International Journal of Computational Intelligence Systems
Vol. **12(2)**, 2019, *pp.* 1232–1244
DOI: https://doi.org/10.2991/ijcis.d.191016.003; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Face Inpainting with Deep Generative Models

Zhenping Qiang[1,*], Libo He[2], Qinghui Zhang[1], Junqiu Li[1]

[1] *College of Big Data and Intelligent Engineering, Southwest Forestry University, No. 300, Bailong Road, Kunming, Yunnan Province, China*
[2] *Information Security College, Yunnan Police College, No. 249, Jiaochang North Road, Kunming, Yunnan Province, China*

ARTICLE INFO

Article History
Received 18 Jul 2019
Accepted 14 Oct 2019

Keywords
Face inpainting
Structural loss
Semantic inpainting
Deep generative models
GANs

ABSTRACT

Semantic face inpainting from corrupted images is a challenging problem in computer vision and has many practical applications. Different from well-studied nature image inpainting, the face inpainting task often needs to fill pixels semantically into a missing region based on the available visual data. In this paper, we propose a new face inpainting algorithm based on deep generative models, which increases the structural loss constraint in the image generation model to ensure that the generated image has a structure as similar as possible to the face image to be repaired. At the same time, different weights are calculated in the corrupted image to enforce edge consistency at the repair boundary. Experiments on different face data sets and qualitative and quantitative analyses demonstrate that our algorithm is capable of generating visually pleasing face completions.

1. INTRODUCTION

The goal of face inpainting, also known as face completion, is to produce a more legible and visually realistic face image from an image with a masked region or that has missing content. As faces play the most substantial role in depicting human characters [1], face inpainting becomes the basis of face verification and identification when an occlusion or damage exists in the facial part of an image. These applications make face completion very important in today's computer vision. Face images have obvious high-level semantics that include many special objects. For face images missing unique pattern objects, the original images may not be restored successfully using traditional inpainting methods. Figure 1 shows how the traditional popular TV [2] and PatchMatch [3] methods fail at face inpainting. The TV method is a total variation-based approach, while PatchMatch is a content aware fill method implemented in Adobe Photoshop.

Recently, deep learning neural networks have been extensively studied and have demonstrated their capability to capture abstract information contained in images at high resolution [4]. One of the feedforward neural networks, the convolutional neural network (CNN) [5], is effective because each of its artificial neurons responds only to some of the neurons connected to it, making the application of deep learning neural networks in large-scale image processing possible by avoiding the over-fitting phenomenon. On the other hand, extensive research [6–10] on generative adversarial networks (GANs) [11] has shown that the visual effect of generated images can be enhanced by adversarial training. Based on this research background, CNN-based image completion methods with adversarial training strategies have already significantly improved image completion performance [12–14].

The early proposed deep learning-based semantic image restoration methods were implemented by training an encoder–decoder CNN (a context encoder) [12] which is closely related to the self-encoder [15–17], to predict the unavailable content in an inpainting image. That is, the network is trained to fill in an image's unknown content based on the known content. However, the context encoder considers the structure of missing regions only during training but not during inference, which will inevitably cause ambiguity and error in the results. Based on the context encoder network, the authors in [18] added a local adversarial loss and a semantic parsing loss to train the model to ensure pixel faithfulness and local content consistency, but ambiguity is still present in the results when missing regions have arbitrary shapes [13].

When considering GANs, if some constraints can be provided in the generating process, such as forcing the generated image to be similar enough to the corresponding part of the known region of the inpainting image, then we can find the best matching latent space representation closest to the natural image manifold without specifying any explicit distance based loss. Then, the image can be restored by fusing the image generated by the GAN and the known region of the inpainting image. In other words, the image can be completed by blending the image generated by the GAN and the known regions of the inpainting image [13] (known as "DIP", Image Inpainting with Deep Generative Models). In spite of this success, some challenges still exist in face inpainting. Firstly, human faces have a definite geometric distribution, and

* *Corresponding author. Email: qzp@swfu.edu.cn*

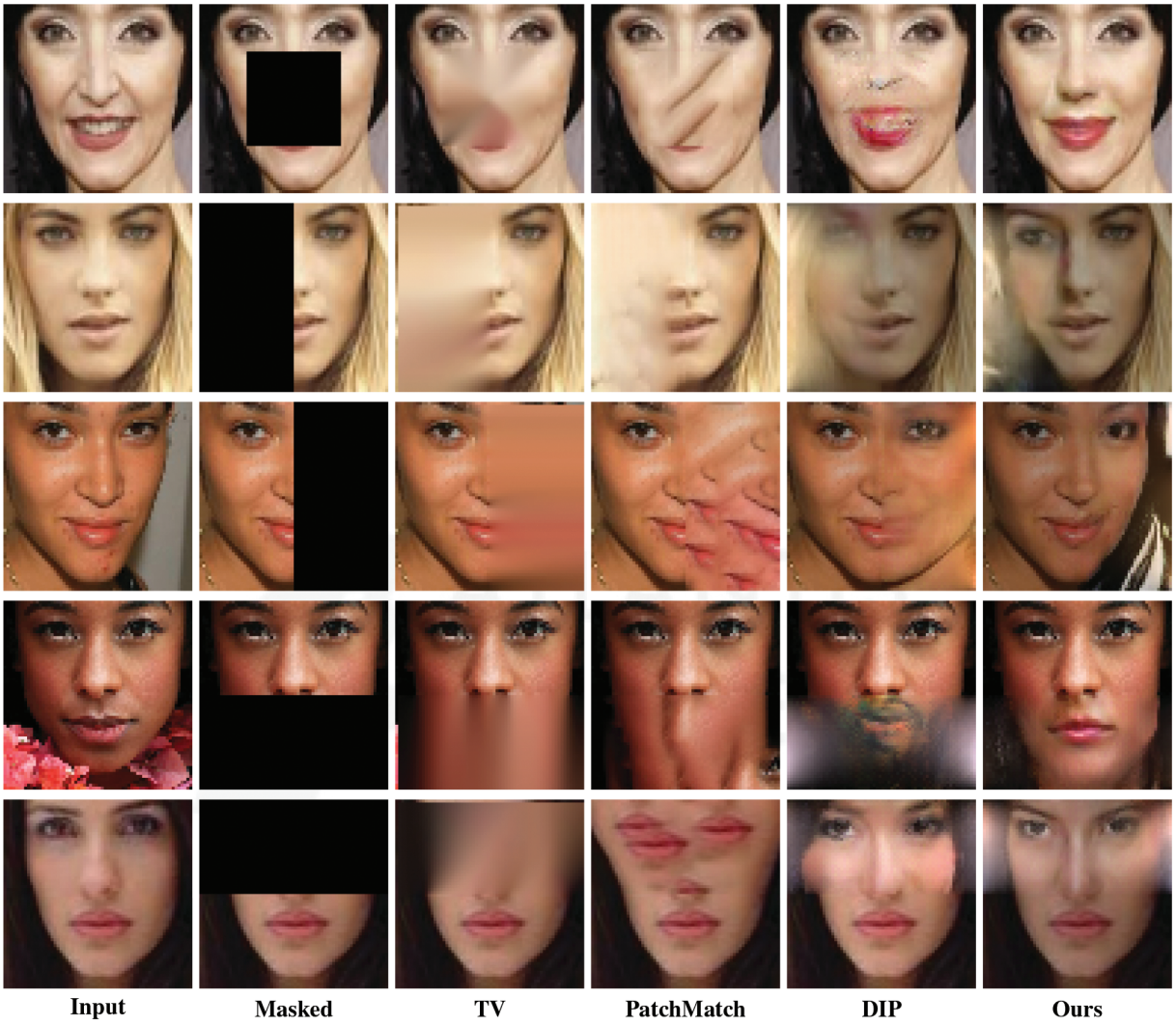


Figure 1 Face inpainting results obtained by TV, PatchMatch, DIP and the proposed method.

hence any face inpainting method based on deep learning must consider the geometric structure loss in the process of restoration. Secondly, coherence is very important in face inpainting and must be considered in the process of face image completion.

To address these two concerns, this study develops a face inpainting network that promotes content continuity and structural consistency. On the one hand, we apply the experience gained from traditional image inpainting methods in our method. That is, more attention should be paid to the continuity of the inpainted region boundary, so we increase the weights for content loss and structural loss at the boundary portions of the region to be repaired, which can ensure the continuity of content in the repair results. On the other hand, for the face completion problem, the rationality of the overall structure of the repair results is very important. Accordingly, we add the structural loss in the generation process to ensure that the generated image has a structure that is as similar as possible to the face image to be repaired. The procedure of the proposed method is as follows. In the first stage, a deep generative model is trained using face samples. In the second stage, a face image is iteratively generated that is "closest" to the input face image. For the iteratively generated image with a combination of adversarial loss, content loss and structural loss, the loss weights near the repair border regions are increased. In the last stage, the image blending method is used to fuse the known region content of the corrupted image and the corresponding generated content to the unavailable region in the original damaged face image. We evaluate our method on the CelebA and SiblingsDB datasets with different shapes of the missing area in an image. Results demonstrate that compared to the traditional methods, our method can implement semantic restoration, and compared to the benchmark DIP method, our method can obtain more realistic and reasonable results (as shown in Figure 1).

