Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **6**, Number **3**, August **2021**

Facial Expression Recognition Based on Double-Channel Facial Images with Robust Occlusion

Hong Tang

College of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
No. 2, Chongwen Road, Chongqing, China
tangh@cqupt.edu.cn

Jun-Ling Xiang*

College of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
No. 2, Chongwen Road, Chongqing, China
Corresponding Author: 390098758@qq.com

Hong-Yu Wei

College of Computer Science and Technology
Chongqing University of Posts and Telecommunications
No. 2, Chongwen Road, Chongqing, China
1104930199@qq.com

Received May 2021; revised July 2021

Abstract. *Occlusion can change the facial appearance significantly and seriously affect facial feature extraction lead to reduce the accuracy of facial expression recognition. In order to obtain better facial expression recognition performance, a double-channel convolutional neural network model to occlusion is proposed. First, after the backbone network based on the residual network feature extraction, the model is designed to include the occlusion perception module and geometry-aware module. The occlusion perception module can perceive the occlusion area, and the geometry-aware module can deeply explore the relationship between facial components, the two modules have a synergistic effect to reduce the impact of occlusion and obtain clean features. Furthermore, since clean features may lack some facial information, a contrastive loss is utilized to learn a similarity metric for image pairs to make sure that the samples with the same expression have similar representations, and at the same time, those with different expressions are far away in the feature space. Finally, a large number of experiments have been carried on two public facial expression datasets (CK+, RAF-DB) for facial occlusion synthesis and the field dataset (SFEW). The results show that the proposed method is significantly better than the traditional convolutional neural network model.*
Keywords: face expression recognition; occlusion robust; double-channel; measure expression; residual network;

1. **Introduction.** In recent years, with the rapid development of information technology and the wide application of computers, facial expression recognition in images has attracted more and more attention [1, 2]. Facial expressions, as a non-verbal means to effectively convey emotional states and intentions in human communication, are also of great value in scientific research. Typical applications include driving fatigue detection [3],

emotion analysis, medical diagnosis [4], public safety, and other fields [5]. This technology has been widely studied by scholars due to its extremely high use value and research significance. For real-world applications of facial expression recognition, how to deal with the effects of uncontrolled lighting, different postures and occlusion are crucial.

Face expression recognition (FER) is to design a model that can automatically and accurately identify the categories of facial expressions contained in a face image or face video sequences. Most of the early studies on facial expression recognition are data sets collected in the experimental environment. The samples collected are all standard positive faces, such as CK+ [6], MMI [7], Jaffe [8] datasets. But to use the technology in real-world Settings, researchers collected large data sets of unconstrained faces in natural environments, such as RAF-DB [9], AffectNet [10], SFEW [11] datasets. A large number of facial expression images provide data support for the algorithm, which greatly promotes the development of facial expression analysis. However, in real scenes, facial occlusion and non-frontal head posture are the two main problems, making facial expression recognition still challenging, because these problems can lead to significant changes in facial appearance, and the occlusion objects and occlusion locations are uncertain. Occlusion can interfere with the extraction of facial expression features and affect the accuracy of facial expression recognition. Researchers believe that a truly robust recognition method should be able to solve the problem of facial expression recognition under occlusion.

Early studies on the occlusion of facial expressions mainly conducted partial occlusion processing on datasets in the laboratory environment and used traditional machine learning algorithms to conduct research. For example, blocking blocks are added to the key parts of the face in CK+ and JAFF datasets to study which areas are the most important for human perception of facial expressions [12]. Based on the analysis of Gabor features, Kotsia et al. found that, compared with other parts, shielding the mouth and eyes has a greater impact on facial expression recognition. Batista et al. [13] proposed a new classification method to accurately recognize the facial expression categories under obscuration, namely sparse representation classifier (SRC). The face image was divided into equal-sized regions, and then the SRC was used to classify the facial expressions in each region. Happy and Routray [14] extracted several significant facial areas by using face marker points, and used the appearance features of the selected facial areas for expression recognition. Wright et al. [15] reconstructed the occluded area of the face by applying robust principal component analysis and used the reconstructed face to carry out expression recognition. With the collection of large-scale data, the use of the convolutional neural network to improve facial expression recognition is gradually increasing. Batista et al. designed a multi-task convolutional network based on facial region to solve the self-occlusion problem caused by the change of head posture. Subsequently, local and global methods have been widely used in facial expression recognition.

In view of the problem that some areas of the face are blocked, resulting in the lack of key expression features, thus affecting the accuracy of facial expression recognition, facial expression can be judged according to the symmetrical area of the face or other highly related facial areas. It is impractical to explicitly remove occlusion because occlusion in a real scene is difficult to detect. We can first determine the occlusion area, filter the occlusion features, minimize the influence of occlusion, and give higher weight to the un-occlusion area of the face, so as to improve the accuracy of the facial expression recognition task. Based on the above analysis, a twin convolutional neural network model with perceptual occlusion and reduced occlusion effect is proposed in this paper. The model mainly includes the geometric relation module and occlusion perception module. Humans recognize facial expressions based on certain areas of the face. The geometric relationship module can mine geometric features of the face (such as symmetry, proximity,

and positional relationship), and recognize facial expressions by using the relationship between five facial components when the face is incomplete. The occlusion sensing module can automatically perceive the occluded face area, filter the occlusion features, output a clean weight probability map, and combine with the geometric feature module to get the face feature map without the influence of occlusion and concentrate the weight in the barrier-free area of the face image. The main framework of this paper adopts the Siamese Network [16] to map the expressions into the feature space, shorten the intra-class distance between the same expressions and extend the inter-class distance between different expressions, to reduce the intra-expression difference and increase the inter-expression difference. Even in the case of incomplete facial organs, the accuracy of expression classification can be improved according to the remaining features retained. Because there is no mature and open occlusion facial expression data set at present, this paper does occlusion synthesis processing on the open datasets.

The work of this paper is as follows:

(1) A new occlusion adaptive deep double-channel network is proposed to solve the problem that partial occlusion affects the accuracy of facial expression recognition. The double-channel model is applied to the image classification method to assist facial expression recognition in the absence of features.

(2) An occlusion perception module is proposed, which can automatically perceive the occlusion area of the face by using only a small amount of convolutional layers to provide prior knowledge for subsequent filtering of occlusion features;

(3) A novel face geometric feature extraction module is proposed. The combination of occlusion perception module and face geometric feature module can obtain clean face feature representation, which greatly eliminates the influence of occlusion.

