Facial Expression Recognition Based on PCD-CNN with Pose and Expression

Hongbin Dong(✉), Jin Xu, and Qiang Fu

Computer Science and Technology College, Harbin Engineering University,
Harbin, China
donghongbin@hrbeu.edu.cn

Abstract. In order to achieve high recognition rate, most facial expression recognition (FER) methods generate sufficient labeled facial images based on generative adversarial networks (GAN) to train model. However, these methods do not estimate the facial pose before passing the images to the generator, which affects the quality of generated images. And mode collapse is prone to occur during the training process, leading to generate a single-style facial images. To solve these problems, a FER model is proposed based on pose conditioned dendritic convolution neural network (PCD-CNN) with pose and expression. Before passing the facial images to the generator, PCD-CNN was used to process facial images, effectively estimating the facial landmarks to detect face and disentangle the pose. In order to accelerate the training speed of the model, PCD-CNN was based on the ShuffleNet-v2 framework. Every landmark of facial image was modeled by a separate ShuffleNet-DeconvNet, maintaining better performance with fewer parameters. To solve the mode collapse during image generation, we theoretically analyzed the causes, and implemented mini-batch processing on the discriminator in the model and directly calculated the statistical characteristics of the mini-batch samples. Experiments were carried out on the Multi-PIE and BU-3DFE facial expression datasets. Compared with current advanced methods, our method achieves higher accuracy 93.08%, and the training process is more stable.

Keywords: Pose estimation · Mode collapse · Expression recognition

1 Introduction

FER is a biometric recognition technology that uses computers to obtain and analyze facial expressions so that computers can recognize and even understand human emotions, thereby achieving the purpose of human-computer interaction [1]. However, the existing FER is based on frontal facial images and videos. The recognition rate of FER in the case of non-frontal faces is not very ideal. Sariyanidi [2] pointed out that the current research on FER faces five major problems: head deflection, illumination changes, registration errors, face occlusion, and identity differences. In these problems, head deflection is an important cause of registration errors and face occlusion. In addition, the lack of sufficient training samples can cause overfitting during the learning process. To solve these problems, Zhang [3] proposed a FER model by jointing pose

J. Zeng et al. (Eds.): ICPCSEE 2020, CCIS 1257, pp. 518–533, 2020.
https://doi.org/10.1007/978-981-15-7981-3_38

and expression. This model can simultaneously synthesize facial images and recognize facial expressions. The model of Zhang made an ideal recognition accuracy under the condition of non-frontal facial expressions. However, before conveying the facial image to the generator, Zhang used the lib face detection method with 68 facial landmarks for face detection [3], which could not effectively estimate the facial pose and affected the generation of facial images. And the model is based on GAN, the training process is unstable, mode collapse is prone to occur, resulting in only a certain style of facial images generated. Kumar [4] proposed a PCD-CNN model based on CNN, which can conditionally estimate the facial landmarks to disentangle the facial pose of the image. The PCD-CNN is a dendritic structure model, can effectively capture shape constraints in a deep learning framework, but due to the many model parameters, the running time is very long.

In response to the above problems, we improved method of Zhang and proposed a FER model based on PCD-CNN with pose and expression. It has three advantages: (1) we use the PCD-CNN to preprocess facial images before conveying the facial images to the generator. Compared with Zhang's method, ours can not only accurately detect the facial area, but also effectively estimate facial pose, (2) the PCD-CNN in our model is based on the ShuffleNet-v2 [5] architecture, which makes our model maintain better performance with fewer parameters, and (3) we perform mini-batch processing on the discriminator in the model and directly calculate the statistical characteristics of the mini-batch samples, making the model more stable during training, avoiding the occurrence of mode collapse as much as possible. Our model is trained and tested on the Multi-PIE [6], BU-3DFE [7] facial expression datasets. Compared with Zhang's and the latest methods, our improved model has higher recognition accuracy and achieves 93.08%.

2 Related Work

In non-frontal FER, disentangling the facial pose can improve the extraction of facial information features by the classifier, thereby improving the accuracy of expression recognition [4]. Common facial pose estimation methods include geometric method, tracking method and nonlinear regression method [8]. The geometric method focuses on the extraction of facial landmarks information and improves the correlation between facial landmarks and facial poses. The tracking method mainly focuses on tracking a person's head in a video. Literature [9] proposed a method of target tracking based on joint probability. Firstly, the target area was located by graph structure method, then the probability model of target tracking was built, and the particle filter was used to track the target of a single frame of image. Finally, the face pose was estimated. The nonlinear regression method focuses on establishing the mapping relationship between poses. Different from existing methods, Kumar [4] proposed PCD-CNN model to disentangle facial pose in unconstrained facial image alignment. PCD-CNN follows the Bayesian formula and effectively disentangles the facial pose of the image by adjusting the facial landmarks.

