Neural Computing and Applications (2022) 34:9717–9731
https://doi.org/10.1007/s00521-022-06961-8

ORIGINAL ARTICLE

Reference-guided face inpainting with reference attention network

Jiazuo Yu[1,2] · Kai Li[1,2] · Jinjia Peng[1,2]

Received: 16 September 2021 / Accepted: 13 January 2022 / Published online: 3 February 2022

Abstract
Face inpainting is a significant problem encountered in many image restoration tasks, in which various methods based on deep learning are explored. Existing methods cannot restore enough structure details as the masked input only provides limited information. In this paper, a novel reference-guided face inpainting method is proposed to generate inpainting results more similar to people themselves, which restores the missing pixels by referring to a reference image besides an original masked image. Concretely, another reference image with the same identity as the masked input is utilized as a conditional input to constrain the generated coarse result of the first inpainting stage. Furthermore, a reference attention module is designed to restore more textural details by computing the similarity between the pixels of the coarse result and the reference image. The similarity is further represented by the similarity maps, which are deconvolved to reconstruct the pixels of the missing regions. Extensive experimental results on CelebA datasets and LFW datasets demonstrate that our proposed method can generate an image with more similar features to people themselves and achieves superior performance to the state-of-the-art methods quantitatively and qualitatively.

Keywords Face inpainting · Image inpainting · Reference image · Reference attention

1 Introduction

Image inpainting, an essential task in image processing, aims to fill the missing or masked regions in images with plausibly synthesized contents. It is a fundamental problem in low-level vision as it can be utilized for repairing damaged images or reconstructing the pixels of the missing regions. With the development of deep learning technology, digital image inpainting based on deep learning has aroused widespread interest in computer vision for the past years.

There are two broad approaches for image inpainting in computer vision: patch matching using low-level image features and feed-forward generative models with deep CNNs, respectively. The former is a traditional approach [1–5], in which the best matching texture patches are sampled from a source image and then pasted into a target image to reconstruct the missing area. For instance, Wilczkowiak et al. [5] proposed a technique that automatically adjusts, clones large image patches, and optimizes the search areas to find the most appropriate patches. PatchMatch [1] achieved real-time image editing by a randomized patch search algorithm. These methods employ the low-level features of the given context and propagate the local image appearance around the target holes to fill the holes. However, since low-level information is insufficient to infer semantical contents in missing regions, the above methods may produce artifacts, blur, and other problems when they complete images with complex textures. In recent years, the continuous development of CNNs [6, 7] has made new progress in the field of vision. The latter deep CNNs-based approach [8–14] overcomes the defect that only low-level features cannot predict high-level semantics in the holes. The methods with Generative Adversarial Networks(GAN) [15] and CNN formulate inpainting as an image generation problem. Furthermore, high-level recognition and low-level pixel synthesis are formulated into an Encoder-Decoder network, which is trained with adversarial networks to encourage the coherency between generated and existing pixels. Notably, some

✉ Kai Li
likai@hbu.cn

[1] School of Cyber Security and Computer, Hebei University, Baoding 071000, China

[2] Hebei Machine Vision Engineering Research Center, Baoding 071000, Hebei, China

works [12] demonstrate that dilated convolutions can be utilized to enlarge receptive without consuming extra computational cost to optimize image inpainting. To better complete images with irregular holes, Yu et al. [11] proposed the gated convolutions with dynamic feature gating mechanism based on partial convolutions introduced by Liu et al. [13]. In addition, Yu et al. [9] introduced a contextual attention layer to borrow the related feature patches from distant spatial locations explicitly. These works can generate new content from highly structured images. However, the input of the above methods is mainly only a single image, and the information for reference is not enough for the face inpainting tasks with complex textures. Therefore, it is a challenging task to restore high-quality images that are more similar to people themselves.

In recent years, a reference-guided image processing method incorporating a reference image to assist the image processing task has been proposed and becomes a new branch. It has been proven to be effective in various visual tasks, especially in Super-Resolution(SR) [16–21]. However, these methods cannot be directly applied to face inpainting. That is because face inpainting needs to learn the feature representation of the holes from existing pixels, and then generate the missing pixels of the holes through texture and structure matching. In this process, the original holes are empty. While in Super-Resolution, low-resolution images generally are obtained by convolution of clear images and blur kernels.

Therefore, it is important and interesting to restore a high-quality face image by incorporating a reference image into the network. In this paper, we propose a reference-guided deep face inpainting method for irregular holes, in which the masked facial image is restored by referring to people's own image. Our proposed method pays more attention to the reference image information in addition to the masked input information. Therefore, our network can get the output with more precise textures and more similar content. The main contributions of this paper are as follows:

This paper proposes an end-to-end reference-guided face inpainting method to generate inpainting results more similar to people themselves, converts the traditional generative adversarial repair model to a conditional generative model. The generated pixels refer to the reference image besides referring to the original masked image.

