Received: 22 June 2023 | Revised: 5 September 2023 | Accepted: 16 November 2023

IET Image Processing

DOI: 10.1049/ipr2.13005

ORIGINAL RESEARCH

An image inpainting method based on generative adversarial networks inversion and autoencoder

Yechen Wang[1,2] | Bin Song[1,2] | Zhiyong Zhang[1,2]

[1]Information Engineering College, Henan University of Science and Technology, Luoyang, Henan, China

[2]Henan International Joint Laboratory of Cyberspace Security Applications, Henan University of Science and Technology, Luoyang, Henan, China

Correspondence
Bin Song, Information Engineering College, Henan University of Science and Technology, Henan, Luoyang 471023, China.
Email: songbin@haust.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61972133; Project of Leading Talents in Science and Technology Innovation in Henan Province, Grant/Award Number: 204200510021; Program for Henan Province Key Science and Technology, Grant/Award Numbers: 222102210177, 232102211060; Henan Province University Key Scientific Research Project, Grant/Award Number: 23A520008; Henan Province Natural Science Fund, Grant/Award Number: 232300420148

Abstract
Image inpainting aims to repair the damaged region according to the known content in the damaged image. Recently, image inpainting methods have poor effects on high-resolution damaged images, and the research on the inpainting of large-area damaged images is limited. Therefore, this paper proposes an image inpainting method based on Generative Adversarial Networks (GAN) inversion and autoencoder. This work consists of two phases: first, the authors design an autoencoder-based GAN, which learns the mapping from noise to low-dimensional feature maps by training a generator, and then converts the generated feature maps into high-resolution images. Thus, the difficulty of learning the mapping relationship is reduced. Second, the authors adopt the learning-based GAN inversion to infer the closest latent code. The trained GAN is then used to reconstruct the complete image. Finally, the authors compare their method with other classical methods on the CelebAMask-HQ, Flickr-Faces-HQ, and ImageNet datasets. According to the quantitative comparison, when the mask range is large, in other words, when the image has a large area of damage, the authors' method is superior to the comparison methods. According to the qualitative comparison, the structure of the high-resolution image inpainted by the authors' method is more reasonable and the texture details are more realistic.

1 | INTRODUCTION

Image is a common form of information carrier in our life. The integrity of information transmission can only be ensured if the image is complete. Many computer vision tasks are based on the analysis and processing of complete and clear images. However, the required image files are often damaged or obscured. These phenomena can bring a serious impact on the implementation of these vision tasks. In order to ensure the integrity of image information transmission, researchers have proposed a series of methods, and image inpainting is an important research direction in the field of digital image processing and computer vision. It mainly achieves the purpose of inpainting the damaged information in the image through computer vision and other technologies.

Current image inpainting methods can be roughly divided into two categories: traditional methods and deep learning methods. Traditional methods include: structure-based image inpainting [1], texture-based image inpainting [2], and image inpainting based on sparse representation [3]. Among them, the structure-based image inpainting is implemented by partial differential equations. However, the model robustness of such methods is poor, and there are problems such as blurring after image inpainting. The texture-based method uses the texture of the known region to construct the damaged information, which can effectively avoid the blurring problem of the inpainting region. However, the ability to obtain high-level semantic information is poor, and the performance is poor when dealing with challenging images such as complex textures. The method based on sparse representation can effectively represent the known information of the image, but when the inpainting area is large, the method is restricted by the limited known information, and the inpainting effect is still not ideal. With the rapid progress of deep learning theory in computer vision and other fields, image inpainting methods based on deep learning in recent years mainly include two categories: (1) Image inpainting methods based on deep convolutional (DC) neural networks proposed by Liu [4] and Pathak [5] et al. (2) The image inpainting method based on generative adversarial networks proposed by Yeh et al. [6]. In 2018, Liu et al. [4] proposed to apply the partial convolution algorithm to image inpainting, so as to repair the damaged part of the image. This method only inputs valid pixels in the uncorrupted regions of the image, and it replaces the classical convolutional layers with partial convolutional layers. This method achieves image inpainting that is independent of the size of the initial damaged part and does not require any post-processing, which is the first time that CNN is used for image inpainting with irregular shape damage. However, it is unreasonable to regard all pixels as valid pixels when updating the Mask. In 2018, Yu et al. [7] used Gated Convolution to optimize partial convolution, so as to better realize the image inpainting of incomplete parts with irregular shapes. However, excessive smoothing and blurring may occur in the actual processing of images. Yeh et al. [6] proposed an optimization-based GAN inversion image inpainting method, which finds the closest latent code by iterative inference. Even though it has high accuracy, it has fatal drawbacks such as long inference time due to multiple optimizations.

In the process of image inpainting, the above methods based on deep neural networks often have poor inpainting effects on high-resolution damaged images, and the research on the inpainting of large-area damaged images is also limited. There are several differences between high-resolution image inpainting and regular-size image inpainting: First, high-resolution images have more pixels, thus capturing richer details and textures. In contrast, the low resolution of regular-size images may suffer from large information loss in terms of details. Secondly, the inpainting of high-resolution images faces a greater computational burden because the number of pixels increases significantly, making the inpainting process of high-resolution images require more time and computational resources. Based on this, we propose an image inpainting method based on GAN inversion and autoencoder. Compared with other methods, our method mainly has the following contributions: (1) The proposed method improves the training effect of high-resolution image inpainting by reducing the difficulty of mapping learning and making GAN easier to train. (2) Our method adopts a learning-based GAN inversion strategy and designs a novel encoder to achieve the prediction of latent code from corrupted images, resulting in more reasonable semantics and high-fidelity results. (3) The quantitative comparison results show that our method has higher inpainting quality than other classical methods when the inpainting task involves large areas of damage. The qualitative comparison results show that the images generated by our method have clear boundaries and are visually more reasonable.

In this paragraph, we introduce the work of this paper. Firstly, in the related work section of this paper, we survey the image inpainting methods in recent years, and then give a comprehensive introduction to GAN and GAN Inversion. In the Methods section, we introduce and illustrate our proposed method in two stages. In the Experiment section, we first explain the three datasets used, and then introduce our experimental environment and data preprocessing. We compare and analyze with other classical methods from both qualitative and quantitative aspects of our experiments. Finally, in the Conclusion section, we summarize the work content of this paper, and expound on the existing shortcomings and the next work in the future.