In addition, the public dataset of non-frontal FER lacks sufficient training samples. In this situation, it is difficult to effectively train non-frontal FER model based on deep neural networks, which may cause overfitting. To solve this problem, we generally generate enough labeled training samples through the model. Since the GAN proposed by Goodfellow [10] has wide applications in image generation, this inspired us to use GAN to generate labeled samples to enrich the training dataset. As a generative model with excellent performance, GAN has two outstanding advantages [11]: (1) does not rely on any prior assumptions; (2) the method of generating samples is very simple (Fig. 1).

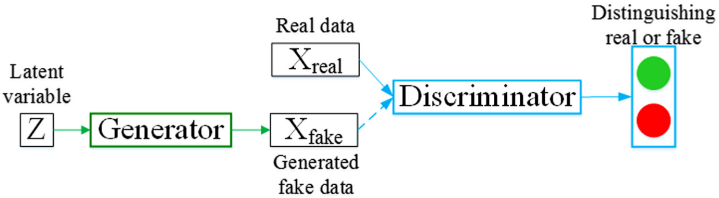


Fig. 1. Basic architecture of GAN. It is mainly composed of generator and discriminator. The task of generator is used to generate "fake data", and the task of the discriminator is to distinguish whether the data is "real data" or "fake data".

However, mode collapse is very common in general GAN model. In simple terms, mode collapse means that the generator produces a single or limited pattern. Common solutions are experience replay and using multiple GAN. Experience replay refers to minimizing the jumps between patterns by displaying previous generated samples to the discriminator at regular intervals. Using multiple GAN is to achieve the purpose of covering all patterns in the sample by merging multiple GAN, avoiding the occurrence of a single pattern. Zhang [3] proposed a model based on GAN for non-frontal FER. The generator in this method is based on the encoder-decoder structure, and can learn generative and discriminative representations in facial images. It can simultaneously synthesize facial images and recognize facial expressions in different poses and expressions. Due to it based on GAN, mode collapse may occur during the training, resulting in only a single style of facial images generated or the effect of generated images is not ideal.

In view of the superior performance and shortcomings of the model proposed by Zhang, we improved it and proposed a FER model based on PCD-CNN with pose and expression. The experimental results prove that the recognition accuracy of improved algorithm is higher than existing advanced algorithms, and the training process is more stable.

3 Our Method

We first introduce the improved PCD-CNN for face detection and facial pose estimation in this section. Then we describe the overall architecture of our model and how to synthesize labeled facial images for training. Finally, the improvement of the discriminator in the model is introduced to improve the stability of the model during training.

3.1 Improved PCD-CNN

In order to accurately locate the facial landmarks and estimate the pose, Kumar [4] proposed PCD-CNN following the Bayes formula. The facial pose of image is accurately disentangled by conditionally adjusting the facial landmarks based on PCD-CNN. Inspired by this, we use PCD-CNN to perform face detection and pose estimation on the facial image before conveying the image to the generator. According to Kumar's settings [4], following the natural hierarchy between image I, head pose P and keypoint C variables, the joint and conditional probability are represented as Eq. (1) and Eq. (2) respectively.

$$p(C,P,I) = p(C\,|\,P,I)p(P\,|\,I)p(I) \tag{1}$$

$$p(C,P\,|\,I) = \frac{p(C,P,I)}{p(I)} = \underbrace{p(P\,|\,I)}_{CNN}\underbrace{p(C\,|\,P,I)}_{PCD-CNN} \tag{2}$$

In Eq. (2), in order to achieve $p(P\,|\,I)$, we train a CNN based on facial image to roughly predict the facial pose. In order to achieve $p(C\,|\,P,I)$, we combine convolutional neural network with multiple deconvolutional neural network arranged in a dendritic structure. The convolutional neural network maps the image to a lower dimension, and then the output of the deconvolutional neural network is used to form facial landmark heatmap. As shown in Fig. 2, PCD-CNN includes a PoseNet and a KeypointNet, which are used for the pose and keypoint estimation, respectively. The different keypoints coordinate position is determined according to the mutual relationship between the keypoints. In order to capture the relationship between different keypoints, the nose is used as the root node. The correlation between different keypoints is modeled by special functions $f_{i,j}$ [4], which can also be achieved by convolution. When adding responses corresponding to neighboring nodes, the low confidence for specific keypoints is strengthened [4]. The experimental results of Kumar prove that the model can effectively capture the shape constraints in the deep learning framework, conditionally estimate the facial landmarks to disentangle the facial pose of the image.

However, the PCD-CNN is based on the Squeezenet-11 architecture. Because more parameters are more likely to cause overfitting during training, and the computing performance of the machine is also higher, the processing time is longer. In order to maintain the performance with fewer parameters, we base the PCD-CNN model on the ShuffleNet-v2 [5] architecture. Due to the special structure of ShuffleNet-v2, the calculation cost is greatly reduced, not only the calculation complexity is very low, but also the accuracy is very high. ShuffleNet-v2 deprecated the 1 × 1 group convolution operation and directly used 1 × 1 ordinary convolution with the same number of input/output channels. A new type of channel separation operation is proposed. The input channel of the module is divided into two parts, one part is directly passed down and the other part is used for true backward calculation. At the end of the module, the output channels on the two branches are directly connected in series. Then perform random array operation on the output feature map of the final output, so that the information between the channels is communicated. It also provides a variant of the module that requires downsampling. In order to ensure that the total number of output channels is increased during downsampling, it cancels the random array operation at the beginning of the module, so that the channels are processed separately and then stitched together to double the final output channel number.