The main technical contributions of the proposed method are summarized as follows:

- A novel network is developed to complete semantic face image inpainting from masked face images. This method generates samples that are more similar to the inpainting face image by adding structural loss and applying an adaptive weight strategy to the face generation model.
- A novel structural loss measurement method based on structural similarity index (SSIM) values is introduced, which includes the SSIM value calculation method for the image to be

repaired and the generated image, and a normalization method of these SSIM values is used to define the structural loss.

- Our method guarantees the consistency of the repaired boundary in the repaired result by implementing an adaptive weight strategy, that is, larger weight values are applied to the structure and content loss of pixels closer to the repair boundary.

2. RELATED WORK

2.1. Image Generation

Owing to the good high-level semantic capture capabilities of the variational auto-encoder (VAE) [19] and GAN, a large number of image generation methods [12,20,21] have been proposed recently. The VAE methods usually use the pixel-wise L2 distance (Euclidean) loss between the generated image and original image to train the network. However, because the Euclidean distance is used to minimize the average value of the difference between all input and output pixels, it will inevitably cause ambiguity. By contrast, GANs are known to generate sharper images compared to VAE. Especially for a particular type of image, GANs can generate samples that are difficult to distinguish between true and false [7]. The DIP method [13] is a semantic face inpainting algorithm based on this idea. In the DIP method, image inpainting is performed by adding the content loss between the image's available information and the corresponding generative samples to constrain the iterative generation of the GAN; specifically, the L1-norm is used to define the content loss. Owing to its very good repair results, the DIP method has become a contemporary benchmark method. In [21] , the authors proposed an improved DIP method. They present a semantically conditioned GAN, which increases the conditional information to constrain the GAN, to map a latent representation to a point in the image manifold based on the underlying pose and semantics of the scene. This method has been successfully applied to face restoration in video sequences.

Our method also represents an improvement to the DIP method. Unlike [21], we do not use the facial semantic map as a condition for the face generation network and instead emphasize the importance of the facial structure for face completion by increasing the structure loss weight in the face generation network directly. In addition, we focus on the repair of a single image, while the method in [21] extracts the facial semantic map based on a video sequence. However, because these methods are based on image generation, they are all based on deep convolutional GAN (DCGAN), which is a network that adds a deep convolutional network structure to GAN. Technical details follow.

GANs consist of two separate neural networks, where one of the neural networks is referred to as "*G*", which stands for the Generator, and the other neural network is called "*D*", which is a Discriminator. The objective function of GAN is a zero-sum or minimax two-player game, where the players are *G* and *D*. Numerous recent studies have proposed improvements to the original GAN for image generation, for example, DCGAN [7], Improved GAN [22], Conditional GAN [6], LAPGAN [23], iGAN/GVM [24], pix2pix GAN [25], StackGAN [26,27], PPGN [28] and so on. These different methods mainly focus on adjusting the network architecture and the loss function used to train the network, and make the final result satisfy the corresponding application.

DCGAN [7] is widely used in image generation applications because it can generate much sharper images. We use a pre-trained DCGAN, which greatly improves the stability of GAN training, to generate the face in our proposed face inpainting network. The DCGAN architecture is described in Figure 2, where the discriminator network *D* takes in both the prediction of *G* and ground truth samples and attempts to distinguish between true and false samples, while *G* attempts to mislead the discriminator *D* to the greatest extent possible. This goal can be represented mathematically as follows:

$$\begin{aligned} &\min_G \max_D V(D,G), \\ &V(D,G) = E_{x\sim p_{data}(x)}\left[log\left(D\left(x\right)\right)\right] \\ &\qquad + E_{z\sim p_z(z)}\left[log\left(1-D\left(G\left(z\right)\right)\right)\right] \end{aligned} \tag{1}$$

where $p_{data}(x)$ and $p_z(z)$ represent the distributions of real data x and noise variables z.

2.2. Image Loss Measurement

Although DCGAN can generate a sample image that appears to be a real image, DCGAN cannot be directly applied to the image inpainting task. This is mainly because of the fact that the image generated by DCGAN may not be related at all to the provided corrupted image. Therefore, an effective method for applying DCGAN to image inpainting is to constrain the image generation of DCGAN through the information in the image to be repaired, so that the generated image is sufficiently similar to the image to be repaired. To accomplish this, the measurement of the similarity between the generated image and the provided corrupted image must be considered.

Image similarity measurement methods include image content-based methods (e.g. sum of squared differences (SSD)), image pixel statistics-based methods (e.g. variation of square error (VSE)), image structure-based methods (e.g. the SSIM [29]), and information theory-based methods, such as normalized cross-correlation, Kullback–Leibler(K–L) divergence, and others. For example, the PatchMatch method uses the L2 distance to measure image patch matching. Considering that pixel-wise metrics tend to reflect the overall difference between the two images, this type of measurement does not involve the direct correlation between images. It is well known that face images have clear structural correlations. According to the well-known visual psychology theory (Gestalt theory), the human eye is particularly sensitive to structural information in images. Therefore, in the generation process of the proposed method, in addition to using the content loss used in DIP to measure the face difference, an image difference metric based on the structural information of the face is added. Specifically, we adopt SSIM, which can measure image similarity based on brightness, contrast and structure, to measure the structural loss of the generated face image.

SSIM was proposed by the Laboratory for Image and Video Engineering at the University of Texas at Austin [29]. Given two images I_1 and I_2, the SSIM of the two images can be calculated as follows:

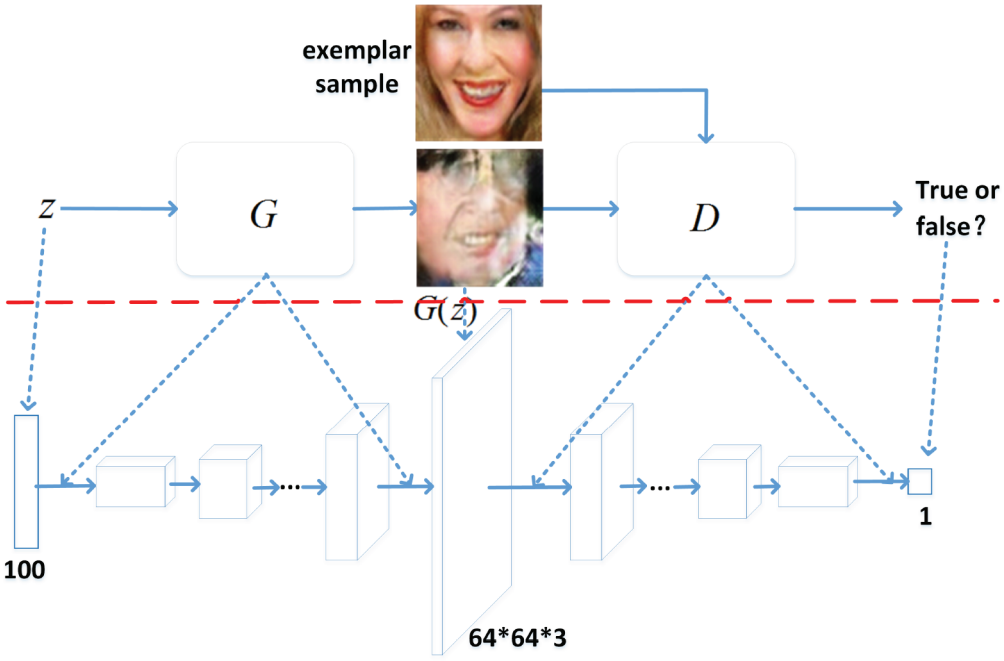


Figure 2 | Deep convolutional generative adversarial network (DCGAN) framework overview. The figure consists of two parts (divided by a red dotted line), the upper part is a schematic diagram of DCGAN training, and the lower part is the overall architecture of DCGAN.

$$SSIM(I_1, I_2) = \frac{\left(2\mu_{I_1}\mu_{I_2} + c_1\right)\left(2\sigma_{I_1 I_2} + c_2\right)}{\left(\mu_{I_1}^2 + \mu_{I_2}^2 + c_1\right)\left(\sigma_{I_1}^2 + \sigma_{I_2}^2 + c_2\right)} \tag{2}$$

where μ_{I_1} is the average of the brightness of the image pixels in I_1, μ_{I_2} is the average of the brightness of the image pixels in I_2, $\sigma_{I_1}^2$ is the variance of I_1, $\sigma_{I_2}^2$ is the variance of I_2, and $\sigma_{I_1 I_2}$ is the covariance of I_1 and I_2. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are constants used to maintain stability. L is the dynamic range of the pixel values in the images. From past experience, $k_1 = 0.01$ and $k_2 = 0.03$.