2 | RELATED WORK

2.1 | Image inpainting

In recent years, deep neural networks have made breakthroughs in image inpainting. In this paper, the authors surveyed image inpainting methods. In order to improve the degree of coordination between the repair area and the surrounding area and global, so that the image inpainting model can meet the tasks of the high-resolution, irregularly damaged, area and multi-scene, Iizuka et al. [8] proposed a global–local consistent image inpainting algorithm (GL) in 2017 by combining generative adversarial networks and convolutional neural networks. The main architecture of GL is an inpainting network and two discriminators (a global discriminator and a local discriminator) networks, which greatly improves the quality of image inpainting. However, the processing effect is not good when the area to be inpainted is large, and it is difficult to deal with images with complex backgrounds. Sagong et al. [9] divided the image inpainting task into the coarse network and the fine network for parallel processing, which improved the efficiency of inpainting. However, this algorithm is only suitable for inpainting small mask regions. Yu et al. [10] proposed a coarse-to-fine inpainting framework with a contextual attention module. Zheng et al. [11] achieved a variety of image inpainting through VAE [12] probabilistic network, and achieved good results. Nazeri et al. [13] used structural edges as auxiliary information to improve the image inpainting effect. Yi et al. [14] proposed a super-resolution image inpainting network with context residual aggregation technology, which filled the gap for super-resolution image inpainting and had efficient inference speed as a lightweight model. The autoencoder network developed by [15] extracts the self-representation information of the target modality and guides the generation model to fuse the target information from multiple modalities. Through this network, it effectively improves the cross-modal consistency with the desired modality, thus greatly improving the performance of image synthesis. The authors in [16] developed a dual-path inpainting network with feedforward path and inversion path, and designed a novel deformable fusion module to align the feature maps of the two paths. Finally, the feedforward path fused the semantic features of the inversion path to realize the image inpainting work. In order to solve the problems existing in the above methods, we propose an image inpainting method based on GAN inversion inference and autoencoder.

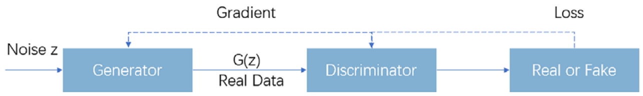

FIGURE 1 Generative adversarial networks architecture.

2.2 | Generative adversarial networks (GAN)

GAN is a deep learning method proposed by Goodfellow et al. [17] in 2014, which consists of two parameterized deep neural networks. The network structure consists of two parts: generator and discriminator, and the game between the generator and the discriminator is used to produce better output results. The Generator extracts the latent distribution of data features from the real training data samples through unsupervised learning, and provides the generated data samples to the discriminator. The Discriminator produces a probability that estimates whether the sample belongs to real training data or generated data. Moreover, the discriminator feeds back the parameters that need to be adjusted in the generated data to the generator. The generator performs parameter tuning after receiving the signal of the loss function that needs to be tuned. The loop is repeated until the discriminator cannot tell whether the image transmitted to it is from the generator output or the original image. In this case, the output of the generator is very close to the original sample image, while the output of the discriminator will approximate a fixed probability value.

Figure 1 illustrates how GAN works. In Gan, the input z to the generator is a random noise vector, which is usually a vector of random numbers drawn from some distribution (usually uniform or Gaussian). During the training of the generator, this random noise vector is continuously sampled at each iteration and fed into the generator. The goal of the generator is to map this random noise vector into the data space to generate realistic data samples, such as images. This process typically involves the transformation and manipulation of multiple neural network layers in order to capture features in the training data from the generated data. The task of the discriminator is to decide whether the input data is real (from the real data distribution) or generated (from the generator). Through adversarial training, the generator gradually learns to generate more realistic data, making it difficult for the discriminator to tell the difference between generated data and real data.

The optimization formula of GAN is as follows:

$$\min_{G}\max_{D} V\ (D, G) = E_{x\sim P_x}\left[\log D(x)\right] + E_{z\sim P_z}\left[\log\left(1 - D\left(G\left(z\right)\right)\right)\right] \quad (1)$$

In this formula, $V(D, G)$ represents the objective function, E represents the expectation, x is the real image sample, z represents the Gaussian noise input to the generator, $G(z)$ represents the image produced by the generator, and $D(x)$ represents the probability judged by the discriminator. More and more studies [18, 19] have shown that GAN has an excellent performance in generating images, which has become the basis of many image inpainting algorithms.

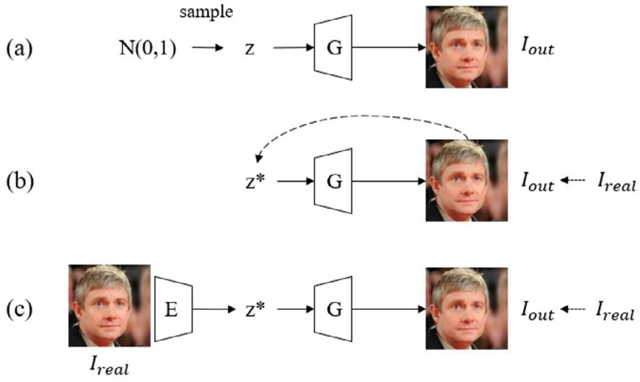

FIGURE 2 Illustration of GAN inversion. GAN, generative adversarial networks.

2.3 | GAN inversion

In recent years, deep learning methods have made significant progress in image inpainting, among which the deep learning image inpainting method based on single feedforward inference is the mainstream method. Although image inpainting methods with feedforward inference produce excellent enough results, they are still less effective at producing reasonable semantic structure, and some methods are insufficient for large areas of corrupted images. GAN inverse image inpainting methods, which have received less attention in the past, provide a new perspective to solve problems in feedforward inference.

Given a trained generator G of a GAN model, it can generate a realistic image from a randomly sampled latent vector z (Figure 2a). The GAN inversion method aims to find the latent code z^* that best matches the damaged image, and then invert the latent code z^* to the image by the pre-trained GAN. Thus, a reconstructed image that is semantically similar to the corrupted image is produced.

$$z^* = \operatorname*{argmin}_{z} L\left(M \odot G\left(z\right), M \odot I\right) \quad (2)$$

In this formulation, M represents the mask of the corrupted image, which is a binary matrix with 0 representing the damaged region and 1 representing the observed region. It obtains the damaged image by Hadamard (⊙) operation with the original image. The Hadamard operation is a per-element operation used for the element-by-element multiplication of two matrices of the same size. It should be noted that it is different from the general matrix multiplication operation, but it multiplies the elements of the corresponding position one by one. For example, in Figure 9, we can obtain the corrupted image $\mathrm{I_M}$ below Figure 9 by performing Hadamard operation on the uncorrupted original image I above and the binary matrix mask M. $L(\cdot)$ represents all the loss functions.