2. **Related work.**

2.1. **Network structure.**

Deep learning is a sub-field of machine learning, which uses a hierarchical architecture to learn high-level abstract features in data. A convolutional neural network (CNN) is a deep learning method. The early CNN was introduced by LeCun in the 1980s, and the basic neural network architecture was convolution and subsampling. With the continuous development of neural networks, deeper CNN architectures are being developed to handle larger image sizes and more complex pattern recognition problems. In 2012, Krizhevsky et al. [17] achieved great success in the ImageNet classification challenge competition. Since then, deep learning methods have been widely applied to computer vision problems such as facial expression recognition [18], face detection, biomedical image analysis, image classification, age estimation, target detection, face component segmentation, etc. And the performance of these applications can be greatly improved.

The CNN structure consists of a convolutional layer, an activation layer, a pooling layer, and a fully connected layer. The convolution layer, which convolves their inputs with a set of learned filters and generates feature maps, is usually followed by a non-linear activation layer. The pooling layer reduces the space size of their input, reduces parameters and computation, and controls overfitting. The full connection layer connects all the activation layers of the upper layer and performs classification tasks in the network. The deep model is designed as an end-to-end learning approach, leveraging powerful hierarchies and thousands or millions of network parameters to learn features and classification capabilities, such as 16 and 19 layers for VGGNets and 22 layers for GoogleNets. Nowadays, the research of neural networks is deeper and deeper, with the deepening of network depth, the performance of the model will increase to a certain extent, but He et al. [19] found in

the experiment, the network layer increases to a certain extent, accuracy of the network will be saturated, can appear even gradient disappeared and gradient explosion problem, leading to more training. Therefore, He et al. proposed the residuals convolutional neural network. Even though the maximum number of network layers is as high as 152 layers, jump structures are added to the convolutional network to realize identity mapping and improve accuracy. For a heap base structure, when the input is x, the learned feature is denoted as $H(x)$, and the residuals $F(x)$ actually learned is:

$$F(x) = H(x) - x \tag{1}$$

The original learning features is $F(x) + x$, since the actual residual will not be 0, it is easier to learn the residual than the original feature, so it has better performance.

2.2. **Facial expression recognition.** Facial expression recognition has always been a challenging problem in the field of sentiment analysis. In essence, it is a multi-classification problem, that is, facial image or video frame is divided into independent expression categories. In recent years, many studies have focused on facial expression recognition from static images. The existing methods can be divided into shallow methods and deep struct-based methods. Facial expression recognition system mainly includes three stages: preprocessing, feature extraction, and expression recognition. Preprocessing can help to extract high-quality features, including face detection, face alignment, image normalization, and other steps. In face detection, some face detectors, such as MTCNN [20] and Dlib [21], are used to locate faces in complex scenes, and face expression images without background influence can be obtained by clipping the located faces.

Facial expression feature extraction is the most important part of the facial expression recognition system. In recent years, researchers have designed a variety of methods to capture facial geometric features and appearance features caused by facial expressions, which are mainly divided into traditional manual feature extraction methods and deep learning-based feature extraction methods. For traditional manual methods, typical features are usually extracted by geometric relations between facial organs or landmark points, mainly including extraction based on local texture features, such as SIFT, HOG, LBP. Shan and Gong et al. [22] improved the performance of facial expression recognition by enhancing LBP features. To improve the comprehensive representation ability, Majumder et al. [23], fused the features of different types before classification. Although the traditional machine learning algorithm improves the accuracy of recognition to a certain extent, these features only have a good performance in a specific environment, with poor generalization ability and low robustness. The experimental results show that the convolutional neural network has better facial expression recognition performance and generalization ability than the traditional method. Tang et al. used a deep convolutional neural network to win the FER2013 challenge. Liu et al. [24] proposed a CNN architecture based on facial action units for expression recognition. Xie et al. [25] proposed to solve the problem of facial expression recognition by using deep comprehensive multi-chip aggregation convolutional neural network. Finally, the purpose of expression classification is to judge the similarity between the features of the test image and a certain type of expression features in the training set, and select the type with the greatest similarity as the output result. Support vector machines and classification based on sparse representation are two popular traditional machine learning methods, or classification using Softmax layer at the last layer of neural network.

In the real environment, the facial area is easily blocked by objects such as scarves, hats, and glasses, which seriously affects the accuracy of expression recognition. Compared with the experimental environment, there are few studies on facial expression recognition in the

real environment. Since facial expressions are mainly concentrated in five facial regions, an attention mechanism is widely applied in facial expression recognition to better locate the parts related to facial expressions. Sun et al. [26] combined shallow features with deep features and added an attention mechanism, proposed a deep integration model of attention, and verified the effectiveness of added attention mechanism on multiple data sets. To improve the expression recognition rate under the natural environment data set, Literature [27] proposed the loss of deep attention center by improving the loss function and integrating the attention mechanism, to realize the adaptive selection of important feature element subset. Li et al. [28] and Wang et al. [29] also adopted the method of adding attention mechanism to deal with the occlusion problem, adaptively adjusting the importance of various parts of the face and focusing more on the unoccluded facial expression area, to extract effective features for expression recognition. Although this method pays higher attention to more important features, damaged features not only affect the current position, but may contaminate the remaining unshaded parts as features are extracted. Lu et al. [30] repaired the damaged face images by GAN and then extracted the features. However, the restoration of face images requires a priori knowledge of occlusion, such as the precise location of the occlusion. Therefore, the method to repair the occlusion position does not applicable to facial expression recognition with arbitrary occlusion.

3. **Double-channel network model with robust occlusion.**

3.1. **Framework outlined.** A double-channel convolutional neural network model with robust occlusion is proposed for partial occlusion facial expression recognition. To solve the occlusion problem effectively, the last residual unit of ResNet-50 is modified into the occlusion adaptive module proposed in this paper. The overall framework is shown in Figure 1. Many existing deep learning methods usually only use single-channel CNN, which is inspired by [31], different from the traditional convolutional network structure, we propose two images are input into the network at the same time, and two networks are established for training. The backbone networks used by the two networks are the same, and the weight sharing is realized. The specific structure is shown in Figure 2. The network module is composed of the geometry-aware module and occlusion perception module, and it mainly has three branches. Firstly, the feature map from previous residual learning blocks is fed into the geometry-aware module and occlusion perception module, and the geometric relation information and occlusion probability matrix of the face are obtained respectively. Then, the output of the two modules is combined into the full connection layer, and the final expression classification result can be obtained by the Softmax classifier. The network optimizes the whole model jointly through cross-entropy loss and contrastive loss of the two networks.