In terms of the implementation of image structural similarity theory, SSIM defines structural similarity from the perspective of image composition and independent brightness and contrast. In this fashion, SSIM can well reflect the differences in the structural properties of objects in different images.

3. METHOD

3.1. Network Architecture

For the semantic image inpainting methods based on deep generative models, the issue of image restoration is not considered at the network training stage. That is, adversarial training takes place in GAN using non-damaged real samples and generated samples, so that "G" has the capability to generate as many real samples as possible, and "D" has the capability to distinguish true samples from false. Then the repair of a corrupted image I can be transformed into the process of generating a new sample S that has very similar content corresponding to the known part of the inpainting image generated by G. In order to achieve this goal, we need to iteratively modify the newly generated image S, and the basis for the modification is to reduce the difference between the regions of S corresponding to the known regions of the corrupted image I. Figure 3 shows the framework for image inpainting based on deep generative models.

According to the framework of Figure 3, the image inpainting process is to generate a new sample S from the noise signal z firstly, and then the element value in signal vector z is updated using backpropagation based on the discriminator loss of S and the different losses between S and I. After modifying z iteratively, we can recover the encoding $\hat{z}$ that is "closest" to the corrupted image.

3.2. Objective Function

We first introduce the discriminator loss L_d, which is different from the GAN optimizing function given by Formula (1), because in image restoration, there is no real sample image input to the network. That is, only the discriminator loss of the new generated samples is available. The discriminator loss L_d is thus defined as follows:

$$L_d(z) = log(1 - D(G(z))) \tag{3}$$

By using L_d, the network can continuously update z to fool D, which ultimately leads to a more realistic corresponding generated image.

Regarding the reconstruction loss L_r with respect to the generator, in DIP, the masked L1 distance between the network G output and the original image are used. By using additional discriminators, DCGAN can generate clearer sample images. However, if only L_r is added, the structure consistency between the generated image and the original image cannot be guaranteed, which will inevitably lead to errors in the repair results, as shown in Figure 1. Especially when filling large missing regions, we must take greater advantage of the

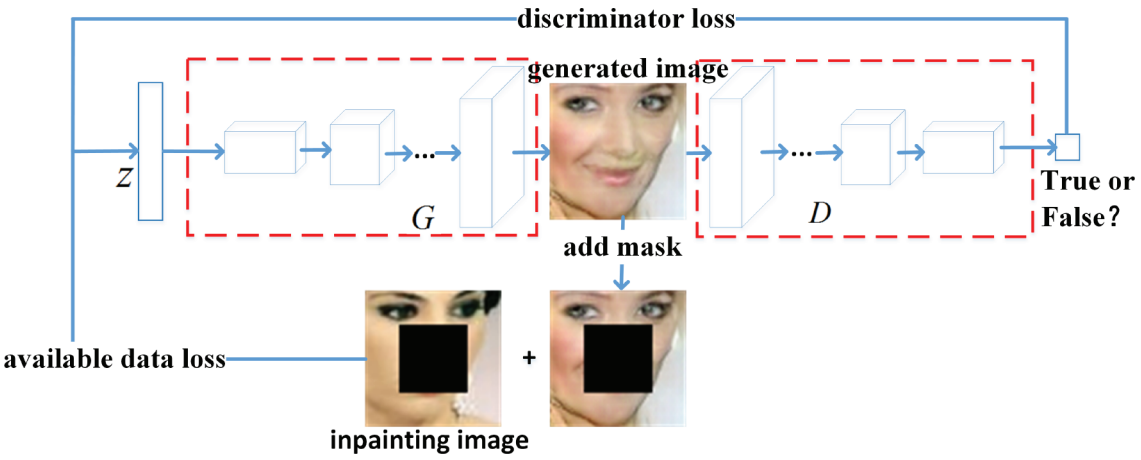


Figure 3 Framework for image inpainting based on deep generative models. The face image is iteratively generated by the DCGAN network with face generation capability, and finally the face image most similar to the known portion of the face image to be repaired can be generated.

remaining available data in the image to be inpainted. In practice we apply the context loss and structure loss to capture such information. In the DIP method, the authors found that the $L1$-norm performs slightly better than the $L2$-norm in the semantic inpainting method based on deep generative models. For the context loss L_c, we also use the pixel-wise $L1$-norm in our method. In order to ensure the consistency of the face structure after inpainting, we introduce the structural loss L_s based on SSIM, and introduce the adversarial loss L_d. The overall loss function is thus defined by:

$$L = \lambda_1 L_d + \lambda_2 L_c + \lambda_3 L_s \tag{4}$$

where λ_1, λ_2 and λ_3 are the weights used to balance the effects of the different losses.

To set reasonable weight values, an effective way is to normalize the various loss values. For example, although SSIM can well describe the structural similarity between two images, its range of values is 0 to 1; when the two images have greater similarity, its value is larger. The L_1-norm value on the other hand is smaller when the two images are more similar. Therefore, the SSIM value cannot be used for structural loss of the inpainting network directly, and the SSIM value representing the similarity of two images must be converted into a structural difference representation between two images. The simplest conversion method is that using the value of $1 - SSIM$ to express the structural loss. In order to unify all losses (the L_c loss and the L_s loss in particular are not at the same levels generally), we need to normalize both the L_c loss and L_s loss. The method we use is to generate a sufficient number of samples from a specific data set used to train G, and then use a number of randomly selected test samples to calculate their SSIM values and L1 values, and finally the normalized parameters are selected according to the statistical results to achieve normalized processing adapted to the training data set. Figure 4 shows scatter plots of SSIM values and L1-norm values between 64 test set images and 12800 images generated by DCGAN trained on the CelebA data set.

We calculate the mean μ_s and standard deviation σ_s of the SSIM values for 64 test samples and all generated images, and use $\mu_s - 3 * \sigma_s$ as the minimum value s_{min} of SSIM values and $\mu_s + 3 * \sigma_s$ as the maximum value s_{max} of SSIM values; the normalized value of L_s can then be calculated by Formula (5). The normalization of L_c can be achieved by the same method, namely, we calculate the mean μ_c and standard deviation σ_c of the $L1$-norm values for 64 test samples and all generated images, such that $c_{min} = \mu_c - 3 * \sigma_c$ and $c_{max} = \mu_c + 3 * \sigma_c$. The formula is given in (6).

$$L_s(I_1, I_2) = \begin{cases} 1, & D_{ssim}(I_1, I_2) < s_{min} \\ \dfrac{(s_{max} - D_{ssim}(I_1, I_2))}{s_{max} - s_{min}}, & others \\ 0, & D_{ssim}(I_1, I_2) > s_{max} \end{cases} \tag{5}$$

$$L_c(I_1, I_2) = \begin{cases} 1, & \| I_1 - I_2 \|_1 > c_{max} \\ \dfrac{\| I_1 - I_2 \|_1}{c_{max} - c_{min}}, & others \\ 0, & \| I_1 - I_2 \|_1 < c_{mim} \end{cases} \tag{6}$$

In the equations, $D_{ssim}(I_1, I_2)$ represents the SSIM values of images I_1 and I_2 and $\| I_1 - I_2 \|_1$ represents the $L1$-norm of the difference between images I_1 and I_2.

In addition to the L_d loss calculated by the discriminator network D based on the generated sample, L_c and L_s are used to retain the available data of the input corrupted images. For each inpainting image I, assume that the missing pixels are indicated by a "mask" matrix M, and each element in M encodes the pixel status, namely "1" for an existing pixel and "0" for a missing pixel. In the traditional exemplar-based inpainting methods, it is considered that the more information that is known around a pixel, the higher the repair priority of the patch where the pixel is located. This inspired us to increase the penalty for the loss in areas close to the missing region border. In other words, the closer a pixel is located to the repair boundary, the greater the penalty assigned to the loss.

To achieve this goal, we propose a weighted loss matrix W for the L_c and L_s losses. W is used to indicate the weight of the pixel loss penalty around the contour of the repair area. The element values in W can be obtained by M, similar to the confidence term [30], which has been used in exemplar-based image inpainting methods.

