IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 30, 2019, accepted May 20, 2019, date of publication May 27, 2019, date of current version June 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919169

Facial Image Inpainting With Deep Generative Model and Patch Search Using Region Weight

JINSHENG WEI, GUANMING LU, HUAMING LIU, AND JINGJIE YAN

College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Corresponding author: Guanming Lu (lugm@njupt.edu.cn)

This work was supported in part by the Key Research and Development Program of Jiangsu Province under Grant BE2016775, in part by the Key Projects of Natural Science Research in Anhui Colleges and Universities under Grant KJ2018A0345, in part by the National Natural Science Foundation of China under Grant 61501249, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150855, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX18_0901, and in part by the Project funded by China Postdoctoral Science Foundation under Grant 2018M632348.

ABSTRACT Facial image inpainting is a challenging task because the missing region needs to be filled by the new pixels with semantic information (e.g., noses and mouths). The traditional methods that involve searching for similar patches are mature but it is not suitable for semantic inpainting. Recently, the deep generative model-based methods have been able to implement semantic image inpainting although inpainting results are blurry or distorted. In this paper, through analyzing the advantages and disadvantages of the two methods, we propose a novel and efficient method that combines these two methods by a series connection, which searches for the most reasonable similar patch using the coarse image generated by the deep generative model. When training model, adding Laplace loss to standard loss accelerates model convergence. In addition, we define region weight (RW) when searching for similar patches, which makes edge connection more natural. Our method addresses the problem of blurred results in the deep generative model and dissatisfactory semantic information in the traditional methods. Our experiments, which used the CelebA dataset, demonstrate that our method can achieve realistic and natural facial inpainting results.

INDEX TERMS Facial image inpainting, deep generative model, similar patch, region weight.

I. INTRODUCTION

Image inpainting, which can fill in the missing region caused by human or inhuman factors, is a challenging and important branch in the field of computer vision. According to the characteristics of the missing region, we can divide image inpainting into semantic image inpainting and un-semantic image inpainting. Semantic image inpainting is more difficult than un-semantic one as it needs to fill the missing region with semantic information that does not exist in the input image. Facial image inpainting [1], [2], which is a type of semantic image inpainting, also has the same difficulty.

For the implementation of image inpainting, the deep learning methods and traditional methods have their own shortcomings. Some traditional methods that use the available pixels in the input image to fill the missing region, including Total Variation (TV) [3], Low Rank (LR) [4], PatchMatch (PM) [5], Fast Matching Method (FFM) [6] and so on, can be applied in object removal [7] and texture inpainting [8], [9]. For instance, PM searches for similar patches in the input image and replaces the missing region, which is significantly effective in texture inpainting. However, these methods are completely unavailable for those tasks that need to repair semantic information.

Some early studies used traditional methods to solve the task of semantic image inpainting. These studies [10], [11] used the information around the missing region to find a similar patch from the Internet, before replacing the whole missing region with this similar patch. This method effectively solves the irreparable problem of semantic information but because there is no prior semantic information in the missing region, their results are prone to discontinuous edges and can produce incorrect semantic information.

Recently, with the rapid development of deep learning, the deep generative models, including the original Generation Adversarial Network (GAN) [12] and VAE-based Generation Adversarial Network (VAE-GAN) [13], have greatly advanced image inpainting. The GAN based methods can

The associate editor coordinating the review of this manuscript and approving it for publication was Peng Liu.

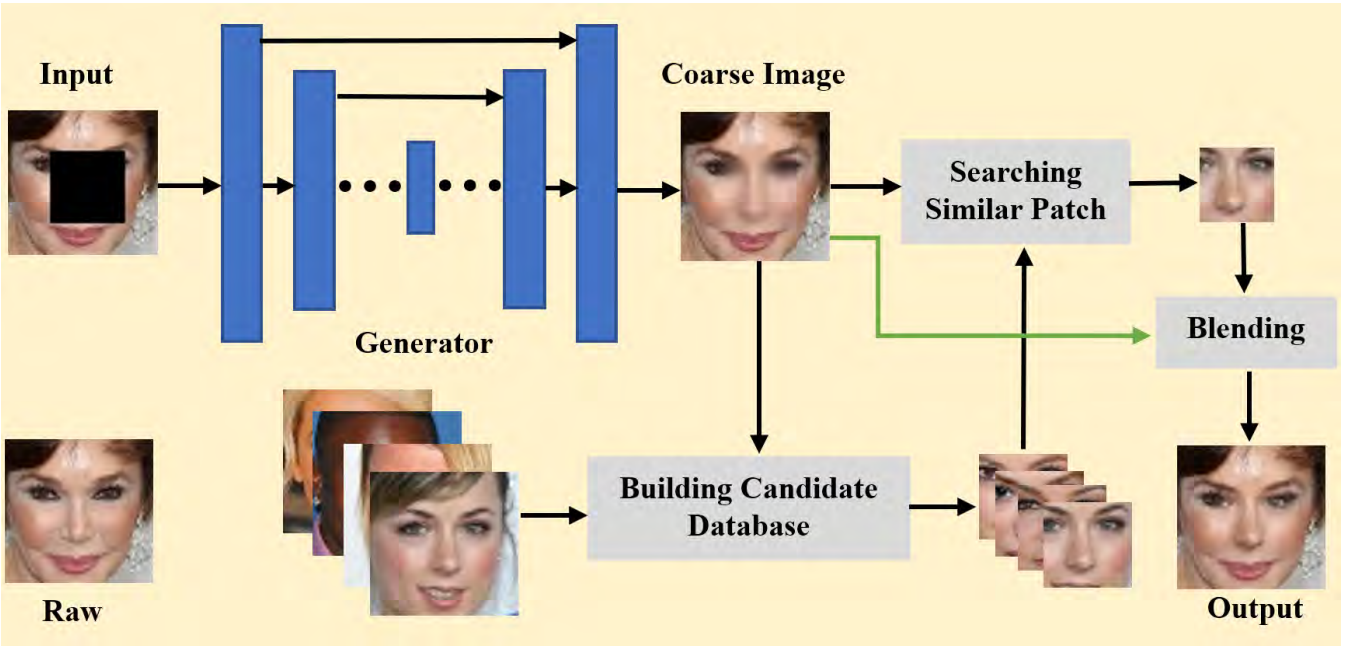

FIGURE 1. Overview of our inpainting framework. The generator is a trained "U-Net" generator. The incomplete image is input into the generator to get the coarse image X that is blurry in the missing region and the candidate database is established by using the image X and facial image dataset. After this, the best similar patch is searched from the candidate database. Finally, Poisson blending is carried out against the background of the image X.

produce clear results but is prone to distortion. In contrast, the VAE-GAN based methods can generate more reasonable edges and semantic information, but unfortunately, their results are relatively blurry.

As shown in Fig.1, we propose a novel method that combines deep generative model with searching for similar patches. The proposed method first trains a "U-Net" [21] generator using the Pix2Pix [22] model whose architecture is similar to VAE-GAN and the generator generates a coarse image whose patch of the missing region is blurry with semantic information. After this, our method searches for a similar patch in a large facial image dataset by utilizing this coarse image. Finally, Poisson blending [14] is used to connect the similar patch and the coarse image. The combination of the two methods solves their respective shortcomings, including the blurry results from the deep generative model and the lack of prior semantic information when using the method of searching for similar patches. Experiments indicate that our method has more realistic and natural results than the two methods.

