Research on Face Image Restoration Based on Improved WGAN

Fugang Liu(✉), Ran Chen, Songnan Duan, Mingzhu Hao, and Yang Guo

Heilongjiang University of Science and Technology, Harbin 150022, China
liufugang_36@163.com

Abstract. This article focuses on the face recognition model in real life scenarios, because the possible occlusion affects the recognition effect of the model, resulting in a decline in the accuracy of the model. An improved WGAN network is proposed to repair occluded facial images. The generator in the improved WGAN network is composed of an encoder-decoder network, and a jump connection is used to connect the bottom layer with the high-level feature information to generate missing facial images. The low-level feature information is connected with the deep-level feature information, and the network's ability to extract features and generate pictures is enhanced at the same time. The paper also uses a global discriminator and a local discriminator, taking all the restored pictures as input to measure the overall authenticity, and taking the restored part of the pictures as input to judge whether the content structure is reasonable. After comparison and analysis of experiments, the improved face image has a complete structure and clear content, which is helpful for face recognition with partial occlusion.

Keywords: WGAN · Face recognition · Face image inpainting

1 Introduction

Face recognition model has high recognition effects, such as Deep face [1], FaceNet [2], DeepID [3], etc. However, problems such as jewelry, illumination and hand occlusion may occur in complex environments, resulting in poor recognition effect of face recognition system and difficult to authenticate identity information. As early as the 20th century, the research on image restoration has begun. Image restoration mainly removes the occluded part or restores the missing part of the image, and the semantics and structure of the repaired image are consistent. The repaired image is reasonable and realistic, and it is difficult for the observer to see the repair trace or find that it has been damaged.

Bertalmio and Sapiro [4] proposed a digital image restoration technology for small-scale missing images, simulating the way that professional painters repair damaged or missing parts of images. The main idea of decomposition based image algorithm [5] is to divide the image into two parts: structure and texture for image processing, and finally add the two sub parts to reconstruct the repaired image. The texture synthesis algorithm

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 134–146, 2022.
https://doi.org/10.1007/978-3-031-04409-0_13

based on image block [6] as a whole is to find a pixel according to the texture feature information at the image position to be repaired, select the image block centered on the pixel, search the similar sample block in the unobstructed area, and replace the most similar image information with the area to be repaired.

Traditional face restoration methods based on texture and structure basically learn the image block information close to the junction of occluded and non occluded parts of the face, and then fill in the missing or occluded parts of the image. The disadvantage is the lack of context structure information. How to use the information associated between occluded and unobstructed areas is the key of research. The face restoration algorithm based on deep learning solves this problem. For example, the face restoration method based on context coding [7], but this method has some problems, such as the repaired face image is not clear enough and it is difficult to repair large-area occlusion.

Therefore, this paper proposes a face image restoration algorithm based on deep learning, which uses the improved WGAN network to repair the partially occluded face image. The generator network is built by encoder and decoder, and the jump connection is used to integrate the bottom features and deep features. The discriminator adopts a global discriminator and a local discriminator, Finally, the repaired image generated by the model has clear texture and reasonable structure, and the repaired face image is used for subsequent face recognition tasks, which can effectively improve the accuracy of partially occluded faces.

2 Related Models and Algorithms

2.1 Revolutionary Neural Networks

Revolutionary Neural Networks [8] is one of the important branches of deep learning model. Especially in the field of image processing, CNN is more widely used, and CNN has better image processing ability. The processing of two-dimensional images by revolutionary neural network is invariant to displacement, scale size and rotation. At the same time, CNN has the characteristics of local connection and weight sharing. There is a certain correlation between local images. The weight sharing uses the same convolution kernel in each layer. These two characteristics can improve the generalization ability of model feature extraction and reduce the amount of parameters.

Generally, CNN network usually includes the following parts: convolution layer, activation function, pooling layer and full connection layer. The basic structure of CNN is shown in Fig. 1.