Our improved PCD-CNN model is based on the ShuffleNet-v2 structure and is implemented by combination of a convolutional neural network and multiple deconvolutional neural network. As shown in Fig. 3, we call it ShuffleNet-DeconvNet. We perform a convolution operation on the eighth pooling layer of PoseNet in the model, and then feed it back to the eighth pooling layer of KeypointNet. We perform ReLU nonlinearity and batch normalization after each convolution layer. Every landmark in the improved PCD-CNN is modeled by a separate ShuffleNet-DeconvNet, so that the parameters in the model can be effectively reduced. Before conveying the facial image to the generator, we use the improved PCD-CNN to preprocess it first. Compared with the lib face detection algorithm with 68 facial keypoints used by Zhang, our method can not only accurately detect human faces, but also effectively estimate facial poses.

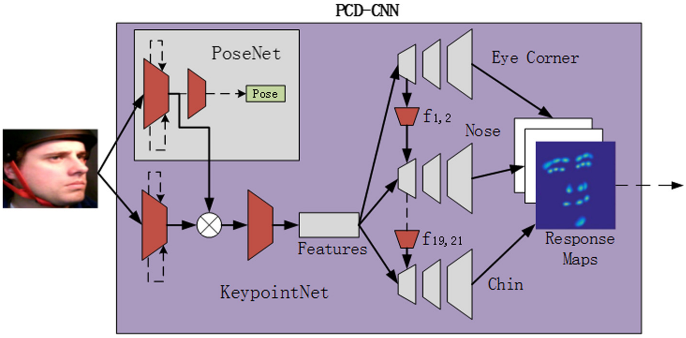


Fig. 2. The overall architecture of the PCD-CNN.

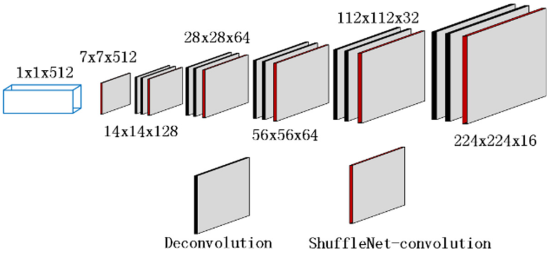


Fig. 3. The architecture of ShuffleNet-DeconvNet.

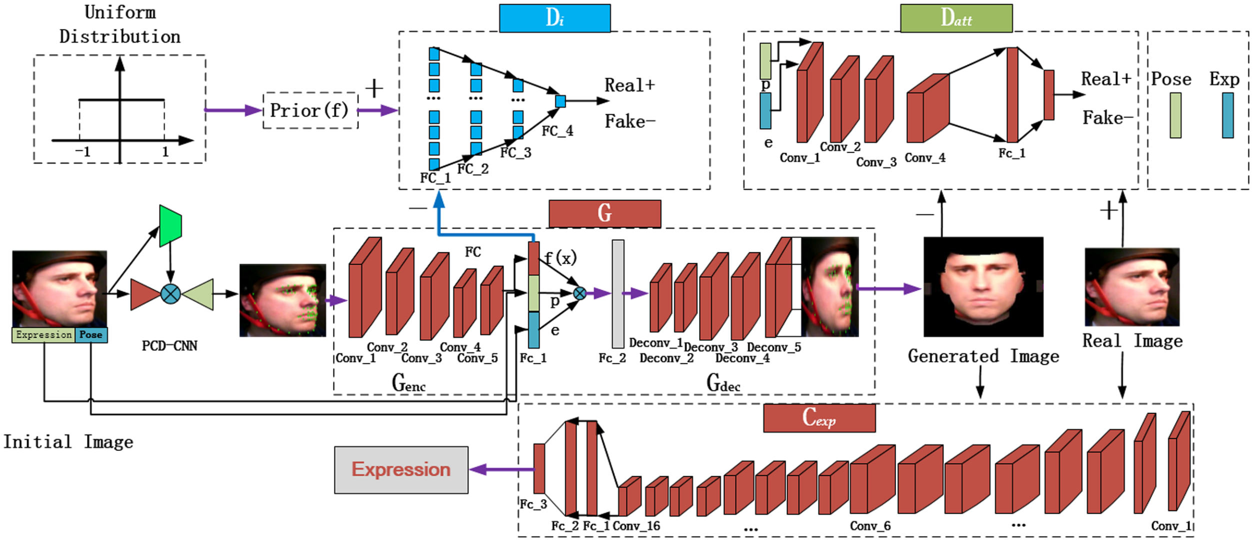


Fig. 4. The overall framework of our model. First, we use improved PCD-CNN model to detect human face and estimate pose. Image is then entered into the encoder-decoder structure generator G for facial image synthesis in different poses and expressions. At last, the generated images and the real images are used for training of the classifier C_{exp} to realize non-frontal FER.