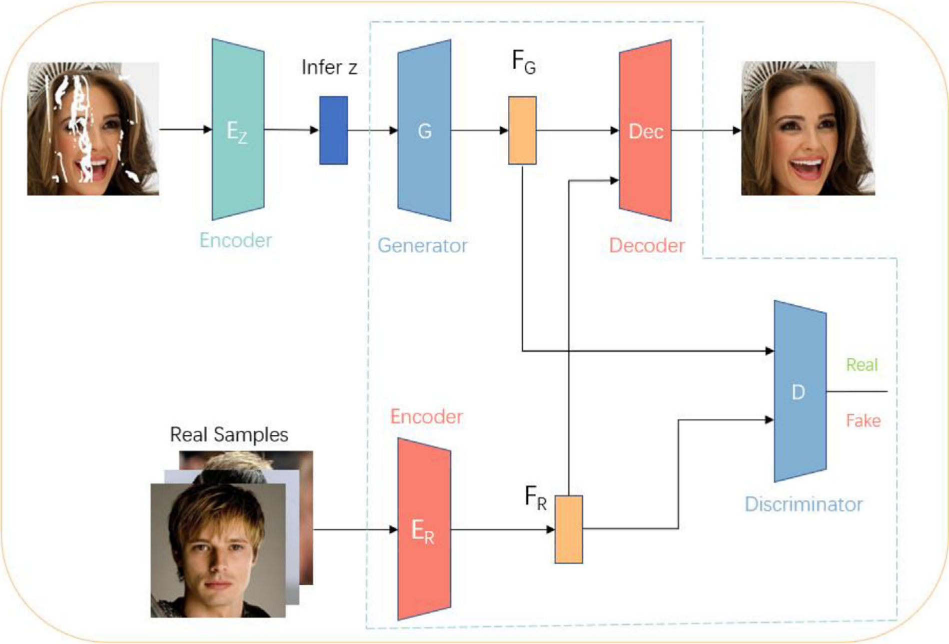

FIGURE 3 Our overall model framework.

z^* denotes the latent code that best matches the damaged image.

GAN inversion can be mainly divided into three categories, one is learning-based GAN inversion, the other is optimization-based GAN inversion, and the third is GAN inversion which combines the first two types of methods. The learning-based strategy (Figure 2c) is to train an encoder that can predict the latent code from damaged images. The optimization-based strategy (Figure 2b) is to iteratively optimize the latent code by backpropagation to minimize the pixel-wise reconstruction loss. In addition, some works combine these two strategies. An encoder is first used to predict the latent code, which is subsequently optimized using an optimization-based strategy.

In addition to image inpainting tasks, GAN inversion has been widely used in other computer vision tasks, such as image editing [20, 21], super-resolution [22, 23], and style transfer [24, 25].

3 | PROPOSED METHOD

In this section, we present our entire network architecture.

Our work consists of two phases, as shown in Figure 3. In the first stage, we design an autoencoder-based GAN (inside the dashed box in Figure 3), which learns a mapping from random noise to low-dimensional feature maps by training a generator. Then, the feature maps generated by the generator are converted into high-resolution images, so as to map the noise distribution to the real image. In the second stage, we adopt the learning-based GAN inversion strategy, fix the trained GAN, and then design to train an encoder network to predict the closest latent code from a given damaged image. When in the actual test, we first obtain the closest latent code for a given damaged image and reconstruct the full image using the trained GAN.

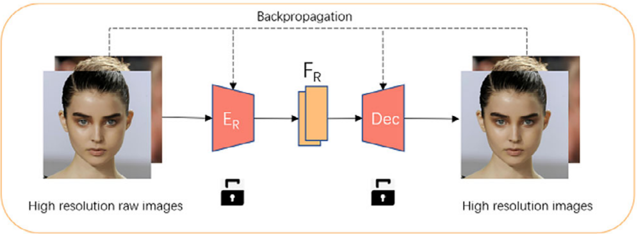

FIGURE 4 Training of the autoencoder. The 'Locked' symbol indicates that the training has been completed, 'Unlocked' symbol indicates that the training is in progress, the same figure below.

3.1 | Autoencoder-based GAN

Compared with other image inpainting methods, the GAN-based method improves the texture and colour processing effect of the image, and reduces the dependence on the information around the damaged area, which is more suitable for image inpainting with large damaged areas. However, the shortcomings of GAN are also obvious. Due to the complexity of the mapping relationship from Gaussian noise to high-dimensional images, it is very difficult to train GAN. With the improvement of the resolution of the generated image, the difficulty of training gradually increases, and the phenomenon of mode collapse is easy to occur, resulting in the lack of authenticity and single images.

Accordingly, the GAN part of this paper is designed to be composed of four parts: encoder, decoder, generator, and

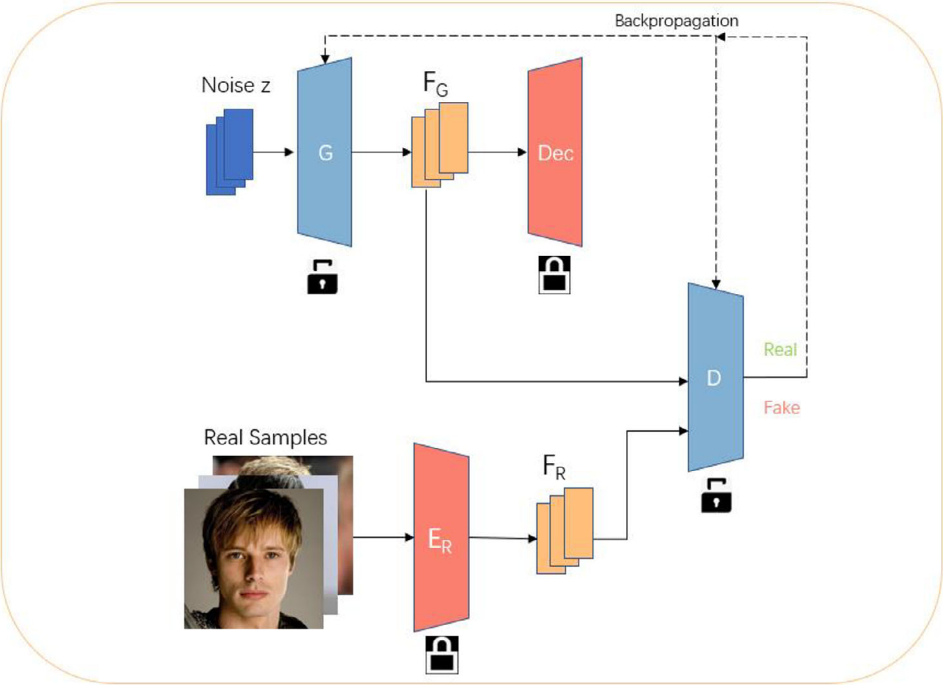

FIGURE 5 Autoencoder-based GAN training. Fixing the autoencoder part, training the generator, and discriminator. GAN, generative adversarial networks.