The method proposed in this paper contains the following contributions: (1) We propose a new method that combines deep learning with traditional methods. Using prior semantic information generated by the deep generative model is beneficial when searching for similar patches, and moreover, the method also solves the problem that the results are blurry in the deep generative model, which improves the results of facial image inpainting. (2) We add a new Loss, Laplace Loss, to the Loss Function of Pix2Pix for accelerating the convergence rate. (3) During searching for similar patches, in order to accommodate the new case where the missing region contains prior information, different regions are given different weights by defining a new strategy called Region Weight (RW) and edge distance is added to the calculation of distance, which improves edge connection.

II. RELATED WORKS

Image inpainting can be divided into many categories, including texture inpainting, semantic inpainting and so on. As this paper deals with facial image inpainting with semantic information, we will mainly introduce the methods for semantic image inpainting by the two aspects of deep learning and traditional methods.

A. TRADITIONAL METHOD

Early image inpainting mainly relies on the information from the existing region in the input images. The TV based methods [15], [16] that considers the smoothness property are a basic algorithm to denoise image by solving the extreme value of a function. By constructing a model that fuses prior low rank matrix and solving the model, LR based methods [4], [17] can effectively improve the results of denoising and deblurring tasks. Criminisi [18], which searches for similar patches from the non-missing region of the input image, is a classical algorithm, but it is limited to the inpainting of the texture and background. Many methods [19], [20] that are similar to Criminisi cannot be applied to semantic image inpainting by matching and copying similar patches from a single image.

References [10], [11] introduced a method that can be used for inpainting semantic images. This method is based on massive images from the Internet and uses the information around the missing region to find the best similar patch to fill the missing region. However, due to the absence of prior

information, the results of the method are prone to erroneous semantic information and often produce discontinuous edges.

B. DEEP LEARNING

The deep learning method based on a large number of samples involves the use of the deep generative model to find the potential distribution characteristics of missing regions and breaks through the bottleneck of traditional methods in semantic image inpainting.

The deep generative model-based methods mainly use GAN and VAE-GAN for semantic image inpainting tasks. On the one hand, the GAN based method [23] involves training an image generator and then using trained generators to generate the most suitable patch for the missing region by adjusting the input random vector. Although the image generated by the generator is sharp, it can be easily distorted so the results of this method, which depends on the performance of the generator, can be easily distorted. On the other hand, VAE-GAN, which combines variational Auto-encoding (VAE) [24] and GAN, can encode the input of an incomplete picture into a vector automatically, before decoding the vector into a complete image. Recently, multiple studies have utilized this model. The Context Encoder (CE) [25] adopts the basic structure of VAE-GAN but the results are blurry and unrealistic. By replacing the original discriminator network with a global context discriminator network and a local context discriminator network, Iizuka *et al.* [26] proposed a model that considers both global and local information, which improves the inpainting results. For this task, Li *et al.* [2] introduced semantic parsing networks to add a semantic parsing loss to the three losses mentioned in [26]. In addition, Isola *et al.* [22] combined the VAE generator with skips and Markovian discriminator [27] to get the Pix2Pix model which can be applied in many fields apart from image inpainting, such as style transfer [28] and super-resolution [29]. Compared with the method [26], its results are better in skin color matching and more robust. Yu *et al.* [32] took the attention mechanism to refine the coarse result using surrounding available pixels. Aiming at irregularly masked images, Liu *et al.* [34] proposed partial convolutions, where the convolution is masked and renormalized is conditioned only on valid pixels. Although this model is well applied to irregular mask and its results are sharp, the results are distorted for continuous masks of a large region. Yang *et al.* [35] presented a block-wise procedural training scheme to address the difficulty of training a very deep generative model and adversarial loss annealing to improve inpainting result. Wang *et al.* [33], designed a Laplacian-pyramid-based convolutional network framework to predict missing regions under different resolutions and adopted modified residual learning model to matching color, which works well on facial image inpainting. The above methods continuously improve the performance of inpainting result, however, they cannot address the artifacts including the lack of detail and blurry or distorted results.

Our work combines Pix2Pix with searching for similar patches and uses the coarse image generated by Pix2Pix to search for similar patches from a large-scale facial image dataset. In addition, the proposed strategies (RW and Laplace Loss) are used to enhance the performance of this combination. We aim to solve the problem of the blurry results in the deep learning methods and semantic information errors in the traditional methods, which makes inpainting results more realistic and reasonable.

III. METHOD

By emphasizing the importance of edges, we accelerate the convergence of the original model (Pix2Pix). After this, the outputs of the improved model are used to search for similar patches. Furthermore, we propose a novel searching algorithm when searching for similar patches, which utilizes Region Weight (RW) and edge distance.

A. IMPROVED MODEL

Pix2Pix mainly consists of two networks: generator G and discriminator D. The generator G uses ''U-net'' which is similar to VAE in architecture and both ''U-net'' and VAE contain an encoder and a decoder. Unlike VAE, ''U-net'' uses a skip between the encoder and the decoder, and the loss function also is different. In our work, we use the Pix2Pix's loss function and add the Laplace Loss to it only for accelerating the convergence rate.

The training of the model is alternately carried out in the form of a game. G tries to generate a fake image that is more similar to the real image. In contrast, D tries to distinguish the real and fake images. Correspondingly, G tends to minimize the loss function:

$$L_G = E(||y - G(x)||_1). \tag{1}$$

D tends to minimize the loss function:

$$L_D = \min_G \max_D E(logD(y) + log(1 - D(G(x)))). \tag{2}$$

The complete loss function for original model is:

$$L_{orig} = \alpha L_G + \beta L_D. \tag{3}$$

where y, G(x) and D(*) respectively denote the real image, the output of G when x is the input and the output of D. Furthermore, α and β are proportion coefficients.

According to the loss function of the generator, the generator has the same bias for each pixel of the generated image. However, the human is very sensitive for the edge of the facial image in evaluating the visual effects. In order to make the model learn the edge information faster, the edge penalty is added to the loss function (Fig.2). Thus, Laplace Loss is defined as:

$$L_L = E(||\Delta y - \Delta G(x)||_1). \tag{4}$$

where Δ denotes the second derivative implemented by the Laplace operator.

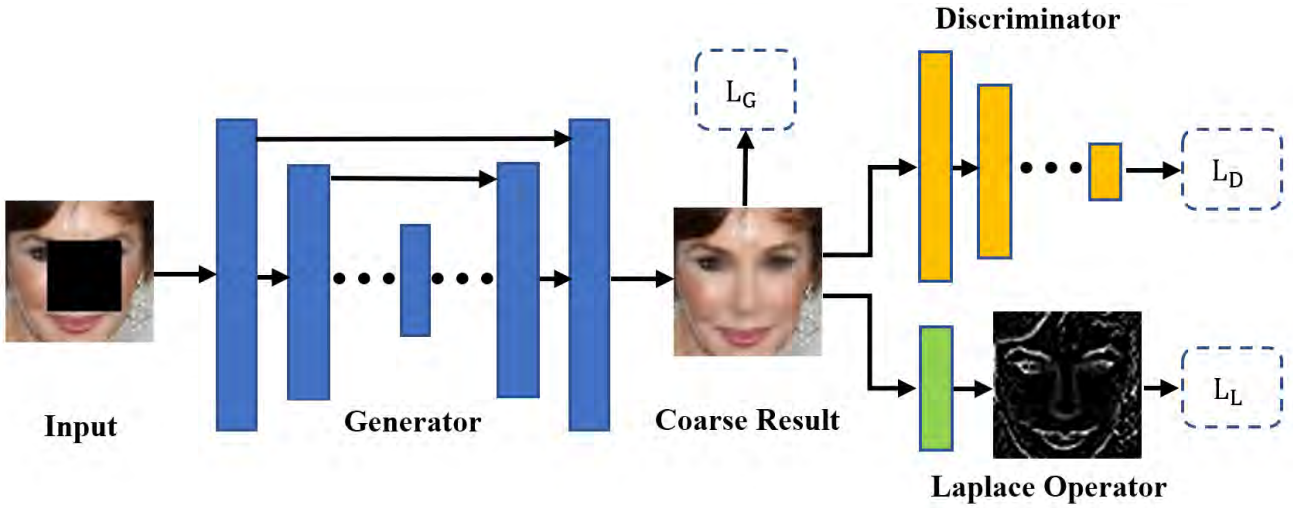

FIGURE 2. Improved model framework. The generator uses "U-Net" [21] and Discriminator is a two-class network in Pix2Pix [22]. We produce Laplace Loss L_L using the Laplace features of the raw images and that of the coarse result generated by the generator.

