+, JAFFE, FER2013, RAF-DB datasets, outperforming many advanced traditional algorithms as well as deep learning models.

II. RELATED WORK

The current existing expression recognition methods generally rely on the overall features of the face for research, without considering some local features of the face, resulting in unsatisfactory accuracy of facial expression recognition. Happy and Routray [5] and Majumder et al. [6] proposed that expression changes usually occur in some prominent five sensory regions, such as near the mouth, nose, and eyes, and the research direction also shifted from feature extraction of the whole face to focus on expression-related regions [7]. The attention mechanism is able to selectively focus on part of the information. Sun et al. [8] used Region of Interest (ROI) to improve the accuracy of expression prediction. Hu et al. [9] proposed Squeeze and Excitation Networks (SE-Net), which enables the capture of local information by increasing the spatial encoding quality. In contrast, Woo et al. [10] added attention maps based on channel and space separately to obtain the Convolutional Block Attention Module (CBAM), but ignored the location information. While Hou et al. [11] proposed Coordinate Attention Networks (CA-Net) can locate the region of interest more accurately by embedding location information into channel attention, which is a plug-and-play lightweight module. While adding an attention mechanism helps to focus on the key parts of expressions, extracting expression features from the whole image alone would overemphasize the completeness of facial expressions and tend to ignore much local detail information. Wang et al. [12] used face keypoints tagging to perform random clipping of faces and then fed into a model with an attention mechanism for training, which considered local detail information but had too many repetitive regions. Jiang et al. [13] realized facial expression recognition by combining block attention module with multi-feature fusion. Gan et al. [14] used multiple attention networks to extract key region features from different regions and achieved better results in facial expression recognition.

Besides, the design of loss function is also important for facial expression recognition. Many loss functions are designed in face recognition, and these loss functions are also applicable to face expression recognition. Meng et al. [15] aggregated similar features in the aggregation category and different features in the estrangement category by contrast loss. Liu et al. [16] proposed Large-Margin Softmax Loss (L-Softmax Loss) by adding constraints m on top of Softmax to make $\|W_1\| \|x\| \cos(m\theta_1) \geq \|W_2\| \|x\| \cos(\theta_2)$, increase the difficulty of learning features thus making the classification boundary strict. Liu et al. [17] also normalized the weights on the basis of L-Softmax Loss: $\|W\| = 1$, resulting in the mapping of points on the features to the unit hypersphere, such that the prediction of the model depends only on the angle between W and X, which can further reduce the interclass distance. The L_2-constrained Softmax Loss proposed by Wang et al. [18] maximizes the classification bounds directly in the cosine space without considering the angular variation of the weight vectors. The Additive angular margin loss (Arcface Loss) proposed by Deng et al. [19] fixes

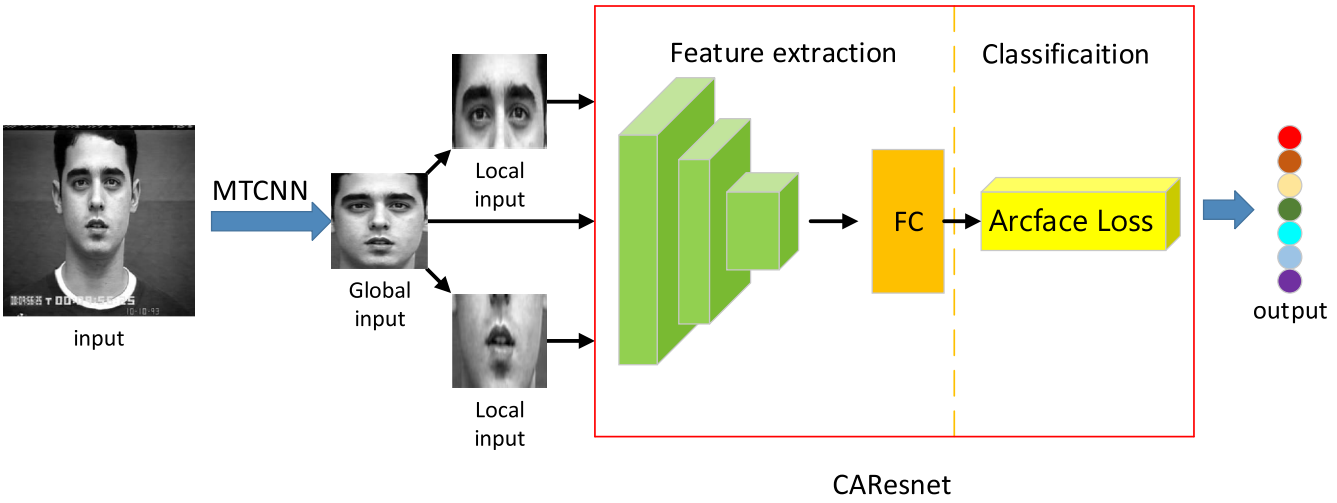


FIGURE 1. MrCAR framework. CAResnet represents the coordinate attention residual network.

the input feature $\|x_i\|$ by normalizing it by L_2, and rescales it to s, With an additive Angle margin penalty m is applied between x_i and W_{yi} to enhance both intra-class tightness and inter-class difference.

III. PROPOSED METHOD

In order to solve the problem of difficulty in obtaining information about key regions of faces in complex environments, as well as to reduce the cases of misclassification among similar negative expressions. In this paper, we propose a face expression recognition model based on multi-region coordinate attention residuals (MrCAR), and the main framework of the model is shown in Figure 1. The algorithm is divided into three steps: 1) detect the face region of the input image using MTCNN, cut and align the global input image, and then perform a crop on the eye and mouth areas to obtain two local input images. 2) feed the global input image and the two local input images into the coordinate attention residual network for feature extraction, and fuse the extracted features through the fully connected layer. The extracted features are fused. By this multi-region image input and coordinate attention residual network, we can effectively locate face expression related areas and extract key feature information and suppress non-key areas. 3) We use Arcface Loss classifier to classify expressions and reduce the misclassification rate among negative expressions.