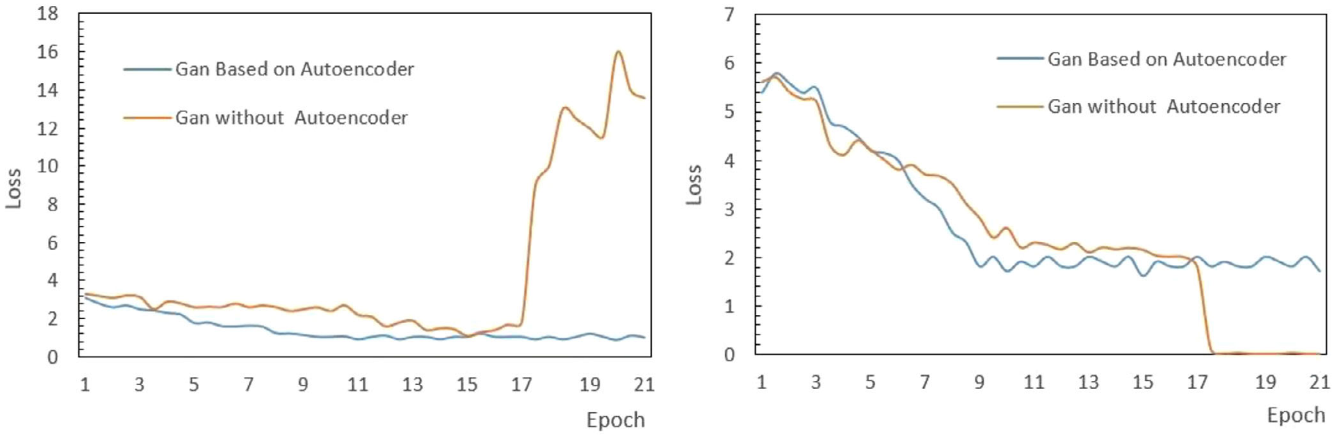

FIGURE 6 Training loss comparison between Gan with autoencoder and Gan without autoencoder. Left is the Loss change curve of the generator, right is the Loss change curve of the discriminator.

discriminator. In addition to the traditional generator and discriminator of GAN, an autoencoder network is added. The autoencoder-based GAN training can be divided into the following two steps:

The first step is to train the autoencoder independently (Figure 4). The input of the encoder is a high-resolution original image sample I, and the output is a low-dimensional feature map F_R extracted from the image. The input of the decoder was the feature map F_R, and the output was the restored high-resolution image.

We use the more robust L1 loss to train the encoder E_R for dimensionality reduction and the decoder Dec for dimensionality increase. We want to make sure that the image input to the encoder and the image output from the decoder are as consistent as possible:

$$L_a = \|Dec\,(E_R\,(I)) - I\|_1 \tag{3}$$

In this formula, I denotes the undamaged original image.

Step 2 (Figure 5) After the autoencoder training, we fix the encoder E_R and the decoder Dec and start training the generator G and the discriminator D. The generator input is Gaussian Noise z and the output is the generated feature map F_G. The discriminator input has two parts, one is the feature map F_R

TABLE 1 Encoder E_Z structure.

Module name	Kernel	Filters	BatchNorm	Non-linearity
Conv1	7 × 7	64	–	ReLU
Conv2	5 × 5	128	Y	ReLU
Conv3	5 × 5	256	Y	ReLU
Conv4	3 × 3	512	Y	ReLU
Conv5	3 × 3	512	Y	ReLU
Conv6	3 × 3	512	Y	ReLU
Conv7	3 × 3	512	Y	ReLU
Fully Connected1	–	4096	Y	ReLU
Fully Connected2	–	100	–	–

extracted by the encoder from the high-resolution raw image, and the other is the feature map F_G generated by the generator. The output of the discriminator is a score between [0,1]. To ensure the authenticity of the final generated results, we use the most used adversarial loss [17] in GAN to constrain the discriminator D:

$$L_D = -E_{I\sim P_I}\left[\log D\left(E_R\left(I\right)\right)\right] \\ -E_{z\sim P_z}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \quad (4)$$

For the generator G:

$$L_G = \; -E_{z\sim P_z}\left[\log D\left(G\left(z\right)\right)\right] \quad (5)$$

In this formula, I denotes the undamaged original image. $E_{z\sim P_z}$ means that z is random noise with Gaussian distribution.

In our autoencoder-based GAN, the encoder is used to reduce the dimensionality of the high-resolution image to the low-dimensional feature space, and the decoder is used to up dimension the low-dimensional feature map to the high-resolution image. The generator is used to generate the low-dimensional feature map, and the discriminator is used to determine whether the low-dimensional feature map is directly extracted from the original image by the encoder E_R or generated by the generator. In this way, the effect of high-resolution image inpainting is improved. By learning low-dimensional feature maps instead of going directly to high-dimensional images, we make the GAN easier to train and reduce the likelihood of mode collapse, making the resulting images more realistic. The generator and the discriminator of our GAN follow the same architecture as the widely used DC-GAN [26]. The encoder E_R used for dimensionality reduction consists of three downsampling layers, all with 4 × 4 convolution kernels, and the number of convolution kernels is 64, 128, and 16, respectively. LeakyReLU, LeakyReLU, and Tanh are used as activation functions in turn. The decoder Dec for the raised dimension consists of three upsampling layers, with the size of the convolution kernels all being 4 × 4 and the number of convolution kernels being 128, 64, and 3, respectively. The activation functions are successively ReLU, ReLU, and Tanh.