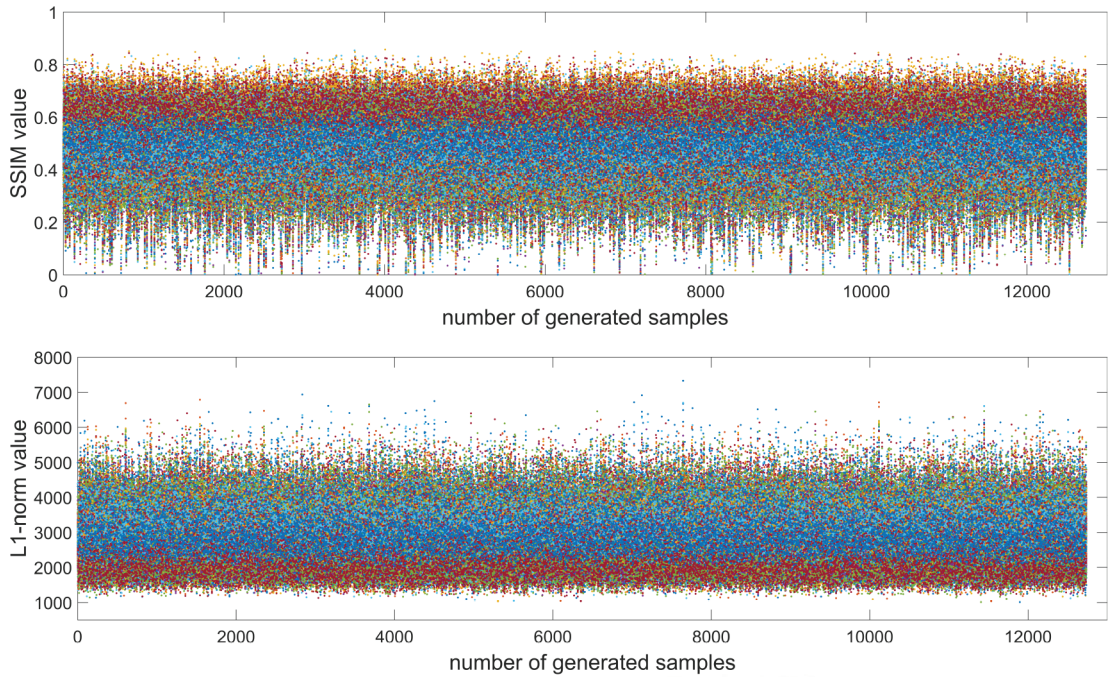


Figure 4 Scatter plots of the structural similarity index (SSIM) values and -norm values between generated samples and test samples.

We also set a block size parameter k. Then W can be calculated as follows:

$$W(p) = 1 + \alpha\left(\frac{\sum_{q\in\Phi_p}(O - M)}{k^2}\right) \tag{7}$$

where O is an all-ones matrix of the same size as M, Φ_p is a patch centerd at pixel p with size $k * k$, and α is a constant.

Through the calculation of Formula (7), the weight values of the pixels far from the repair contours (where the distance is greater than k) will be 1, and for the pixels near the repair contours, the weight values will increase according to the number of unavailable pixels nearby, which can ensure that greater loss penalty weight values are set for pixels closer to the repair border.

Thus, for each input corrupted face image I, the weighted multiple constraints loss function of our method is:

$$\begin{aligned} L(I) \;&= \lambda_1 Log(1 - D(G(z))) + \lambda_2 L_c(W \odot M \odot G(z), \\ &\quad W \odot M \odot I) + \lambda_3 L_s(W \odot M \odot G(z), W \odot M \odot I) \end{aligned} \tag{8}$$

where $\odot$ represents an element-wise product.

Figure 5 shows the results for image inpainting using each of the three losses independently. Figure 5(c) shows that the inpainted regions in the results using only the L_d loss constraint produce clear face structure information; however, as there is no constraint added to the repair process by using other information available in the images to be repaired, the generated face images are not related to the samples to be repaired, and the repaired results are correspondingly unreasonable. In the inpainting results using only the L_c loss constraint (Figure 5(d)), the repaired regions have a certain correlation with the input corrupted face images, but there is obvious blurring. We were surprised to find that by using only the L_c. loss as a constraint, face image inpainting can produce very good repair results (Figure 5(e)), which demonstrates that considering the structural loss is critical for image completion of images with obvious high-level semantic features.

3.3. Blending Results

By using a back-propagation algorithm based on the total loss of L_d, L_c and L_s, we can update the input signal z iteratively and thus obtain the closest $\hat{z}$ in the latent representation space of the corrupted image. However, unlike the traditional inpainting methods in which repair processes are carried out by diffusing the information from the repair contours in the missing region gradually, the inpainting methods based on deep generative models must stitch the generated loss blocks into the corrupted images. Therefore, it is necessary to use a blending method to ensure the natural transition of the image blocks from different images. Poisson blending [31] is used in our method to reconstruct the final results.

For an input corrupted image I, it is assumed that the most similar generated image is $T = G(\hat{z})$ by iterative updating. By using the Poisson blending method, the reconstruction process of the final result image, F, follows four steps: The first step is to calculate the gradient field of I and T, the second step is to merge the gradient fields of I and R according to the mask M, that is, replace the gradient field of the corrupted region in I with the corresponding region gradient field in T. The third step is to calculate the divergence of the merged gradient field. Finally, according to the Poisson reconstruction equation, the coefficient matrix is solved to obtain the reconstructed image regions corresponding to the input corrupted image. In Figure 6, we used gray scale images to illustrate the process clearly, and an illustration diagram of the

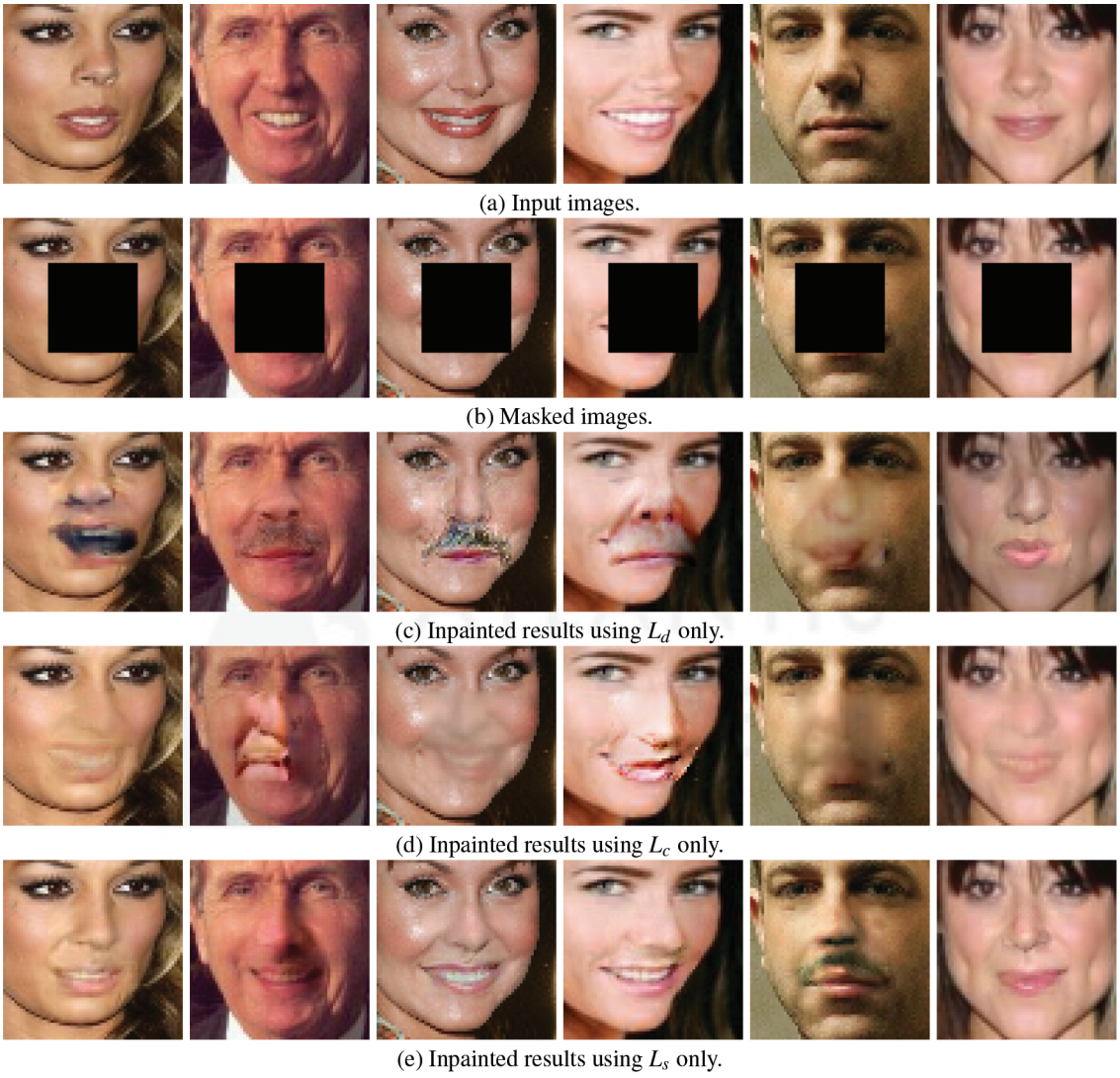


Figure 5 | Face image inpainting results using only loss constraints, respectively.

blending process is shown. In Figure 7, we present a comparison of several sets of inpainted results obtained with and without the blending method. The experimental results demonstrate that the final inpainting quality is improved significantly by the blending process.