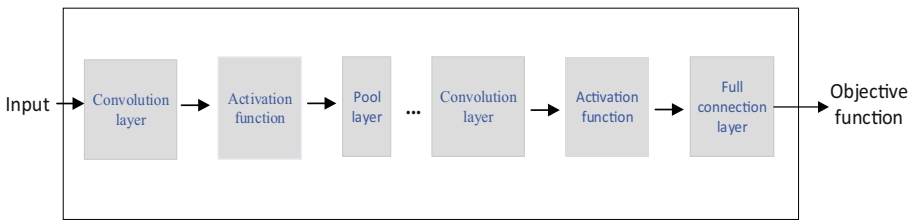

Fig. 1. CNN network structure

2.2 Generative Countermeasure Network

The generic countermeasure network is mainly composed of a generator (G) and a discriminator (D) [9]. As shown in Fig. 2, the uniform distribution Z is trained by the generator G to generate a false sample G(z), judge whether it is true by the discriminator D, and measure the difference between the generated data and the true data distribution, so as to optimize the generator G. The two start to iterate and update each other. Generator G hopes that the generated result can deceive D, and discriminator D hopes to judge G(z) as false, so that generator G can be continuously optimized to generate target samples through each round of confrontation optimization.

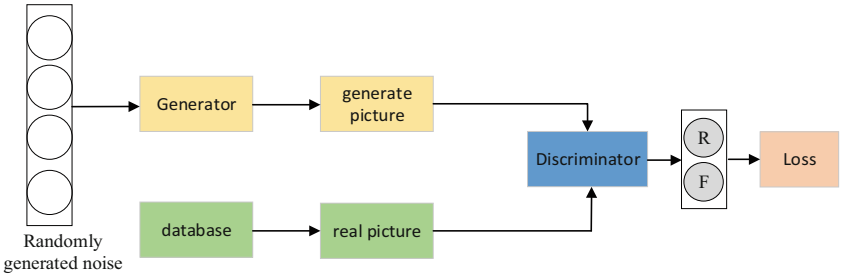

Fig. 2. Generative confrontation network structure

The loss function of Gan network is shown in formula (1):

$$\arg\min_{G}\max_{D} V(G,D) = E_{x\sim P_{\text{data}}}[\log(D(x))] + E_{x\sim P_G}[\log(1 - D(G(\text{x})))] \quad (1)$$

In formula (1), $\arg\min_{G}\max_{D} V(G,D)$ means maximizing the discriminator loss function and minimizing the generator loss function; When x obeys the distribution of real data P_{data} is input into the discriminator, the expectation of $\log(D(x))$ is $E_{x\sim P_{\text{data}}}[\log D(x)]$; When x obeys the distribution of generated data P_{G}, the expectation of $\log(1-D(G(x)))$ is $E_{x\sim P_G}[\log(1 - D(G(\text{x})))]$.

According to formula (1), the essence of the model is to optimize the discriminator D first, then to optimize generator G. So given a generator, maximize $V(G, D)$, the optimal solution D^* of the discriminator is obtained.

2.3 DCGAN Network

Deep convolution neural network (DCGAN) combines convolution neural network (CNN) with generation countermeasure neural network. Using the powerful feature extraction ability of convolution neural network, we can strengthen the ability of network generator to generate pictures and improve the ability of discriminator to distinguish the authenticity of images. DCGAN uses transpose convolution to generate complete image data. In 2010, Zeiler [9] first proposed the concept of transposed convolution. Transpose convolution and standard convolution can be regarded as mutual inverse processes. Convolution describes a many to one process in which the features of the input

image are extracted through the convolution kernel, while transpose convolution, on the contrary, is a one to many process. The low-dimensional feature vector is mapped to the high-dimensional feature vector by transpose convolution to generate a picture.

As shown in Fig. 3(a), in the standard convolution process, the size of the input characteristic image is 3 × 3, convolution kernel of 2 × 2. The step size is 1, no filling, and the size of the output feature map is 2 × 2.