3.2. **Geometry-aware module.** Element addition and element multiplication are common methods for aggregating multiple output features. In the residual network, the connection of the residual blocks adopts the element addition method, and the element multiplication rule is used to estimate the multiple expressions of the feature mapping. Both of these two methods are convolutional local calculation, ignoring the spatial relationship of features, and they are in no order. The matrix cross product calculation proposed in this module is a non-local operation, which can obtain the response at a certain position by calculating the weighted sum of the row and column features in the input feature map. Wang et al. [32] indicated that the traditional convolution operation can only simulate the relationship between local neighborhoods. Although the dependency relationship of each position of an image can be obtained through repeated convolution,

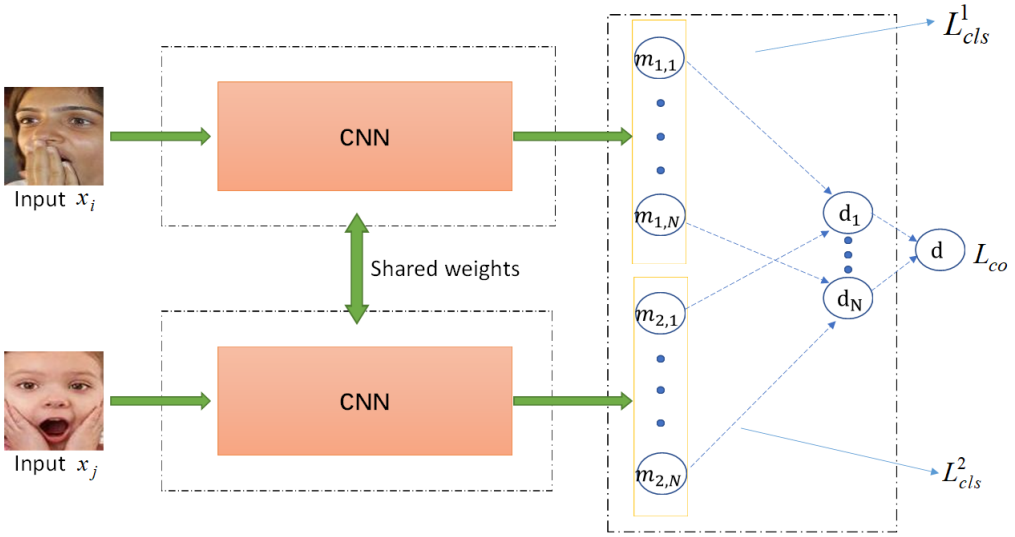

FIGURE 1. Overall model framework

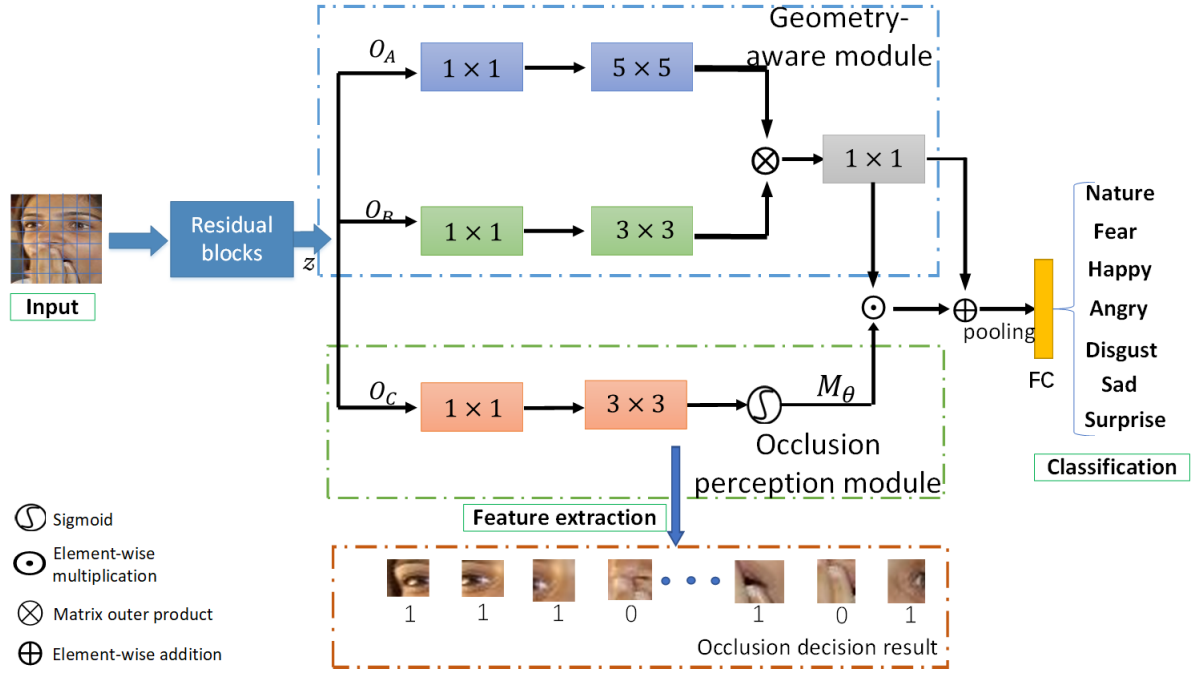

FIGURE 2. Facial expression recognition network structure based on occlusion perception

the computational efficiency is low. However, due to the highly dependent relationship between the different facial features, when the face is partially obscured, an expression can be judged based on the asymmetrical part of the face or other highly related facial areas. Zhu et al. [33] proposed to use the cross-product of two matrices to obtain the correlation between feature channels and to mine the features of facial geometric relations. Inspired by their work, we propose a geometrical relation module, which uses the cross-product of the matrix to capture the geometrical relation between five senses.

As shown in Figure 2, the geometry-aware module consists of two branches, which are O_A and O_B. In order to obtain multi-scale features, O_A and O_B select different sizes of convolutional layer filter sizes, where the convolutional layer filter sizes are $1 * 1$, $5 * 5$

and $1 * 1$, $3 * 3$ respectively. These two branches will go through the BN layer and ReLU nonlinear activation function after each processing of a convolution layer, in which the 1*1 convolution layer can increase the nonlinearity of the decision function without affecting the receptive field of the convolution layer. In Figure 2, the output features of the two branches are multiplied, and geometric feature mapping is formed through the matrix cross product of the corresponding channel, so as to realize the coding of geometric relations between different facial components. Finally, the geometric feature mapping is input into the $1 * 1$ convolution layer to obtain the final geometric representation, which can be expressed as:

$$f(x_i) = f_A^T f_B \tag{2}$$

f_A^T represents the output characteristics of the branch O_A, and O_B represents the output characteristics of the branch O_B. The Figure 3 is a schematic diagram of the matrix cross product.