3.2 The Overall Framework of the Model

After improved PCD-CNN preprocesses the facial images, we follow the work of Zhang to put the processed images into a generator based on the encoder-decoder structure, and learn the identity of the images at the same time, as shown in Fig. 4. The model includes a generator G, two discriminators D_{att}, D_i, and a classifier $C_{\exp}$. The generator G is divided into an encoder G_{enc} and a decoder G_{dec}. G_{enc} learns the mapping $f(x)$ from the facial image to the identity, and then $f(x)$ connect the pose and expression into the decoder G_{dec} to generate the corresponding pose and expression facial image. The task of the discriminator D_{att} in the model is to distinguish whether the entered facial images are generated images or real images, which is a binary classification problem. In the training process, the distribution of training data is $P_{d(x,y)}$, the label of the training image is y, and following the image generation principle of GAN, we can train our generator G and discriminator D_{att} by Eq. (3).

$$\begin{aligned}\min_G \max_{D_{att}} & E_{x,y \sim p_d(x,y)}[\log D_{att}(x,y)] \\ & + E_{x,y \sim p_d(x,y)}[\log(1 - D_{att}(G(x,y),y))]\end{aligned} \tag{3}$$

$$\begin{aligned}\min_G \max_{D_i} & E_{f^* \sim Prior(f)}[\log D_i(f^*)] \\ & + E_{x \sim p_d(x)}[\log(1 - D_i(G_{enc}(x)))]\end{aligned} \tag{4}$$

$$\begin{aligned}L_c(G,C) = E_{x,y^e}[&-y^e \log C(G(x),y^e) \\ &-y^e \log C(x,y^e)]\end{aligned} \tag{5}$$

The role of the discriminator D_i is mainly to be able to improve pose smooth and expression conversion. It can be seen from Fig. 4 that $f(x)$ is an identity mapping of facial image. It needs to be clear that when the generator generates a facial image, the pose and expression are changed, and the identity characteristics of the image remain unchanged. $Prior(f)$ is a prior distribution, and $f^* \sim Prior(f)$ represents a random sampling process from $Prior(f)$ [3]. The generator G and discriminator D_i are trained by the min-max objective function, as shown in Eq. (4). The role of the classifier $C_{\exp}$ is mainly to classify all the images in different poses to achieve expression recognition, include generated images and real images. The loss function of the classifier is softmax, as shown in Eq. (5).

3.3 Improvements to the Discriminator

In order to avoid the mode collapse as much as possible and improve the stability of the model during the image generation process, we improved the discriminator model of Zhang [3]. The model has two discriminators D_i and D_{att}. The role of the discriminator D_i is mainly to improve pose smooth and expression conversion. The task of the discriminator D_{att} is not only to distinguish whether the samples come from the real samples or the generated samples, but also to feedback the generator information to generate samples with dissimilar styles as much as possible. When mode collapse occurs during the image generation process, the generator G often maps different hidden variables Z to the same pattern X. After updating the discriminator D_{att}, the discriminator will soon find that the pattern X is abnormal. Therefore, the degree of trust in the pattern X is reduced, that is, the probability of the sample in the training dataset to generate pattern X, so the generator G will map many different hidden variables Z to another pattern Y. Similarly, the discriminator will find that the pattern Y is also abnormal, so the discriminator and generator enter a meaningless loop [22]. Regarding the feedback of the discriminator, the response of generator is excessive. Ideally, the generator should map some hidden variables Z to pattern Y and map some hidden variables Z to pattern X. The reason for the mode collapse is related to the way we train the generator. The objective function is shown in Eq. (6), The generator G generates m samples $\{x_1, x_2, \ldots, x_i, \ldots, x_m\}$, and then sends these x_i to the discriminator D_{att} separately to obtain gradient information. Because the discriminator can only process one sample independently at a time, the gradient information obtained by the generator on each sample lacks unified coordination and points in one direction [23]. And there is no mechanism that requires that the output results of the generator differ greatly from each other. For example, due to the discriminator does not trust the pattern X and trusts the pattern Y very much, the discriminator will guide generator to approach the pattern Y for any randomly generated sample: $G(z) \rightarrow Y$. That is, for any sample, the gradient direction passed by the discriminator D_{att} to the generator G is the same, the generator updates the parameters according to the gradient direction, and it is very easy to transfer all the hidden variables Z to the pattern Y, thereby generating a single-style facial images. To solve this problem, Goodfellow [12] proposed a mini-batch discriminator. For each sample x_i of a mini-batch, the result of an intermediate layer $g(x)$ of the discriminator is led out. It is an n-dimensional vector, which is multiplied by a learnable $n \times p \times q$-dimensional tensor T, the $p \times q$-dimensional feature matrix M_i of the sample x_i is obtained, which can be regarded as $p \times q$-dimensional features. Calculate the sum of the r-th feature difference between each sample x_i and other samples in the mini-batch, as shown in Eq. (7), where $M_{i,r}$ represents the r-th row of the matrix M_i, and the difference between the two vectors is represented by the $L1$ norm.