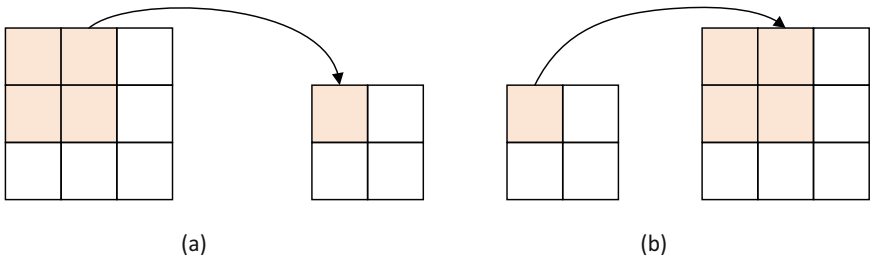

Fig. 3. Standard convolution process and transposed convolution process

2.4 WGAN Network

WGAN solves the problems of gradient disappearance and model collapse of KL divergence and JS divergence in the process of GAN network model training by introducing Wasserstein distance [10] into GAN network model. Wasserstein distance is shown in formula (2):

$$W(P, Q) = \inf_{\gamma \in \prod(P,Q)} E_{(x,y)\sim\gamma}[\|x - y\|] \tag{2}$$

In formula (2), $W(P, Q)$ means the lower bound of all possible distance expectations of joint distribution γ;$\prod(P, Q)$ means a set of joint distributions between P distribution and Q distribution; $E_{(x,y)\sim\gamma}[\|x - y\|]$ represents the expectation of the distance between (x, y) samples with joint distribution γ.

3 Network Design

3.1 Improved WGAN Network

The improved WGAN network structure consists of two parts (three networks), one is the generator model G, which is used to generate images; The other part is composed of two discriminator model D, which is used to distinguish whether an image is a real image or a generated image. The discriminator consists of a global discriminator and a local discriminator. The overall structure of the network is shown in Fig. 4.

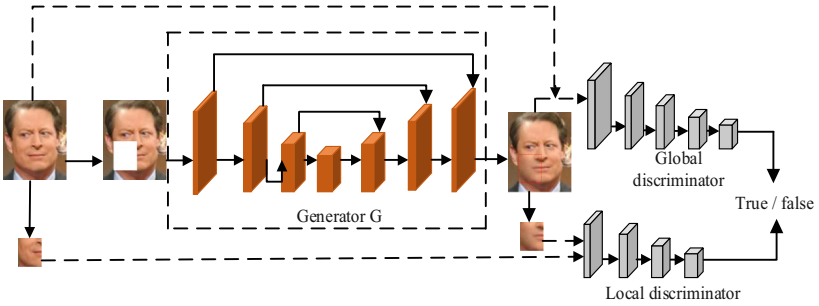

Fig. 4. Improved network structure

The generator is composed of encoder and decoder, which is composed of convolution layer and deconvolution layer. At the same time, the convolution layer of the coding layer is mapped to the corresponding deconvolution layer through jump connection. The coding layer uses the activation function Leaky-Relu to prevent gradient thinning; The rewind layer uses the ReLu activation function, and the last output layer uses the Tanh activation function. The specific structure of the generator is shown in Fig. 5.

