Facial Image Inpainting Algorithm Based on Attention Mechanism and Dual Discriminators

Jianchu She and Ying Liu
Center for Image and Information Processing, Xi'an University of Posts & Telecommunications, Xi'an, China
Email: 617895826@qq.com, liuying_ciip@163.com

***Abstract*—In recent years, deep learning technology is widely used in the field of facial image inpainting. The existing methods are prone to structural distortion, semantic capture blur and repair result distortion when dealing with large occluded areas. This paper proposes a facial image inpainting algorithm based on the Generative Adversarial Networks (GAN) framework, which combines the self-attention mechanism and the global local discriminator. First, the important features of the original image are captured by the attention layer in the generation network to obtain a larger receptive field and discard irrelevant information. In the discriminator network, the global and local discriminators are combined, and the good adaptability of the two discriminators to the overall semantics and local edge details is used to perform feedback training on the repair network and images. The final experimental results show that the algorithm in this paper has better repairing effect and higher training efficiency than other algorithms, and is superior to other existing algorithms in subjective visual perception and objective evaluation index.**

***Index Terms*—facial image inpainting, generative adversarial networks, attention mechanisms, global and local discriminator, deep learning**

I. Introduction

Facial image recognition technology [1] is widely used in various fields in today's society, but the recognition efficiency will be greatly reduced when the processed image is blurred and occluded. After using the face de-occlusion technology on the occluded image, the overall structure and specific details of the missing area can be effectively restored, and a generated image with higher similarity to the original image can be generated. In the field of criminal investigation, face analysis, comparison and retrieval is one of the important methods for public security personnel to analyze criminal investigation cases [2], [3].

Early algorithms generate repaired images through overall semantics. In order to ensure the continuity of the occluded and unoccluded areas, M. Bertalmio [4] *et al.* proposed a new digital inpainting algorithm for static images, which uses the overall semantics to associate the part to be filled, which fully improves the integrity of the image inpainting. Barnes [5] *et al.* proposed an interactive image editing tool based on a random algorithm to quickly determine the approximate match between image patches. Ballester [6] *et al.* introduced a joint interpolation algorithm based on image gray level and gradient direction, which can process all missing areas of data at the same time. Based on the deep learning model, D. Pathak [7] and others created an unsupervised visual feature learning algorithm based on contextual content, which is one of the first representative algorithms to apply deep learning to image inpainting. J. Yu [8] *et al.* proposed a feedforward generation network with a content-aware layer using a contextual content encoder, which can extract features similar to the area to be repaired from a far area. Aiming at the problem of de-occlusion and repair of facial images, Iizuka S. [9] uses global and local discriminators to ensure that the generated image is in the overall semantic structure conforms to the original image. But when repairing boundary regions, structural distortions are often caused. J. Yu *et al.* [8] includes a feedforward generation network with a content-aware layer which can extract features similar to the region to be repaired from the far region. Nazeri K. [10] and others proposed a two-order adversarial edge connection model, which successfully solved the problem of edge blur. In order to select the most consistent semantic structure from multiple inpainting images at the same time, Zheng C. [11] proposed a multi output image inpainting method, which can generate a variety of reasonable solutions for image inpainting. Yang [12] and others combined the image de-occlusion repair generation sub-network with the prediction key point sub-network, which can more effectively capture the specific positions of facial organs on different facial structures.

In order to effectively solve the problem that the local repair structure is blurred and the key features around the repair target cannot be accurately extracted in the face image de-occlusion repair, this paper proposes a face de-occlusion algorithm combining the self-attention mechanism and the global local discriminator. This algorithm adds a local discriminator to the overall generative confrontation network to cooperate with the global discriminator to make the final repair result more accurate regardless of the overall or local details. Adding the self-attention mechanism layer can more efficiently and accurately extract features such as texture, color and

Manuscript received August 12, 2021; revised December 14, 2021.

doi: 10.18178/joig.10.1.43-49

structure around the occluded block area, so that the repaired image has a reasonable structure and conforms to the subjective visual perception of humans.