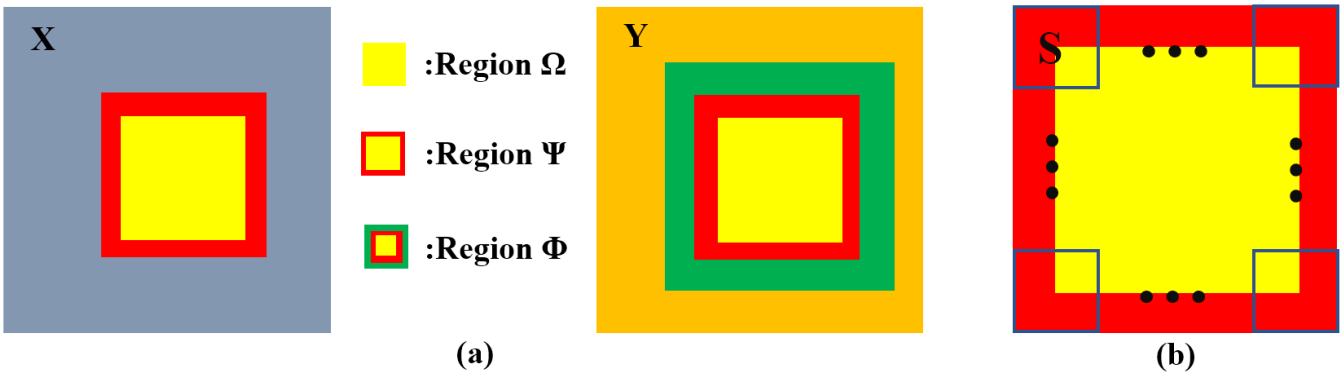

FIGURE 3. Division of regions. (a) X is the coarse image generated by the generator and Y is one of these images in the image dataset, The patch of the region Ψ in the image X is used to search for similar patches from the patch of the region Φ in the image Y. (b) division of n small regions around the missing region Ω.

Therefore, the total loss of the improved model is L_{total}:

$$L_{total} = \alpha L_G + \beta L_D + \gamma L_L. \tag{5}$$

where γ is a proportion coefficient.

The output of the trained generator is used to search for similar patches. Since subsequent operations only need it to provide skin color information and semantic information, it need not be extremely clear. Thus, our method is robust for the output of the generator.

B. SEARCHING FOR SIMILAR PATCH

After the incomplete image pass through the improved model, we obtain the coarse image X whose patch in the missing region, containing semantic information, is blurry. Blurring has a significant visual impact on an image so the results should be further optimized. The image X is used to search for a clear similar patch R and the missing region is replaced with the patch R to obtain a clearly complete image.

In this paper, searching for similar patches includes two parts: building the candidate database and searching for the best similar patch. These two parts are described in detail below.

1) BUILDING CANDIDATE DATABASE

These methods [10], [11] only use the information of the non-missing region to search for similar patches so they easily produce errors in the semantic information of the selected patches. For example, the hair close to eyes is considered as a sunglass to search for the facial patch with sunglasses, which is shown in Fig.10. Our method builds the candidate image database DB according to the coarse image X generated by the improved model. The semantic information of the patches in DB is correct as it considers both the pixels of the missing region and its surrounding regions in the image X.

The image X is divided into different regions, which is shown in Fig.3(a). The image X is divided into a missing region Ω and a small region around the region Ω. Both of them constitute a region Ψ. Furthermore, in the image Y from facial image dataset, the region Ψ and its surrounding regions constitute a region Φ in addition to defining the region Ω and Ψ. After intercepting the patch of the region Ψ in the image X as a patch P, preliminary similar patches are searched

from the image Y according to the patch P. In order to reduce the computational cost of the search and further reduce the probability of semantic information errors, we only search for the nearest patch in the region Φ of the image Y (not in the whole image Y) according to the similarity of the facial image in facial image dataset. The searching method also enhances the robustness of our method in terms of facial position. An image Y_i is extracted from the facial image dataset. A patch Q_{ij} with the same size as the patch P is slipped and intercepted in the region Φ of the image Y_i. Based on the square Euclidean distance between the patch P and the patch Q_{ij}, the patch Q_{imin} that is most similar to the patch P is obtained in the image Y_i. The formula is as follows:

$$Q_{imin} = arg \min_{Q_{ij}} d^2(P, Q_{ij}). \tag{6}$$

where d(x,y) denotes the Euclidean distance between x and y; and the patch Q_{ij} represents the patch j in the region Φ of the image Y_i

For each image Y_i in the facial image dataset, we obtain a preliminary similar patch Q_{imin}, whose Euclidean distance with the patch P is d_i, and sort these patches from small to large according to the d_i value. After this, the first N patches are added as candidate patches to DB.

As the region Ω in the image X is blurry but contains semantic information, the candidate patches searched by our method are more suitable for our follow-up needs compared to the previous works [10], [11].

2) REGION WEIGHT

Before searching for the best similar patch, we first introduce the concept of RW.

The region with edges is more important than the region without edges for the edge connection of the boundary. As shown in Fig.4, the connection in the eye region is more important than that in the skin region. A more important region has a greater edge feature response. Thus, in order to make the edge connection more natural, different weights are assigned to different regions according to the edge feature map. As shown in Fig.3(b), the image X is divided into n small regions S_i around the missing region Ω. Because the edges of the region other than the n region S_i are continuous and our method searches for the best similar patch only according to these n small regions, weights only are assigned to these n small regions. the second derivative of the patch P is calculated to obtain an edge feature map F, before summing the pixel values in a small region S_i of the feature map F as the weight of the small region S_i. In detail, the weight of the small region S_i is determined as follows:

$$w_i = \sum_{k \in S_i} \widehat{P}(k) \qquad i = 1, 2 \ldots n. \tag{7}$$

$$\widehat{P}(k) = \begin{cases} m * \Delta P(k) & if\ k \in \Omega \\ \Delta P(k) & if\ k \in \Psi - \Omega \end{cases}. \tag{8}$$

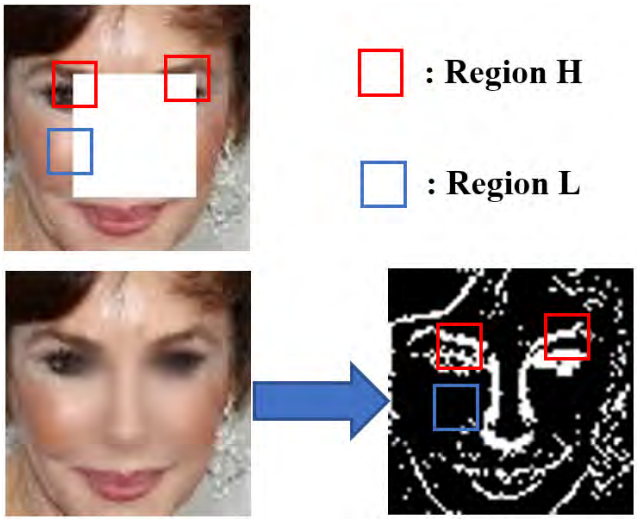

FIGURE 4. The importance of different regions. Region H is more important than Region L.

where k stands for the pixel point and P(k) represents the pixel value at the pixel point k. Because blurring weakens the edge response, the second derivative value in region Ω is multiplied by m with a value that is greater than 1 to enlarge its value. It is important to note that the weight is only related to the patch P from the input image but not to the patches in DB.

