Lightweight Facial Emotion Recognition Network Based on Inception V3

JIN Jing*[a], YANG Dan[b], WEI Chunyan[c]

([a]School of electronic and Information Engineering，Lanzhou Jiao tong University，Lanzhou，China 730070; [b]Tianjin International Engineering Institute，Tianjin University，Tianjin ,China 300000； [c]Gansu Provincial Hospital，Lanzhou，China 730070)

ABSTRACT

[Objective]Facial emotion recognition is mainly used in psychological medicine, driving monitoring, human-computer interaction and so an. In recent years, Convolutional Neural Network is a common method for facial emotion recognition. With the extensive and in-depth application, the performance requirement of facial emotion recognition is increasing day by day. Augmenting the depth and width of the network is more direct method to improve the recognition performance, but it will cause the parameters and computation to increase rapidly and lead to a long training time. In order to solve this problem, a lightweight face emotion recognition network based on Inception V3 is proposed. The Shortcoming of long training time due to too many parameters in original Inception V3 is improved. [Method(detailed)] The network consists of two convolutional layers, two asymmetric Inception modules, two pooling layers and two fully connected layers combined with Dropout. The total parameters of proposed model is 70% of Inception V3. The FFA-NET filter image is used to enhance the image and the Gaussian noise is used to balance the dataset. The accuracy of the model is improved by dynamic attenuation learning rate of the loss value in validation set. [Result] RAF-DB, CK+ and JAFFE data sets are used to train the model. The training time of a single image is less than 16ms. The average recognition accuracy on testing set is 77.89% and 91.82% respectively. [Conclusion]Compared with the algorithms with superior performance in face emotion recognition in recent years, the accuracy has been improved and the training time is shorter, which can be better applied to various scenarios.

Keywords: Facial emotion recognition; Inception V3; Lightweight Network;

1 INTRODUCTION

Emotion is a state that synthesizes a person's thoughts, affection and behaviors. It can reflect the healthy degree of a person's psychology [1]. At present, automatic facial emotion recognition by image processing and machine learning methods has become an important research in computer vision.

The research on facial emotion recognition is mainly divided into two directions: traditional methods and deep learning. Traditional methods use global or local methods to extract facial features related to emotions and then classify the features based on methods such as Bayesian networks [2]. Convolutional Neural Networks and Generative Adversarial Network in deep learning for facial emotion recognition has become a popular method in recent years [3]. Chen Tuo et al proposed a deep neural network that combines temporal and spatial features to analyze and understand facial expression information in video sequences to improve the performance of emotion recognition. The acceleration module is composed of continuously stacked small convolutional cores to replace the shallow feature extraction module of the network [4].Wang Xiaohong et al proposed a convolutional neural network based on the idea of Inception, using 1×1convolutional kernels of equal size. It can classify facial emotions in small datasets effectively and achieve good results on both CK+ and JAFFE datasets [5]. Zhang Hongli improved the GoogLeNet network by optimizing pruning to accelerate training speed and improve the accuracy of facial expression

Foundation items: Natural Science Foundation of Gansu Province (No.21JR11RA062); the University Innovation Foundation of Gansu Province (No.2022A-047)

* JIN Jing. 28089092 @qq.com; phone 13919416582

recognition with an accuracy of over 80% on JAFFE, CK+ and Cohn Kanada datasets [6]. Cheng Weiyue proposed a deep convolutional neural network algorithm that fuses global and local features. Two improved convolutional neural network branches extract global and local features respectively and use the weighted fused features for classification [7]. Zhang Wei et al proposed facial expression recognition network based on Attention Mechanism [8], which combines multiscale feature extraction with spatial attention and improves recognition accuracy by inputting weighted feature maps of both channels and spaces into subsequent networks to continue feature extraction and classification. Zhang et al constructed an end-to-end depth model based on Generative Adversarial Networks (GANs) for simultaneous facial expression recognition and facial image synthesis. Expanding the dataset improves the generalization ability of the model while improving recognition accuracy [9]. Delphine et al proposed an expression recognition algorithm for partially occluded faces. Based on the texture or geometric shape of the face, a method based on jumped AutoEncoder is used to reconstruct the occluded part of the face in the optical flow field [10].

Currently, the main goal of facial expression recognition is to improve classification accuracy, but higher accuracy usually corresponds to deeper and wider network structures, which still requires relatively low training time and high hardware requirements. Therefore, the paper proposes a lightweight facial emotion recognition network based on the Inception module, which significantly improves the time efficiency of network training, and also improves the classification and recognition accuracy of the network model.