II. Algorithm Description

A. Overall Network Architecture

The overall face image inpainting network architecture of this article consists of two parts: an image inpainting network based on a fully convolutional neural network and a discriminator network. As Fig. 1 shows, the discriminator network is divided into two parts, a local discriminator and a global discriminator, which are used to improve the overall consistency of structure and clarity of details. A self-attention layer is added to the image restoration network, which can capture more important semantic information from the occluded image during the restoration process. This algorithm model generation network part uses Weighted Mean Squared Error (MSE) [13] as the loss function to calculate the difference between the original image and the generated image pixels. The expression is shown in (1):

$$L(x, M_c) = \| M_c \cdot (C(x, M_c) - x) \|^2 \tag{1}$$

where $\cdot$ is the pixel multiplication, $\| \|$ is the Euclidean norm.

The discriminator network uses the GAN loss function, and its goal is to maximize the probability of similarity between the generated image and the original image. The expression is shown in (2):

$$\min_C \max_D E[\log D(x, M_d) + \log(1 - D(C(x, M_c)), M_c)] \tag{2}$$

where D is the discriminator and G is the generator. Finally, combining these two loss functions, we get:

$$\min_C \max_D E[L(x, M_c) + \alpha \log D(x, M_d) + \alpha \log(1 - D(C(x, M_c), M_c))] \tag{3}$$

This paper uses the random occlusion block of the CelebA face data set as the training data of the network model. The face de-occlusion inpainting model is divided into the following steps which Fig. 1 has already shown: First, the original image and the randomly occluded face image are used as the input of the network, the face image generation confrontation inpainting network model is trained to make the repair network generator generate repairs after the face image. Then the output inpainting image is sent to the local and global discriminators for true and false judgment. Finally, the true and false weights in the range of 0-1 obtained by weighting the 2 discriminators are fed back to the image inpainting network for retraining.

After the training of network repairing for face de-occlusion, the trained model can be used for face image de-occlusion test. In the test process, only the original test image needs to be input at the input end, and then the model will automatically generate random occlusion. Finally, the final repaired face image is generated through the generation of repair network.

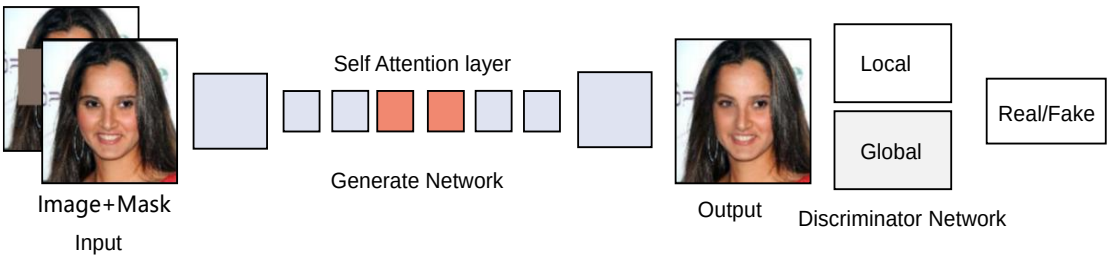

Figure 1. Network structure of face de-occlusion based on attention mechanism and global local discriminator.

B. Generative Adversarial Networks

In 2014, Goodfellow [14], [15] *et al.* proposed Generative Adversarial Network (GAN) based on the idea of zero-sum game, which is an emerging technology of semi-supervised and unsupervised learning. As shown in Fig. 2, GAN provide a way to perform deep learning without in-depth annotation of training data. The representations it learns can be used for image synthesis, semantic image editing, style conversion, image super-resolution, etc. The two major models that make up the GAN network are the generator model D and the discriminator model G. In the generative network G, the input z is pure random noise sampled from the prior distribution p(z). It is usually selected as Gaussian distribution or uniform distribution, and the input z is as realistic as possible. The input of the discriminator network D is a real sample or a generated sample, and the discriminating network is trained and the network parameters are updated by generating the image x to improve the discriminating ability of the discriminating network for true and false pictures. Finally generator's generated data is infinitely close to the real data, so that the discriminator cannot at the moment accurately identify and generate data. The objective function formula (4) is as follows:

$$\max_G \min_D V(G, D) = E_{P_{data}(x)} \lg D(x) + E_{p_g(x)} \log(1 - D(x)) \tag{4}$$