Considering that noise affects the calculation of the edge feature map F and influence the weight, we set a threshold before calculating (7) and after calculating (8). The pixel value smaller than the threshold is set to 0 while the pixel value larger than the threshold is unchanged. Above calculation results for the feature map F are illustrated as shown in Fig.5.

3) SEARCHING FOR BEST SIMILAR PATCH

This section involves determining the best similar patch from the N patches in the candidate database DB. Although these patches in DB are close to our needs in skin color and semantic information, we still need to further align the edges to achieve more natural edge connections.

Using RW, each patch B_i in DB is scored and the patch with the highest score is what we need. The purpose of this section is mainly to search for the patch that can make edges more continuous at the boundary, hence the edge distance is added to the general distance. As the score is inversely proportional to the distance, the reciprocal of distance is used as the score for these n small regions. After this, these n scores are weighted according to RW of the n regions. The formula for selecting patch is as follows:

$$B_{best} = arg \max_{B_i} \sum_j w_j * [\sum_{k \in S_j} \hat{d}(P(k), B_i(k))]^{-1}. \tag{9}$$

$$\hat{d}(P(k), B_i(k)) = q * d^2(\widehat{P}(k), \Delta B_i(k)) + d^2(P(k), B_i(k)). \tag{10}$$

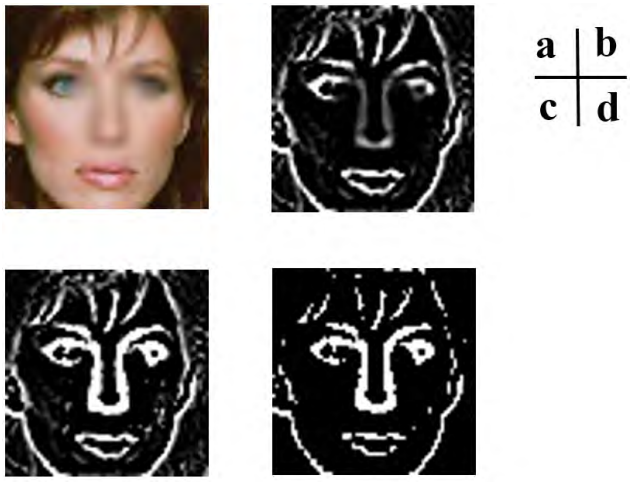

FIGURE 5. Some calculation results for Laplacian feature map. a: the coarse image X generated by generator, b: Laplacian feature map of the image X, c: the enhanced feature map in the blurry region, and d: feature map after noise is removed.

where q is a constant coefficient; $i = 1, 2, \ldots, N; j = 1, 2, \ldots$, n; i and j denote the index of the N patches in DB and the index of the n region S in a patch, respectively; w_j is the weight of region S_j and is independent of i; and $\widehat{P}(k)$ is calculated by (8).

C. INPAINTING

Once the best similar patch B_{best} is obtained, the pixels in the region Ω of the image X can be replaced with the pixels in the counterpart region of the best similar patch B_{best}. We name the patch of the region Ω in B_{best} as patch R. After this, our method only need to replace the patch $\widehat{R}$ of the region Ω in the image X with the patch R. However, as the patch R is visually incompatible with the surrounding regions, Poisson blending is used to address this problem. The image X and the patch R are the background and foreground respectively. Poisson blending has been used in many previous methods [23], [26] but unlike these methods, the image X contains the correct information of skin color and continuous edges, which is more conducive for computing the gradient and produces more natural skin color for the blending results.

IV. EXPERIMENT

In this section, we evaluate the proposed method on the CelebA facial image dataset [30]. First, we compare the original model (Pix2Pix) with the improved model (Pix2Pix with Laplace Loss) in terms of the convergence rate of each model. Second, we compare edge connections with and without Region Weight (RW). Third, we compare the inpainting results of some traditional methods, original model, and other four methods [23], [26], [32], [34] with our results. In addition, we extend our method to high resolution image inpainting. At the end of this section, we present and analyze the limitations of our method and mention future work.

A. DATASET, PARAMETER AND MASK

Since both deep generative model and searching for similar patches require a large number of facial image samples, we chose the CelebA dataset, which contains 202,599 face images and their landmark locations. We cropped and resized these images into 64*64*3 facial images with similar facial positions. Searching for similar patches requires high-quality and semantically correct image sources so we artificially deleted mislabeled and low-quality facial images. Finally, approximately 185,000 face images were selected as the experimental dataset and the last 60 images were chose as the testing set.

In the training, the learning rate is 0.001 and α, β, γ, used to balance L_G, L_D, and L_L, are 10, 1, and 1, respectively. We used a 5*5 Laplacian operator to overcome the difficulty of edge extraction in blurry images. In addition, we set the following parameters: N = 50, m = 5, and q = 20, where N is the number of patches in the candidate database DB, m is mentioned in (8) and q is mentioned in (10).

For better testing, we masked at least two facial organs, with each mask containing edge connections on their boundary. We mainly tested three types of masks: 1) masking mouth, nose, and eyes; 2) masking nose and eyes; 3) masking mouth and nose. In addition, the other three different shapes and sizes of masks were also employed for testing.

B. ABLATION STUDY

We investigate the effectiveness of Laplace Loss and RW. The experimental results show that Laplace Loss can accelerate the convergence of the model, and RW can make the edge connection more continuous and natural.

1) CONVERGENCE RATE

We compare the convergence rate between the original model and the improved model. Fig.6 qualitatively shows the changes in the inpainting results with epoch changes in two methods. As shown in the first epoch from Fig.6, the improved model already generates the obvious contour of noses and eyes whereas the original model's results are strongly distorted. The improved model begins to stabilize at the third epoch while the original model begins to stabilize at the ninth epoch. Furthermore, our careful observation reveals that the original model can easily generate a striped texture, such as the mouth part of from the first epoch to the ninth epoch.

A comparison of PSNR curves is shown in Fig.7. The improved model can quickly achieve a relatively high PSNR value and overall, the PSNR values of the improved model are also slightly higher than that of the original model.

2) EDGE CONNECTION

We compare the edge connection with and without RW. Fig.8 shows the results with and without RW before blending. The results with RW are more continuous and natural generally than ones without RW, such as edge connection of the eyes, nose, and facial wrinkles.

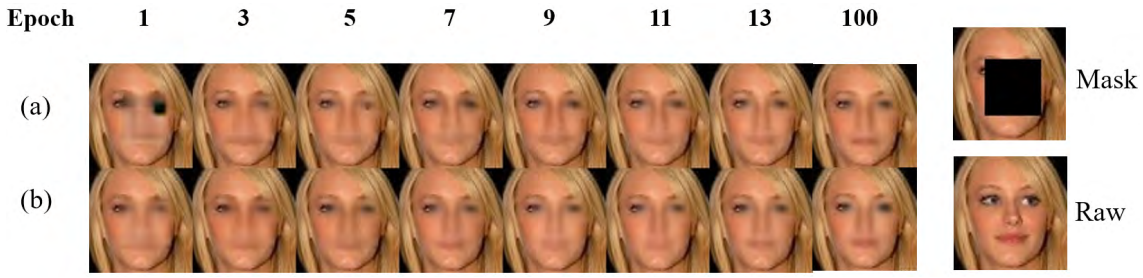

FIGURE 6. Comparison of convergence rate. (a) the original model; (b) the improved model.