A. MULTIZONE INPUT

In this paper, multi-region input is adopted, and MTCNN [3] algorithm is used to preprocess the image before feature extraction. MTCNN is a cascading deep learning algorithm that can quickly complete face detection. Its structure mainly consists of three sub-networks: Proposal Network (P-Net), Refine Network (R-Net) and O-Net (Output Network). The input image is first processed by three convolutions of P-Net and a classifier to obtain several approximate face boundary regions. Then continue to use R-Net to screen these areas, remove many areas of inaccurate positioning, output high reliability of the face region so as to get more accurate positioning of the face region; Finally, O-Net is used for accurate face feature localization, and the face area and 5 key feature points are framed. The multi-region input in MrCAR model in this paper is mainly realized through MTCNN network. Firstly, face detection, alignment and clipping are performed by MTCNN algorithm to obtain global input. Then further cut the eye and mouth related areas to get local input; Finally, the global input picture and the local detail input picture are sent to the network to train the feature extraction.

B. COORDINATE ATTENTION RESIDUAL NETWORK

The coordinate attention residual network model proposed in this paper is shown in Figure 2, which consists of a convolutional layer, a maximum pooling layer, three coordinate attention residual modules, three Down-sampling modules, a global average pooling layer, a fully connected layer, and an Arcface Loss classification layer. The CARblock and DSblock in Figure 2 represent the coordinate attention residual module and represent the Down-sampling module. Let the size of the input image be44 $\times$ 44. First, the image size is changed to 22 $\times$ 22 by a convolutional layer and a maximum pooling layer, and then the feature information of the image

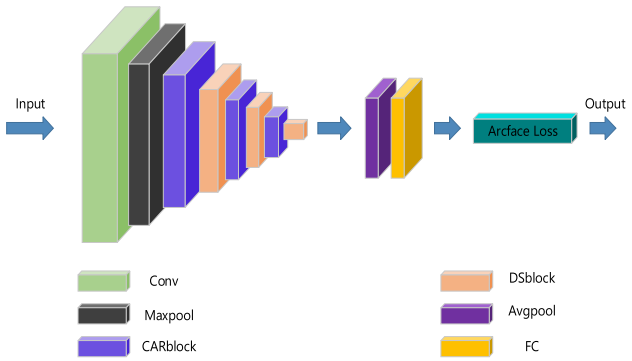


FIGURE 2. Structure of coordinate attention residual network model.

is processed by the coordinate attention residual module and the Down-sampling module, and then it is processed by the global average pooling layer and the fully connected layer, and finally the classification output is made directly by the Arcface Loss layer. The parameters of this network structure are shown in Table 1.

TABLE 1. Network structure parameters.

Layer name	Ouput size	Detail
Conv1	(44,44,64)	(3,3) stride 1
Maxpool	(22,22,64)	(3,3) stride 1
CARblock1	(22,22,128)	stride 1
DSblock1	(11,11,128)	stride 2
CARblock2	(11,11,256)	stride 1
DSblock2	(6,6,256)	stride 2
CARblock3	(6,6,512)	stride 1
DSblock3	(3,3,512)	stride 2
Avgpool	(1,1,512)	
FC, Arcface Loss	(7)	

1) CA-NET

CA-Net mainly decomposes the 2-dimensional global pooling operation into two 1-dimensional codes, so that not only cross-channel information can be effectively obtained, but also direction sensing and position sensitive information can be obtained. Its specific structure is shown in Figure 3.

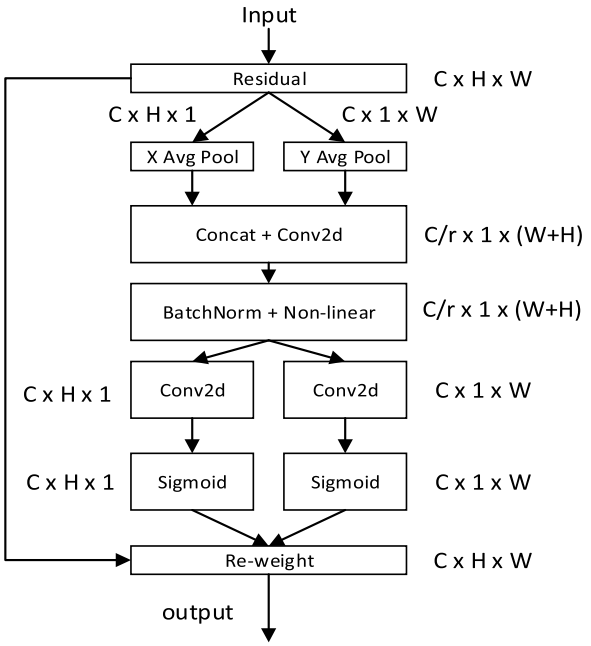


FIGURE 3. CA-Net structure.

Where, variable C represents the number of channels, variable represents H height and variable W represents width. Channel relationships and long-term dependencies are encoded by precise location information to form a feature map sensitive to direction and position to enhance feature representation of facial expressions. The implementation of coordinated attention module consists of two steps: coordinate information embedding and coordinate attention generation.

Coordinate information embedding: To motivate remote spatial interactions where the attention module can capture more precise location information, the global pooling is decomposed into a pair of one-dimensional feature encoding operations by decomposing it in the form of Formula (1).