2 INCEPTION V3 DEEP NETWORK

Szegedy et al [11] proposed the Inception network module, and on this basis, successively proposed Inception V1, V2, and V3 network. The core idea of the Inception module is combining different convolutional layers in parallel and the resulting matrices processed by different convolutional layers are spliced together in the depth dimension to form a deeper matrix. Inception modules can be stacked repeatedly to form larger networks, effectively expanding the network depth and width and improving the accuracy while preventing the overfitting. Inception module first performs dimensionality reduction processing on larger matrices while aggregating visual information on different sizes and facilitating feature extraction from different scales. The network first extracts the image features initially through three convolution layers with 3×3 convolution kernel, and then extracts the main features and reduces the size of the image through a 3×3 pooling layer. Then it applies 3×Inception、5×Inception and 2×Inception structure to further extract image features. The extracted image is pooled to 1×1×2048 structure by 8×8 pooling layer, followed by classification using linear activation function and Softmax function. The network extracts facial features through multi-layer convolution series and parallel structures. The network has a large number of layers and parameters, requiring a large amount of training time. At 299×299 size face image, the training time for a single image is about 250ms and the recognition accuracy is 76.6%.

Inception V3 improves the utilization of computing resources compared to traditional CNN and can deepen the width and depth of the network without changing the computational budget, thereby further improving the accuracy of the network. However, the increase of network width and depth will be accompanied by increased networks complexity, more parameters and higher hardware requirements. In addition, overfitting problems due to small sample are inevitable.

3 LIGHTWEIGHT NETWORK BASED ON INCEPTION MODULE

3.1 Network structure

This paper proposes a lightweight facial expression recognition network based on Inception V3 network, which classifies and recognizes seven basic emotions: "angry", "bored", "fear", "happy", "neutral", "sad" and "surprised".

The main difference between the classification network in the paper and the Inception V3 network mainly includes four parts: First, the Inception V3 network structure is simplified with fewer network layers and smaller parameter sizes in this paper; Secondly, the Dropout method with loss rate is used to randomly disconnect the connections between neurons, further avoiding the overfitting in the training process while reducing the parameter size of the network; Thirdly, a nonlinear double fully connected structure is used for classification, preserving main features while effectively reducing training time; At last, the cosine annealing algorithm is compared with the method of dynamic attenuation learning rate based on the loss value of verification set, and the latter is selected to train the algorithm in this paper, so as to improve the accuracy of the model and control the training times. The structure of

facial expression recognition network proposed in this paper is shown in Figure 1.

The network model in this paper consists of two convolutional layers and two different Inception modules, namely, 3 × Inception and 5 × Inception, two pooling layers and two fully connected layers combined with Dropout. Compared with Inception V3 network, the convolutional layer in this paper is reduced from six to two and a 2×Inception module is reduced. At meanwhile, the total parameter quantity of this model is 70% of the Inception V3. The network parameter is shown in Table 1 and modified parameters are marked in bold.