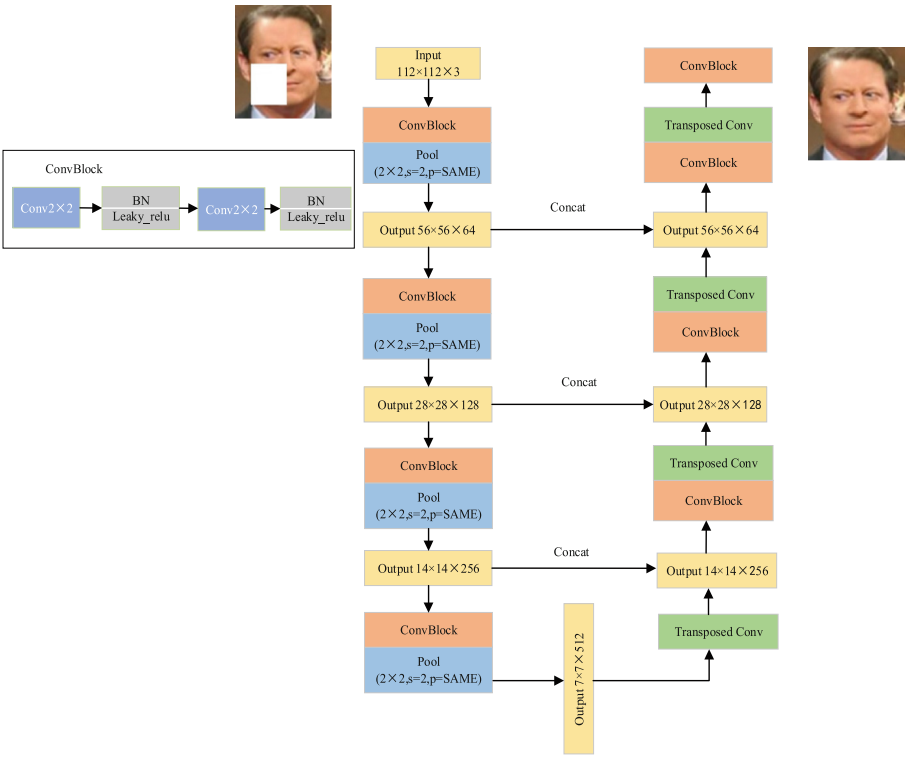

Fig. 5. Generator network structure

The discriminator is composed of Convolutional neural network. By constantly learning against the generator, the ability of the model to distinguish real pictures and generate pictures is improved. The local discriminator inputs the repaired complete image to identify whether the global image is consistent; The local discriminator inputs the image of the repaired part to judge whether the parts are consistent. The specific structure of the discriminator is shown in Fig. 6.

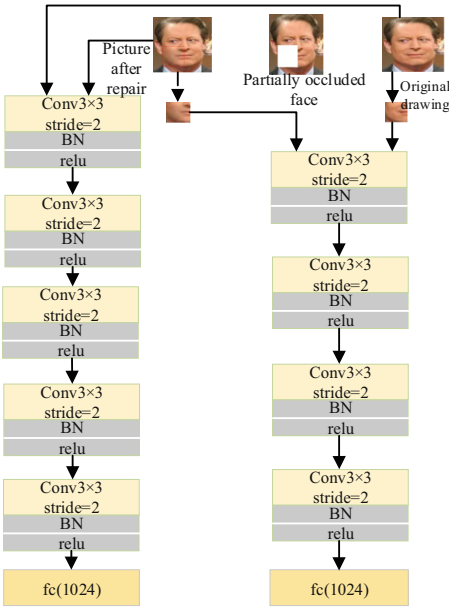

Fig. 6. Discriminator network structure

In the previous section, the loss function of the generator is formula (3), while the generator in this paper is composed of coder and decoder. The coder extracts the picture hidden features and inputs them to the decoder to generate face pictures. Therefore, the loss function generated in this paper is:

$$L_G = -E_{z\sim P(x,y)}[D(G(Z))] \tag{3}$$

In formula (3), L_G represents the Generator loss function, $P(x, y)$ is the feature extracted from the image to be repaired by the encoder; $D(G(Z))$ is the image generated by the generator.

The purpose of discriminator is to distinguish the authenticity of real data and generated data. In this paper, both global and partial discriminators use Wasserstein distance. Therefore, the loss function of the discriminator is mainly composed of two parts. The global discriminator uses the Wasserstein distance to measure the distance between the generated face and the face to be repaired, and some discriminators use the Wasserstein distance to measure the distance between the repaired part and the sample occlusion

area. Therefore, the loss function of the discriminator is listed as follows:

$$\begin{aligned} L_D &= L_g + L_l \\ &= E_{z\sim Pg(x,y)}[D(G(z))] - E_{z\sim Pgdata}[D(x)] \\ &+E_{z\sim Pl(x,y)}[D(G(z))] - E_{z\sim Pldata}[D(x)] \end{aligned} \tag{4}$$

In formula (4), L_D means the discriminator loss function; L_g represents the global discriminator loss function takes the image repair part as the input to calculate its expected value; L_l represents the local discriminator loss function takes the whole repaired picture as the input to calculate its expected value.

3.2 Face Recognition Network

Convolutional neural network learns the characteristics of images through convolution layer and pooling layer, becomes more and more intelligent, and can well liberate human beings from repetitive work. MobileNetV3 [11] combines the important network structure modules of MobileNetV1 [12], V2 [13] and SeNet [14], further improving the effect of the model. This paper mainly studies the improvement of MobileNetV3_small, including adjusting the network structure, loss function and optimizer, applying the improved model to face recognition, and finally carrying out sufficient experiments and result analysis.