$$\min_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)}))) \tag{6}$$

$$o(x_i)_r = \sum_j \exp\left(-\| M_{i,r} - M_{j,r} \|_{L1}\right) \tag{7}$$

Then each sample will calculate and get a corresponding vector, as shown in Eq. (8).

$$o(x_i) = \left[o(x_i)_1, o(x_i)_2, \ldots, o(x_i)_p\right]^T \tag{8}$$

$$o(x_i) = \frac{1}{n} \sum_{i=1}^{n} (\sigma_i) \tag{9}$$

$$\sigma_i = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(g\left(x_j\right)_i - \hat{g}_i\right)^2} \tag{10}$$

Finally, $o(x_i)$ is connected to the next layer of the corresponding intermediate layer. We simplified the calculation method of the mini-batch discriminator to make the calculation process easier. For the input sample x_i of the discriminator, we extract an intermediate layer as the n-dimensional feature $\{g(x_1), g(x_2), \ldots, g(x_i), \ldots, g(x_m)\}$, calculate the standard deviation of each dimension and average the values, as shown in Eq. (9) and Eq. (10). Similarly, $o(x_i)$ is connected to the output of the corresponding intermediate layer as a feature map. The improved mini-batch discriminator does not contain the parameter T to be learned, and directly calculates the statistical characteristics of the batch samples, which is more concise. The idea of the two methods is basically the same, no longer let the discriminator process only one sample at a time, but process all samples of a mini-batch at the same time [13]. The specific implementation is based on the original discriminator intermediate layer add a mini-batch layer, whose input is $g(x_i)$ and the output is $o(x_i)$. The difference is that method of Goodfellow also includes a learning parameter T, and the calculation process involves norm, which is more complicated. We will extract an intermediate layer of D_{att} as the n-dimensional feature $\{g(x_1), g(x_2), \ldots, g(x_i), \ldots, g(x_m)\}$, calculate the standard deviation of each dimension and average it. When mode collapse occurs and the generator needs to be updated, generator G first generates mini-batch of samples $\{G_1, G_2, \ldots, G_i, \ldots, G_m\}$. Since these samples are in pattern X, The D_{att} will determine how close one sample is to other samples in mini-batch, so as to distinguish these samples that lack diversity. And the mini-batch discriminator will not simply give all samples x_i the same gradient direction. Thereby avoiding the occurrence of mode collapse, and improving the stability during model training.

4 Experimental Results and Analysis

In order to prove the effectiveness of improved model, we conducted experiments on the Multi-PIE and BU-3DFE standard datasets, respectively, then compared with Zhang and the latest algorithm.

4.1 Experimental Datasets

The datasets used in our experiments are Multi-PIE and BU-3DFE, as shown in Table 1. These are two standard facial expression datasets with various poses, and gradually become an important test set in the field of FER.

Table 1. Details of each dataset used in the experiment.

Datasets	Poses	Expressions	Samples
Multi-PIE [6]	5	6	7655
BU-3DFE [7]	35	6	21000

The Multi-PIE is developed on the basis of the CMU-PIE facial dataset, and contains more than 75,000 multi-pose, illumination and expression facial images of 337 volunteers. The pose and illumination change images were also collected under strict constraints. According to the work of Zhang [3], we selected the facial images of 270 volunteers and captured 1,531 images at five pan angles of $(\pm 30^\circ, \pm 15^\circ, 0^\circ)$ respectively, so we have a total of $1531 \times 5 = 7,655$ facial images in our experiments. The expressions of the images were divided into six categories: disgust, neutral, scream, smile, squint and surprise. Similarly, we use 5-fold cross-validation. So we have 6,124 training samples and 1,531 testing samples respectively [3]. Due to our method could generate facial images with different style, the generated images and the real images together are $6124 \times 5 \times 6 + 6124 = 189,844$ images to train our classifier.

The BU-3DFE facial dataset is a sequence of 606 facial expressions obtained from 100 volunteers. It contains 6 expressions of anger, disgust, fear, happy, sad, and surprise, and is mostly used for 3D facial expression analysis [7]. Similarly, we follow the work of Zhang, the poses of the used facial images include 7 pan angles $(\pm 45^\circ, \pm 30^\circ, \pm 15^\circ, 0^\circ)$ and 5 tilt angles $(\pm 30^\circ, \pm 15^\circ, 0^\circ)$ [7]. We randomly divided 100 volunteers into 80 as the training dataset and 20 as the testing dataset. Therefore, in our experiments, there are $100 \times 6 \times 5 \times 7 = 21,000$ facial images, including 16,800 training samples and 4,200 test samples.