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

Which is associated with the c channel output. For the input x, each channel is coded along horizontal and vertical coordinates using pooled cores of sizes $(H, 1)$ and $(1, W)$, respectively. Therefore, the output of the c channel with height h can be expressed as:

$$Z_c^h = \frac{1}{W} \sum_{0 \le i \le W} x_c(h, i) \tag{2}$$

Similarly, the output of channel c of width w can be written as:

$$Z_c^w = \frac{1}{H} \sum_{0 \le j \le H} x_c(j, w) \tag{3}$$

By aggregating facial expression features along two spatial directions, horizontal and vertical, respectively, the attention mechanism module can capture long-term dependencies through these two spatial directions while preserving more accurate location information.

Coordinated attention generation: The coordinate information embedding is performed by Formula (2) and (3) to obtain the aggregated feature mapping, and then the stitching operation is performed, and the 1×1 convolutional transform function F_1 is used to operate on it to obtain:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{4}$$

where $[\cdot, \cdot]$ denotes the join operation along the spatial dimension; δ is the nonlinear activation function; $f \in R^{C/r(H+W)}$ is the intermediate feature map that encodes the spatial information in the horizontal and vertical directions, and r denotes the Down-sampling ratio values. Next, f is divided into two independent tensors $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$ along the space dimension; By using the other two 1×1 convolution transformations F_h and F_w to transform f^h and f^w respectively into tensors with the same number of input x channels, we get

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{5}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{6}$$

where σ is a *sigmoid* function; the outputs g^h and g^w denote attention weights. Finally, the output of our coordinate attention block Y can be written as:

$$Y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

2) COORDINATED ATTENTION RESIDUALS MODULE

The coordinated attention residual module constructed in this paper is shown in Figure 4. The structure with stride=1 represents CARblock and the structure with stride=2 represents DSblock. Our improvements mainly introduce multi-scale convolutional blocks and coordinate attention mechanisms based on the traditional residual structure. Firstly, the convolution of 1×1 is downscaled and processed by BN layer and ReLU activation function, and then divided into two branches for feature extraction by convolution layers of 1×1 and 3×3, respectively, to expand the perceptual field of the network so as to obtain richer feature information of different sizes of image local regions.

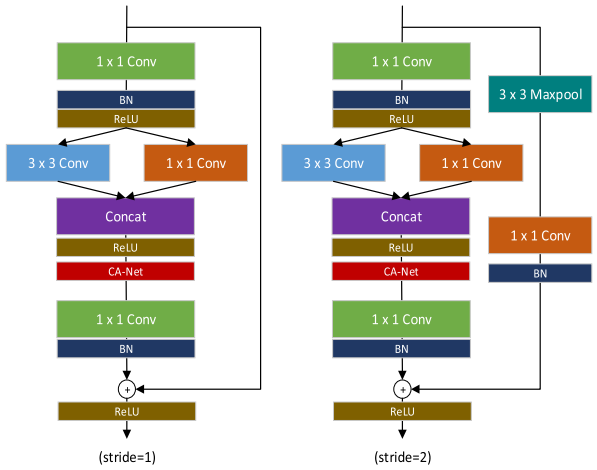


FIGURE 4. Coordinate attention residual block structure.

The feature information obtained from these two branches is then fused by Concat, and then the local key information is located and obtained more accurately by CA-Net. It is then up-dimensioned by a 1×1 convolution to reach the same dimensionality as at the input, and finally summed with the constant mapping branch to form the residual structure and output by the ReLU activation function. In addition to this, the coordinate residual module in this paper also improves the Down-sampling module at step size 2 by replacing the 1×1 convolution at step size 2 with a 1×1 convolution at step size 1 and adding a maximum pooling of size 3×3 at step size 2. Because the original 1×1 convolution with step size 2 causes much information loss, the improved Down-sampling module helps classification and reduces information loss by combining a 3×3 convolution with a 3×3 maximum pooling. It also does not increase the complexity of the network.

3) ARCFACE LOSS

Softmax is a loss function applied to the classification problem and is located in the last layer of the convolutional neural network layer, transforming the problem of N classification into an $N \times 1$ dimensional vector with a probability sum of 1. where $x_i \in R^d$ denotes the depth feature of the i-th sample, which belongs to class y_i; $W_j \in R^d$ denotes the weight of column j-th of $W \in R^{d \times n}$. $b_j \in R^n$ is the bias term. N represents the number of samples and n represents the number of classifications. The Softmax Loss expression is:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{yi}^T x_i + b_{yi}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{8}$$

However, Softmax Loss does not require intra-class compactness and inter-class separation. However, inside the face expressions, negative expression facial features have certain similarity and confusion is difficult to distinguish, and the commonly used Softmax Loss classification is not effective. In order to ensure the separability, while making the feature vector classes as compact as possible within the class and as separate as possible between classes. Therefore, Arcface Loss [17] is used instead of Softmax Loss in the network model in this paper.