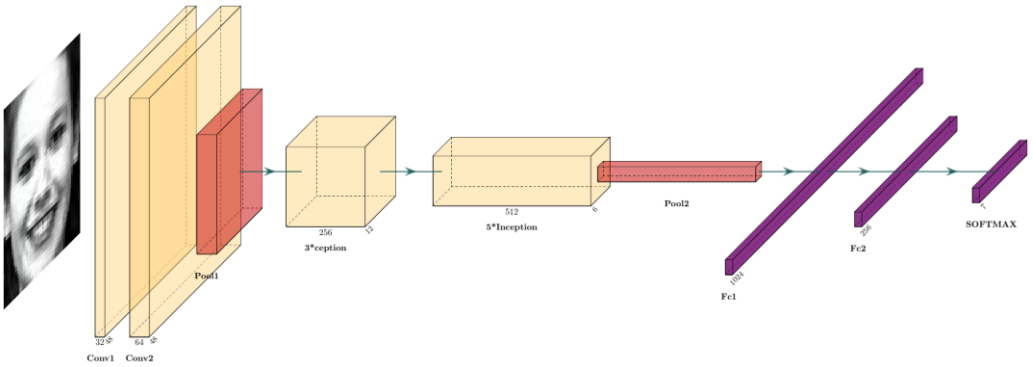


Figure.1 Proposed lightweight facial expression recognition network

Table 1 Parameter list of simplified Inception V3 network

layer	channels	Kernel size	step	fill	Dropout	Feature map size
Input	0	0	None	None	0	(48,48,1)
Conv padded1	32	3×3	**1**	2	0	(48,48,32)
Conv padded2	**64**	3×3	1	2	0	(48,48,64)
Maxpooling1	0	**2×2**	**2**	**0**	**25%**	(24,24,64)
3×Inception	3×Inception					(12,12,256)
5×Inception	5×Inception					(6,6,512)
Maxpooling2	0	**2×2**	**2**	**0**	**0**	(3,3,512)
Flatten	**None**	**None**	**None**	**0**	**0**	(1,1,4608)
Fc1	**None**	**None**	**None**	**0**	**25%**	(1,1,1024)
Fc2	**None**	**None**	**None**	**0**	**25%**	(1,1,256)
Output(Softmax)	**None**	**None**	**None**	**0**	**0**	(1,1,7)

3.2 Algorithm flow

The specific process of facial emotion recognition using this model is as follows:

Begin

Input: 48×48 face image in grayscale

Step1: Extracting the features initially by two-channel 3×3 convs

Step2: Using2×2 Max Pooling and 25% Dropout rate, reducing parameters while retaining key features

Step3: Further extracting facial features by 3×Inception

Step4: Extracting emotional features again by 5×Inception

Step5: 2×2 Max Pooling to retain key features

Step6: Entering the fully connected layer：Fc1 and Fc2，

Output: Obtaining classification results through Softmax

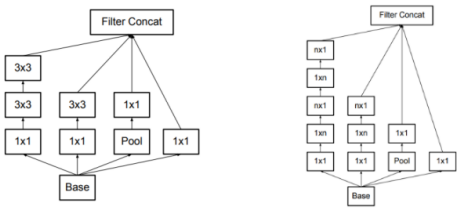
End

In Step 1,3×3 convolution kernel with 32 channels performs convolution operations by filling value 2 and step size 1 to perform preliminary feature extraction. The extracted facial feature image is then convolved by the 3×3 convolution kernel with 64 channels, the filling value is set as 2 and the step size is also set as 1. When performing convolution operations, it chooses activation function ReLU which adds nonlinear features. After convolution operation, it adds a Batch Normalization to prevent vanishing gradient problem. In Step 2, 2×2 Maximum pooling is used to remove redundant features and extract maximum features. The Dropout packet loss rate function is used to accelerate the training speed. After the pooling layer operation is completed, some nodes are temporarily discarded from the network with probability 25%. Through this stage, some unimportant features have been eliminated, facilitating further feature extraction by the Inception module in the next stage.

In Step3, the 3×Inception structure is used to further extract the image feature, where the 1×1 convolution kernel is mainly used to reduce the dimension and increase the nonlinear features. While 5×5 kernel results in a larger receptive field, it also comes with huge parameters. Experiment shows that replacing the 5×5 convolution kernel with two consecutive 3×3 convolution kernels neither affects the size of the receptive field nor increases the parameters. After each convolution kernel we use the regularization method Batch Normalization to speed up the training. The paper executes multiple convolution kernels and pooling windows in parallel to extract facial features. See Figure 2(a) for the 3×Inception structure used in the paper.

After the Step3, the feature map size is 12×12. When the size of the feature map is between 12 and 20, Step 4 uses another asymmetric 5×Inception structure (Figure 3(b)) for further feature extraction, where the value of n is 3. This module also uses the 1×1 kernels for dimensionality reduction and nonlinearity enhancement, but the difference is that the 1×3 and 3×1 asymmetric kernels are used for the convolution operation. The asymmetric convolution kernels not only reduce the parameters, speed up the operation and reduce overfitting, but also increase the feature diversity of the model. When the feature map is in the moderate size (12~20), asymmetric convolution is better on feature extraction.

In Step5, 2×2 Max pooling is used to deal with the redundant information of the model. The dual fully connected structure combined with 25% packet loss ratio is used for feature classification in Step6. Finally, the recognition and classification of seven types of facial emotions are completed by Softmax.



(a)3×Inception **(b)5×Inception**

Figure.2 Inception Structure

4 EXPERIMENT AND RESULT ANALYSIS

The experiment is based on Python3 TensorFlow library and uses Keras framework to build the network. The AutoDL server is adopted. The hardware platform is 7-core Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and the GPU is RTX 3060.