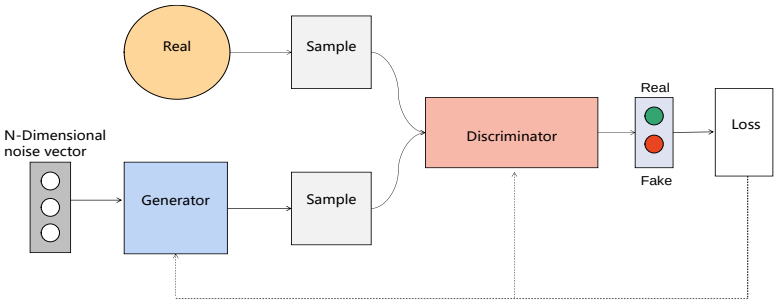

Figure 2. Generative adversarial network structure.

C. Global Local Discriminator

Pathak [7] *et al.* proposed an unsupervised visual feature learning algorithm based on a generative adversarial network. The discriminator network part can only adjust and repair the overall image structure through a unique discriminator. When it comes to repairing high-resolution images or details around obscured blocks, the repaired images are prone to not meet the semantics or blur. Iizuka S. [9] and others subdivided the discriminator network into a global discriminator network and a local discriminator network. The global discriminator takes the complete image as input, recognizes the scene consistency of the restored image, and makes the overall image structure reasonable and consistent with actual semantics. The local discriminator only observes and discriminates on a quarter-size area of the original image centered on the filled area block, and is used to identify local consistency, so that the edge part and the edge detail part of the repaired area are more accurate. These two network modules are based on the convolutional neural network [16], which compresses the input inpainting image into a feature vector. Finally, the outputs of the two networks are fused together through the connection layer to obtain a probability that the predicted image is true or false. The overview of discriminator network is shown in Table I, II, III:

TABLE I. LOCAL DISCRIMINATOR NETWORK STRUCTURE

Type	Kernel size	Stride	Output
Convolution	5×5	2×2	64
Convolution	5×5	2×2	128
Convolution	5×5	2×2	256
Convolution	5×5	2×2	512
Convolution	5×5	2×2	512
Full-Connect	-	-	1024

TABLE II. GLOBAL DISCRIMINATOR NETWORK STRUCTURE

Type	Kernel size	Stride	Output
Convolution	5×5	2×2	64
Convolution	5×5	2×2	128
Convolution	5×5	2×2	256
Convolution	5×5	2×2	512
Convolution	5×5	2×2	512
Convolution	5×5	2×2	512
Full-Connect	-	-	1024

TABLE III. CONNECTION LAYER NETWORK STRUCTURE

Type	Kernel size	Stride	Output
Contact	-	-	2058
Full-Connect	-	-	1

The global discriminator consists of 6 convolutional layers and 1 fully connected layer. It takes a 256*256 image as input and outputs a 1024dimensional vector. All convolutional layers use 2×2 pixel steps to reduce image resolution Compared with the repair network, all convolutions use a filter of size 5*5. The local discriminator follows the same pattern, the difference is that the input is a 128×128 pixel patch, which surrounds the repair area block. Since the initial input resolution is half of the global discriminator, the first convolutional layer is omitted, and the output is still a 1024-dimensional vector, which represents the local context information around the repair area block. Finally, the outputs of the global and local discriminators are connected together to form a 2048-dimensional vector, which is processed by a fully connected layer to output a continuous value. The fully connected layer uses the Sigmoid [17] activation function to make the output value in the range of [0, 1], which represents the probability that the image is real instead of being repaired.