On this basis, first fix the deviations $b_j = 0$; $W_j^T x_i = \|W_j\| \|x_i\| \cos\theta_j$, where θ_j is the angle between W_j and x_i; normalize the weights W_j : $\|W_j\| = 1$; The input features $\|x_i\|$ are fixed by normalizing L_2, and rescaled to S. The learned embedding features are distributed on a hypersphere of radius S. As shown in Formula L_1:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{9}$$

Since the embedded features are distributed around the center of each feature on the hypersphere, an additional boundary penalty m is added to the angle between $\|W_j\|$ and $\|x_i\|$ on top of L_1 to enhance both intra-class closeness and inter-class variability. As shown in Formula L:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{10}$$

IV. EXPERIMENT AND DISCUSSION

A. DATASETS AND EXPERIMENT SETTINGS

1) DATASET

In this experiment, four public datasets commonly used in facial expression recognition were selected: CK + [20] dataset, FER2013 [21] dataset JAFFE [22] dataset and RAF-DB [23] dataset. CK+ dataset contains 593 image sequences of 123 individuals, among which 327 are labeled image sequences. Images with strong expressions are selected as experimental data. A total of 981 images are selected, with the following 7 expressions: angry, disgust, fear, happy, sadness, surprise and contempt, cropped the face of the image to a fixed size of 48×48. The FER 2013 facial expression dataset is mainly derived from a variety of Internet images. There are a total of 35887 facial expression photos in this dataset. The FER2013 facial expression dataset consists of three parts: Training, PublicTest and PrivateTest. The number of facial expression photos corresponding to these three

parts was 28709,3589 and 3589 respectively. All of them contain the following seven expressions: angry, disgust, fear, happy, sadness, surprise, neutral; The images in this dataset are all grayscale images of size 48 × 48. JAFFE dataset contains 213 facial expressions of 10 Japanese women, including: happy, sadness, fear, angry, surprise, disgust and neutral. Here, we also cut the size of the image to 48 × 48. The RAF-DB dataset contains approximately 30,000 facial images, including seven basic expressions: anger, disgust, fear, happy, sadness, surprise, and neutral. In this paper, 15,339 images with expression classification labels were selected for experiments, including 12,271 training sets and 3,068 test sets. The partial image of the four datasets is shown in Figure 5.

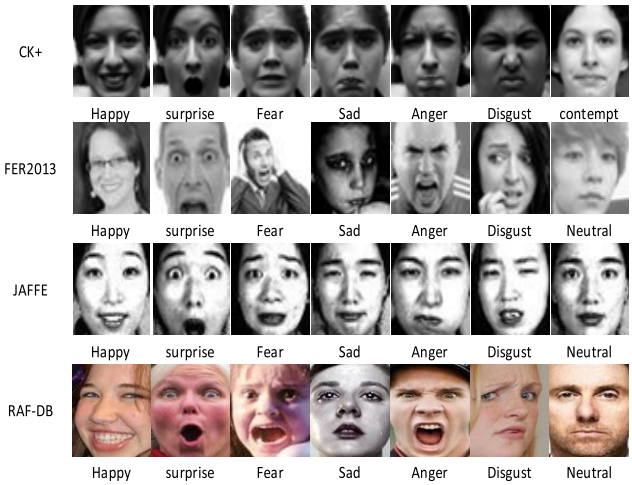


FIGURE 5. Sample images of four datasets.

The images were preprocessed before the training, and the original images of CK+ dataset, FER2013 dataset and JAFFE dataset were randomly clipped to the size of 44 × 44, while the original images of RAF-DB dataset were clipped to the size of 98 × 98. Then, random rotation, shrinkage, mirroring and other geometric processing are performed to increase the dataset. The original CK+ dataset and JAFFE dataset had a small number of pictures, and data enhancement was carried out on them. MTCNN network was used to detect and align faces in FER2013 dataset, so as to remove parts without faces in the dataset. In the testing phase, the initial images are first center-cropped and then mirrored so that the data can be increased to 10 times the original dataset. After loading these photos into the model, the average probability is taken and the maximum output classification corresponds to its expression classification. The advantage of this operation is that it can better reduce the cases of misclassification.

2) EXPERIMENT SETTINGS

The experimental environment is Windows 10, 64-bit OS, NVIDIA GeForce RTX 2060 GPU, 11th Gen Intel (R) Core (TM)i7-11700@2.50GHz CPU, code written in python, and deep learning environment built by Pytorch.

For CK+ dataset and JAFFE dataset, the number of images in the two datasets after data enhancement was 7881 and 6930, respectively. In this paper, 10 -fold cross validation method is selected for testing. Here, the images of the expanded CK+ dataset and JAFFE dataset are divided into ten parts, nine of which are the training set and one of which is the test set. Therefore, the number of training set images for the CK+ dataset is 7093 and the number of test set images is 788; the number of training set images for the JAFFE dataset is 6237 and the number of test set images is 693. Then, to reduce the chance, we also conducted several iterations of the validation method, while repeatedly training and testing through the generated random subsamples, and finally taking the average of the mean probability. For both the CK+ and JAFFE datasets, 100 epochs are set, the batch size is set to 64, and the learning rate is set to 0.01. When the model runs through 20 epochs on the CK+ dataset, the learning rate decays to 0.9 times that of the previous epoch after every 2 epochs, while after 20 epochs on the JAFFE dataset, the learning rate decays to 0.9 times that of the previous epoch after every 1 epoch. For FER2013 and RAF-DB data sets, the experimental parameters were set differently. In both experiments, 200 epochs were set, the initial learning rate was set to 0.01, and the batch size was set to 32. However, after the model nuns 80 epochs on the FER2013 dataset, the learning rate of every 5 epochs becomes 0.9 times of the last epoch. When the model runs through 50 epochs on the RAF-DB dataset, the learning rate of each interval of 5 epochs becomes 0.9 times of the previous epoch.