4.1 Experiment Dataset

RAF-DB is a large-scale facial expression database, including 29,762 images from the real world [12]. The paper

selects a calibrated set in the database. It is composed of 100×100 color images with a total of 15,339 images, 12,271 images in the training set and 3,068 images in the test set. The JAFFE database [13] includes 213 gray-scale images of 256×256 size. The samples number is small, but the collection environment is in a laboratory with uniform lighting and brightness conditions, which is suitable for training simple CNN. The CK+ dataset was published in 2010 [14] and is an extension of the CK dataset. The CK+ dataset includes 981 gray-scale images of 48×48 with more samples than JAFFE dataset. RAF-DB data set is larger, but the model performance will be affected due to complex acquisition environment, multiple acquisition angles and other factors. Therefore, the paper combines RAF-DB data set with JAFFE and CK+ data set to construct multi-angle and multi-environment samples, avoiding the problem of unbalanced distribution of RAF-DB data set effectively, which is helpful for model adjustment.

4.2 Data Preprocessing

There are four steps for data preprocessing. The first step is extracting the face region by Haar feature extraction classifier; the second step is filtering and enhancing the image by FFA-NET; the third step is to further ensure the quality of the data set by image clipping and data cleaning; the last step is data set balance by Gaussian noise.

The face image in JAFFE data set is 256×256 size, which is inconsistent with the input image size of network. At meanwhile, there are some factors affecting facial feature extraction, such as hair, background, etc. So, Haar feature detection classifier is used to extract face region .Then the size of the extracted face image is cut to 48×48.

After face region extraction, some data samples are not clear enough. Therefore, the paper adopts FFA-NET (Feature Fusion Attention network) proposed by Qin et al. [15] to filter and enhance the sample images. The FA feature attention module can flexibly deal with different feature regions and set different weights for each channel and pixel feature. These weights will combine with local residual learning to generate enhanced images. Compared with the image only processed by gray processing, the gray distribution of the filtered enhanced image is more uniform and the enhancement effect is better.

In general, the sample distribution of the data set is uneven. For example, the number of fear samples in the RAF-DB dataset is only 6% of happy samples. If we simply expand the quantity of the small samples, it is easy to overfitting in the training process. Compared with the traditional method of replicating datasets, the method of adding noise can effectively avoid overfitting while expanding the data. In the paper, Gaussian noise is used to expand the data set. Specifically, Gaussian noise is randomly generated by setting the mean to 0 and the variance to 5.The data of the classification with small samples is expanded by two or three times.

4.3 Model training

RAF-DB, JAFFE, CK+ data sets are divided into training set, verification set and test set. JAFFE and CK+ data set are randomly split and the ratio of training set and test set is 4:1. 10% face images are randomly selected from RAF-DB training set, split JAFFE data set and original training set of CK+ as the verification set and the remaining 90% face images in the data set is training set.

The training parameters are shown in Table 2. Through the training, the average accuracy of the training set reached 99.59%, the average accuracy of the verification set reached 94.62%. The average time of each iteration is less than 1s.

Table 2 Training parameters

Parameters	Value
batch size	64
epochs	30
learning rate	0.001

A method of dynamically attenuating the learning rate according to the loss value is adopted in the training. This method can ensure that the model is always in the learning stage and avoid the increase of training time caused by ineffective iteration, so as to further optimize the model and improve the model performance. Whether to attenuate the learning rate is considered by judging the attenuation of the loss value every five iterations. If the loss value is on the rise, then the attenuation learning rate will be 5% of the original learning rate; otherwise, the original learning rate will be maintained. When the learning rate attenuates to 0, the learning rate attenuation is stopped. In each iteration of the model, the change of the loss value is checked by the callback function, so that the learning rate can be dynamically adjusted. Specifically, the initial learning rate is set to 0.001. Fig. 4 shows the dynamic change of learning rate by the loss value of the verification set. After experimental comparison, the accuracy of attenuated

learning rate can be improved by about 2% compared with fixed learning rate.