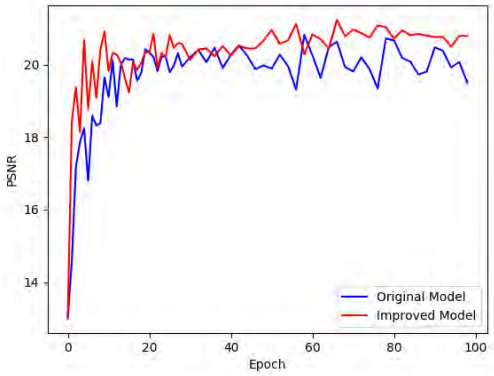

FIGURE 7. Comparisons of PSNR curve between the improved model and original model.

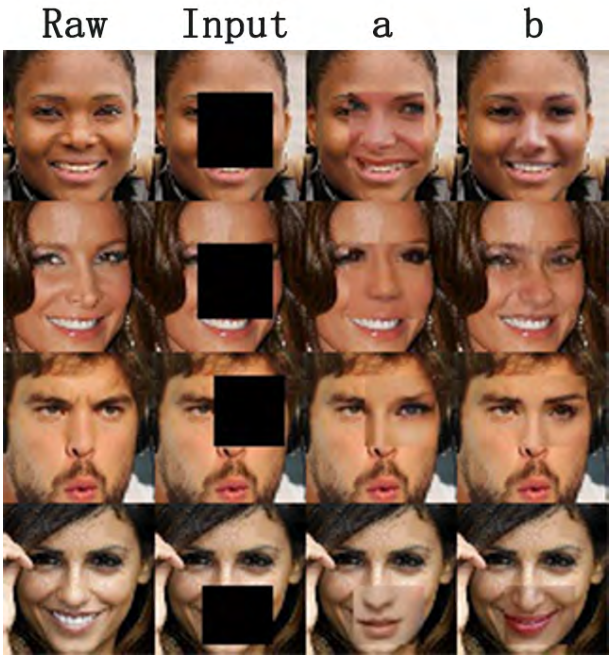

FIGURE 8. Comparisons of the edge connection with or without RW before blending. a: the results without RW, b:the results with RW.

C. INPAINTING RESULTS

In this section, through some contrast experiments, we qualitatively and quantitatively evaluate and analyze the results generated by the proposed method.

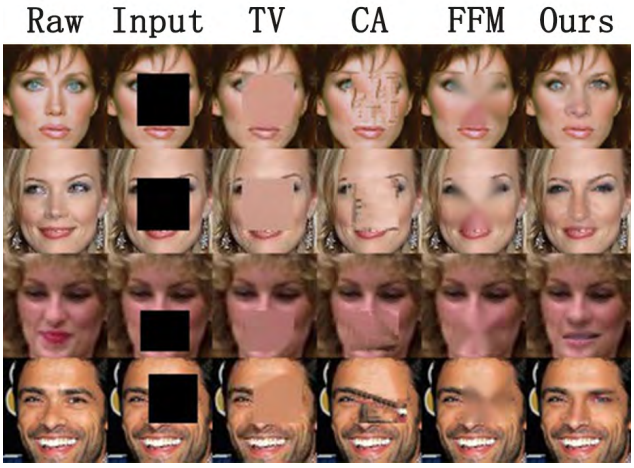

FIGURE 9. Comparisons with some methods, including TV, CA and FFM.

1) QUALITATIVE RESULTS

First, we compare our method with TV, Criminisi Algorithm (CA), and FFM. As shown in Fig.9, these methods are totally ineffective in semantic inpainting. Compared with these methods, our method can obtain correct and natural semantic information.

Second, we compare our method with the nearest neighbor (NN) based methods [10], [11]. Unlike NN, our method searches for the similar patches from the experimental dataset. In the first and third rows, our results are more continuous than that of NN in terms of edge connection. In the second and fourth rows in Fig.10, NN's results are prone to obvious semantic errors and our results are more prominent in semantic matching.

Third, in Fig.11, two methods, the original model and the method [26] (GL), are compared with our method. The results of the original model are excellent in terms of skin color and semantic information but they are blurry, which makes the overall vision much-poorly. Furthermore, GL's clarity is higher than the original model's but lower than ours while there are frequently errors in its skin color. In addition, as the gradient calculation of the GL's results at the boundary is limited in both foreground and background, the results is not ideal and even worse in visual effects after Poisson blending. Following the original paper, we used first FFM then Poisson blending, with the results compared and shown

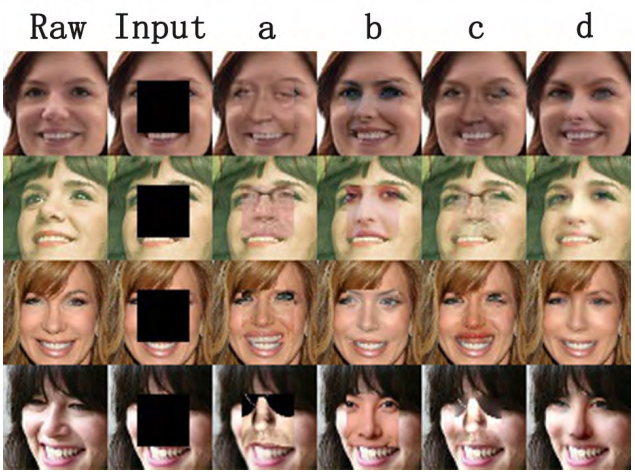

FIGURE 10. Comparisons with NN. a: the results of NN before blending, b: our results before blending, c: the results of NN after blending, d: our results after blending.

FIGURE 11. Comparisons with the original model(OM) and GL.

in Fig.16. Compared with the two methods, our method has more realistic results with more details.

Fourth, we compare our method with two methods [32] (CIICA) and [34] (PC). As shown in the Fig.12, the results of CIICA have higher clarity than that of the original model,

FIGURE 12. Comparisons with CIICA and PC.

as well as the skin color information is correct. However, its results are still slightly blurry and lack details, which makes the results have artifacts. In addition, PC's results are clear but tend to be distorted and unrealistic. PC is well applied to irregular masks, but its results are not ideal for the continuous masks of a large region. Compared with the two methods, our method has more realistic and natural results.

Fifth, we compare our method with the method [23] (DGM) based on GAN. This method is similar to our method in that both methods fill the missing region by finding the most similar patch. The difference is that our method involves searching for similar patches from real images in the experimental dataset while DGM finds similar patches from the fake images generated by DCGAN [31]. These fake images are prone to distortion and the image from the experimental dataset is more realistic. Thus, this determines that our method is superior to DGM in the quality of similar patches, which leads to our results being more realistic and natural than DGM. The results are shown in Fig.13.