4.2 Experiment Introduction

The overall architecture of our method is shown in Fig. 4. Firstly, we detected faces and estimated facial poses based on the improved PCN-CNN model. Then according to the settings of Zhang [4], the facial image is cropped to size 224 × 224, and the image intensity is linearly scaled to the range of [1,1]. In the model, the generator is implemented based on the encoder-decoder structure. The encoder and decoder are connected through $f(x)$ that identifies the characteristics. $f(x)$ associates the pose p with the expression e and outputs it through the fully connected layer in the network. We use fractionally-strided convolution to transform the cascaded vector into a generated image of pose p and expression e with the same size as the real image. The main role of D_{att} is to be able to distinguish whether the samples are generated samples or real samples, discriminator performs batch normalization after each convolutional layer. The classifier network $C_{\exp}$ is implemented based on VGGNet-19. In our model, the classifier $C_{\exp}$ is trained by using both generated samples and real samples. Our model is implemented based on TensorFlow [14] and is trained with the ADAM optimizer [15]. It has a learning rate of 0.0002 and momentum 0.5. All weights are initialized from a zero-centered normal distribution with a standard deviation of 0.01.

4.3 Experimental Results

Experimental Results on the Multi-PIE Dataset. The red part represents the experimental results of our improved model, and the blue part represents the results of Zhang, as shown in Fig. 5. We can observe that except disgust (DI), the accuracy rate of the other five expressions are higher than the model of Zhang, and the accuracy of four expressions exceeds 93%. Figure 6 shows the each pose accuracy rate in two methods. It can be seen that our model has higher recognition accuracy in any pose than the model of Zhang, and the average recognition accuracy is 93.08%. Because before the facial image is passed to the generator, we use the improved PCD-CNN model to replace the lib face detection model. The improved model can capture facial landmarks well. Thus, the face area is accurately detected, and the facial pose is estimated, which improves the quality of synthetic facial image in different poses and expressions, and makes the classifier easier extract facial features in the image. In order to compare the convergence rate of the two methods, we selected N real images. Through the generative model, we can get $5 \times 6 \times N = 30N$ samples for training the classifier. From Fig. 7, we can see that our model has a faster convergence rate than model of Zhang, indicating that improved PCD-CNN has high computational efficiency. Because improved PCD-CNN is based on the ShuffleNet-v2 framework, it is able to maintain better performance with fewer parameters.

In addition, we also compare the improved model with the current state-of-the-art algorithms [16, 17]. Detailed results are shown in Table 2, we can clearly see the expression recognition accuracy in different poses and average accuracy of each algorithm. The highlighted results in the table are marked in bold. Obviously, our model has the best recognition accuracy at the pan angles +30°, +15°, 0°, and the average accuracy is 93.08%, which is better than all current methods. By careful observation, it can be found that other methods cannot achieve better recognition accuracy when the facial image is on the front. Because our method could synthesize different style facial images to make the training sample more sufficient, so that the classifier can get better performance.

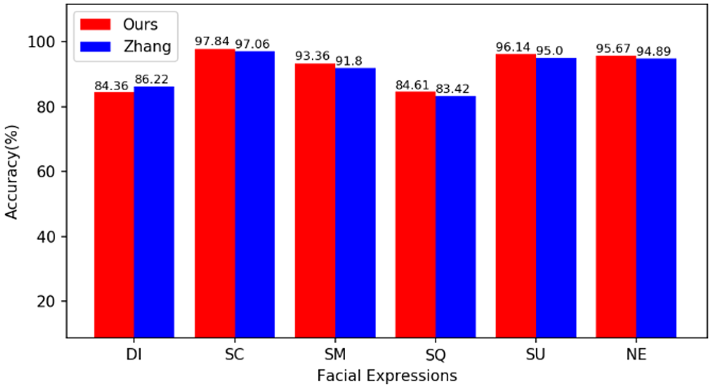


Fig. 5. Expression recognition rate in two methods on the Multi-PIE dataset.

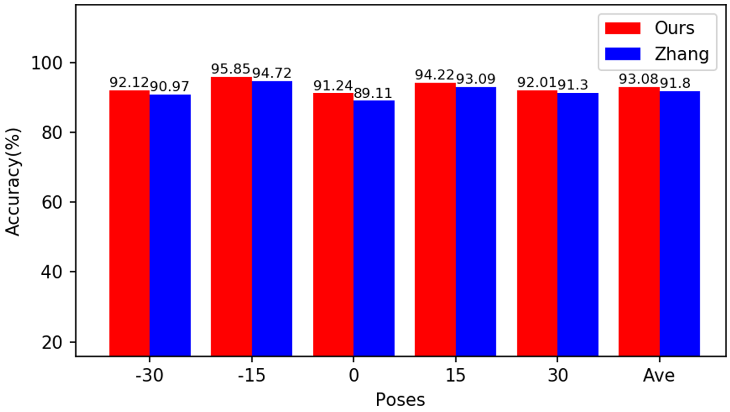


Fig. 6. Expression recognition rate of different poses in two methods on the Multi-PIE dataset.