D. Attention Mechanism

In neural networks, convolution kernel is usually local, so in order to increase the receptive field, the method of stacking convolution layer is often used, but in fact, this method is not efficient. In order to capture more semantic information, we add the self-attention mechanism [18], that is to let the system network use its own attention to ignore irrelevant information and focus on key information like human beings. Through a series of attention distribution weight parameters to emphasize or select the important information of the target processing object, and suppress some irrelevant details. The essence of the self-attention mechanism function can be described as a query to a series of key-value pairs. The calculation is mainly divided into three steps as shown in (5), (6), (7): The first step is to calculate the similarity between the query and each key to obtain the weight, where Q and K are query-key mapping pairs:

$$f(Q,K_i)=\begin{cases} Q^T K_i \\ Q^T W_a K_i \\ W_a[Q;K_i] \\ v_a^T \tan(W_a Q + U_a K_i) \end{cases} \tag{5}$$

The second step uses the softmax function to normalize these weights:

$$a_i = soft\max(f(Q,K_i)) = \frac{\exp(f(Q,K_i))}{\sum_j \exp(f(Q,K_j))} \quad (6)$$

In the last step, the weight and the corresponding key value are weighted and summed to obtain the final attention value.

$$Attention(Q,K,V) = \sum_i a_i V_i \quad (7)$$

The calculation formula of the self-attention mechanism used in the algorithm in this paper is as shown in (8): use the score of the self-attention mechanism to first calculate the self-attention map according to the features f_d of the middle layer of the decoder;

$$\alpha_{j,i} = \frac{\exp(X_{ij})}{\sum_{i=1}^{N} \exp(X_{ij})} \quad (8)$$

where $X_{ij} = Q(f_{di})^T Q(f_{dj})$, N is the number of pixels, $Q(\mathrm{f_d}) = W_q \mathrm{f}_d$, W_q which is a convolution filter with a size of 1*1.

$$e_{dj} = \sum_{i=1}^{N} \alpha_{j,i} f_{di}, \quad z_d = \gamma_d e_d + f_d \quad (9)$$

As shown in (9), the algorithm in this paper uses a proportional parameter γ_d to balance the e_d and f_d. e_d is the final self attention result score.

With self-attention layer, details can be generated using cues from all feature locations. And moreover, the discriminator can check that highly detailed features in distant portions of the image are consistent with each other. The attention mechanism gives more power to both generator and discriminator to directly model the long-range dependencies in the feature maps. Compared with residual blocks with the same number of parameters, the self-attention blocks also achieve better results. The training is not stable when we replace the self-attention block with the residual block. Using the self attention mechanism, the global reference can be realized in the process of model training and prediction. So our model has a good bias variance trade-off, so it is more reasonable.

E. Image Inpainting Network

The image inpainting network in the overall network architecture is based on a fully convolutional neural network. The input of the repair network is an RGB image with a binary channel. The binary channel represents the image repair mask (1 pixel to be repaired) and the output is an RGB image. The generation network uses a 12-layer convolutional network to encode the original image (removing the part that needs to be filled) to obtain a grid of 1/16 of the original image. Then the grid is decoded using a 4-layer convolutional network to obtain the restored image. During the entire repair process, it is not hoped that any changes will occur in parts other than the repaired area, so the output pixels outside the repaired area are restored to the original input RGB value. The overall structure allows to reduce the amount of memory and calculation time by reducing the resolution before further processing the image, and then restore to the original resolution through deconvolution, which is composed of an expanded convolutional layer with small steps. Dilation convolution uses a decentralized convolution kernel, allowing a larger input area to be used to calculate each output pixel. Because the contextual content is essential for realism, by using dilated convolution, it can integrate multi-scale contextual information without loss of resolution and no need to analyze the rescaled image. Compared with the standard convolutional layer, it can effectively "see" a larger range of input images when calculating each output pixel, and capture more accurate context information. And after using hole convolution [19], the inpainting network model can calculate occlusion blocks larger than 99*99 pixels, effectively dealing with large-area occlusion problems.

III. Experimental Results

According to the face image inpainting algorithm proposed in this paper, through a large number of experiments and data set verification, it is concluded that this method has certain effectiveness and superiority in processing and repairing face image.

A. Data Set and Parameter Settings

The overall experiment is based on the Ubuntu16.04 platform, using python3.6 and pytorch framework as the programming environment, the CPU information is Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, the graphics card model is NVIDIA TITAN Xp, and the graphics card memory is 12G.