Finally, as shown in Fig.14, the same images and six types of masks are used to compare on all deep learning methods. Overall, the results of CIICA and original model are blurry, which makes their results have artifacts. The performance

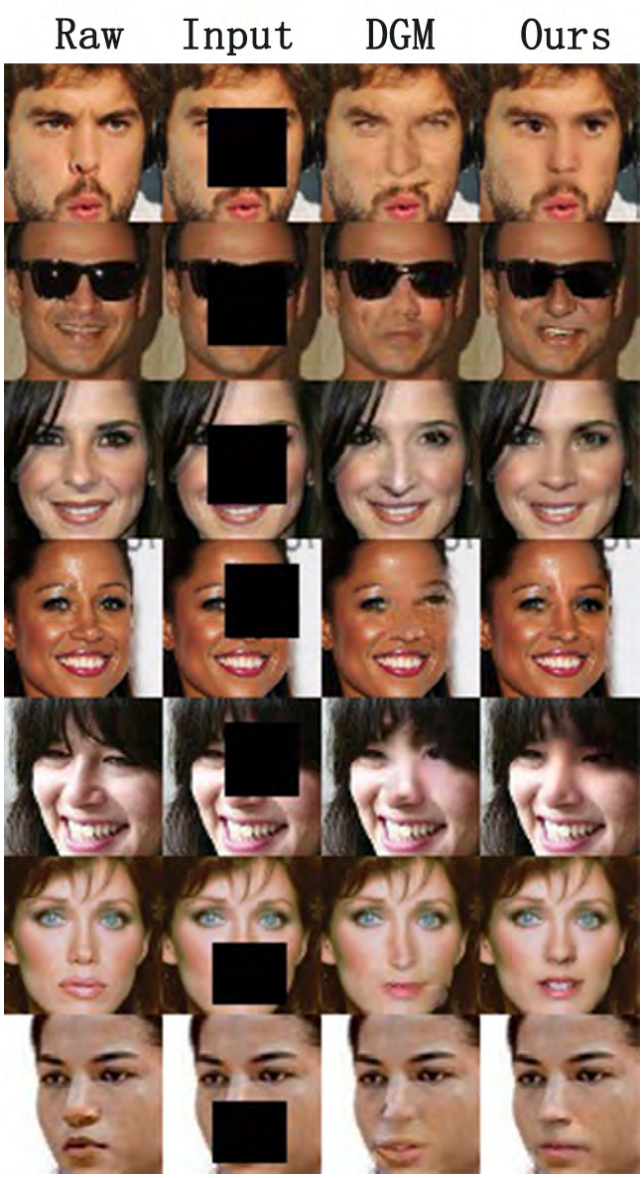

FIGURE 13. Comparisons with DGM.

of GL is drastically reduced in the case of a large mask, and color inconsistency is a defect as well. The PC's results are sharp and excellent when the mask is relatively small or irregular, but its results are not realistic with the continuous masks of a large region. The DGM's results are prone to distortion, but there are also a few expect results. In addition, compared with the original model's results, the blurriness for CIICA's results is reduced greatly, but CIICA's results still lack details. Compared with other five methods, our method has clear results with more correct semantic information and more details.

In addition, in order to observe the clarity of the results clearly, we locally enlarged the results for those methods whose results are not distort. As shown in Fig.15, compared with the results of these methods(original model, GL and CIICA), our results, similar to the raw image, contain more details, which make the missing region consistent with the surrounding regions.

By above comparing with some traditional methods (TV, CA, FFM, and NN) and deep learning methods (original model, GL, CIICA, PC and DGM), these results confirm that our method achieves better inpainting result.

TABLE 1. Comparisons of quantitative results. A/B results are by other method/our method. GL_1 and GL_2 are the results of GL before and after blending, respectively.

Method	PSNR	SSIM
Original Model	**28.67**/27.02	**0.91**/0.90
CIICA	**29.39**/27.02	**0.92**/0.90
GL_1	20.43/**27.02**	0.83/**0.90**
GL_2	24.05/**27.02**	0.85/**0.90**
DGM	24.16/**27.02**	0.84/**0.90**
PC	26.18/**27.02**	0.88/**0.90**

2) QUANTITATIVE RESULTS

As TV, LR, CA, FFM, and NN have obvious shortcomings in qualitative results, quantitative results are no longer given.

PSNR and SSIM are evaluation metrics based on the differences with the raw image. However, the intention of image inpainting is to fill the missing region and make the image more realistic and natural as a whole. As a raw image only is one of many possibilities, many works [23], [32] mentioned that PSNR and SSIM are imperfect in semantic inpainting in terms of reconstruction errors. However, similar to previous works, we compare our method with others on PSNR and SSIM values of the results.

Table 1 shows the comparison between the original model, CIICA, GL, DGM, PC and our method. Except for the original model and CIICA, the PSNR value of other methods are lower than ours. The results of the original model and CIICA are contrary to the previous qualitative results. We analyze this contrary phenomenon and found that their outputs are extremely consistent with the raw image in terms of overall skin color and semantic information while the blurring greatly affects the visual effect, which is what our method is trying to address. In addition, our results have more details which are obtained from the image in the dataset and may differ greatly from the raw image, which reduces the PSNR value. An example with the error images is shown in Fig.17. From the figure, our method has more realistic results than two other methods(CIICA and the original model) and the blurring results of these two methods have visible artifacts. However, PSNR values of CIICA and the original model are higher than ours. Judging from the error images, our higher error results are mainly from the follow reasons: 1) Our results have more details which differ from the raw image, for example, skin. 2) The eyebrows of the raw image are sparse and the eyebrows of the results are blurry in CIICA

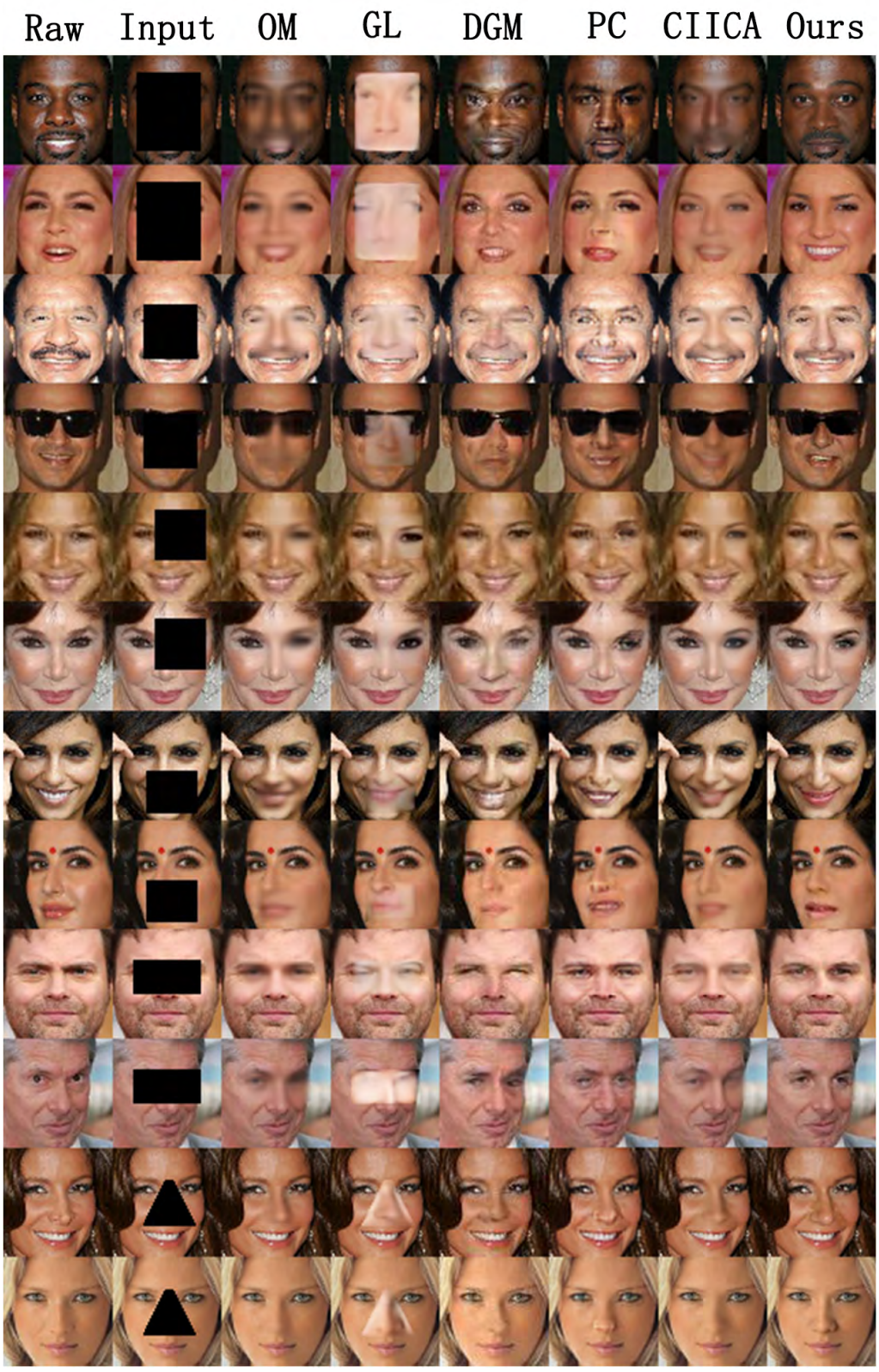

FIGURE 14. Comparisons with original model(OM), GL, DGM, PC, CIICA and our method.