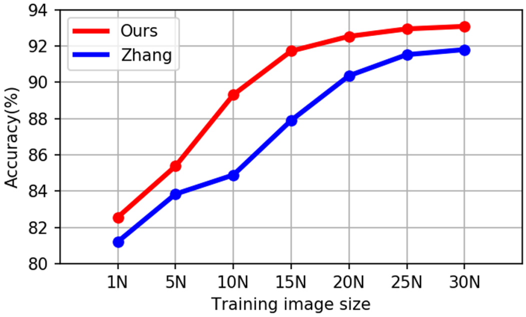
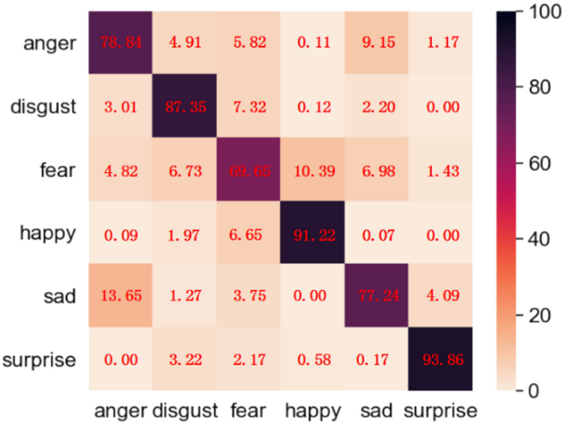


Fig. 7. Comparison of the convergence rate in two methods on the Multi-PIE dataset.

Table 2. Comparison with the latest methods on the Multi-PIE dataset.

Methods	+30	+15	0	−15	−30	Average
DS-GPLVM	90.11	89.97	82.42	**96.96**	**93.55**	90.60
D-GPLVM	86.04	85.96	78.70	93.51	91.65	87.17
GMLPP	87.36	87.22	78.16	94.13	91.86	87.74
GMLDA	85.72	86.64	76.60	94.18	90.47	86.72
GPLRF	86.01	85.66	77.59	93.77	91.65	86.93
MvDA	87.89	87.10	77.51	94.22	92.49	87.84
LDA	87.47	87.07	77.21	94.37	92.52	87.72
kNN	74.78	75.03	68.36	81.74	80.88	76.15
Zhang	91.30	93.09	89.11	94.72	90.97	91.80
Ours	**92.01**	**94.22**	**91.24**	95.85	92.12	**93.08**

Experimental Results on the BU-3DFE Dataset. Figure 8 shows the confusion matrix of our model for FER on the BU-3DFE dataset. It shows that recognition accuracy of surprise reaches the highest at 93.86%, followed by happy at 91.22%. Because when people are surprised or happy, the texture of the facial muscles is obvious. Generator is relatively easy to generate this kind of facial images, and it is easier to extract facial features for classifier to achieve expression recognition than other expressions. From Fig. 9, we can see our improved model convergence speed faster than the model of Zhang, and the training process is more stable. Because we perform mini-batch processing on the discriminator in the model and simplify the calculation method, thereby improving the rate of convergence and stability of the model.



	anger	disgust	fear	happy	sad	surprise
anger	78.84	4.91	5.82	0.11	9.15	1.17
disgust	3.01	87.35	7.32	0.12	2.20	0.00
fear	4.82	6.73	69.65	10.39	6.98	1.43
happy	0.09	1.97	6.65	91.22	0.07	0.00
sad	13.65	1.27	3.75	0.00	77.24	4.09
surprise	0.00	3.22	2.17	0.58	0.17	93.86

Fig. 8. The confusion matrix of our model for FER on the BU-3DFE dataset.

Similarly, we performed many experiments on the BU-3DFE dataset with improved model and compared the results with the latest methods. As shown in Table 3, our average expression recognition rate reached the highest, at 83.03%, which is higher than the method of Zhang (81.20%) and other methods. Because we have improved model of Zhang, not only generating facial images with various poses and expressions, but before passing the facial images to the generator, we use a more superior PCD-CNN to detect face and estimate facial pose. This method not only has fewer parameters, but also has higher calculation efficiency. In order to avoid the mode collapse as much as possible, we use a mini-batch discriminator to process and directly calculate the statistical characteristics of the mini-batch samples, which is more concise. These measures make generator generate facial images with different styles and greatly to improve the quality and quantity of training samples, facilitate the extraction of feature information for classifier, and improve the accuracy of non-frontal FER.

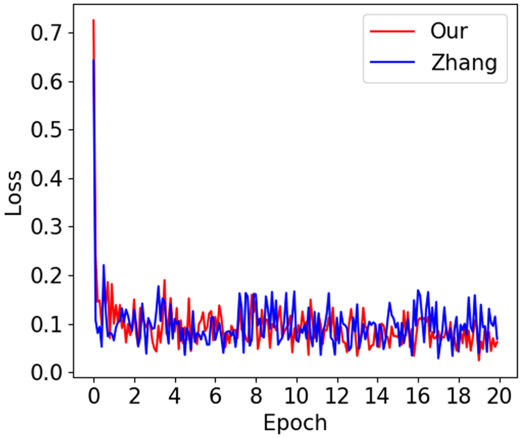


Fig. 9. Comparison of the convergence speed in two methods on the BU-3DFE dataset.