In this paper, CelebA [20] face data set is used for model adaptive training of the face image restoration network, and it takes 90 hours to complete the training of the entire data set. Input a 256*256 size image at the network input, and a 128*128 fixed regular occlusion block will be generated in the center of the image as the occlusion image and enter the repair network together with the original image. The network model optimizer adopts the Adam strategy to optimize the parameters, the Batch-Size is set to 16, and the training learning rate is 0.00001.

B. Experimental Results

The contrast algorithms used in the experiment are the image inpainting algorithms based on deep learning in recent four years, which are the context attention mechanism (Attention) method; the context-encoder method; the global and local discriminators (Globally and Locally) method. These three methods are all tested on the CelebA face dataset used by the algorithm in this paper. In order to verify the superiority of this algorithm, this paper uses the peak Signal-to-Noise Ratio (PSNR) [21] and Structural Similarity (SSIM) [22] to measure the final experimental restoration effect from an objective point of view. PSNR is the most common and widely used image objective evaluation index. It is usually used

to evaluate the quality of an image after restoration compared with the original image. The higher the PSNR, the smaller the distortion after compression. The general value range is between 20-40, the larger the value, the better the image quality. SSIM is an index to measure the structural similarity of two images. The larger the value, the better, and the maximum is 1. PSNR evaluates image quality based on error sensitivity, while SSIM compares the difference between the restored image and the original image in terms of brightness, contrast, and overall image structure.

After using the CelebA dataset for model training for each algorithm, 5 face images are randomly selected from the dataset, and a fixed rule occlusion of 128*128 is generated in the middle of the image. The repaired image generated by different algorithms is shown in Fig. 3.

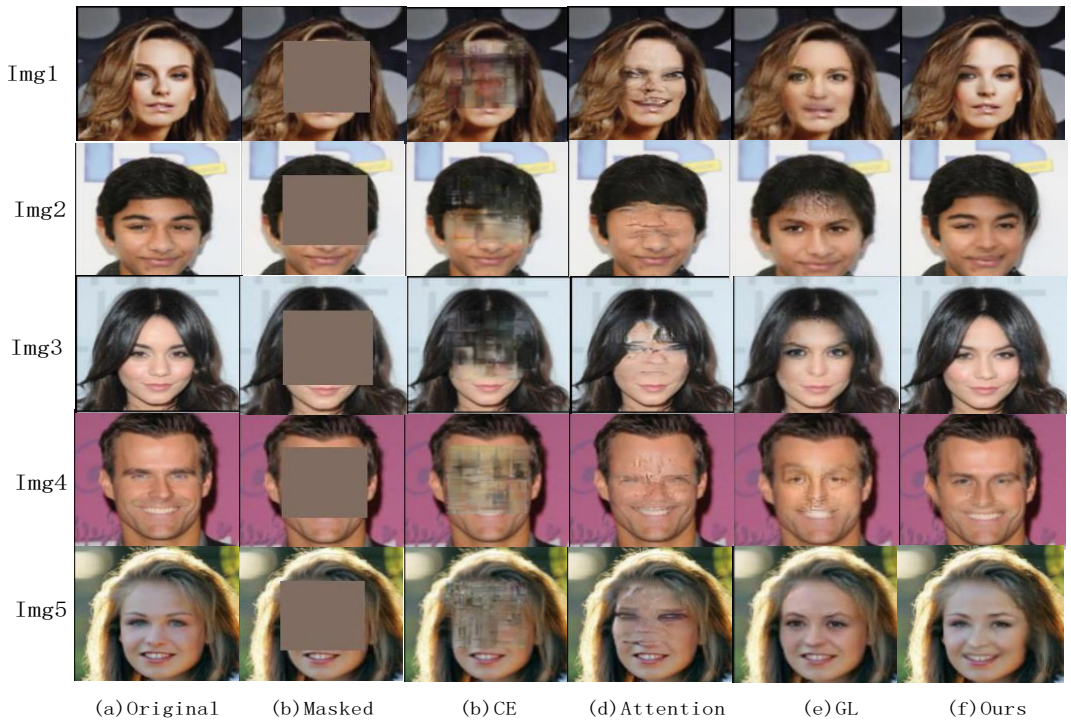

Figure 3. The repair results of different algorithms when the center is occluded.