FIGURE 3. Matrix outer product

3.3. **occlusion perception module.** In the training stage of the model, the occluded face area is easy to interfere with feature extraction, and the errors generated will pollute the unoccluded area, resulting in the failure of convergence. In order to reduce the sensitivity to occlusion, the occlusion perception module proposed can adaptively identify the occlusion region in the image, and then filter the features of the occlusion region. All the images input to the network are preprocessed to detect the coordinates of face key points and realize face alignment. All the images input to the network are preprocessed to detect the coordinates of face key points and realize face alignment. As shown in Figure 3, the aligned face image is divided into $N * N$ non-overlapping regions, represented as $\{b_i\}_{i=1}^{N*N}$, with the purpose of detecting the probability of each region being occluded. In this paper N is set to 7 according to the size of the input image, so that the key points of the facial features can have appropriate corresponding relationship with the location of each region.

The O_C branch contains 2 convolution layers, 2 pooling layers and a logistic regression function. The same size as the convolution layer filter of the O_B branch is $1 * 1$ and $3 * 3$, when the input image is I, the occlusion perception module outputs the occlusion probability matrix M_θ , which can be expressed as:

$$M_\theta = \sigma(g(\Omega : I)) \tag{3}$$

Where $g(.)$ is the convolution operation, Ω represents all parameters after the convolution operation, and σ represents the sigmoid function.

The occlusion judgment will be realized by setting a threshold value. If the occlusion probability of the b_i region is greater than the threshold value α, it will be identified as occlusion, and the occlusion function m_i will be determined to be the specific element value of the M_θ matrix. The ideal result can be expressed as a classification function:

$$m_i = \begin{cases} 1 & if\ occ < \alpha \\ 0 & else \end{cases} \tag{4}$$

Where occ represents the occlusion ratio of the region, 1 represents the occlusion ratio is less than α, and M_θ represents the occlusion ratio is greater than the threshold. In the optimization process, the occlusion probability matrix M_θ is sparse by L_1 regularization.

The occlusion probability matrix M_θ is integrated into the output features of the geometric-aware module through element multiplication, finally, a clean feature representation of the whole face can be obtained (the weighted feature mapping of the output features). It can be expressed as:

$$\tilde{f}(x_j) = M_\theta f(x_j) \tag{5}$$

Since the result of the multiplication of the two modules will weaken the output of the geometric relation module, the output of the geometric relation module and the result of the multiplication of the two modules can be expressed as:

$$F(x_j) = \tilde{f}(x_j) + f(x_j) = M_\theta f(x_j) + f(x_j) \tag{6}$$

3.4. **Emotional similarity measurement.** Mehdipour et al. [34] indicated that if a large number of occlusion face images were not specially trained, occlusion would lead to larger in-class differences in feature space and higher similarity between classes, which would seriously affect the work of deep convolutional neural network model. Inspired by this, we construct an expression recognition framework composed of two identical CNNs, as shown in Figure 1. Two cross-entropy losses and one contrastive loss are used to learn the expression of facial expression recognition. Using the given expression classification tag, the cross-entropy loss function is used at the end of each network to calculate the classification error and is used to fine-tune the parameters of the lower layer. In addition, the contrastive loss is used to learn the similarity measurement of image pairs to ensure that the samples with the same classification have similar representations, and the samples with different representations are far away in the feature space. Contrastive loss between double-channel is a powerful tool to measure the similarity between images, but it needs some changes to be used as an efficient classifier. In this framework, training datasets need to be processed. Firstly, a reference image is selected as the benchmark in each expression category (face image without no occlusion), and then the dataset is divided into two pairs of positive sample pairs, in which the positive sample pair includes the benchmark image and other samples of the same category, while the negative sample pair includes the benchmark image and other samples of different categories. At this point, the training set can be expressed as $\{[x_i, y_i]\}_{i=1}^{n}$, $y_i \in \{1, 2, ..., P\}$ where P is the number of categories of facial expressions and represents the corresponding labels. Positive and negative sample pairs are randomly selected from the training set, and the label of the sample pair can be obtained by the following formula:

$$l([x_i, x_j]) = \begin{cases} 1, y_i = y_j \\ 0, y_i \neq y_j \end{cases} \tag{7}$$

where l represents the sample pair label.

$m_{1,i}$ and $m_{2,i}$ are the one-dimensional eigenvectors of the double-channel. By calculating the Euclidean distance between $m_{1,i}$ and $m_{2,i}$, the similarity of the two branches can be measured:

$$d = \|m_{1,i} - m_{2,i}\|_2^2 \tag{8}$$

Then, the similarity between the two eigenvectors is calculated by sigmoid:

$$s = \sigma\left(\sum_j \alpha_j d^{(j)}\right) \tag{9}$$

In the above formula, α_j measures the importance of each d is represented as $d^{(j)}$, and the output is between 0 and 1. As shown in Figure 4, in the feature space, the distance between positive sample pairs is represented by d^2 to shorten the feature distance between the same classes, while the distance between negative sample pairs is represented by d^1 to push the feature distance between different classes.

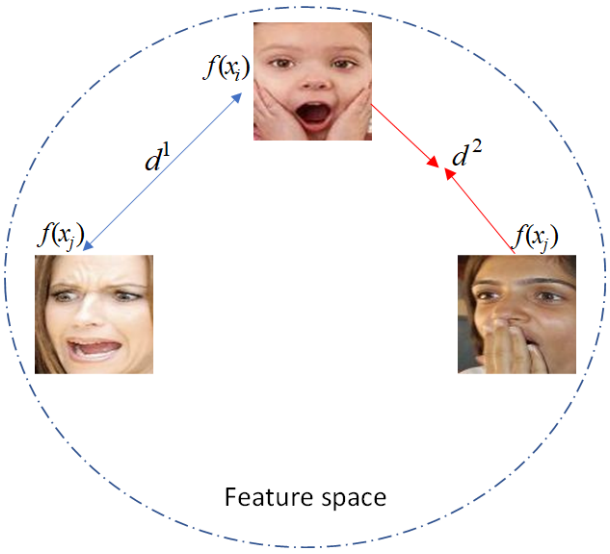

FIGURE 4. The distance between different expressions in the feature space

The contrastive loss between two channels can be expressed as:

$$\begin{aligned} L_{co} &= \sum_{i=1}^{N} L(y_i, f(x_i), f(x_j)) \\ &= \tfrac{y_i}{2} \cdot D(f(x_i), f(x_j)) + \tfrac{1-y_i}{2} \cdot \max(0, \gamma - D(f(x_i), f(x_j))) \end{aligned} \tag{10}$$

where N represents the total number of input samples, y_i represents whether the input samples are of the same category, positive samples are compared to $y_i = 1$, negative samples are compared to $y_i = 0$, where γ is a threshold value set. Only Euclidian distance between $0 \sim \gamma$ is considered. When the distance exceeds γ, the loss is regarded as 0.