3.4. Implementation Details

To generate sample images which are very close to the input face image, we make full use of the available information in the input image to restrain the generation process of the deep generative models (G and D). In our experiments, DCGAN is trained using the CelebA dataset, which consists of 202,599 face images (of which 2000 face images were reserved as a dataset for testing). After training, a $64 * 64 * 3$ image can be generated from a 100–dimensional vector using model G. For discriminator model D, the input is an image with dimensions of $64 * 64 * 3$ and its output is fed to a two-class softmax. The full architecture details can be found in [7]. The normalized parameters are set as $s_{min} = 0.1198$ and $s_{max} = 0.8431$. (For $64 * 12800$ SSIM values, the mean $\mu_s = 0.4836$, the standard deviation $\sigma_s = 0.1198$, $s_{min} = \mu_s - 3 * \sigma_s$, $s_{max} = \mu_s + 3 * \sigma_s$), and $c_{min} = 739$, and $c_{max} = 4861$ are set. (For $64 * 12800$ $L1$-norm values, the mean $\mu_c = 2800$, the standard deviation $\sigma_c = 687$, and $c_{min} = \mu_c - 3 * \sigma_c$ and $c_{max} = \mu_c + 3 * \sigma_c$). The difference loss weights are set as: $\lambda_d = 0.003$, $\lambda_c = 0.2$, and $\lambda_s = 0.797$ in all experiments. The values for these loss weights are set based on experience. In fact, through Figure 5, we can clearly see that the maximum weight should be given to the structural loss to ensure good repair results. The parameters in Formula (7) are set as: $k = 7$ and $\alpha = 2$, and their values are also set based on experience. Note that k cannot be too small; otherwise it cannot ensure the consistency of the repair contours. In all experiments, the number of iterations is set to 1000 to ensure convergence of the generated face image.

4. EXPERIMENTAL RESULTS

We carry out extensive experiments to demonstrate the capabilities of our proposed method compared to DIP, the state-of-the-art method for semantic face inpainting. Specifically, it includes various standard mask repair comparison experiments, arbitrarily

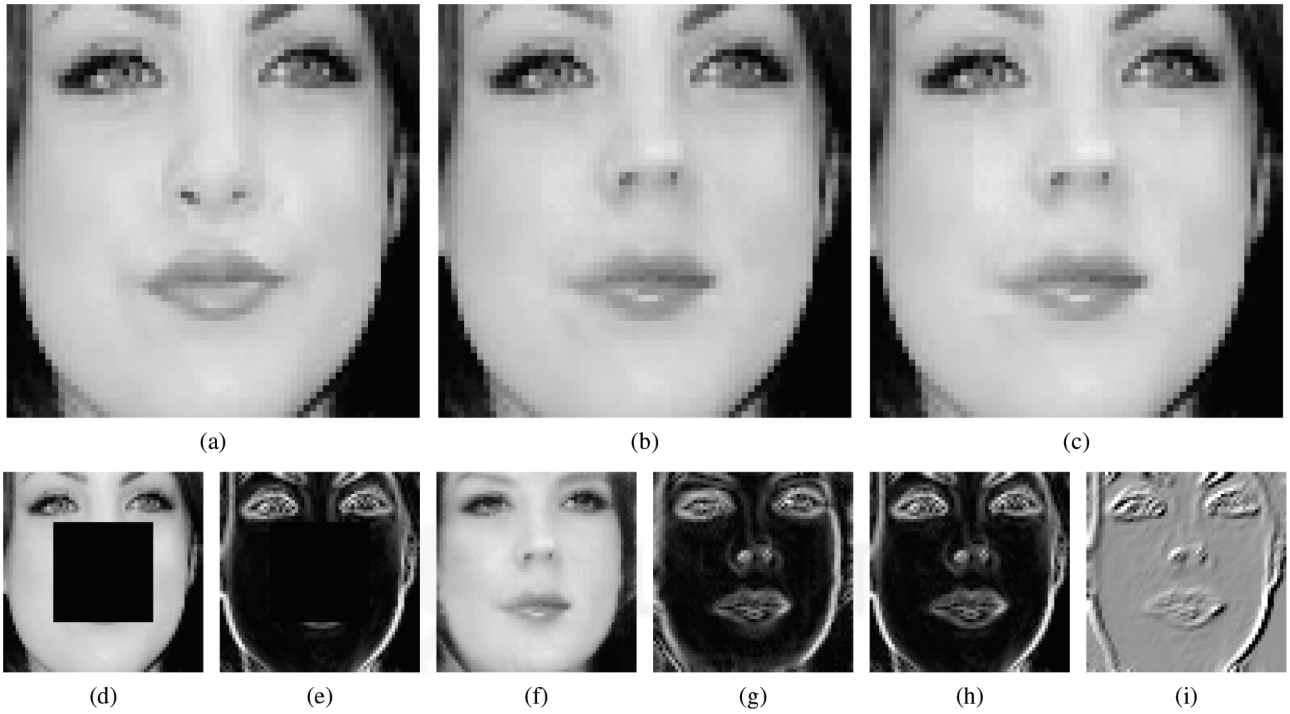


Figure 6 | Image blending processing illustration; (a) original image; (b) inpainting result with blending process; (c) inpainting result using the direct stitching method; (d) the masked input image; (e) the gradient field of (d); (f) iterative generated sample image; (g) the gradient field of (f); (h) merged gradient field by combining (e) and (g); and (i) divergence of the merged gradient field.