and the original model, while our results have clear eyebrows. This indicates that quantitative results can not evaluate the performance of different methods well in semantic inpainting as mentioned in [23], [32].

We used Laplacian clarity to evaluate the clarity of the results quantitatively. Our method is compared with two other methods (the original model and CIICA) whose PSNR and SSIM are higher than ours. In addition, we also show the

TABLE 2. Comparisons of clarity.

Original Model	GL_1	GL_2	CIICA	Our Method
291	968	591	683	**1139**

FIGURE 15. Comparisons of local clarity.

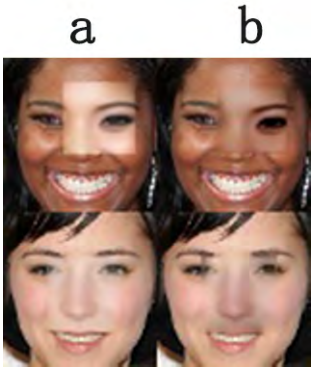

FIGURE 16. Comparisons of the results before and after blending for GL. a: the results before blending, b: the results after blending.

clarity of GL's results. As shown in Table 2, compared with the results of the original model, CIICA, and GL, our results have the highest clarity, which is consistent with the qualitative results.

3) HIGH-RESOLUTION RESULTS

We extended our method by maping low-resolution similar patches to high-resolution ones directly. In this way, our method can be directly applied to high-resolution facial image inpainting without retraining the high-resolution generator and the computational complexity is unchanged in searching for similar patches. First, our method needs a facial image dataset with a high resolution. After this, the images in the dataset and the input image with high-resolution are unified to 64*64*3. our method then is used to find the best similar patch and according to this best similar patch, the corresponding high-resolution patch is obtained. Finally, the missing region in the input image is filled by the best high-resolution similar patch. Fig.18 and Fig.19 show some examples with higher resolution.

D. LIMITATIONS

Although our method can obtain promising results, there are still some limitations. Our method relies on a large facial image dataset. In order to accommodate more posed faces, the sample size of facial image dataset needs to be expanded. However, a larger sample size means that it will take longer to build the candidate database. Thus, the time complexity of

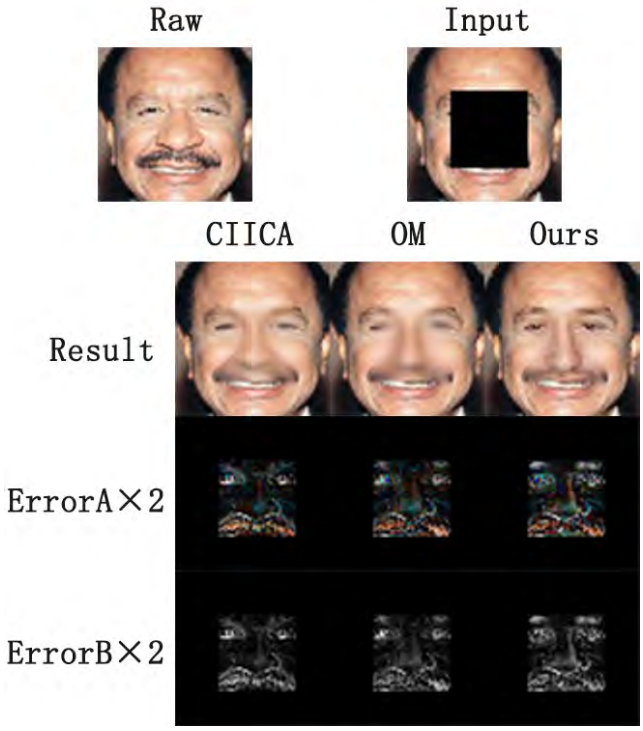

FIGURE 17. Comparisons of the error images between CIICA, original model(OM), and our method. ErrorA and ErrorB are the error images with RGB and Gray, respectively. the error images are magnified twice to facilitate display. The PSNR values for CIICA, OM, and our method are 24.76, 24.67, and 23.11 respectively. Our results have the lowest PSNR, and SSIM is consistent with PSNR. However, our results are the most realistic visually.

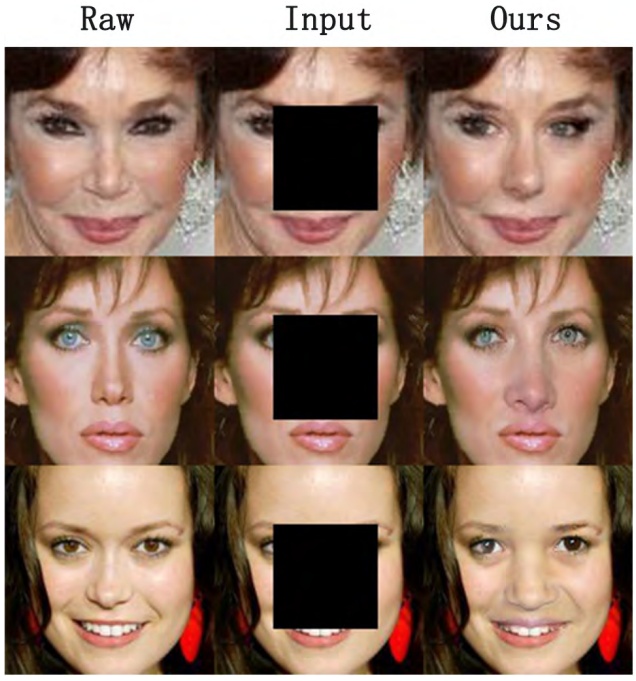

FIGURE 18. Inpainting results with 128*128 resolution.

our method is relatively large and implementing our method approximately takes about 97s for one image. As shown in table 3, compared with other deep learning methods, our time

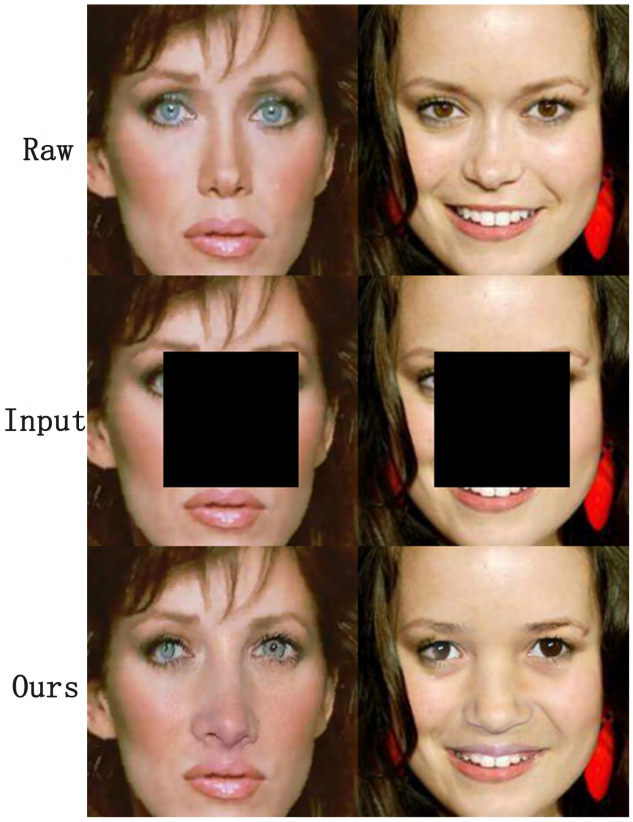

FIGURE 19. Inpainting results with 256*256 resolution.