As can be seen from Figure 6, our autoencoder-based Gan starts from Epoch = 9, the loss values of the generator and discriminator remain basically stable, and the network training has converged. However, for Gan without autoencoder, starting from Epoch = 17, the generator loss suddenly increases sharply and the discriminator loss suddenly decreases. It can be seen that the Gan training undergoes mode collapse. This training process comparison proves that our method makes the training of GAN easier by reducing the difficulty of mapping learning, thereby improving the training effect of high-resolution image inpainting.

3.2 | Image inpainting network based on GAN inversion

In the GAN inversion path, our goal is to find the closest latent code of the damaged/masked image. Although the inversion strategy based on optimization used by Yeh et al. [6] can find the exact latent code, it almost has an unacceptable inference time. In order to ensure the efficiency of inference, we choose the learning-based inversion strategy for the inversion inference of GAN.

In this stage, we fix the trained GAN, and then design to train an encoder E_Z to predict the most appropriate latent code z from the damaged image, and then use the GAN trained in the previous stage to reconstruct the full image with the latent code z as input

$$\mathrm{z} \; = E_z \; \left(I_M\right) \quad (6)$$

In this formulation, I_M is the damaged image in the input encoder E_Z.

Table 1 shows the specific structure of our encoder E_Z with a total of nine layers. Among them, convolutional layers are used in the first seven layers, fully connected layers are used in the last two layers, and batch normalization operation and activation function are not used in the last layer to ensure that the final output conforms to the P_Z distribution.

In the inversion path training (Figure 7), our input images are damaged/masked images to simulate the situation in the real environment to improve the effect in the future practical scene. In this paper, we follow the photo-realism loss L_P and reconstruction loss L_r used in Lahiri [27] to train the encoder E_z to extract the latent code of the image. L_P loss and L_r loss can ensure that the repaired output image is located near the true data manifold, which greatly improves the quality of the repaired image

$$L_p = \log\left(1-D\left(G\left(E_z\left(I_M\right)\right)\right)\right) \quad (7)$$

$$L_r = \|I - I_{out}\|_1 \quad (8)$$

In the above formula, where $I_{out} \; = \; Dec(G(E_z(I_M)))$, $\mathrm{I_M}$ is the damaged image, E_Z is the encoder that extracts the latent code of the damaged image, G is the generator, and D is the discriminator.

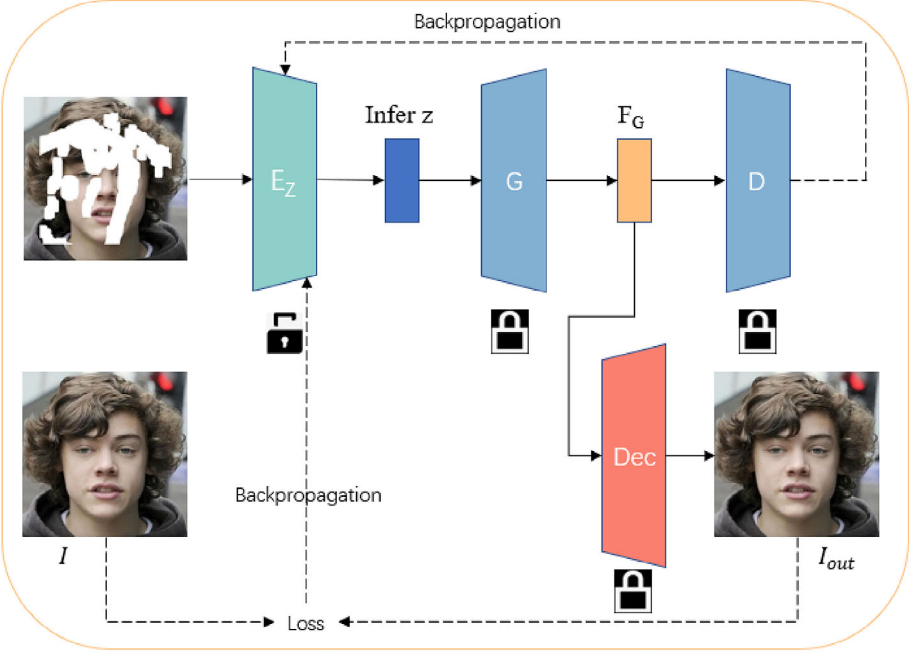

FIGURE 7 GAN Inversion path training. GAN, generative adversarial networks.

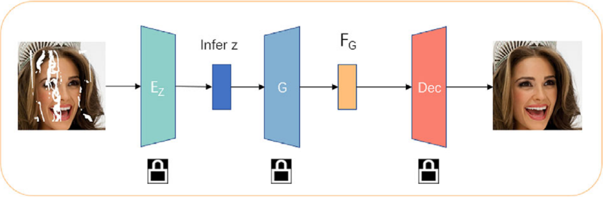

FIGURE 8 The process of our proposed method in the actual image inpainting.

The total loss function for the inversion part can be summarized as follows:

$$L = L_p + \lambda L_r \tag{9}$$

The best results were obtained when λ was set to 20 according to the experience in the experiment.

Figure 8 shows the whole process of image inpainting by our method in the actual inpainting task. Instead of iteratively optimizing the latent code z for each test image during inference, we take a learning-based GAN inversion approach, resulting in more reasonable semantics and high-fidelity results.

4 | EXPERIMENT

4.1 | Datasets

In order to verify the robustness and generalization ability of our method, we adopt the public test datasets CelebAMask-HQ [28], Flickr-Faces-HQ (FFHQ) [18], and ImageNet [29]. CelebAMask-HQ is composed of 30,000 high-resolution face images of 1024 × 1024 pixels collected from CelebA [30], which can be said to be the HD version of CelebA. Each image has corresponding annotation data, which can be used for the training and testing of generative adversarial networks such as face recognition and image generation. We randomly split 27,000 images from them for training and the remaining 3000 images for testing. FFHQ is a high-quality image dataset of faces scraped from Flickr. It consists of 70,000 high-quality face images of 1024 × 1024 pixels with large differences in age, ethnicity, and image background, from which we randomly divide 65,000 images for training and the remaining 5000 images for testing. ImageNet is a large-scale manually annotated dataset for visual object recognition research, containing tens of millions of images with over 20,000 categories. From the Person category, we randomly selected 30,000 face images, randomly divided 27,000 images for training, and the remaining 3000 images for testing.