The occlusion area is 128*128 regular rectangular occlusion, which makes the comparison experiment done by the Context-Encoder algorithm, Attention algorithm, Globally and Locally algorithm and the algorithm in this paper more convincing. In the experiment, the effect of the CE algorithm is the worst among the four algorithms compared, and the gap with the original image is large. The repaired image generated by this algorithm is seriously blurred in the occlusion matrix area, and the overall facial structure has not been successfully repaired. In contrast, the Attention algorithm, GL algorithm and the algorithm in this paper are more optimistic about the completion of face restoration. These three algorithms can all repair and generate face structures that conform to people's objective vision, and the generated facial features and hair are better in line with the real face. For the Attention algorithm, because the algorithm ignores the detail occlusion area, large or small detail distortion will occur at the edge of the detail. This algorithm suppresses the generation of useless information through the attention mechanism, as shown in 2(d), the repair of facial hair is closer to the original image. However, because the GL algorithm is not able to capture and emphasize the key information points in the face, the overall semantics in the generated images are often consistent but the details are biased, as shown in 2(e). Compared with the first three algorithms, the algorithm proposed in this paper can accurately restore the semantic status of the original image, give the observer a good sense of subjective discrimination, and the content of the repaired image has continuity, and the distortion of image tearing is weak. As shown in 1(f), this algorithm has better adaptability for more detailed and richer occlusion images. Through the overall repair effect diagram, it can be seen that the algorithm in this paper is more reasonable to restore the shape of the face, mouth and eyes, and the repair of the size and color of the eyes is more in line with the actual situation.

PSNR and SSIM are used to compare the similarity between the repaired image and the original image. The results are shown in the Table IV, V below.

TABLE IV. COMPARISON OF PSNR EVALUATION INDICATORS OF DIFFERENT ALGORITHMS

	Globally and Locally	Context Encoder	Attention	Ours
Img 1	22.50	18.47	24.33	27.57
Img 2	22.65	22.20	23.26	26.51
Img 3	25.68	25.64	26.11	26.28
Img 4	25.44	24.35	26.48	27.09
Img 5	23.91	17.80	22.92	24.72
Avg.	24.03	21.69	24.62	26.43

TABLE V. COMPARISON OF SSIM EVALUATION INDICATORS OF DIFFERENT ALGORITHMS

	Globally and Locally	Context Encoder	Attention	Ours
Img 1	0.672	0.822	0.889	0.892
Img 2	0.805	0.864	0.879	0.910
Img 3	0.834	0.886	0.894	0.913
Img 4	0.701	0.866	0.853	0.873
Img 5	0.693	0.831	0.860	0.891
Avg.	0.741	0.853	0.875	0.896

It can be clearly seen from the two tables that the algorithm evaluation index in this paper is the best under the same test dataset. Among them, the CE algorithm has poor adaptability to facial images, so the values in the two evaluation indicators are low, and the repair effect appears serious distortion, and the overall semantics of the repaired regional image does not match. In the GL method, a local discriminator is added to the details of the repair area, so that the semantics of the eyes and nose parts match the original image, and the overall structure is correct. However, some key information is not extracted due to the lack of attention mechanism. The attention method uses a single global discriminator, which leads to local distortion under the condition of normal global semantics. Because the algorithm in this paper focuses on the extraction of key facial details and the accurate capture of local regional semantics and structure, it can be seen that it can produce the best results both subjectively and objectively, indicating that the performance of the algorithm proposed in this paper is the best.

IV. CONCLUSION

In this paper, we propose a face image de-occlusion algorithm which combines attention mechanism and global local discriminator. The whole algorithm network model is composed of a generative confrontation network which integrates global and local discriminators and the attention layer in the generative network. The results show that the proposed method can generate the repair image with complete structure and subjective vision, and has a high value of objective evaluation index. The experimental results show that the proposed algorithm has good repair effect for face image, good adaptability and restoration effect for large block occlusion, and can repair more details without semantic change.