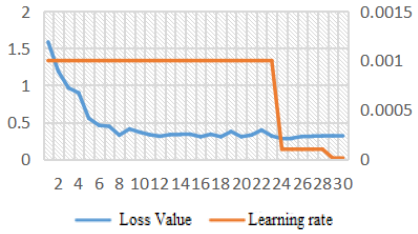


Figure.3 Loss value and learning rate curve of validation

4.4 Model testing result on datasets

The proposed model is used to test the CK+ dataset. Table 3 shows the Confusion Matrix of the test results. The average accuracy of the proposed model on the CK+ dataset is 91.82%. The table shows that the accuracy is as high as 100% on "happy". The accuracy is higher on "bored", "neutral", and "surprised" (90.43%, 90.54% and 97.6%). The accuracy is lower on "fear" and "sad" (74.72%, 75.19%). The main reason is that part features of "fear" and "sad" show the tendency of "angry". The "fear" sample is too few, imbalance of dataset leads to low accuracy.

Table 4 shows the Confusion Matrix of test results on JAFFE dataset. The average accuracy of JAFFE dataset is 77.89%, which is lower than CK+ dataset. The accuracy of "happy" and "surprised" is high, reaching 93.33% and 95%. The accuracy of "bored" is 68.67%. The accuracy of "neutral", "sad" and "angry" is 78.67%, 62.67% and 70.00%, respectively. The "fear" is lower, reaching 62.00%. The main reason is that some data of the "fear" category showed other characteristics such as "sad" and "surprised", so the recognition rate decreased.

Table 3 Test Confusion Matrix of CK+

	angry	**bored**	**fear**	**happy**	**neutral**	**sad**	**surprised**
angry	0.86	0.11	0.00	0.00	0.00	0.00	0.03
bored	0.05	0.90	0.00	0.02	0.02	0.00	0.00
fear	0.09	0.00	0.75	0.11	0.00	0.00	0.06
happy	0.00	0.00	0.00	1.00	0.00	0.00	0.00
neutral	0.00	0.04	0.00	0.00	0.91	0.02	0.04
sad	0.13	0.04	0.01	0.05	0.01	0.75	0.00
surprised	0.00	0.00	0.00	0.00	0.00	0.02	0.98

JAFFE data set shows a significant decrease in accuracy compared with CK+ data set. In particular, classification result such as "angry," "sad" and "bored" show a downward trend. There are several reasons. Firstly, JAFFE dataset are collected earlier, and there are similarities between different categories. Secondly, the distribution of various types in JAFFE is unbalanced. Although the paper carries out Gauss noise balancing on the dataset, there will also be problems of overfitting and low generalization ability.

Table 4 Test Confusion Matrix of JAFFE

	angry	**bored**	**fear**	**happy**	**neutral**	**sad**	**surprised**
angry	0.70	0.15	0.00	0.00	0.05	0.10	0.00
bored	0.25	0.69	0.00	0.00	0.00	0.07	0.00
fear	0.03	0.08	0.62	0.00	0.03	0.12	0.11
happy	0.00	0.00	0.00	0.93	0.00	0.03	0.03
neutral	0.00	0.00	0.00	0.00	0.79	0.00	0.21

sad	0.11	0.07	0.05	0.00	0.15	0.63	0.00
surprised	0.00	0.00	0.00	0.00	0.05	0.00	0.95

4.5 Comparison on time efficiency and accuracy

In this part, the proposed method is compared with literature [5], literature [6], literature [11] and literature [16] in terms of training time and accuracy. The Inception V3 is proposed in literature [11]. The detailed results are shown in Table 5.The experiment results show that greatly improvement in time efficiency is the main contribution of this paper.

Table 6 compares the classification accuracy (%) of different algorithms on the CK+ dataset. Compared with the Inception V3 model, the paper has achieved improvement in "happy" and "neutral", and the average accuracy is higher than the traditional model. Compared with literature [5], literature [6] and literature [11], the average accuracy of this paper is the highest, especially the accuracy of "angry", "bored", "happy", "neutral" and "surprised" has been greatly improved. However, the accuracy on "sad" and "fear" is lower than other methods.

As shown in Table 7, compared with the Inception V3 model, the proposed model has achieved improvement in the "surprise" class in JAFFE dataset and the average accuracy is better than the Inception V3 model. However, there is a larger gap between the results of JAFFE dataset and CK+ dataset. The reason is that the data feature of the JAFFE dataset are not obvious and the naked eye cannot identify the emotion accurately. There is a large deviation between the label and the data。So it is necessary to re-label and clean the dataset to further improve the accuracy.