Table 3. Comparison with the current state-of-the-art methods on the BU-3DFE dataset.

Methods	Pan (°)	Tilt (°)	Total	Average
Ekweariri 2017 [8]	(−45, +45)	(−30, +30)	35	75.40
Zhou 2019 [7]	(−45, +45)	(−30, +30)	35	78.64
Jiao 2019 [1]	(−45, +45)	(−30, +30)	35	81.06
Zhang 2018[3]	(−45, +45)	(−30, +30)	35	81.20
Ours	(−45, +45)	(−30, +30)	35	83.03
Ozdemir 2019 [18]	(0, +90)	–	5	80.21
Gacav 2018 [19]	(0, +90)	–	5	78.90
Kim 2019 [20]	(0, +90)	–	5	80.46
Kaleekal 2019 [21]	(0, +90)	–	5	81.32

5 Conclusion

In this paper, we propose a FER model based on PCD-CNN with pose and expression. Firstly, before conveying facial images to the generator, we use the PCD-CNN model to detect faces and estimate facial poses; secondly, in order to accelerate the training speed of the model, the PCD-CNN is based on the ShuffleNet-v2 framework; finally, to avoid the mode collapse in training process, we carry out mini-batch processing on the discriminator in the model and simplify the calculation method. Compared with other methods, our method achieves better results and the training process is more stable. In future work, we will further improve the stability of model training, improve the quality of generated facial image, and consider the impact of other factors on FER, such as illumination changes, identity differences.

Acknowledgment. We would like to acknowledge the support from the National Science Foundation of China (61472095).

References

1. Jiao, Y., Niu, Y., Zhang, Y., Li, F., Zou, C., Shi, G.: Facial attention based convolutional neural network for 2D + 3D facial expression recognition. In: IEEE Visual Communications and Image Processing, pp. 1–4 (2019)
2. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
3. Zhang, F., Zhang, T., Mao, Q., Xu, C.: Joint pose and expression modeling for facial expression recognition. In: Conference on Computer Vision and Pattern Recognition, pp. 3359–3368 (2018)
4. Kumar, A., Chellappa, R.: Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment. In: Conference on Computer Vision and Pattern Recognition, pp. 430–439 (2018)
5. Dong, J., Yuan, J., Li, L., Zhong, X., Liu, W.: An efficient semantic segmentation method using pyramid ShuffleNet V2 with vortex pooling. In: IEEE 31st International Conference on Tools with Artificial Intelligence, pp. 1214–1220 (2019)
6. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–8 (2008)
7. Zhou, W., Zhao, C., Lu, L., Zhao, Q.: Dense correspondence of 3D facial point clouds via neural network fitting. In: IEEE International Conference on Image Processing, pp. 3731–3735 (2019)
8. Ekweariri, A.N., Yurtkan, K.: Facial expression recognition using enhanced local binary patterns. In: 9th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 43–47 (2017)
9. Yang, J., Zhang, F., Chen, B., Khan, S.U.: Facial expression recognition based on facial action unit. In: Tenth International Green and Sustainable Computing Conference (IGSC), pp. 1–6 (2019)
10. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
11. Yin, R.: Multi-resolution generative adversarial networks for tiny-scale pedestrian detection. In: IEEE International Conference on Image Processing (ICIP), pp. 1665–1669 (2019)
12. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs (2016)
13. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation (2017)
14. Zeng, Z., Gong, Q., Zhang, J.: CNN model design of gesture recognition based on Tensorflow framework. In: IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 1062–1067 (2019)
15. Zhang, Z.: Improved adam optimizer for deep neural networks. In: IEEE/ACM 26th International Symposium on Quality of Service, Banff, pp. 1–2 (2018)
16. Kim, S., Kim, H.: Deep explanation model for facial expression recognition through facial action coding unit. In: IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 1–4 (2019)
17. Divya, M.B.S., Prajwala, N.B.: Facial expression recognition by calculating euclidian distance for eigen faces using PCA. In: International Conference on Communication and Signal Processing (ICCSP), pp. 0244–0248 (2018)
18. Ozdemir, M.A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., Akan, A.: Real Time Emotion Recognition from Facial Expressions Using CNN Architecture. In: Medical Technologies Congress (TIPTEKNO), pp. 1–4 (2019)

19. Gacav, C., Benligiray, B., Özkan, K., Topal, C.: Facial expression recognition with FHOG features. In: 26th Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2018)
20. Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. **10**(2), 223–236 (2019)
21. Kaleekal, T., Singh, J.: Facial Expression recognition using higher order moments on facial patches. In: 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–7 (2019)
22. Huang, R., Dong, H., Yin, G., Fu, Q.: Ensembling 3D CNN framework for video recognition. In: International Joint Conference on Neural Networks (IJCNN), pp. 1–7. Budapest, Hungary (2019)
23. Fu, Q., Wang, X., Dong, H., Huang, R.: Spiking neurons with differential evolution algorithm for pattern classification. In: IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, pp. 152–157 (2019)