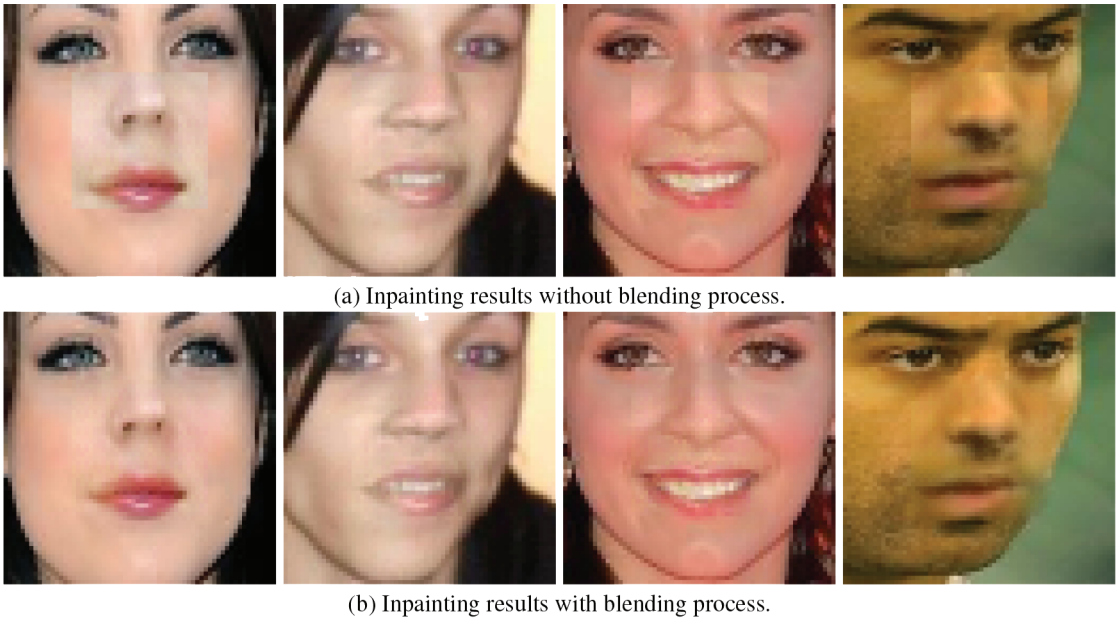


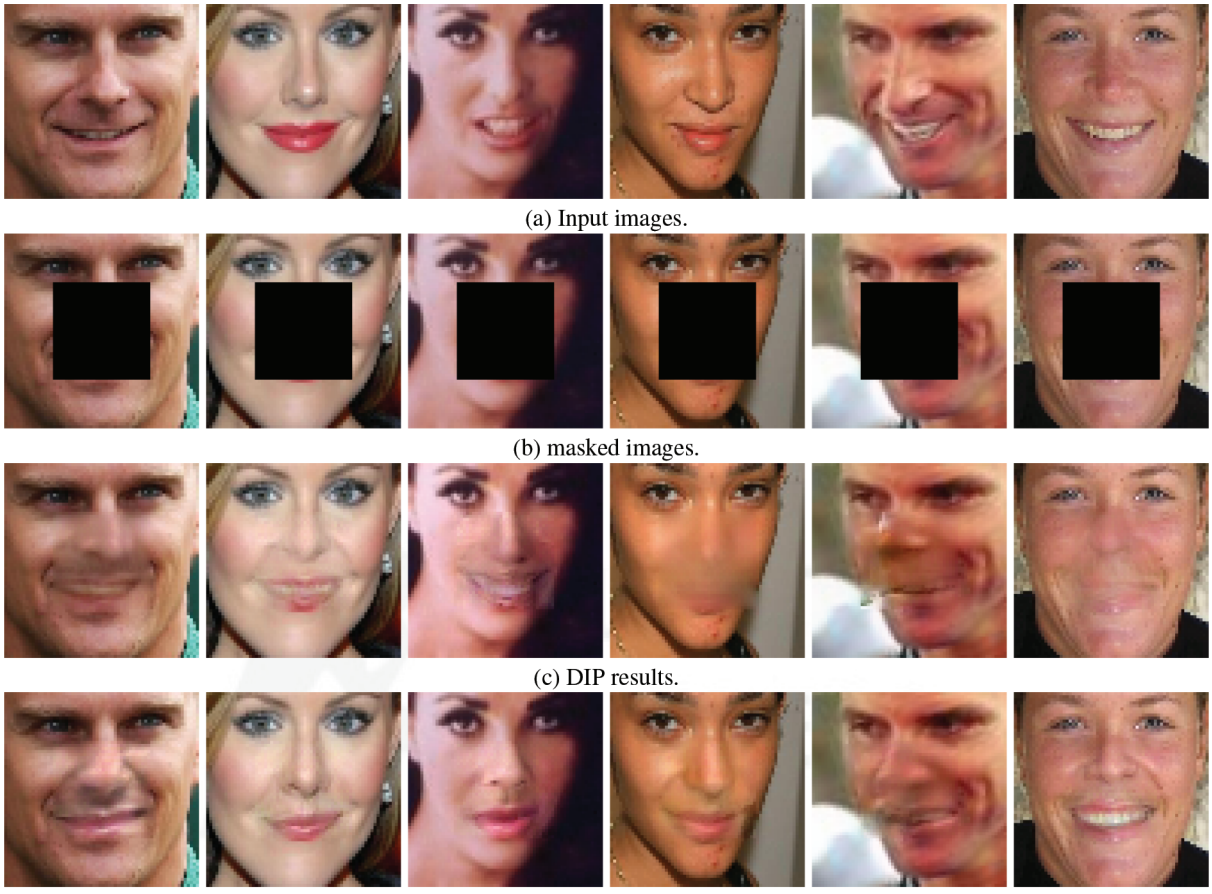
Figure 7 | Comparison of results between inpainting with and without blending.

masked face image repair comparison experiments, different face database comparison experiments, and a large number of pixel loss completion experiments.

4.1. Qualitative Analysis

In order to qualitatively analyze the performance of the proposed method, we compare the method with the DIP method. Figure 8 shows the repair results of the DIP method and our method in the case of center-masked images. Figure 9 shows the repair results for various standard masked images, and Figure 10 shows the repair results for arbitrarily masked face images.

In the six groups of experimental results given in Figure 8, our method achieved better repair results than the DIP method. There are richer face structures and more coherent edges in the proposed method; these properties make the overall repair results more



(a) Input images.

(b) masked images.

(c) DIP results.

(d) Our results.

Figure 8 | Comparison between DIP method and our method for center-masked images.

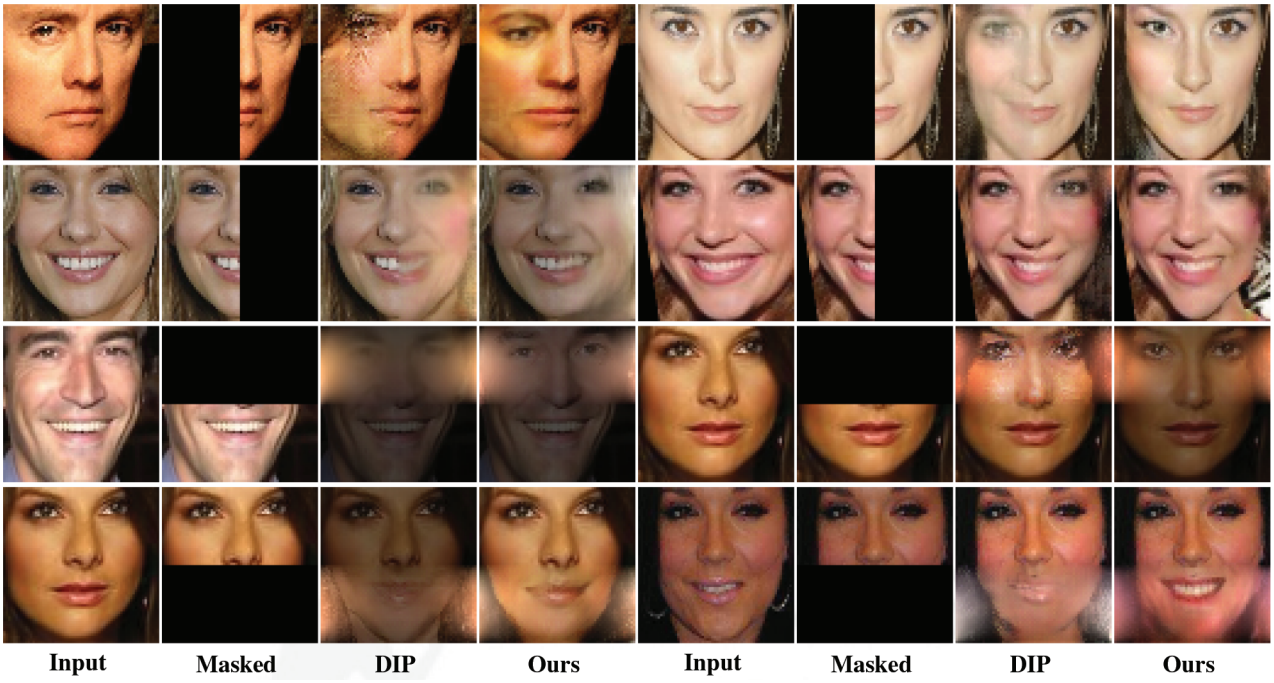


Figure 9 | Comparison between DIP method and our method for standard masked images.

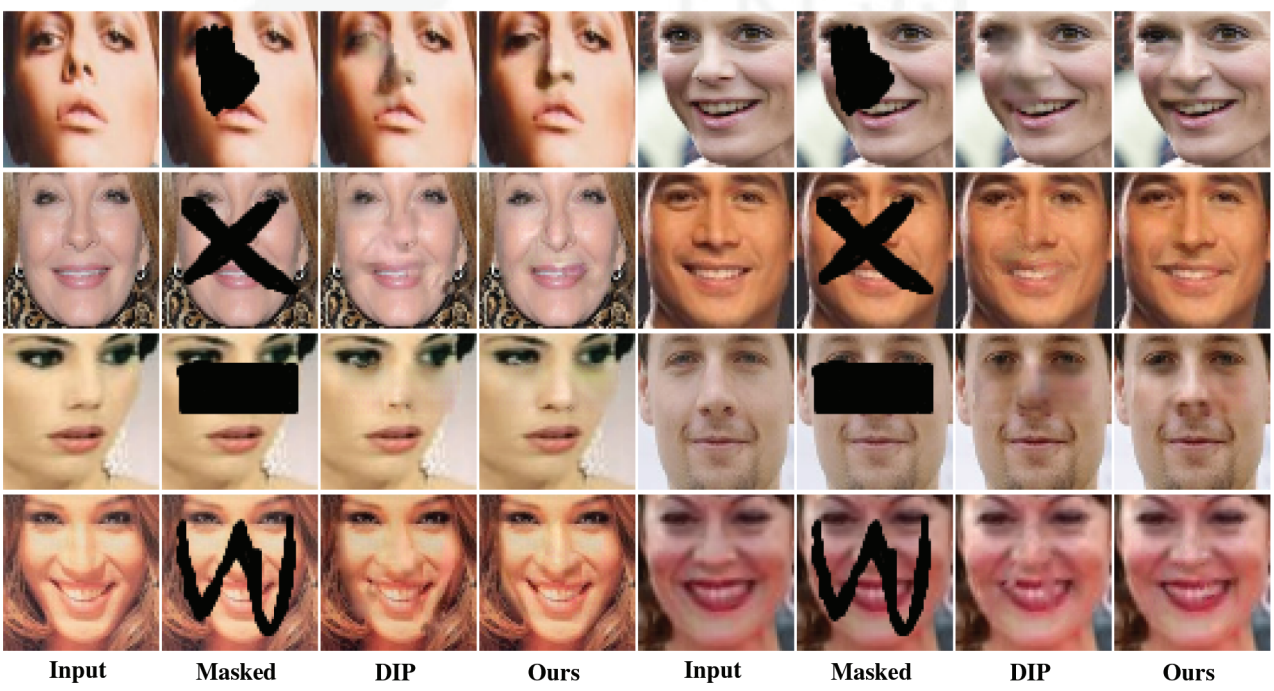


Figure 10 | Comparison between DIP method and our method for arbitrarily masked images.