At the end of the network, the softmax classifier is connected after the full connection layer. It maps the output of the previous layer to the expression class and normalizes the probability to 1, making it easier for data processing. The formula is as follows:

$$y_j = \frac{e^{z_j}}{\sum_i e^{z_i}} \tag{11}$$

Each element of Equation (11) corresponds to a unique expression class. The class with the largest element value is regarded as the prediction class, and the whole network is optimized by cross-entropy loss. The calculation formula is as follows:

$$L_{cls}(y_i, y_j) = -\sum_{i=1}^{n} y_i \log y_j \tag{12}$$

where y_i is the prediction label, y_j is the true label, and n represents the total number of categories.

The total loss. The overall loss function of the model proposed in this paper can be expressed as:

$$L = \lambda_1 L_{co} + \lambda_2 L_{cls}^1 + \lambda_3 L_{cls}^2 \tag{13}$$

λ_1, λ_2 and λ_3 are the weights of expression similarity loss and expression classification loss respectively. All relevant parameters of the method are updated by the stochastic gradient descent method (SGD) through back propagation. In the test, only one branch of the network was required to give the results.

Table 1. Detailed setup of experimental dataset

Database	CK+		RAF-DB		SFEW	
Attribute	training	testing	training	testing	training	testing
Happy	75	9	5136	1306	198	73
Sad	225	24	2471	506	172	73
Surprise	48	6	1301	308	96	57
Fear	189	21	297	69	98	47
Disgust	66	9	754	152	66	23
Angry	123	12	741	164	178	77
Nature	0	0	2607	677	150	86
Contempt	159	8	0	0	0	0
Total	885	89	13307	3182	958	436

4. **Experimental process.**

4.1. **Datasets.** Experiments were conducted on three public datasets, namely CK+, RAF-DB and SFEW datasets, to demonstrate the effectiveness of the proposed method. CK+ is the dataset in the experimental environment, RAF-DB and SFEW are the datasets in the real scene.

(1) The CK+ dataset consists of 593 video sequences taken by 123 people, in which face images are controlled by the lab. The video sequences lasted from about 10 to 60 frames, and 327 of the 593 sequences were labeled with emotions, and each was classified into one of seven emotional categories: anger, contempt, disgust, fear, happy, sadness, and surprise. Each image sequence changes from neutral at the beginning to a peak. We mainly choose the last three frames of each sequence are selected to compose the training and test sets.

(2) RAF-DB dataset of 29,672 diverse face images downloaded from the Internet. Different from the collection in CK+ experimental environment, RAF-DB is a real-world emotional dataset. Each image has about 40 independent labels, which combines facial expression features such as race, gender, head posture, age, and illumination. Firstly, the dataset was manually labeled and then screened, and the samples were set as 7 basic affective labels and 11 compound effective labels. We only used images with basic emotions were used, and 16,489 images with expression classification labels were selected for the experiment, of which 13,307 were taken as training samples and 3,182 were taken as test samples.

(3) SFEW is a field static facial expression dataset consisting of 1,766 images, including 958 for training, 436 for validation, and 327 for testing. There are seven expression categories in the dataset: anger, disgust, fear, neutral, happy, sad, and surprised. We selected 958 images as training samples, and 436 images were verified. Table 1 shows the details of the dataset selected for this article.

(4) Facial occlusion dataset. For unoccluded facial expression recognition, there are mature and open experimental datasets, but almost no mature and standard datasets for researchers to conduct facial expression recognition research on facial occlusion. Therefore, we decided to use images from CK+ and RAF-DB datasets as prototypes to simulate

occlusion. In real life, the face is easy to be mask, scarf, sunglasses and so on articles for daily use, this paper collected about 300 cartoon images from the Internet, such as glass, glasses, masks, hair, mobile phone, etc., used to generate the shade facial occluder device, through at different positions of the facial expression image randomly add cartoon image to simulate the face cover. The changes of the data set after simulated occlusion are shown in Figure 5. The case that the occlusion exceeds 50% of the face area is not considered, because the facial features are basically covered, and the occlusion filled is of little significance. One half of the original dataset is selected to synthesizing the occluded image, and then the other half of the unoccluded image is mixed and combined to form the processed dataset. The processed datasets are OCC-CK+ and OCC-RAF-DB datasets. Since SFEW is collected in a real scene, it is more challenging, so occlusion is no longer done.

FIGURE 5. The distance between different expressions in the feature space

4.2. **Preprocessing.** An image contains not only the face, but also the background, as well as the attitude and other interference factors. At the same time, to ensure the consistency of the face size and position, the image in the dataset needs to be preprocessed first. The main steps include face detection, face alignment and image size normalization. In this paper, MTCNN algorithm is used for face detection and face alignment. After the standard face image is obtained, the image is cut and scaled to the size.

It may not be enough to use a limited number of images in the facial expression dataset to learn the depth model. To reduce the possibility of overfitting, data enhancement is used to train the convolution model. The images in the data set are randomly cut into size, and the cut images are randomly mirrored before being sent into training. In the test stage, we cut the corners and the center of the image, and then do the mirroring operation, which can expand the database by 10 times. Finally, the probability value obtained is averaged, and the maximum output classification is the corresponding expression. This method can effectively reduce the classification error of expression.

4.3. **Implementation details.** The model proposed in this paper is based on the PyTorch framework and runs on Windows10 operating system, and NVIDIA Quadro RTX 6000 GPU is used for experiments. In the experiment, the SGD method was used for the overall optimization of the model. The initial learning rate was set as 0.01, the momentum was set as 0.9, and the weight attenuation was set as 0.0005. During the training phase, the batch size is set to 64. For CK+ dataset, the learning rate is multiplied by 0.1 after each iteration of 500 times, and the total number of training is set to 1000. For RAF-DB and SFEW datasets, the learning rate is multiplied by 0.1 after each iteration of 5000 times, and the total number of iterations is set to 100,000. In the combined training of contrastive loss and facial expression classification loss, λ_1 is set to 2, λ_2 and λ_3 is set to 1 according to experience [35].