B. EXPERIMENTAL RESULTS AND ANALYSIS

The accuracy and loss of the MrCAR model on the CK+ dataset, JAFFE dataset, FER2013 dataset and RAF-DB dataset are shown in Figure 6. The accuracy of the model when tested on the CK+ dataset has been increasing until 40 epochs with the same number of iterations; although there are some fluctuations in the middle, it starts to converge gradually after 40 epochs, and the loss also starts to converge after 40 epochs. The accuracy and loss of the model on the JAFFE dataset start to converge gradually after the 30th epoch. The accuracy of the model on the FER2013 dataset not only increases until 125 epochs, but also converges after 125 epochs; the loss also converges at 125 epochs. The accuracy and loss of the model on the RAF-DB dataset then converge after the 100th epoch. The analysis of the accuracy and loss curves on the four datasets shows that the network model in this paper has good generalization ability and robustness. However, on the FER2013 dataset, the difference between the results during training and the results from the tests conducted is large. This is because the network learns the expression features of the training set during training, and when the model reaches optimization, there are many mislabeled samples of non-faces on the test set, which makes the model recognize a large difference between the results on the test set and the training set.

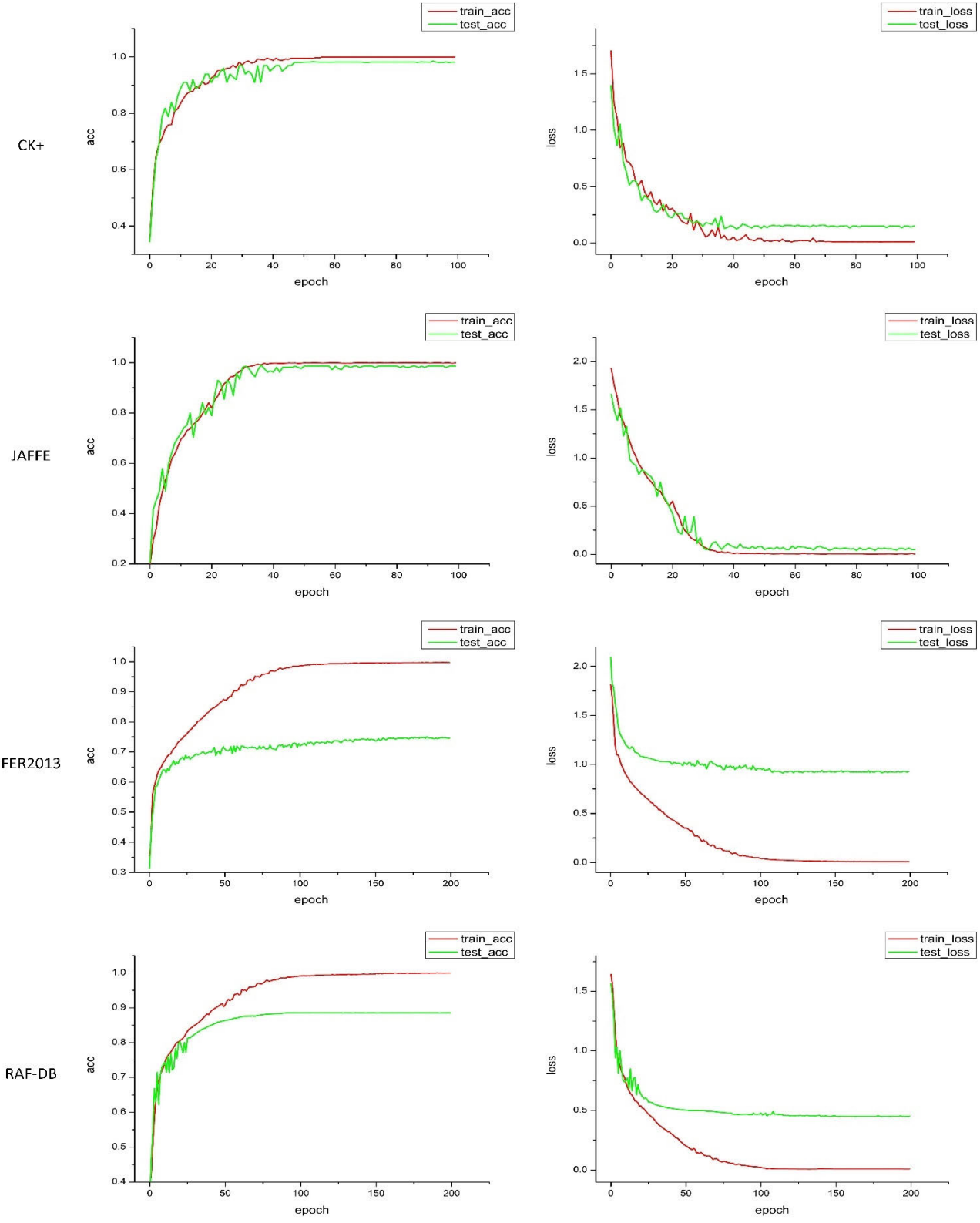


FIGURE 6. Accuracy and loss curves of MrCAR across four datasets.

The confusion matrix of the MrCAR model proposed in this paper on the four datasets CK+, JAFFE, FER2013 and RAF-DB is shown in Figure 7. Here, we choose FER2013 and RAF-DB datasets to verify Arcface Loss. Therefore, two