reasonable. In Figure 9, except for the first group of experiments, which repair the upper half of the whole face image, the two methods are not ideal, while the other experimental results using the proposed method achieve better results. In addition to the advantages shown in Figure 8, more consistent facial skin is achieved in our results, and some experimental results have achieved amazing results in which it is difficult to find the repair traces to human eyes. In Figure 10, four different arbitrarily masked inpainting results are shown, and compared with the original images, there are obvious blurred marks in the results using DIP, and unique objects in the face images are more obvious(as shown by the results in the first line).

The effectiveness of the proposed method is verified by qualitative analysis through the three experiments and the results shown in Figures 8–10. The experimental results demonstrate that the final restoration results are more reasonable when the structural loss and adaptive weight strategy are applied.

4.2. Quantitative Analysis

To further verify the effectiveness of the proposed method, we compare the experimental results obtained by randomly discarding 20%, 40%, 60% and 80% of the pixels in an image. The visual experimental results are shown in Figure 11, where the first row to the fourth row show the inpainting results when discarding 20%, 40%, 60% and 80% pixels, respectively. The advantages and disadvantages between the proposed method in this paper and the DIP method are difficult to distinguish by the human eye in these results. Therefore, quantitative analysis results are presented in Table 1, and the peak signal-to-noise ratio (PSNR) and SSIM values of the inpainted images and the original unmodified images are compared. Note that the proposed method obtains relatively higher PSNR and SSIM values.

Table 1 | PSNR values (dB) and SSIM values comparison between DIP and our method based on different proportions of missing pixels.

	DIP		Ours	
	PSNR	SSIM	PSNR	SSIM
20%	19.8658	0.9517	20.0649	0.9533
40%	19.0177	0.8975	19.9802	0.9089
60%	18.2592	0.8547	19.8783	0.8892
80%	15.7974	0.7823	19.7447	0.8461

4.3. Limitations

All of the above experiments are based on the CelebA data set, and the test set is also from CelebA. On the trained network, we therefore use the data from other data sets to carry out further experiments. Figures 12 and 13 show the experimental results when using images from the SiblingsDB data set, and Figure 14 shows the repair results of the Asian face data set.

Although our method is capable of obtaining semantically pleasing inpainting results, even on images not used to train the network, it has some limitations. Through the results shown in Figures 12–14, we can find similar failure cases for both DIP and our method. This is because the CelebA data set images are roughly cropped and aligned, while the other data sets are not processed in this manner.

The generation model used in all experiments was DCGAN trained using images from the CelebA dataset, and it is difficult to obtain good results for face images that are not rich in the data set. As shown in Figures 14, the results for the two methods on Asian face images are unsatisfactory.

5. CONCLUSION

In this paper, we apply the experience gained from traditional image inpainting methods to semantic face inpainting based on

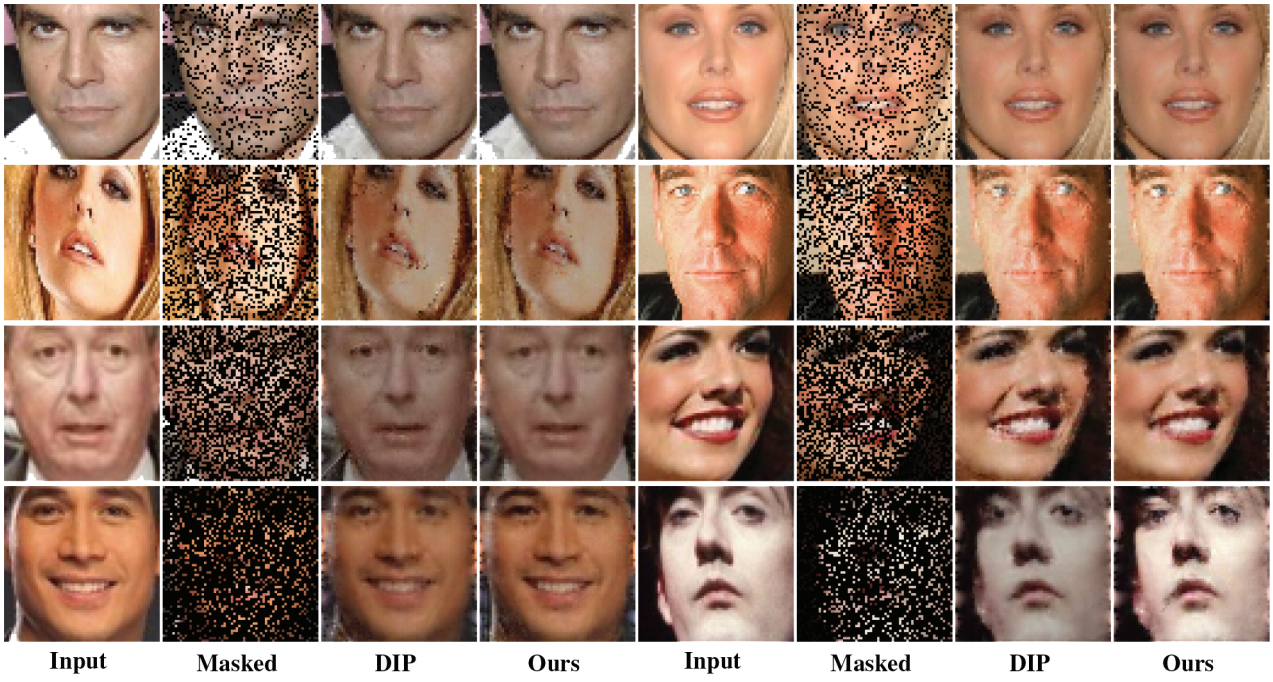


Figure 11 | Comparison between DIP method and our method when randomly discarding of the pixels in an image.

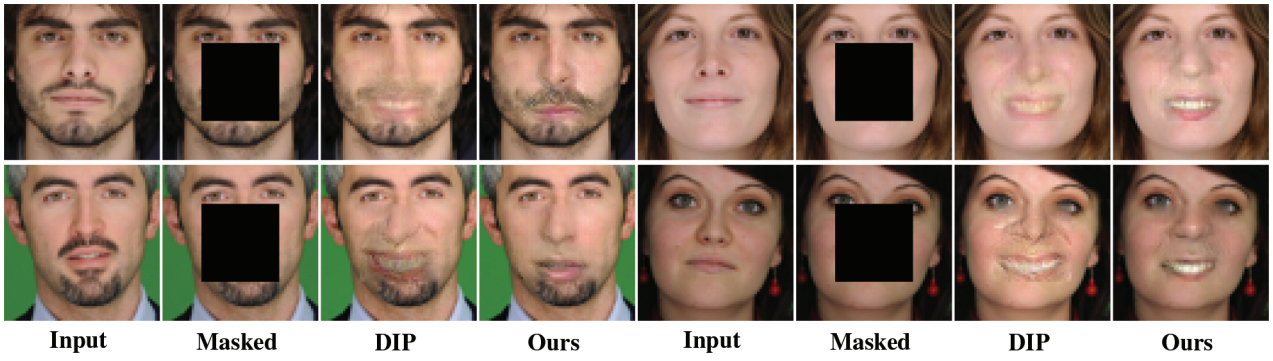


Figure 12 | Comparison between DIP method and our method for images from SiblingsDB with center mask.

Figure 13 | Comparison between the DIP method and our method on SiblingsDB images with arbitrary masks.