As for the mask dataset, we use the mask dataset proposed by Liu et al. [4] in 2018, which is widely used in image inpainting tasks. It contains 12,000 mask test images with different scales, and the mask image size is 512 × 512. Their masks are classified based on their proportion of area relative to the entire image size (1%–10%, 10%–20%, 20%–30%, 30%–40%, 40%–50%, 50%–60%). To validate our method, we resized images to 512 × 512 resolution on FFHQ, and ImageNet for experiments, and we used 1024 × 1024 resolution images on CelebAMask-HQ for experiments.

4.2 | Experimental environment and data preprocessing

Our experiments are implemented in Pytorch, a deep learning framework, and we use mini-batch gradient descent to update the parameters. We used an NVIDIA TITAN Xp GPU(12GB)

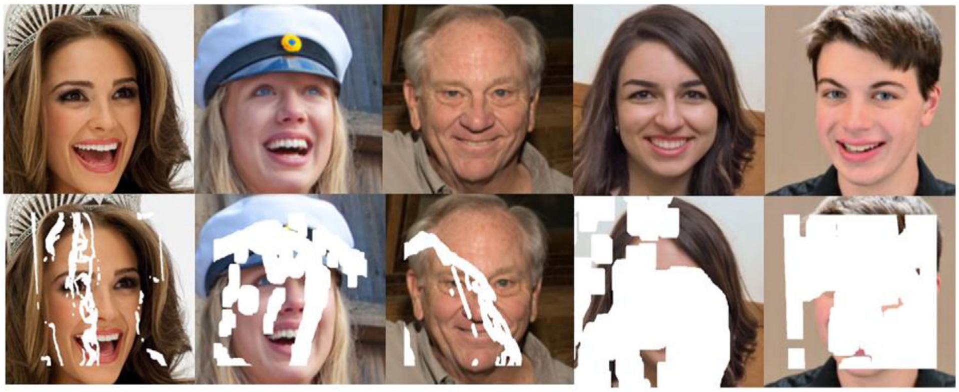

FIGURE 9 Data preprocessing.

with batch_size set to 4, all loss functions were optimized using Adam optimizer, and the learning rate was set to 0.0001 to fine-tune the network until the network converged.

Figure 9 shows the preprocessing results of the data, where the top part of the picture shows the undamaged original image I, and the bottom part shows the damaged image I_M after adding the random mask. In order to verify the ability of the model, each image is randomly superimposed with random mask regions and input to the model for training and testing.

4.3 | Qualitative comparisons

The result comparison of image inpainting mainly includes qualitative comparison and quantitative comparison. Qualitative comparison is mainly to observe whether the colour of the inpainted image is appropriate, whether the grain is consistent, whether the inpainted information is reasonable, and whether the inpainted traces are obvious.

In order to qualitatively evaluate the results after image inpainting, we compare the proposed model with several classical models such as Pconv [4], Gated Conv [7], and CoModGAN [31]. Figure 10 shows the qualitative visual comparison results of each model. It can be seen that the results generated by the two methods Pconv and Gated Conv contain partially distorted content, and there are certain artefact effects and colour differences. The generated results and performance of the CoModGAN model are similar to our effect, but it usually produces discordant content and unmasked regions. Our method is better at mining information from the inside of the image, and can generate more semantic results with the help of the GAN model that has been trained in advance. Therefore, it can better deal with different ranges of mask regions, and produce more realistic and reasonable high-resolution image results.

4.4 | Quantitative comparisons

The more widely used evaluation indicators for quantitative comparison of image inpainting are mainly as follows: structural similarity (SSIM) [32], peak signal-to-noise ratio (PSNR), Frechet Inception Distance (FID) [33], and Learned Perceptual Image Patch Similarity (LPIPS) [34]. SSIM measures the overall similarity of two images from three aspects of image brightness, structure, and contrast, and the value range is [0,1]. The SSIM value is closer to 1, indicating that the similarity of the two images is higher and the image inpainting quality is more effective. PSNR is the most widely used measurement method, which calculates the difference of pixel values between two images to evaluate the quality of image inpainting. The larger the value of PSNR, the more realistic the samples generated by the generation network, and the better the inpainting quality. The FID value calculates the distance between the distributions of two multidimensional variables, and can represent the diversity and quality of the generated image. The smaller the FID value, the better the diversity of the image, and the better the quality. LPIPS considering the factors of human visual perception, is closer to human to judge the quality of the images. A lower value of LPIPS indicates that the two images are more similar, and vice versa, the difference is greater.

In order to quantitatively evaluate the image inpainting results, we compare our method with other methods on three data sets, respectively. The methods we compare in our experiments are all open-source methods, and we use the same experimental conditions to ensure a fair comparison.

From the comparison results of different mask proportions in the FFHQ dataset in Table 2, it can be seen that the image inpainting performance of each model is relatively similar when the mask proportion is small, and the inpainting performance of each model gradually becomes worse with the increase of the mask proportion. When the mask proportion is 1% to 20%, the results of our model are slightly worse than those of CoModGAN, which may be due to the fact that there is no complex structure missing in the image in the face of small mask occlusion, so this situation occurs. When a higher percentage of mask, such as the range of 20% to 60% in Table 2, our model shows more excellent performance; it is also shown that when inpainting task involving a large area missing, our method has a higher quality of inpainting. In terms of the efficiency of the