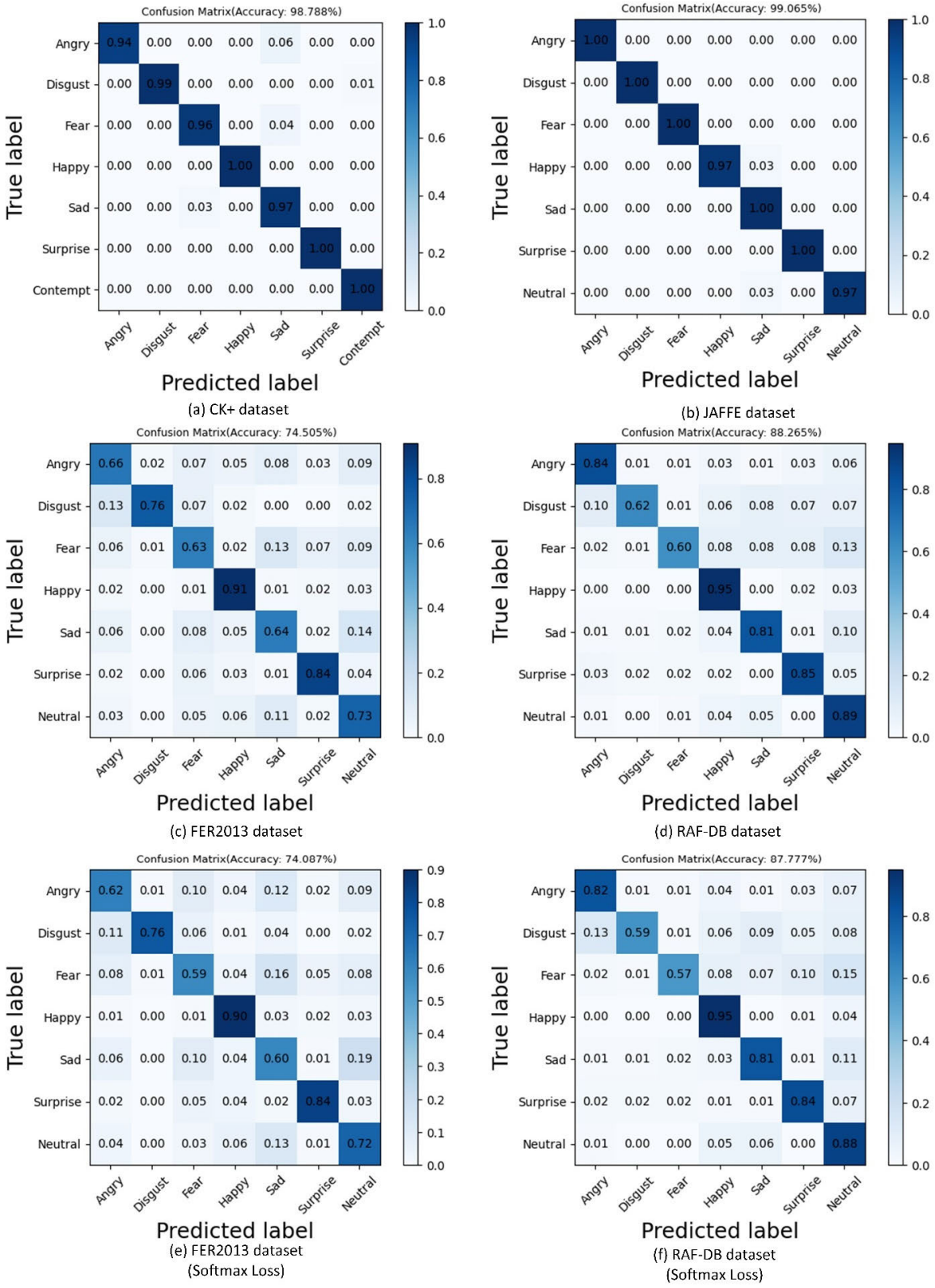


FIGURE 7. Confusion matrix on four datasets.

confusion matrices are obtained on each of the FER2013 dataset and the RAF-DB dataset: where (c) and (d) denote the confusion matrices obtained on the MrCAR model FER2013 dataset and the RAF-DB dataset, respectively, while

in (e) and (f) denote the confusion matrices obtained using Softmax Loss on the FER2013 dataset and the RAF-DB dataset, respectively. The overall accuracies of MrCAR model CK+, JAFFE, FER2013, and RAF-DB on the four datasets were 98.788%, 99.065%, 74.505%, and 88.265%, respectively. The confusion matrices of both CK+ and JAFFE were obtained by ten times repeated cross-validation. the accuracy of various expressions recognition on both CK+ and JAFFE datasets were high, like Happy and Surprise, which reached 100% accuracy. The overall accuracy of various expression recognition on the FER2013 dataset is relatively low because the dataset is sourced from a wide variety of images, there are many mislabeled samples of non-face expressions, there are many obscured, non-frontal faces, and there are many watermarks in the images. Both confusion matrices of the FER2013 dataset reveal that the accuracy of the two categories Happy and Surprise is higher, which is because these two categories of expressions have large amplitude of facial actions that are easy to recognize, and there are many of them in this dataset. In contrast, the accuracy of the model for recognizing the three categories of similar negative expressions, Angry, Fear, and Sad, is lower, which is due to the fact that these three expressions originally have many recognizable facial actions. However, by using the confusion matrix obtained from two different loss functions it is obvious that the Arcface Loss used in this paper's algorithm has a better classification effect than Softmax loss. Arcface Loss can enhance both intra-class closeness and inter-class variability, which improves the accuracy of the model in recognizing three classes of expressions, Angry, Fear and Sad, and at the same time reduces the cases of misclassification of these three classes of expressions by the model. The overall accuracy of various expression recognition on the RAF-DB dataset is higher than that on the FER2013 dataset because there is no artificial noise such as wrong labels and watermarks in the RAF-DB dataset, and only a small number of non-frontal and obscured images exist. However, the recognition rate for two similar negative expressions, disgust and fear, is lower. Similarly comparing the confusion matrix obtained on the RAF-DB dataset using two loss functions, Arcface Loss and softmax loss, it can also be found that using Arcface Loss can reduce the false recognition rate of the model for two similar expressions, disgust and fear.

C. COMPARISON AND ABLATION EXPERIMENT

1) COMPARISON EXPERIMENT

In order to prove the effectiveness of the facial expression recognition algorithm proposed in this paper, we compare with some advanced FER methods, and test on CK+, JAFFE, FER2013 and RAF-DB data sets respectively. The test results are shown in the following tables.