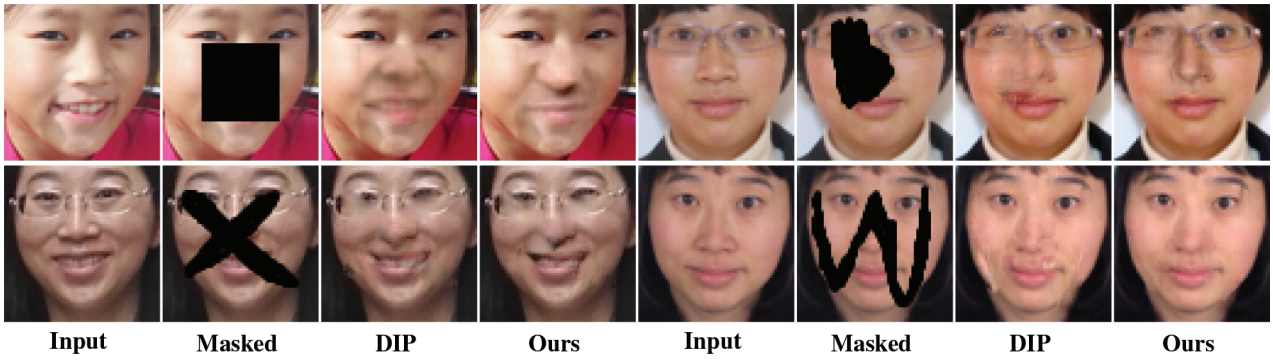


Figure 14 | Comparison between the DIP method and our method on Asian face images with different masks.

deep generative models, and through many experiments, the proposed method is proven effective for face inpainting. Compared with traditional image inpainting methods, the proposed method can achieve semantic face completion and its advantages are easily discerned. Compared with the DIP benchmark method proposed recently for semantic face inpainting, the proposed method can enhance content continuity and structural consistency and yield more reasonable inpainted results.

Although we have made some progress in face inpainting, there remains room for further improvement, and we propose the following promising directions for future work.

- Standard face model and corresponding representation loss function: The importance of a face's structural information for face inpainting has been demonstrated in this paper, and it is well known that the standard face model is a basic feature of a face's structure [32]. How to obtain the standard face model of the corrupted face image and represent the obtained result as a loss function will be a very valuable research contribution [33].
- Symmetric feature and corresponding representation loss function: Another basic feature of the face structure is symmetry [34], which is also a high-level semantic feature of the face. It is highly desirable to be able to represent and apply face symmetry feature(s) to face inpainting effectively.
- High-resolution face inpainting and synthesis: Notwithstanding GANs and other models that have greatly improved the quality of face completion, high-resolution face inpainting remains an open problem [35]. A synthesis approach may be an effective way to solve this problem [36].

In the future, we hope that more applications based on face image inpainting will be developed and applied in real life.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

ACKNOWLEDGMENTS

This work is supported by the project of National Natural Science Foundation of China (11603016), Key Scientific Research Foundation Project of Southwest Forestry University (111827) and the project of Scientific Research Foundation of Yunnan Police Officer College (19A010).

REFERENCES

[1] N. Ersotelos, F. Dong, Building highly realistic facial modeling and animation: a survey, Visual Comput. 24 (2008), 13–30.
[2] J. Shen, T.F. Chan, Mathematical models for local nontexture inpaintings, SIAP. 62 (2002), 1019–1043.
[3] C. Barnes, E. Shechtman, A. Finkelstein, *et al.*, PatchMatch: a randomized correspondence algorithm for structural image editing, TOG. 28 (2009), 1–11.
[4] O. Vinyals, A. Toshev, S. Bengio, *et al.*, Show and tell: a neural image caption generator, in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, 2015, pp. 3156–3164.
[5] Y. Lecun, L. Bottou, Y. Bengio, *et al.*, Gradient-based learning applied to document recognition, Proc. IEEE. 86 (1998), 2278–2324.
[6] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint, arXiv:1411.1784, 2014. https://arxiv.org/abs/1411.1784
[7] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint, arXiv:1511.06434, 2015. https://arxiv.org/abs/1511.06434
[8] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, arXiv preprint, arXiv:1701.07875, 2017. https://arxiv.org/abs/1701.07875
[9] D. Berthelot, T. Schumm, L. Metz, Began: boundary equilibrium generative adversarial networks, arXiv preprint, arXiv:1703.10717, 2017. https://arxiv.org/abs/1703.10717
[10] H. Zhang, V. Sindagi, V.M. Patel, Image de-raining using a conditional generative adversarial network, TCSVT. (2019), 1.
[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, Generative adversarial nets, in Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, 2014, pp. 2672–2680. http://dblp.uni-trier.de/db/conf/nips/nips2014.html#GoodfellowPMXWOCB14
[12] D. Pathak, P. Krahenbuhl, J. Donahue, *et al.*, Context encoders: feature learning by inpainting, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016, pp. 2536–2544.

[13] R.A. Yeh, C. Chen, T.Y. Lim, *et al.*, Semantic image inpainting with deep generative models, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 5485–5683.
[14] B. Dolhansky, C. Canton Ferrer, Eye inpainting with exemplar generative adversarial networks, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, 2018, pp. 7902–7911.
[15] Y. Bengio, P. Lamblin, D. Popovici, *et al.*, Greedy layer-wise training of deep networks, in Proceedings of the 19th International Conference on Neural Information Processing Systems, Canada, 2007, pp. 153–160.
[16] Y. Bengio, Learning deep architectures for AI, Found. Trends® Mach. Learn. 2 (2009), 1–127.
[17] J. Masci, U. Meier, D. Ciresan, *et al.*, Stacked convolutional auto-encoders for hierarchical feature extraction, in International Conference on Artificial Neural Networks (ICANN), Part I, Espoo, 2011, pp. 52–59.
[18] Y. Li, S. Liu, J. Yang, *et al.*, Generative face completion, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 3911–3919.
[19] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint, arXiv:1312.6114, 2013. https://arxiv.org/abs/1312.6114
[20] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, arXiv preprint, arXiv:1605.09782, 2016. https://arxiv.org/abs/1605.09782
[21] A. Lahiri, K. Ayush, P.K. Biswas, *et al.*, Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images: automated vessel segmentation in retinal fundus image as test case, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 42–48.
[22] T. Salimans, I. Goodfellow, W. Zaremba, *et al.*, Improved techniques for training gans, in Advances in Neural Information Processing Systems, Barcelona, 2016, pp. 2226–2234. http://dblp.uni-trier.de/db/conf/nips/nips2016.html#SalimansGZCRCC16
[23] E.L. Denton, S. Chintala, R. Fergus, Deep generative image models using a Laplacian pyramid of adversarial networks, in Advances in Neural Information Processing Systems, Montreal, 2015, pp. 1486–1494. http://dblp.uni-trier.de/db/conf/nips/nips2015.html#DentonCSF15
[24] J.Y. Zhu, P. Krähenbühl, E. Shechtman, *et al.*, Generative visual manipulation on the natural image manifold, in European Conference on Computer Vision (ECCV2016), Amsterdam, 2016, pp. 597–613.
[25] P. Isola, J.Y. Zhu, T. Zhou, *et al.*, Image-to-image translation with conditional adversarial networks, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 1125–1134.
[26] H. Zhang, T. Xu, H. Li, *et al.*, Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks, in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 5907–5915.
[27] H. Zhang, T. Xu, H. Li, *et al.*, StackGAN++: realistic image synthesis with stacked generative adversarial networks, TPAMI. 41 (2019), 1947–1962.
[28] A. Nguyen, J. Clune, Y. Bengio, *et al.*, Plug and play generative networks: conditional iterative generation of images in latent space, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 4467–4477.
[29] Z. Wang, A.C. Bovik, H.R. Sheikh, *et al.*, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004), 600–612.
[30] A. Criminisi, P. Perez, K. Toyama, Object removal by exemplar-based inpainting, in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Madison, 2003, pp. 721–728.
[31] P. Pérez, M. Gangnet, A. Blake, Poisson image editing, TOG. 22 (2003), 313–318.
[32] S.F. Kak, F.M. Mustafa, P.A. Valente, A review of person recognition based on face model, Eurasian J. Sci. Eng. 4 (2018), 157–168.
[33] J. Deng, J. Guo, N. Xue, *et al.*, Arcface: additive angular margin loss for deep face recognition, in 2019 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, 2019, pp. 4690–4699. http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
[34] R. Huang, S. Zhang, T. Li, *et al.*, Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis, in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2439–2448.
[35] B. Jiang, H. Liu, C. Yang, *et al.*, Face inpainting with dilated skip architecture and multi-scale adversarial networks, in 2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), Taipei, 2018, pp. 211–218.
[36] C. Yang, X. Lu, Z. Lin, *et al.*, High-resolution image inpainting using multi-scale neural patch synthesis, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 4076–4084.