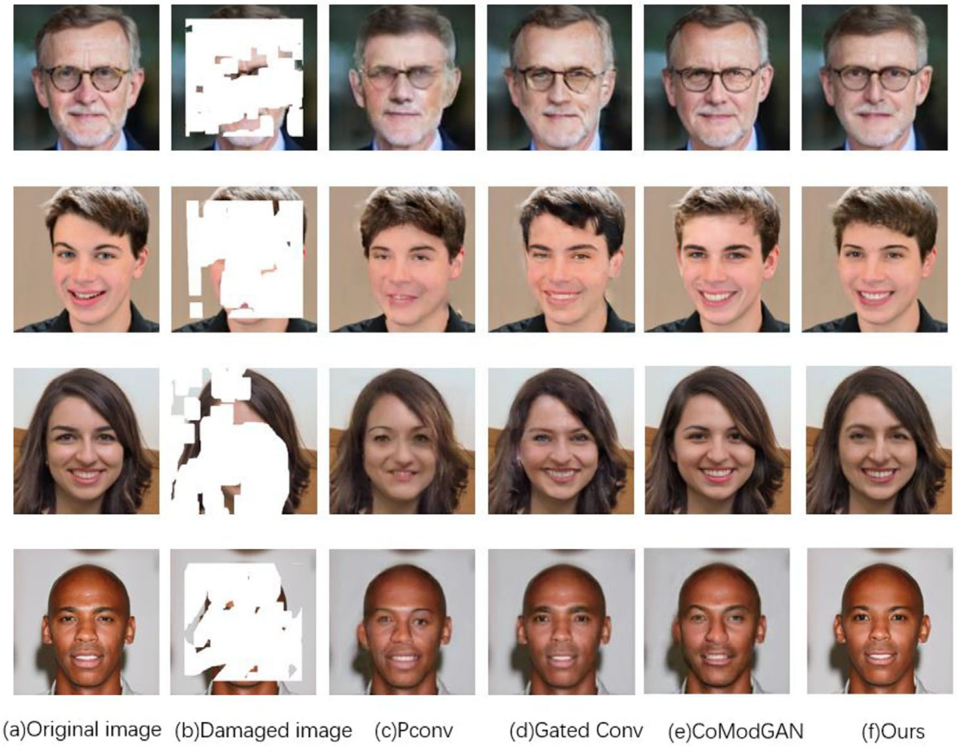

FIGURE 10 Comparison of model qualitative image inpainting results.

TABLE 2 Comparison results of image inpainting of models on FFHQ.

	Pconv				**Gated Conv**				**CoModGAN**				**Ours**			
Mask	**PSNR**	**SSIM**	**FID**	**LPIPS**	**PSNR**	**SSIM**	**FID**	**LPIPS**	**PSNR**	**SSIM**	**FID**	**LPIPS**	**PSNR**	**SSIM**	**FID**	**LPIPS**
1%–10%	32.489	0.956	1.94	0.209	32.646	0.965	1.50	0.203	**33.451**	**0.973**	**1.39**	**0.188**	32.954	0.970	1.46	0.192
10%–20%	27.628	0.896	3.30	0.216	27.958	0.942	2.41	0.212	**28.434**	**0.948**	**2.10**	0.207	28.214	0.947	2.25	**0.205**
20%–30%	23.654	0.816	5.44	0.227	24.423	0.883	3.98	0.224	24.844	0.900	3.92	0.220	**24.972**	**0.912**	**3.88**	**0.218**
30%–40%	21.657	0.774	8.10	0.239	22.163	0.853	6.68	0.230	22.426	0.861	5.69	0.225	**22.450**	**0.862**	**5.22**	**0.222**
40%–50%	18.646	0.623	13.21	0.375	19.374	0.796	9.60	0.312	20.154	0.815	8.42	0.246	**20.312**	**0.819**	**8.23**	**0.235**
50%–60%	15.437	0.583	19.36	0.446	16.842	0.660	16.41	0.425	17.528	0.742	13.34	0.412	**17.814**	**0.759**	**11.58**	**0.356**

FF-HQ, Flickr-Faces-HQ; FID, Frechet inception distance; LPIPS, learned perceptual image patch similarity; PSNR, peak signal-to-noise ratio; SSIM, structural similarity.

TABLE 3 Comparison results of image inpainting of models on ImageNet.

Method	**PSNR**	**SSIM**	**FID**	**LPIPS**
Pconv	23.786	0.788	8.20	0.276
Gated Conv	24.218	0.862	6.53	0.258
CoModGAN	24.672	0.885	5.62	0.232
Ours	**24.894**	**0.907**	**5.31**	**0.216**

FID, Frechet inception distance; LPIPS, learned perceptual image patch similarity; PSNR, peak signal-to-noise ratio; SSIM, structural similarity.

TABLE 4 Comparison results of image inpainting of models on CelebAMask-HQ.

Method	**PSNR**	**SSIM**	**FID**	**LPIPS**
Pconv	19.875	0.754	16.56	0.426
Gated Conv	22.168	0.805	13.20	0.387
CoModGAN	24.594	0.874	5.82	0.246
HiFIll	24.586	0.870	5.35	0.249
Ours	**24.597**	**0.875**	**5.28**	**0.244**

FID, Frechet inception distance; LPIPS, learned perceptual image patch similarity; PSNR, peak signal-to-noise ratio; SSIM, structural similarity.

method, the average running time of our method is 0.081 s per image of the FFHQ dataset on NVIDIA TITAN Xp GPU.

Tables 3 and 4 show the comparison results of each model on ImageNet and CelebAMask-HQ datasets, respectively. In order to fully illustrate the performance of our method in high-resolution images, we specially add HiFIll [14], which focuses on high-resolution image inpainting, to conduct comparative experiments on CelebAMask-HQ. Tables 3 and 4 show that our method is better than the comparison methods in the four indicators of PSNR, SSIM, FID, and LPIPS, which proves the performance of our method.

5 | CONCLUSION

In this paper, an image inpainting method based on GAN inversion and autoencoder is proposed to solve the problem of high-resolution damaged image inpainting and large-area damaged image inpainting. The proposed method is compared with other classical methods on the CelebAMask-HQ dataset, FF-HQ dataset, and ImageNet dataset. From the results of qualitative comparison with other methods, it can be seen that the image structure generated by our method is more reasonable and the texture details are more realistic. Because our method is better at mining information from the inside of the image, it can generate more semantic results with the help of the GAN model that has been trained in advance, thus producing more realistic and reasonable high-resolution image results. From the results of quantitative comparison, it can be seen that when the mask ratio is in the range of 20% to 60%, that is, when the image has large area damage, our model shows extremely excellent performance, which proves that the proposed method has a higher inpainting quality when the inpainting task involves large area damage. In view of the phenomenon that the inpainting results of the proposed model are slightly worse than those of CoModGAN when the mask ratio is small in the quantitative comparison, further work will propose new solutions to this problem. In addition, the image inpainting of complex scenes is also the development trend in the future, and is also the focus of our next work.

Finally, our proposed method can be easily extended to other image tasks, such as image generation, image editing, image denoising, and image super-resolution. Image retrieval [35] has been a very popular and interesting field in recent years, and content-based image retrieval is also a field for future research and exploration of the method proposed in this paper. In addition, we will try to apply model transfer to video inpainting.