Another image with the same identity as the masked input is added to guide the inpainting. And a novel reference attention layer is designed to focus on related patches at the reference image to reconstruct a final output that is more similar to people themselves.

To the best of our knowledge, our method is the first attempt to reference-guided face inpainting. Evaluations on CelebA datasets and LFW datasets further validate the effectiveness of our proposed method, and achieve competitive qualitative results and superior quantitative results over the exsiting state-of-the-art methods.

2 Related work

In the Section, we briefly review the previous work of the traditional image inpainting, deep image inpainting, and guided image inpainting.

2.1 Traditional image inpainting

There are various approaches proposed for the image inpainting task. One of the early image inpainting approaches is diffusion-based image synthesis. This technique mainly used low-level features to fill holes by copying or borrowing surrounding textures [22–24]. These methods can only fill small holes, and noticeable artifacts and noise will appear when dealing with large missing areas or non-stationary image data. Patch-based approaches have been proposed to fill large holes in natural images, which can perform more complicated image completion than diffusion-based techniques. Simakov et al. [25] presented a bidirectional patch similarity-based approach to better model non-stationary visual data for inpainting, object removal, and more applications. However, the calculation of patch similarity is very resource-consuming, so this method is not widely used. Later works used patch matching [1] to iteratively search for the best fit patch, which can fill the holes and produce smooth results. It is seen that these methods can produce plausible texture generation in the holes. But the above methods are inadequate to fill in holes on complicated structures because of their dependence on low-level features such as the sum of squared differences of patch pixel values.

2.2 Deep image inpainting

Recently, deep neural networks including GANs and CNNs [26] exhibit excellent performance in image processing. Existing approaches mainly focus on distilling information from high-level features. Ding et al. [27] showed that by integrating low-level information, performance can be improved with enhanced feature representation and accurately located discriminative regions. Chang et al. [28] showed that it is possible to cultivate subtle details without the need for overly complicated network designs or training mechanisms - a single loss is all it takes. In addition, Deng et al. [29] proposed a novel High-Quality Generative Adversarial Network (HQ-GAN) for controllable editing of multiple face attributes in high-resolution images. It provides ideas for exploring high-resolution image restoration.

As for the use of deep learning methods for image restoration, Pathak et al. [8] introduced a deep CNNs-based Context Encoders(CE) network, which is trained to directly reconstruct the missing region by combining adversarial loss [15] and $\mathcal{L}_2$ loss. CE utilized the encoder to capture the compact latent feature representation and then employed the decoder to reconstruct the missing area. Lizuka et al. [12] improved CE by proposing global and local discriminators as adversarial losses. A local discriminator focuses on a small region centered around the inpainting regions, and a global discriminator is trained on the entire image, where Poisson blending is applied as a post-processing to combine global and local results. Yeh et al. [30] inferred the content of arbitrarily large missing areas in the image based on the semantic information of the image. Yu et.al. [9] introduced a method with a refinement network with the contextual attention layers to replace the post-processing mentioned above. Furthermore, Yu et al. [11] proposed gated convolution for inpainting irregular masked image, which optimized partial convolution [13] by using soft gating instead of the original hard gating, and achieved the state-of-the-art performance for irregular image inpainting. Que et al. [31] presented a unified framework for single image-based rain removal that handles various types of rainy images, which has great significance of reference for inpainting various holes areas. In order to reconstruct more details, Nazeri et al. [32] proposed a two-stage model EdgeConnect, which comprises of an edge generator followed by an image generative network. More recently, Zhang et al. [33] integrated global semantics and local features in a unified image generator with a solid ability to learn and leverage semantic priors. He et al. [34] applied face restoration to recognition to improve visual quality and optimize recognition results. These methods have shown impressive results for generating plausible visual details. However, the input of these methods is only a single image resulting in insufficient information for reference. The generated result may be distorted, and the similarity to themselves is not high when used for complex face inpainting.

2.3 Guided-based image inpainting

There have been many explorations in guided image processing [16–21, 35–40] to solve the problem of insufficient reference information for single-input. Specially, Hays et al. [37] firstly utilized millions of photographs as a database to search for an example image that is most similar to the input, and then completed the image by cutting and pasting the corresponding regions from the matched image.