Table 2 shows the accuracy of the proposed algorithm and other advanced methods [14], [24], [25], [26], [27], [28] on CK+ dataset. The accuracy of the proposed algorithm reaches 98.78%, which is 1.09% higher than that of the method Umer [28] with the highest accuracy.

TABLE 2. Comparison results on the CK+ dataset.

methods	Accuracy (%)
DCMA-CNN [24]	93.46
MRAN [14]	96.28
Yang [25]	97.02
SCAN [26]	97.31
FER-IK [27]	97.59
Umer [28]	97.69
MrCAR	98.78

TABLE 3. Comparison results on the FER2013 dataset.

methods	Accuracy (%)
SHCNN [29]	69.10
ROLs [30]	72.50
Xie [31]	72.67
IcRL [32]	72.97
SNNs [33]	73.00
LHC-Net [34]	73.53
MrCAR	74.50

TABLE 4. Comparison results on the JAFFE dataset.

methods	Accuracy (%)
ELM [35]	83.12
Liu [36]	96.80
Liang [37]	98.48
Li [38]	98.52
Yuan [39]	98.53
MrCAR	99.06

TABLE 5. Comparison results on the RAF-DB dataset.

methods	Accuracy (%)
Ada-CM [40]	85.32
Fang [41]	85.54
LBAN-IL [42]	85.89
IDFL [43]	86.69
DACL [44]	87.78
MrCAR	88.26

Table 3 shows the accuracy of the proposed algorithm and other advanced methods [29], [30], [31], [32], [33], [34] on the FER2013 dataset. The average accuracy of the proposed algorithm is 74.50%, which is 1.5% and 0.97% higher than that of SNNs [33] and LHC-Net [34], respectively. Table 4 shows the accuracy of the proposed algorithm and other advanced methods [35], [36], [37], [38], [39] on the JAFFE dataset. The average accuracy of the proposed algorithm is 99.06%, which is 0.53 higher than that of Yuan [39]. Table 5 shows the accuracy of the proposed algorithm and other advanced methods [40], [41], [42], [43], [44] on RAF-DB dataset. The average accuracy of the proposed algorithm is 88.26%, which is 1.58% and 0.48% higher than that of IDFL [43] and DACL [44].

2) ABLATION EXPERIMENT

In order to verify the validity of the MrCAR model design in this paper, ablation experiments were conducted on CK+, FER2013, JAFFE and RAF-DB datasets. MrCAR represents the final algorithm model (multi-region feature

extraction method is adopted, including CA-Net and Arcface Loss); MrCAR-ALoss was used to represent the algorithm model with Arcface Loss removed (Softmax loss was used). MrCAR-CA is used to represent the algorithm model after removing the CA-Net attention mechanism. MrCAR-Mfe is used to represent the algorithm model without multi-region feature extraction. The ablation results of MrCAR model on the four datasets are shown in Table 6.

TABLE 6. Accuracy of ablation experiments on four datasets (%).

methods	CK+	FER2013	JAFFE	RAF-DB
MrCAR	98.78	74.50	99.02	88.26
MrCAR-Mfe	94.04	72.72	94.58	87.77
MrCAR-ALoss	97.73	74.08	98.01	86.73
MrCAR-CA	95.82	73.41	96.24	85.13

Table 6 shows the accuracy of the proposed algorithm in ablation experiments on four datasets. It can be seen from the table that the accuracy of the algorithm model without multi-region feature extraction, the algorithm model without Arcface Loss and the algorithm model without CA-Net attention mechanism in the four data sets is not higher than that of the MrCAR algorithm model. The ablation results verify the effectiveness of the proposed algorithm. At the same time, it can be found that whether the multi-region feature extraction method has the greatest impact on the algorithm model, CA-Net attention mechanism has the second greatest impact on the algorithm model, and Arcface Loss has the least impact on the algorithm model.

V. CONCLUSION

In this paper, a multi-region coordinate attentional residual facial expression recognition model (MrCAR) is designed. Firstly, the traditional global input image is cut to obtain multi-region input to eliminate complex background noise. Secondly, coordinate attention residual network was used to extract key features of different regions to improve the utilization rate of key features. Finally, Arcface Loss was used to replace the traditional Softmax Loss to improve the classification ability of the model, so as to reduce the misjudgment of negative expressions. The test accuracy rates of CK+, JAFFE, FER2013 and RAF-DB were 98.788%, 99.065%, 74.505% and 88.265%, respectively. The results show that MrCAR is more competitive in natural scenarios. In the following work, the real-time video facial expression recognition will be further studied.

REFERENCES

[1] C. Jiang, Z. Liu, M. Wu, J. She, and W. Cao, "Efficient facial expression recognition with representation reinforcement network and transfer self-training for human–machine interaction," *IEEE Trans. Ind. Informat.*, early access, Jan. 17, 2023, doi: 10.1109/TII.2022.3233650.

[2] R. Mao, R. X. Meng, and R. J. Sun, "Facial expression recognition based on deep convolutional neural network," in *Proc. 3rd Int. Conf. Intell. Comput. Hum.-Comput. Interact. (ICHCI), Jan.*, 2023, pp. 498–503.

[3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[4] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Comput. Sci.*, vol. 7, pp. 1–60, Apr. 2009.

[5] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[6] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 103–114, Jan. 2018.

[7] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, pp. 3046–3062, Apr. 2021.

[8] X. Sun, P. Xia, L. Zhang, and L. Shao, "A ROI-guided deep architecture for robust facial expressions recognition," *Inf. Sci.*, vol. 522, pp. 35–48, Jun. 2020.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[10] 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.