Improvement of MobileNetV3 Network Structure

Combined with the MobileNetV3 network module in the previous section, the input image size is changed to 3 * 3, so as to reduce the amount of network operation. Because too many convolution kernels are stacked in the high layer, the effect of the model is reduced and a large number of redundant and highly similar convolution kernels are caused. Therefore, aiming at the above problems, this paper reduces the number of layers of the network and effectively reduces the amount of parameters. In order to reduce the memory access of the network, this paper adopts the deep separable convolution network, improves the running speed of the network by reducing the channel expansion coefficient in the module, reduces the amount of parameters and redundancy of the model, and enhances the attention of the model channel by using compression and activation blocks.

Convolutional neural networks no longer use global average pooling, but 7×7 globally separable convolution substitution. If global pooling is used for feature extraction of the whole picture, it means that the feature importance of the corner is consistent with that of the middle. This is obviously wrong. The corner part of the image only includes the features of a small part of the face, and the middle part is the important feature of the face. In this paper, global separable convolution will be used instead of global pooling to give different importance to different receptive fields. The improved network structure is shown in Table 1, where n represents the number of repetitions of this operation.

Joint Loss Function Based on CenterLoss and Softmax

The loss function consists of two parts, which are model parameters. The Softmax loss function is used as a classifier, and the network iteratively trains and optimizes the loss

Table1. Improved MobileNetV3 network model structure

Input	Operator	exp size	out	SE	NL	s	n
112 × 112 × 3	conv2d,3 × 3	-	64	-	RE	2	1
56 × 56 × 64	depthwise, 3 × 3	16	64	√	RE	1	1
56 × 56 × 64	Bneck, 3 × 3	128	64	√	RE	2	3
28 × 28 × 64	Bneck, 3 × 3	512	128	-	RE	2	1
14 × 14 × 128	Bneck, 3 × 3	256	128	√	HS	1	3
14 × 14 × 128	Bneck, 3 × 3	512	128	√	HS	2	1
7 × 7 × 128	Bneck, 3 × 3	256	128	√	HS	1	2
7 × 7 × 128	Bneck, 3 × 3	256	512	√	HS	1	1
7 × 7 × 512	LinearGDConv7 × 7	-	512	√	-	1	1
1 × 1 × 512	Linearconv1 × 1	-	128	√	-	1	1

function to obtain the global optimal solution. In the second half, CenterLoss calculates the distance between the sample features and the middle features as the loss function. At the beginning of network training, the feature center is randomly selected. After the network optimizes the loss function, the feature center is updated, and the network model is continuously updated until the network reaches the optimum. When the super parameter is, the loss function is Softmax loss function, and it can play the constraint ability of the loss of the control center.

$$L_{loss} = L_{soft\max} + L_{centerloss}$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{W_{y_i}^T x(i)}}{\sum_{l=1}^{k} e^{W_l^T x(i)}} + \lambda\cdot\frac{1}{2}\sum_{i=1}^{n}\left\|x_i - c_{y_i}\right\|_2^2 \tag{5}$$

4 Experiment and Result Analysis

The training set of face restoration experiment in this paper adopts the large-scale CelebA face data set published by the Chinese University of Hong Kong, which contains a total of 10177 celebrity face images, a total of about 202599 face images, and the size of the original face image is 178 * 218. The images in face data have done a lot of labeling work, including face key point labeling, face rectangle labeling and face attribute labeling information. Therefore, CelebA data set is widely used in various tasks related to face in the field of computer vision, such as face attribute recognition, face detection and key point detection.

The detected partial face images are uniformly scaled to size 112 × 112 × 3. We process the pictures of CelebA data set, add rectangular box occlusion, simulate the occlusion that may occur in the real situation, and then use this data set for network training, including the following examples of facial occlusion (Fig. 7):