Table 5 training time of different algorithms(ms)

	Literature [5]	Literature [6]	Literature [11]	Ours
时间	<47	<200	<156	<16

Table 6 Classification accuracy of different algorithms on CK+

	Literature [5]	Literature [6]	Literature [11]	Literature [16]	Ours
angry	83.33	74.17	**94.82**	71.24	**85.52**
bored	**91.43**	81.23	**98.85**	77.45	90.43
fear	66.67	**84.58**	**97.34**	81.86	74.72
happy	97.56	94.49	**99.51**	86.91	**100.00**
neutral	**96.15**	69.81	42.31	65.29	**90.54**
sad	62.67	**93.04**	**96.07**	85.78	75.19
surprised	96.00	89.96	**98.39**	85.68	**97.6**
Average	84.83	85.09	**89.62**	79.26	**91.82**

Table 7 Classification accuracy of different algorithms on JAFFE

	Literature [5]	Literature [6]	Literature [11]	Literature [16]	Ours
angry	75.00	**81.02**	69.99	**75.02**	70.00
bored	50.00	**81.95**	53.33	**76.39**	68.67
fear	66.67	**75.78**	54.02	**71.88**	62.00
happy	83.33	89.01	**100.00**	84.39	**93.33**
neutral	66.67	61.56	**80.00**	57.84	**78.67**

sad	83.33	**87.34**	**93.33**	82.63	62.67
surprised	83.33	**88.29**	83.33	83.68	**95.00**
Average	**82.09**	**83.84**	75.74	74.19	77.89

5 CONCLUSION

In this paper, a lightweight network model is proposed for face image emotion recognition. Compared with Inception V3 network and classical face emotion recognition methods, the proposed model gets a significant improvement in time efficiency. Classification accuracy in CK+ dataset also is improved greatly, especially in the "happy" and "neutral" emotion. On the JAFFE dataset, the average classification accuracy of the paper is higher than compared literature. At the same time, the proposed model can reduce computation amount and avoid overfitting. In the following work, the strong feature retention model can be adopted to further reduce the training time and improve classification accuracy.

REFERENCES

[1] Lu Y. Research on Emotion Recognition Based on Deep Neural Network. Wuhan: Wuhan University, 2020: 20-28.

[2] Hong H Q, Shen G P, Huang F H. A review of Expression Recognition Technology. Journal of Frontiers of Computer Science and Technology, 2022, 16(08):1764-1778.

[3] Yang D K, Huang S, Wang S L, et al. Facial Expression Recognition Method Based on Generative Adversarial Network and Network Integration. Journal of Computer Applications, 2022, 42(03):750-756.

[4] Chen T, Xing S, Yang W W, et al. Facial Expression recognition based on time-domain feature fusion . Journal of Image and Graphics, 2022, 27(07):2185-2198.

[5] Wang X H, Liang Y C, Ma X C. A Deep Learning Algorithm for Face Expression Classification Based on Inception. Optical Technique, 2020, 46(03):347-353.

[6] Zhang H L, Bai X Y. Facial Expression Recognition method using Optimized Pruning GoogLeNet. Computer Engineering and Applications, 2021, 57(19):179-188.

[7] Cheng W Y, Zhang X Q, Lin K Z, et al. Deep Convolutional Neural network Algorithm integrating global and local Features. Journal of Frontiers of Computer Science and Technology, 2022, 16(05):1146-1154.

[8] Zhang W, Li P. Facial expression recognition network based on attention mechanism. Journal of Tianjin University (Natural Science and Engineering Technology), 2022, 55(07):706-713.

[9] Zhang X , Zhang F , Xu C . Joint Expression Synthesis and Representation Learning for Facial Expression Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, PP(99):1-1.

[10] D. Poux, B. Allaert, N. Ihaddadene, Dynamic Facial Expression Recognition Under Partial Occlusion With Optical Flow Reconruction[J]. IEEE Transactions on Image Processing, 2022(31): 446-457.

[11] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[12] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[13] Shih F Y , Chuang C F , Wang P . Performance Comparison Of Facial Expression Recognition in JAFFE DATABASE[J]. International Journal of Pattern Recognition & Artificial Intelligence, 2008, 22(3):445-459.

[14] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]//2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010: 94-101.

[15] Qin X, Wang Z, Bai Y, et al. FFA-Net: Feature fusion attention network for single image dehazing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11908-11915

[16] Du L, Hu H. Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy[J]. Computer Vision and Image Understanding, 2019, 186: 13-24.