[11] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[12] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[13] M. Jiang and S. L. Yin, "Facial expression recognition based on convolutional block attention module and multi-feature fusion," *Int. J. Comput. Vis. Robot.*, vol. 13, no. 1, pp. 21–37, Oct. 2022.

[14] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020.

[15] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 558–565.

[16] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," 2016, *arXiv:1612.02295*.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6738–6746.

[18] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[21] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, Sep. 2013, pp. 117–124.

[22] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[23] 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.

[24] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211–220, Jan. 2019.

[25] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2018.

[26] D. Gera and S. Balasubramanian, "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition," *Pattern Recognit. Lett.*, vol. 145, pp. 58–66, May 2021.

[27] Z. J. Cui, T. F. Song, Y. R. Wang, and Q. Ji, "Knowledge augmented deep neural networks for joint facial expression and action unit recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 14338–14349.
[28] S. Umer, R. K. Rout, C. Pero, and M. Nappi, "Facial expression recognition with trade-offs between data augmentation and deep learning features," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 2, pp. 721–735, Feb. 2022.
[29] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78000–78011, 2019.
[30] X. Sun, S. Zheng, and H. Fu, "ROI-attention vectorized CNN model for static facial expression recognition," *IEEE Access*, vol. 8, pp. 7183–7194, 2020.
[31] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE Trans. Cybern.*, vol. 51, no. 5, pp. 2787–2800, May 2021.
[32] Y. Chen and H. Hu, "Facial expression recognition by inter-class relational learning," *IEEE Access*, vol. 7, pp. 94106–94117, 2019.
[33] W. Hayale, P. S. Negi, and M. Mahoor, "Deep Siamese neural networks for facial expression recognition in the wild," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1148–1158, May 2021.
[34] R. Pecoraro, V. Basile, and V. Bono, "Local multi-head channel self-attention for facial expression recognition," *Information*, vol. 13, no. 9, pp. 419–436, Aug. 2022.
[35] M. Wafi, F. A. Bachtiar, and F. Utaminingrum, "Feature extraction comparison for facial expression recognition using adaptive extreme learning machine," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, pp. 1113–1122, Feb. 2023.
[36] J. Liu, Y. Feng, and H. Wang, "Facial expression recognition using pose-guided face alignment and discriminative features based on deep learning," *IEEE Access*, vol. 9, pp. 69267–69277, 2021.
[37] H. G. Liang, Y. Bo, Y. X. Lei, Z. X. Yu, and L. H. Liu, "A CNN-improved and channel-weighted lightweight human facial expression recognition method," *J. Image Graph.*, vol. 27, no. 12, pp. 3491–3502, Sep. 2022.
[38] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020.
[39] D. R. Yuan, Y. Zhang, Y. J. Tang, B. Y. Li, and B. L. Xie, "Multiscale residual attention network and its facial expression recognition algorithm," *J. Chin. Comput. Syst.*, vol. 2022, pp. 1–8, Nov. 2022.
[40] 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.
[41] 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.
[42] H. Li, N. Wang, Y. Yu, X. Yang, and X. Gao, "LBAN-IL: A novel method of high discriminative representation for facial expression recognition," *Neurocomputing*, vol. 432, pp. 159–169, Apr. 2021.
[43] Y. Li, Y. Lu, B. Chen, Z. Zhang, J. Li, G. Lu, and D. Zhang, "Learning informative and discriminative features for facial expression recognition in the wild," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3178–3189, May 2022.
[44] 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.

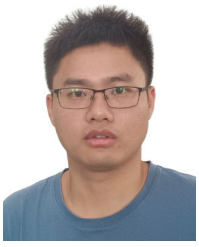
JIANGHAI LAN was born in Sichuan, China, in 1995. He received the B.S. degree from the Sichuan University of Science and Engineering, in 2021, where he is currently pursuing the degree with the School of Automation and Information Engineering. His current research interests include image processing and deep learning.

XINGGUO JIANG received the Ph.D. degree from the Institute of Optics and Electronics, Chinese Academy of Sciences, in 2007. He is currently an Associate Professor with the School of Automation and Information Engineering, Sichuan University of Science and Engineering. His current research interests include image processing, intelligent information processing, and deep learning.

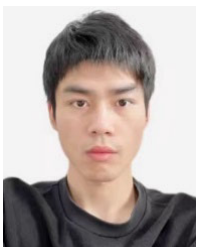
GUOJUN LIN received the degree from the Zhejiang University of Technology, in July 2001, the master's degree from Southwest Jiaotong University, in March 2008, and the degree from the University of Electronic Science and Technology of China, in December 2014. He was a Software Engineer with Shenzhen Tianpai Electronics Company Ltd., in September 2008. Since January 2015, he has been a Lecturer with the Sichuan University of Science and Engineering.

XU ZHOU was born in Sichuan, China, in 1998. He received the bachelor's degree from the Sichuan Institute of Technology, in 2021, where he is currently a graduate student with the School of Automation and Information Engineering. His current research interest includes occlusion face recognition.

SONG YOU was born in Sichuan, China, in 1999. He received the B.S. degree from the Sichuan University of Science and Engineering, in 2021, where he is currently pursuing the degree with the School of Automation and Information Engineering. His current research interest includes cartoon style migration.



ZHEN LIAO was born in Sichuan, China, in 1994. He received the B.S. degree from Jiamusi University, in 2017. He is currently pursuing the degree with the School of Automation and Information Engineering, Sichuan University of Science and Engineering. His current research interest includes face avatar generation.



YANG FAN received the B.S. degree from the Sichuan University of Science and Engineering, in 2021, where he is currently pursuing the degree with the School of Automation and Information Engineering. His current research interests include image processing and deep learning.