4.4. **Experimental Results.**

4.4.1. *Occlusion threshold analysis.* Different values are set for verification on the three datasets, among which CK+ and RAF-DB are the processed occlusion datasets, and the values set are 0.3, 0.4, 0.5, 0.6, 0.7 respectively. The results are shown in the Figure 6. In the three datasets, 0.5 is better than other values, so the α value is set to 0.5.

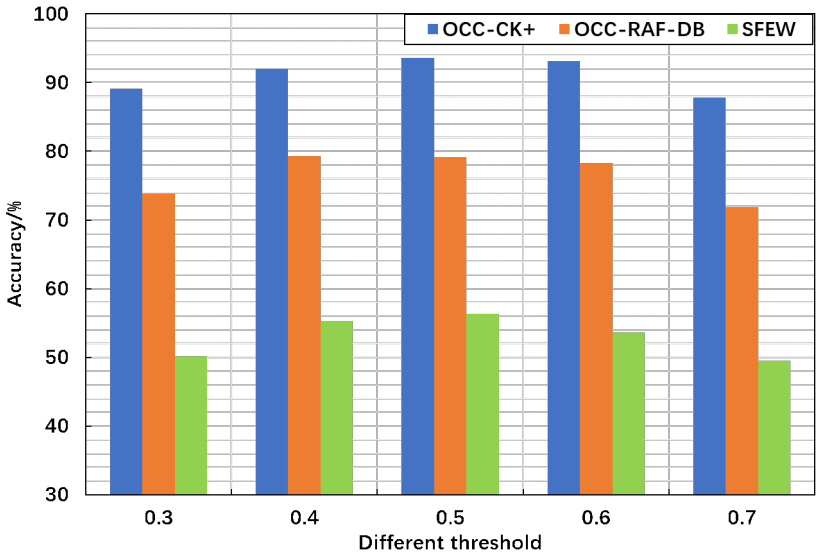

FIGURE 6. Evaluations of α weight on different datasets

4.4.2. *Based on CK+ dataset experiments.* The CK+ dataset is further divided into eight subsets, where the categories in any two subsets are mutually exclusive. Eight cross-validation strategies were then adopted, in each run, data from 6 subsets were used for training, and data from the remaining two subsets were used for validation and testing respectively. The final identification result is the average of eight verifications. The average recognition rate of this model in the OCC-CK+ dataset reaches 93.21%. Table 2 shows the results of confusion matrix, which describes in detail the recognition accuracy of each expression and the proportion of being misclassified as other expressions, where the diagonal term represents the recognition accuracy of each expression. As can be seen, except for disgust and contempt, the recognition rate of all other expressions is above 90 percent, but disgust and contempt are relatively low. The reason is that the range of happy, angry and other emotional expressions is relatively large, which is easier to identify. Even in the case of partial facial shielding, there is still a good recognition accuracy. Fear and surprise, both of which have similar expressions, such as the mouth and eyes opening wide. Table 3 shows the comparative evaluation results of CK+ dataset experiments. The experimental results of the proposed model are compared with the widely used basic classification models, AlexNet, VGG-16 and ResNet-50, respectively, and the results before and after processing occlusion on CK+ dataset are shown. The experimental environment and parameter design of the basic classification model are all the same as this model. It can be seen that AlexNet, VGG-16 and ResNet-50 have a good classification effect on the unoccluded CK+ dataset, reaching more than 92%. However, for the OCC-CK+ dataset, the overall recognition rate decreases by 4%∼5%. Compared with these classification models, the recognition rate before and after occlusion decreases by 4.15%. At the same time, Table 3 compares the recognition results between the single-channel model and the double-channel model. It can be found that the single-channel

Table 2. Facial expression recognition confusion matrix based on OCC-CK+ database

Expression	Expression						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Contempt
Angry	0.92	0	0.02	0	0.05	0.01	0
Disgust	0	0.81	0.13	0	0.01	0.02	0.03
Fear	0	0	0.95	0.02	0	0.01	0
Happy	0	0	0.01	0.99	0	0	0
Sad	0.03	0.01	0.02	0	0.93	0	0.01
Surprise	0	0	0.01	0	0.01	0.97	0.01
Contempt	0.04	0	0	0.01	0.03	0.06	0.86

Table 3. Comparison results of different methods based on CK+ database

Method	CK+	OCC-CK+
AlexNet [36]	94.40%	90.22%
VGG-16 [37]	92.05%	87.59%
ResNet50 [19]	92.12%	88.16%
Single-model	96.58%	92.69%
Double-model	97.36%	93.21%

Table 4. Facial expression recognition confusion matrix based on OCC-RAF-DB database

Expression	Expression						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Nature
Angry	0.81	0.03	0.02	0.04	0.06	0.03	0.02
Disgust	0.06	0.59	0.01	0.07	0.10	0.14	0.03
Fear	0.02	0.05	0.52	0.03	0.04	0.12	0.23
Happy	0.01	0.02	0.01	0.89	0.04	0.02	0.01
Sad	0.01	0.02	0.00	0.04	0.83	0.07	0.03
Surprise	0.01	0.01	0.02	0.03	0.04	0.85	0.03
Nature	0.03	0.04	0.01	0.03	0.03	0.02	0.84

model is obviously a little lower than the double-channel model in accuracy. In addition, the recognition rate of this model on OCC-CK+ is higher than that of other classification models on CK+ datasets, which fully proves the robustness of the double-channel model proposed in this paper to occlusion.

4.4.3. *Based on RAF-DB dataset experiments.* In order to study the classification performance of each expression category on the OCC-RAF-DB dataset, the confusion matrix of this model is shown in Table 4. The average recognition result of this model on the OCC-RAF-DB dataset reaches 78.85%, which obviously achieves the highest and lowest classification accuracy in the categories of happy and fear respectively. In addition to fear and disgust, which had the lowest recognition rate, the recognition rate of the other five categories of facial expressions reached more than 80 percent. This is because these facial expressions have well-defined regional features, such as "upturning of the corners of the mouth" and "elevating of the cheeks." Confusing categories of expressions include fear versus natural, disgust versus anger. Table 5 shows the experimental comparative evaluation results in the RAF-DB dataset. The recognition rates of the two modes of

Table 5. Comparison results of different methods based on RAF-DB database

Method	RAF-DB+	OCC-RAF-DB
AlexNet [36]	77.48%	73.71%
VGG-16 [37]	80.96%	75.26%
ResNet50 [19]	82.83%	76.40%
gACNN [28]	85.07%	80.54%
OADN [38]	87.16%	-
PG-CNN [39]	83.27%	-
ATL [40]	84.50%	-
Single-Channel	83.32%	78.12%
Double-Channel	84.47%	78.85%