In future research, it is necessary to closely integrate the needs of real life with the facial image restoration algorithm, and specify specific solutions that match each other for various scenes in life. And in future research, focus on collecting Asian faces and adapting the inpainting algorithm to Asian faces images through training, so that the algorithm can be better used in the field of life and criminal investigation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Jianchu She and Ying Liu conducted the research; Jianchu She and Ying Liu revised and edited the paper together; Jianchu She and Ying Liu acquired funding; all authors have approved the final version.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (No. 61802305).

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans on Pattern Analysis & Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, 2006.
[2] R. Chandra, *et al.*, "Texture synthesis with recurrent variational auto-encoder," ArXiv preprint, arXiv: 1712.08838, 2017.
[3] R. Gross, "Face databases," in *Handbook of Face Recognition*, Springer, 2005, pp. 301-327.
[4] M. Bertalmio, *et al.*, "Image inpainting," in *Proc. the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 417-424.
[5] C. Barnes, *et al.*, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. on Graphics*, vol. 28, no. 3, p. 24, 2009.
[6] C. Ballester, *et al.*, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Trans. on Image Processing*, vol. 10, no. 8, pp. 1200-1211, 2001.
[7] D. Pathak, *et al.*, "Context encoders: Feature learning by inpainting," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536-2544.
[8] J. Yu, *et al.*, "Generative image inpainting with contextual attention," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 5505-5514.
[9] S. Iizuka, E. Simo-Serra, snd H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. on Graphics*, vol. 36, no. 4, pp. 1-14, 2017.
[10] K. Nazeri, *et al.*, "EdgeConnect: Generative image inpainting with adversarial edge learning," ArXiv preprint, arXiv: 1901.00212, 2019.
[11] C. Zheng, T. J. Cham, and J. Cai, "Pluralistic image completion," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438-1447.
[12] Y. Yang, *et al.*, "LaFIn: Generative landmark guided face inpainting," arXiv preprint, arXiv:1911.11394, 2019.
[13] M. Greiff, A. Robertsson, and K. Berntorp, "MSE-optimal measurement dimension reduction in Gaussian filtering," in *Proc. IEEE Conference on Control Technology and Applications*, 2020.
[14] I. Goodfellow, *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, pp. 2672-2680, 2014.
[15] X. Sun and X. L. Ding, "Data augmentation method based on generative adversarial networks for facial expression recognition sets," *Computer Engineering and Applications*, vol. 56, no. 4, pp. 115-121, 2020.
[16] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," Advances in Neural Information Processing Systems, vol. 28, pp. 262-270, 2015.
[17] D. P. Kingma and B. J. Adam, "A method for stochastic optimization," arXiv:1412.6980, 2014.
[18] U. Leonards, *et al.*, "Attention mechanisms in visual search—An fMRI study," *Journal of Cognitive Neuroscience*, vol. 12, suppl. 2, pp. 61-75, 2000.
[19] D. M. Vo and S. W. Lee, "Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions," *Multimedia Tools and Applications*, 2018.
[20] J. B. Huang, *et al.*, "Image completion using planar structure guidance," *ACM Transactions on Graphics*, vol. 33, no. 4, p. 129, 2014.

[21] V. V. Voronin, *et al.*, "Video inpainting of complex scenes based on local statistical model," *Electronic Imaging*, pp. 1-6, 2016.
[22] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th International Conference on Pattern Recognition*, Istanbul, Turkey, August 2010.

Jianchu She received his B.E. degree from Xi'an Polytechnic University, China. He is now pursuing the M.Sc. degree at Xi'an University of Posts and Telecommunications. His research focuses on image inpainting and he is a member of the Criminal Investigation Image Processing Team of Xi'an University of Posts and Telecommunications.

Ying Liu received the B.E. degree in School of Information Engineering from Xidian University, China, M.Eng in School of Electrical Engineering from the National University of Singapore, and Ph.D. in School of Computing and Information Technology from Monash University, Australia. She is currently a full professor in the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications (XUPT), China. Her research activities focus on image/video retrieval.