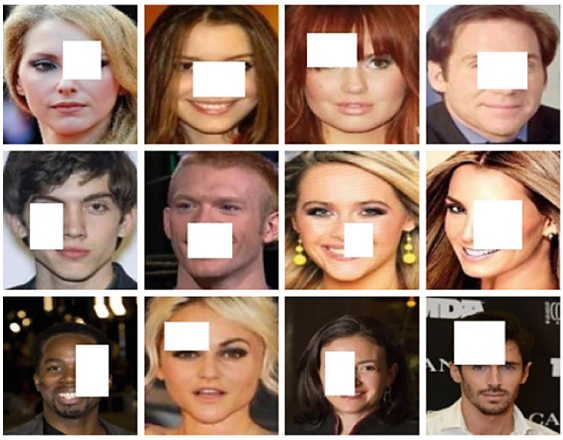

Fig. 7. The processed data set

Peak signal-to-noise ratio measure (PSNR). It mainly calculates the difference between the pixel values of the original picture and the two pictures after occlusion repair. It can measure whether the repaired image has distortion and the difference between images. The calculation process is as follows (6).

$$PSNR = 10 \times \log_{10}[\frac{(2^n - 1)^2}{MSE}]$$
$$MSE = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left\|f(i,j) - f(\hat{i},j)\right\|^2 \tag{6}$$

In formula (6), n is the number of bits of the pixel value (generally is 8), that is, the gray scale of the pixel value is 256; MSE is mean square error between the repaired image and the original image;H and W represent the Height and width of image respectively.

Structural similarity index (SSIM) is an index to calculate whether the structure of the original image is similar to that of the image after occlusion repair. It is mainly measured from three aspects: brightness, contrast and structure。As shown in Eq. (7) for the calculation of SSIM, The larger the calculated SSIM value, the more similar the two images are.

$$SSIM(x,y) = \frac{(2u_x u_y + C_1)(2\sigma_{xy} + C_2)}{(u_x^2 + u_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{7}$$

In formula (7), u_x, u_y represent the mean of image x and y, respectively; σ_x, σ_y represent the variance of image x and y, respectively; σ_{xy} is the Covariance of X and Y.

Experiment 1

The face data set CelebA is divided into training set and test set according to 9:1. Before using the training set to train the network model, the image is processed to add occlusion, and then the original network and the improved network model are trained respectively.

After the model is trained, the same test set is used, and the repaired images and the original images are evaluated by PSNR and SSIM indicators. The calculation results are shown in Table 2.

Table 2. Comparison between the original algorithm and the improved algorithm

Evaluating indicator	Original algorithm	This paper improves the algorithm
SSIM	0.88	0.92
PSNR	27.48	29.58

For the comparison between the repaired images of the original algorithm and the improved algorithm, see Fig. 8. From left to right, there are occluded images, repaired images and original face images. The above three images are the face repaired by the original algorithm, and the lower half is the improved face repair image. It can be directly seen from the figure that the effect of the face repair algorithm repaired in this paper is clearer and more complete than that of the original algorithm.

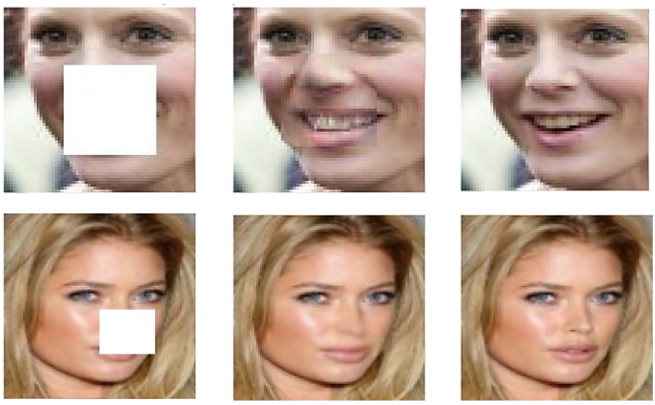

Fig. 8. Comparison of repair results

Experiment 2

It mainly proves the performance of face recognition under occlusion of different area sizes. The experiment uses LFW public data set for evaluation. On the test data set samples, four area sizes of occlusion of 10%, 20%, 30% and 40% are used respectively, as shown in Table 3:

Table 3. Results of face recognition with different occlusion areas

Occlusion area / algorithm	10%	20%	30%	40%
Improved mobilnetv3_ Small + joint loss function	93.86%	91.21%	87.23%	82.3%
Paper algorithm	94.56%	93.65%	92.35%	90.07%

Draw a broken line diagram according to the data in Table 2, as shown in Fig. 9.