Table 6. Facial expression recognition confusion matrix based on SFEW database

Expression	Expression						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Nature
Angry	0.7	0.04	0.07	0.02	0.03	0.08	0.04
Disgust	0.01	0.54	0.05	0.06	0.04	0.09	0.21
Fear	0.2	0.05	0.43	0.02	0.07	0.22	0.01
Happy	0.01	0.03	0.02	0.71	0.03	0.15	0.05
Sad	0	0.03	0.2	0.06	0.46	0.11	0.14
Surprise	0.05	0.08	0.17	0.15	0.01	0.51	0.03
Nature	0.04	0.13	0.01	0.05	0.08	0.1	0.59

this model, the single-channel model and the double-channel model, reached 83.32% and 84.47% in the dataset without occlusion processing, which better than most of the models compared, except gACNN and OADN. This is because OADN explicitly uses the face key point information to suppress the noise information of the occluded region, while gACNN uses the attention mechanism to focus on the more discriminative unoccluded region and combines the local information with the global information, and its recognition accuracy is better than the current extensive model. For OCC-RAF-DB dataset after occlusion processing, the recognition rates of all models are significantly reduced, with an average decrease of 5%∼6%. Both the original dataset and the occluded dataset, the performance of double-channel model is better than that of single-channel model.

4.4.4. *Based on SFEW dataset experiments.* SFEW dataset is created by the field static facial expressions, which contains a large number of occluded and multi-pose facial expressions images, so the facial expression recognition results are generally low, and the occlusion processing is not done in this paper. The average recognition result of this model on SFEW datasets is 56.83%. Table 6 is the confusion matrix of expression recognition in this model, the expression with the highest recognition result is happy, followed by angry. Happy has a high recognition rate in the results of these datasets. The recognition rate of the other expressions was significantly lower, among which the recognition rate of fear was the lowest, only 43%, and the error rate of anger and surprise was higher, reaching 22%. Although the recognition rate of expression except for anger and happiness is significantly reduced, it is still around 50%, and the overall recognition rate is at an average level, indicating that this model has certain fault-tolerant ability and stability. To further evaluate the advantages of this model, a comparison was made between the basic classification model and the model of [41]. It can be seen from Table 7 that the

TABLE 7. Comparison results of different methods based on SFEW database

Method	SFEW
AlexNet [36]	51.59%
VGG-16 [37]	53.13%
ResNet50 [19]	56.07%
Meng et.al [41]	54.30%
Single-model	55.79%
Double-model	56.83%

results of the double-channel model are both better than the recognition rate of the other models. Both the single-channel model and the double-channel model proposed in this paper are higher than their recognition rates, which proves that the proposed occlusion suppression method achieves good results and can reduce the influence of occlusion to a certain extent.

4.4.5. *Contrast experiment of changing occlusion area and position.* In order to study the influence of occlusion on facial expression recognition from many aspects, we will change the location and area of occlusion to conduct experiments.

Fix the occlusion position, select the left eye, right eye, nose and mouth for occlusion, and the area of the occlusion object is within 12*12, which is about a quarter of the size of the image. CK+ dataset is selected for occlusion processing. Three models, namely double-channel model, single-channel model and Resnet50, are used for comparison. As can be seen from Figure 7, among the three models, the double-channel model has the highest recognition rate no matter which part of the occlusion is, and the occlusion of the nose has the least impact on expression recognition, while the occlusion of the mouth has the largest impact. The most important facial area for facial expression is the mouth, followed by the eyes.

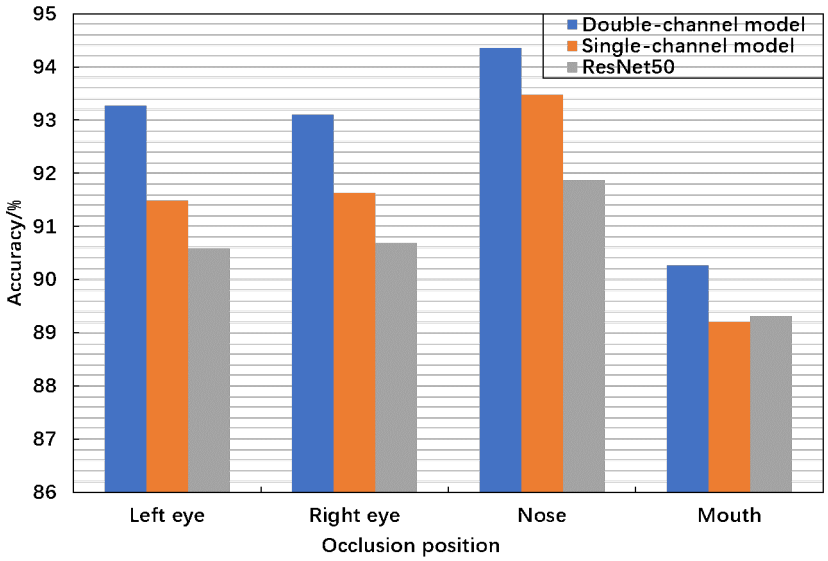

FIGURE 7. The recognition rate under different occlusion parts

For different occlusion areas, the occlusion ratio of 10%∼70% is selected respectively for the comparative experiment, and the occlusion ratio of more than 80% is not considered, because the occlusion area is too large and the facial information was too small. Similarly,

CK+ dataset is selected for occlusion processing in the comparison of double-channel model, single-channel model and ResNet50 model. As can be seen from Figure 8, the recognition rate gradually decreases with the increase of the covered area. When the covered area exceeds 50% of the face area, the recognition rate drops sharply. The double-channel model has the best expression recognition effect among the three models.