TABLE 3. Comparisons of time complexity (Unit: sec.).

OM	GL	DGM	CIICA	PC	Our Method
2.90	4.41	57.89	3.38	6.91	96.67

complexity is not dominant. Moreover, our method easily ignores the symmetry of the face too.

Some failure samples are shown in Fig.20. In the first row, the image quality in the experimental dataset affects the inpainting result seriously. In the second row, the face is located to the right of the incomplete image with rare skin color and facial pose. There is no such type of facial image in the experimental dataset, which makes the result unsatisfactory. In the last row, the result does not match the symmetry of the face but meanwhile, the result has a special effect.

E. FUTURE WORK

Future studies should focus on solving some limitations of our method. First, for shortening the time of image inpainting, we plan to reduce the search range by improving and classifying the facial image dataset. Second, when searching for similar patches, we will add a symmetric penalty to solve the problem of asymmetric inpainting results. Finally, in our method, as the Pix2Pix model can be replaced by other models, we plan to replace it with a better deep generative model, such as [32].

FIGURE 20. Some failure results. a: our results before blending, b: our results after blending.

V. CONCLUSIONS

In this work, we propose a new inpainting method for facial images, which provides a new idea for image inpainting and other image processing tasks. This method combines the deep generative model with searching for similar patches. A deep generative model based on Pix2Pix is adopted and Laplace Loss is considered to accelerate convergence. When searching for similar patches, different weights are given to different regions around the boundary and the calculation of distance takes into account the edge distance. Experimental results show that our theory is validated and our method has remarkable performance in facial image inpainting.

REFERENCES

[1] J. Sulam and M. Elad, "Large inpainting of face images with trainlets," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1839–1843, Dec. 2016.
[2] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5892–5900.
[3] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.
[4] X. Liang, X. Ren, Z. Zhang, and Y. Ma, "Repairing sparse low-rank texture," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 482–495.
[5] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, Jul. 2009.
[6] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graph. Tools*, vol. 9, no. 1, pp. 23–34, 2004.
[7] H. J. Hsu, J. F. Wang, and S. C. Liao, "A hybrid algorithm with artifact detection mechanism for region filling after object removal from a digital photograph," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1611–1622, Jun. 2007.
[8] F. Tang, Y. Ying, J. Wang, and Q. S. Peng, "A novel texture synthesis based algorithm for object removal in photographs," in *Proc. Annu. Asian Comput. Sci. Conf.*, Chiang Mai, Thailand, Dec. 2004, pp. 248–258.
[9] A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 341–346.
[10] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. Graph.*, vol. 26, no. 3, p. 4, 2007.

[11] O. Whyte, J. Sivic, and A. Zisserman, "Get out of my picture! Internet-based inpainting," in *Proc. Brit. Mach. Vis. Conf*, 2009, pp. 1–11.
[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Jun. 2014, pp. 2672–2680.
[13] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. ICML*, New York, NY, USA, Jun. 2016, pp. 1558–1566.
[14] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. SIGGRAPH*, Jul. 2003, pp. 313–318.
[15] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *J. Vis. Commun. Image Represent.*, vol. 12, no. 4, pp. 436–449, Dec. 2001.
[16] W. Zuo and Z. Lin, "A generalized accelerated proximal gradient approach for total-variation-based image restoration," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2748–2759, Oct. 2011.
[17] X. Lin and X. Yang, "Effective exemplar-based image inpainting using low-rank matrix completion," in *Proc. IEEE 7th Int. Conf. Awareness Sci. Technol.*, Sep. 2015, pp. 37–42.
[18] A. Criminisi and P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[19] S. Zhang and X. Zhou, "An improved scheme for Criminisi's inpainting algorithm," in *Proc. 4th Int. Congr. Image Signal Process.*, Shanghai, China, Oct. 2011, pp. 2048–2051.
[20] A. Li, Y. Li, W. Niu, and T. Wang, "An improved criminisi algorithm-based image repair algorithm," in *Proc. 8th Int. Congr. Image Signal Process.*, Shenyang, China, Oct. 2015, pp. 263–267.
[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, Oct. 2015, pp. 234–241.
[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5967–5976.
[23] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6882–6890.
[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Banff, Canada, Dec. 2014, pp. 1–15.
[25] D. Pathak and P. Krähenbühl, J. Donahue, T. Darrell, A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2536–2544.
[26] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, Jul. 2017.
[27] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Sep. 2016, pp. 702–716.
[28] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stereoscopic neural style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6654–6663.
[29] C. Ledig and L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.
[30] Z. Liu, P. Luo, X. Wang, X. Tang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Santiago, Chile, Dec. 2015, pp. 3730–3738.
[31] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, San Juan, PR, USA, Nov. 2016, pp. 1–15.
[32] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 5505–5514.
[33] Q. Wang, H. Fan, G. Sun, Y. Cong, and Y. Tang, "Laplacian pyramid adversarial network for face completion," *Pattern Recognit.*, vol. 88, pp. 493–505, Apr. 2019.
[34] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 89–105.
[35] C. Yang, Y. Song, X. Liu, Q. Tang, and C.-C. J. Kuo, "Image inpainting using block-wise procedural training with annealed adversarial counterpart," Mar. 2018, *arXiv:1803.08943*. [Online]. Available: https://arxiv.org/abs/1803.08943

JINSHENG WEI received the B.S. degree in information engineering from Fuyang Normal University, Fuyang, China, in 2016. He is currently pursuing the Ph.D. degree with the Successive Master-Doctor Program from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China. His research interests include image processing, pattern recognition, computer vision, and machine learning.

GUANMING LU received the B.E. degree in radio engineering and the M.S. degree in communication and electronic systems from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 1985 and 1988, respectively, and the Ph.D. degree in communication and information systems from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is currently a Professor with the College of Communication and Information Engineering, NUPT. His current research interests include image processing, affective computing, and machine learning.

HUAMING LIU received the B.S. degree in computer science and technology from Shangqiu Normal University, Shangqiu, China, in 2005, and the M.S. degree in computer software and theory from Northwest Minzu University, Lanzhou, China. He is currently pursuing the Ph.D. degree with the Nanjing University of Posts and Telecommunications, Nanjing, China. He was a Visiting Scholar with the University of Science and Technology of China, Hefei, China, in 2015. He is currently an Associate Professor with the School of Computer and Information Engineering, Fuyang Normal University, Fuyang, China. His research interests include image processing, pattern recognition, software engineering, and so on.

JINGJIE YAN received the B.E. degree in electronic science and technology and the M.S. degree in signal and information processing from the China University of Mining and Technology, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in signal and information processing from Southeast University, Nanjing, China, in 2014. Since 2015, he has been with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, as a Lecturer. His current research interests include pattern recognition, affective computing, computer vision, and machine learning.