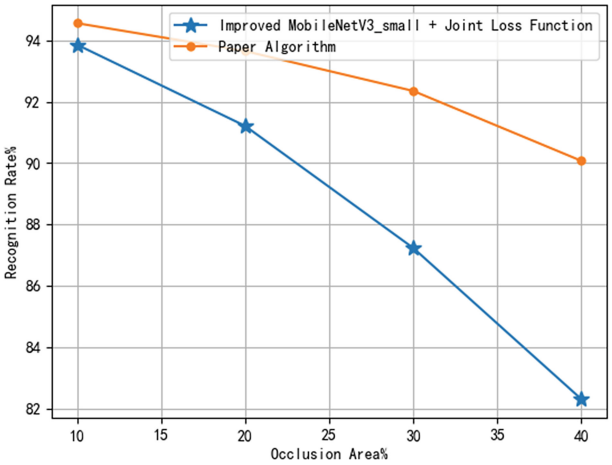

Fig. 9. The recognition accuracy of different occlusion areas

As can be seen from the experimental results shown in Fig. 9, the appearance of occlusion increases the difficulty of face recognition algorithm. With the increasing occlusion area, this algorithm can maintain a good accuracy, and still has a stable accuracy when the occlusion area is greater than 20%. Therefore, this algorithm is more robust to face recognition with partial occlusion.

5 Conclusion

In order to make face recognition still have a high recognition rate in the case of occlusion, this paper uses the improved WGAN network to repair the partially occluded area and increase the image face features, which can effectively improve the accuracy of face recognition.

(1) The process of mutual confrontation learning between GAN network generator and discriminator is applied to the problem of face restoration. The improved

WGAN generator adopts the structure of encoder and decoder. The network learns the global information of the picture by extracting the image features, and the generated picture semantic information is more complete.

(2) At the same time, the generator uses jump connection to map the convolution layer of the coding layer to the corresponding deconvolution layer, and fuse the bottom feature information with the deep feature information, which will make the generated picture clearer.

(3) The discriminator adopts global discriminator and local discriminator, which can ensure the correctness of global information and local repair information at the same time。

(4) Through the experimental comparison and the comparison of the generated repaired images, the images repaired by the face repair algorithm proposed in this paper are clearer and semantically complete.

Acknowledgements. This work has been partially supported by Heilongjiang Science Foundation Project (LH2021F052).

References

1. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014). https://doi.org/10.1109/CVPR.2014.220
2. Schroff, F., Kalenichenko, D., Philbin, J: FaceNet: a unified embedding for face recognition and clustering. IEEE (2015)
3. Ouyang, W., Zeng, X., Wang, X., et al.: Deformable Deep Convolutional Neural Networks for Object. arXiv:1412.5661v2 (2015)
4. Bertalmoí, M., Sapiro, G., Caselles, V., et al.: Image Inpainting. In: Proceedings of Annual Conference on Computer Graphics & Interactive Techniques, pp. 417–424 (2000)
5. Li, P., Wang, H., Li, X., Zhang, C.: An image denoising algorithm based on adaptive clustering and singular value decomposition. IET Image Process. **15**, p. 3 (2021)
6. Drori, I., Cohen-Or, D., Yeshurun, H.: Fragment-based image completion. ACM Trans. Graph. **3** (2003)
7. Angah, O., Chen, A.Y.: Removal of occluding construction workers in job site image data using U-Net based context encoders. Autom. Constr. **119** (2020),
8. Sun, Y., Wang, X., Tang, X: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 891–1898 (2014)
9. Yonghao, M., Ming, Z., Jing, L., Yaguo, L.: Application of an improved maximum correlated kurtosis deconvolution method for fault diagnosis of rolling element bearings. Mech. Syst. Signal Process. **92**, 173–195 (2017)
10. Arjovsky, M., Chintala, S., Bottou, L: Wasserstein GAN. CoRR,abs (2017)
11. Howard, A., Sandler, M., Chen, B., et al.: Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision, IEEE (2020)
12. Sandler, M., Howard, A., Zhu, M., et al.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)

13. Howard, A.G., Zhu, M., Chen, B., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. Int. J. Intell. Sci. **11** (2017)
14. Liu, Y.F., Meng, L., Qiu, D., et al.: Multi-task squeeze-and-excitation networks for pedestrian attributes recognition. Sci. Techno. Eng. **19**(24), 237–241 (2019)