AUTHOR CONTRIBUTIONS

Yechen Wang: Conceptualization; investigation; methodology; resources; software; visualization; writing—original draft. **Bin Song**: Conceptualization; formal analysis; investigation; project administration; supervision; validation; writing—review & editing. **Zhiyong Zhang**: Conceptualization; formal analysis; funding acquisition; validation.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the CelebAMask-HQ dataset at https://doi.org/10.1109/CVPR42600.2020.00559 [28], the Flickr-Faces-HQ dataset at https://doi.org/10.1109/CVPR.2019.00453 [18], and the ImageNet dataset at https://doi.org/10.1007/s11263-015-0816-y [29].

ORCID

Bin Song https://orcid.org/0000-0003-4051-3794

REFERENCES

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: International Conference on Computer Graphics an Interactive Techniques, pp. 417–424 (2000). https://doi.org/10.1145/344779.344972
2. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. 13, 1200–1212 (2004). https://doi.org/10.1109/TIP.2004.833105
3. Xu, Z., Sun, J.: Image inpainting by patch propagation using patch sparsity. IEEE Trans. Image Process. 19, 1153–1165 (2010). https://doi.org/10.1109/TIP.2010.2042098
4. Liu, G., Reda, F.A., Shih, K.J., Wang, T., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: 15th European Conference on Computer Vision (ECCV), pp. 89–105 (2018). https://doi.org/10.1007/978-3-030-01252-6_6
5. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., IEEE: Context encoders: Feature learning by inpainting. In: 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2536–2544 (2016). https://doi.org/10.1109/CVPR.2016.278
6. Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M.: Semantic image inpainting with deep generative models. In: 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6882–6890 (2017). https://doi.org/10.1109/CVPR.2017.728
7. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T., IEEE: Free-form image inpainting with gated convolution. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4470–4479 (2019). https://doi.org/10.1109/ICCV.2019.00457
8. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graphics 36, 1–14 (2017). https://doi.org/10.1145/3072959.3073659
9. Sagong, M., Shin, Y., Kim, S., Park, S., Ko, S.: C. S. IEEE: PEPSI: Fast image inpainting with parallel decoding network. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11352–11360 (2019). https://doi.org/10.1109/CVPR.2019.01162
10. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: IEEE: Generative image inpainting with contextual attention. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5505–5514 (2018). https://doi.org/10.1109/CVPR.2018.00577
11. Zheng, C., Cham, T., Cai, J., C. S. IEEE: Pluralistic image completion. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1438–1447 (2019). https://doi.org/10.1109/CVPR.2019.00153
12. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv:1312.6114 (2013). https://doi.org/10.48550/arXiv.1312.6114
13. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M., IEEE: Edge-Connect: Structure guided image inpainting using edge prediction. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3265–3274 (2019). https://doi.org/10.1109/ICCVW.2019.00408
14. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7505–7514 (2020). https://doi.org/10.1109/CVPR42600.2020.00753
15. Cao, B., Cao, H., Liu, J., Zhu, P., Zhang, C., Hu, Q.: Autoencoder-based collaborative attention GAN for multi-modal image synthesis. IEEE Trans. Multimedia 1–16 (2023). https://doi.org/10.1109/TMM.2023.3274990
16. Wang, W., Niu, L., Zhang, J., Yang, X., Zhang, L., Abdal, R.: Dual-path image inpainting with auxiliary GAN inversion. In: 2022 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11411–11420 (2022). https://doi.org/10.1109/CVPR52688.2022.01113

17. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: 28st Annual Conference on Neural Information Processing Systems (NIPS), pp. 2672–2680 (2014). https://doi.org/10.3156/JSOFT.29.5_177_2
18. Karras, T., Laine, S., Aila, T., C. S. IEEE: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4396–4405 (2019). https://doi.org/10.1109/CVPR.2019.00453
19. Ishaan, G., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein GANs. In: 31st Annual Conference on Neural Information Processing Systems (NIPS), pp. 5767–5777 (2017)
20. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of GANs for semantic face editing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9240–9249 (2020). https://doi.org/10.1109/CVPR42600.2020.00926
21. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain GAN inversion for real image editing. In: 16th European Conference on Computer Vision (ECCV), pp. 592–608 (2020). https://doi.org/10.1007/978-3-030-58520-4_35
22. Chan, K.C.K., Wang, X., Xu, X., Gu, J., Loy, C.C., IEEE, C.S.: GLEAN: Generative latent bank for large-factor image super-resolution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14240–14249 (2021). https://doi.org/10.1109/CVPR46437.2021.01402
23. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C., IEEE: PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2434–2442 (2020). https://doi.org/10.1109/CVPR42600.2020.00251
24. Abdal, R., Qin, Y., Wonka, P., IEEE: Image2StyleGAN: How to embed images into the StyleGAN latent space? In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4431–4440 (2019). https://doi.org/10.1109/ICCV.2019.00453
25. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN++: How to edit the embedded images? In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8293–8302 (2020). https://doi.org/10.1109/CVPR42600.2020.00832
26. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 2016 International Conference on Learning Representations (ICLR) (2016)
27. Lahiri, A., Jain, A.K., Agrawal, S., Mitra, P., Biswas, P.K.: Prior guided GAN based semantic inpainting. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13693–13702 (2020). https://doi.org/10.1109/CVPR42600.2020.01371
28. Lee, C., Liu, Z., Wu, L., Luo, P., IEEE: MaskGAN: Towards diverse and interactive facial image manipulation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5548–5557 (2020). https://doi.org/10.1109/CVPR42600.2020.00559
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
30. Liu, Z., Luo, P., Wang, X., Tang, X., IEEE: Deep learning face attributes in the wild. In: 2015 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3730–3738 (2015). https://doi.org/10.1109/ICCV.2015.425
31. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv:2103.10428 (2021). https://doi.org/10.48550/arXiv.2103.10428
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
33. Hensel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: 31st Annual Conference on Neural Information Processing Systems (NIPS), pp. 6629–6640 (2017)
34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
35. Kelishadrokhi, M.K., Ghattaei, M., Fekri-Ershad, S.: Innovative local texture descriptor in joint of human-based color features for content-based image retrieval. Signal Image Video Process. 17, 4009–4017 (2023). https://doi.org/10.1007/s11760-023-02631-x

How to cite this article: Wang, Y., Song, B., Zhang, Z.: An image inpainting method based on generative adversarial networks inversion and autoencoder. IET Image Process. 1–11 (2023). https://doi.org/10.1049/ipr2.13005