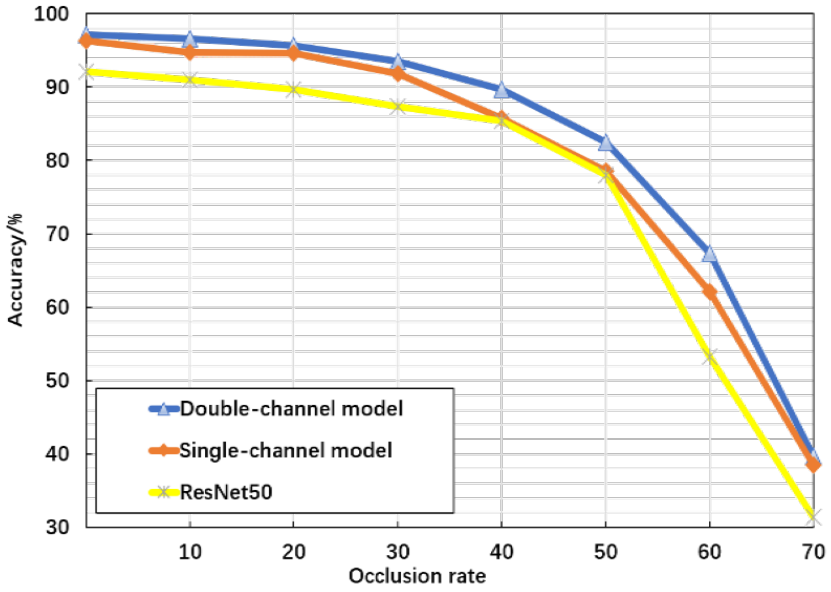

FIGURE 8. The recognition under different occlusion intensities

5. **Conclusion.** In order to reduce the impact of occlusion error on facial expression recognition, we propose a double-channel convolution model with occlusion robust. Clean facial features can be extracted through the geometry-aware module and the occlusion perception module. The double-channel network framework is used to shorten the intra-class distance of the same expression and assist expression recognition when the facial information is incomplete. Through a number of comparative experiments on three datasets, the results show that this model is better than the traditional classification model, but it is difficult to deal with the problems of too large occlusion area and head posture change, and the training time of this model is too long, the model is not light enough, and this model needs to be improved constantly.

Acknowledgment. This work is supported by Program for Changjiang Scholars and Innovative Research Team in University (IRT 16R72).

REFERENCES

[1] M. Pantic, A. Pentland, A. Nijholt, T. S. Huang, Human computing and machine understanding of human behavior: A survey, *Lecture Notes In Computer Science*, vol. 4451, pp. 47–71, 2007.

[2] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: A survey of registration, representation, and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[3] L. Zhao, Z. C. Wang, X. J. Wang, Q. Liu, Driver drowsiness detection using facial dynamic fusion information and a DBN, *IET Intelligent Transport Systems*, vol. 12, no. 2, pp. 127–133, 2017.

[4] J. M. Wu, M. H. Tsai, S. H. Xiao, Y. P. Liaw, A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction, *Journal of Ambient Intelligence and Humanized Computing*, 2020, https://doi.org/10.1007/s12652-020-01826-1.

[5] E. K. Wang, C. M. Chen, M. M. Hassan, A. Almogren, A deep learning based medical image segmentation technique in internet-of-medical-things domain, *Future Generation Computer Systems*, vol. 108, pp. 135–144, 2020.
[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-workshops*, pp. 94–101, 2010.
[7] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, pp. 65–70, 2010.
[8] M. J. Lyons, S. Akamatsu, M. G. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.
[9] S. Li, W, Deng, J. P. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2852–2861, 2017.
[10] A. Mollahosseini, B. Hasani, M. H. A. Mahoor, Affectnet: a database for facial expression, valence, and arousal computing in the wild, *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
[11] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark, *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 2106–2112, 2011.
[12] J. D. Boucher, P. Ekman, Facial areas and emotional information, *Journal of Communication*, vol. 25, no. 2, pp. 21–29, 1975.
[13] J. C. Batista, V. Albiero, O. Bellon, L. Silva, Aumpnet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network, *IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 866–871, 2017.
[14] S. L. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2014.
[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.
[16] S. Chopra, R. Hadsell, Y. Lecun, Learning a similarity metric discriminatively, with application to face verification, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546, 2005.
[17] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
[18] S. Xie, H. F. Hu, Facial expression recognition with FRR-CNN, *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.
[19] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
[20] K. P. Zhang, Z. P. Zhang, Z. F. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, *Electronics Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[21] B. Amos, B. Ludwiczuk, M. Satyanarayanan, others, Openface: a general-purpose face recognition library with mobile applications, *CMU School of Computer Science*, vol. 6, no. 2, pp. 770–778, 2016.
[22] C. F. Shan, S. G. Gong, P. W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
[23] A. Majumder, L. Behera, V. K. Subramanian, Automatic facial expression recognition system using deep network-based data fusion, *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 103–114, 2016.
[24] M. Y. Liu, S. X. Li, S. G. Shan, X. L. Chen, Au-Inspired deep networks for facial expression feature learning, *Neurocomputing*, vol. 159, pp. 126–136, 2015.
[25] S. Y. Xie, H. F. Hu, Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks, *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.
[26] X. Sun, P. P. Xia, F. J. Ren, Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition, *Neurocomputing*, vol. 444, pp. 378–389, 2021.

[27] A. H. Farzaneh, X. J. Qi, Facial expression recognition in the wild via deep attentive center loss, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2402–2411, 2021.

[28] Y. Li, J. B. Zeng, S. G. Shan, X. L. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[29] K. Wang, X. J. Peng, J. F. Yang, D. B. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, *IEEE Transactions on Image Processing*, vol. 29, no. 5, pp. 4057–4069, 2020.

[30] Y. Lu, S. G. Wang, W. T. Zhao, Y. Zhao, WGAN-based robust occluded facial expression recognition, *IEEE Access*, vol. 7, no. 5, pp. 93594–93610, 2019.

[31] X. N. Liu, Y. Zhou, J. Q. Zhao, R. Yao, B. Liu, Y. Zheng, Siamese convolutional neural networks for remote sensing scene classification, *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1200–1204, 2019.

[32] X. L. Wang, R. Girshick, A. Gupta, K. M. He, Non-local neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[33] M. L. Zhu, D. Shi, M. J. Zheng, M. Sadiq, Robust facial landmark detection via occlusion-adaptive deep networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3486–3496, 2019.

[34] G. M. Mehdipour E. H. Kemal A comprehensive analysis of deep learning based representation for face recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34–41, 2016.

[35] Z. Y. Qin, J. Wu, Visual saliency maps can apply to facial expression recognition, *arXiv preprint arXiv:1811.04544*, 2018.

[36] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

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

[38] H. Ding, P. Zhou, R. Chellappa, Occlusion-adaptive deep network for robust facial expression recognition, *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9, 2020.

[39] Y. Li, J. B. Zeng, S. G. Shan, X. L. Chen, Patch-gated CNN for occlusion-aware facial expression recognition, *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2209–2214, 2018.

[40] C. Florea, L. Florea, M. S. Badea, C. Vertan, A. Racoviteanu, Annealed label transfer for face expression recognition, *BMVC*, 104, 2019.

[41] Z. B. Meng, P. Liu, J. Cai, S. Z. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 558–565, 2017.