Recently, image processing, image compositing, and restoration work with guidance have been continuously developed. Wang et al. [41] proposed synthesizing high-resolution photo-realistic images from semantic label maps using conditional generative adversarial networks. Zhang et al. [42] introduced colorization networks, which joined user guidance as additional input. Sangkloy et al. [43] proposed a deep generative network, which added sketched boundaries and sparse color as guidance to synthesize images. Inspired by the above work, Yu et al. [11] employed sketch(or edge) as user guidance to extend the image inpainting network. In addition, Li et al. [44] applied the reference-guided method to image deblurring, which correlated the high-quality reference image into the deep network for a better deblurring effect. More recently, Zhou et al. [45] also carried out work related to the reference-guided image inpainting, which combined a variety of colors and spatial transformations and achieved good results. Moreover, for the face-swapping task, Li et al. [40] proposed FaceInpainter, which implemented controllable Identity-Guided Face Inpainting under heterogeneous domains and showed excellent performance. However, these works cannot be directly used in the irregular masked face restoration work. To the best of our knowledge, the paper is the first face inpainting task, which is guided by another facial image with the same identity as the original input.

3 Reference-guided face inpainting network

In the following subsections, this paper describes our approach in detail. This paper firstly shows the architecture of our proposed method, and then introduces the details of the dilated gated convolution(conv), double attention layers, and the loss functions.

3.1 Previous inpainting method revisit

Based on convolutional and generative adversarial network, inpainting with a single masked input is aimed at repairing missing areas through large-scale data training. Formally, given the masked target I_t^M with the mask M, the network can restore the hole regions with specious content to generate the result I_o. The network can be expressed as the mapping: $G : I_t^M \rightarrow I_o$. Recently, inpainting network was improved to a two-stage network with two discriminators, and used a more complex network to obtain better inpainting results. The formula is expressed as:

$$G_1 : I_t^M \rightarrow I_c, \quad G_2 : I_c \rightarrow I_o, \tag{1}$$

where I_c represents the coarse inpainting results, G_1 and G_2 represent the generator of coarse and refinement inpainting stage, respectively. In addition, the hallucinated edges [32]

as a prior is introduced to the inpainting network, the formula expression of the inpainting model becomes:

$$G_1 : \left\{ I^M_{gray}, C^M_{gt}, M \right\} \rightarrow C_{pred}, \\ G_2 : \left\{ I^M_t, C_{comp} \right\} \rightarrow I_o, \tag{2}$$

where C_{pred} represents the predicted edge map, C_{comp} is calculated from C_{pred}. And I^M_{gray} and C^M_{gt} denote the masked grayscale counterpart and masked edge map, respectively. After that, semantic priors, landmark priors and other methods have been continuously proposed to improve the repair results of the network.

3.2 Overview

The architecture of our proposed method is illustrated in Fig. 1. The generator is mainly designed based on the structure of the Encoder-Decoder. Given a masked input $I^M_t \in \mathbb{R}^{W \times H \times 3}$ and a reference image $I_r \in \mathbb{R}^{W \times H \times 3}$ as the inputs of our network. I^M_t can be expressed as:

$$I^M_t = (1 - M) \bigodot I_t, \tag{3}$$

where $M \in \mathbb{R}^{W \times H \times 1}$ represents a single-channel mask, which indicates the hole regions with value one, and elsewhere with zero. 1 stands for a tensor of the same shape as the mask M, and $\bigodot$ denotes element-wise multiplication. In the stage of coarse inpainting, the inputs I^M_t and I_r are down-sampled twice by gated convolution for encoding. Then, the encoded results are extracted to high-level features by dilated gated convolution. Moreover, two up-sampling layers decode the high-level features to reconstruct the coarse result $I_c \in \mathbb{R}^{W \times H \times 3}$. The coarse inpainting stage can be expressed as the mapping: $G_1 : \left\{ I^M_t, I_r \right\} \rightarrow I_c$. In the stage of refinement inpainting, the coarse result I_c and the reference image I_r are taken as the inputs to obtain the similarity between the generated regions and effective regions. The generated regions represent the pixels generated in the missing areas by the coarse inpainting. The effective regions include the reference image and the remaining regions except for the masked areas. The similarity is further presented by the similarity maps, which guide to reconstruct high-level feature maps. Furthermore, the feature maps are deconvolved by twice up-samplings to better generate the final result $I_o \in \mathbb{R}^{W \times H \times 3}$. The refinement inpainting stage can be expressed as the mapping: $G_2 : \{I_c, I_r\} \rightarrow I_o$. Finally, the mapping of our entire generator is: $G : \left\{ I^M_t, I_r \right\} \rightarrow I_o$. As for the discriminator, it is based on a convolutional network, where the input consists of a generated final result I_o and a reference image I_r. As Fig. 1 shows, five convolutions with kernel size 5 and stride 2 are stacked to capture the feature statistics of Markovian patches. The output is a

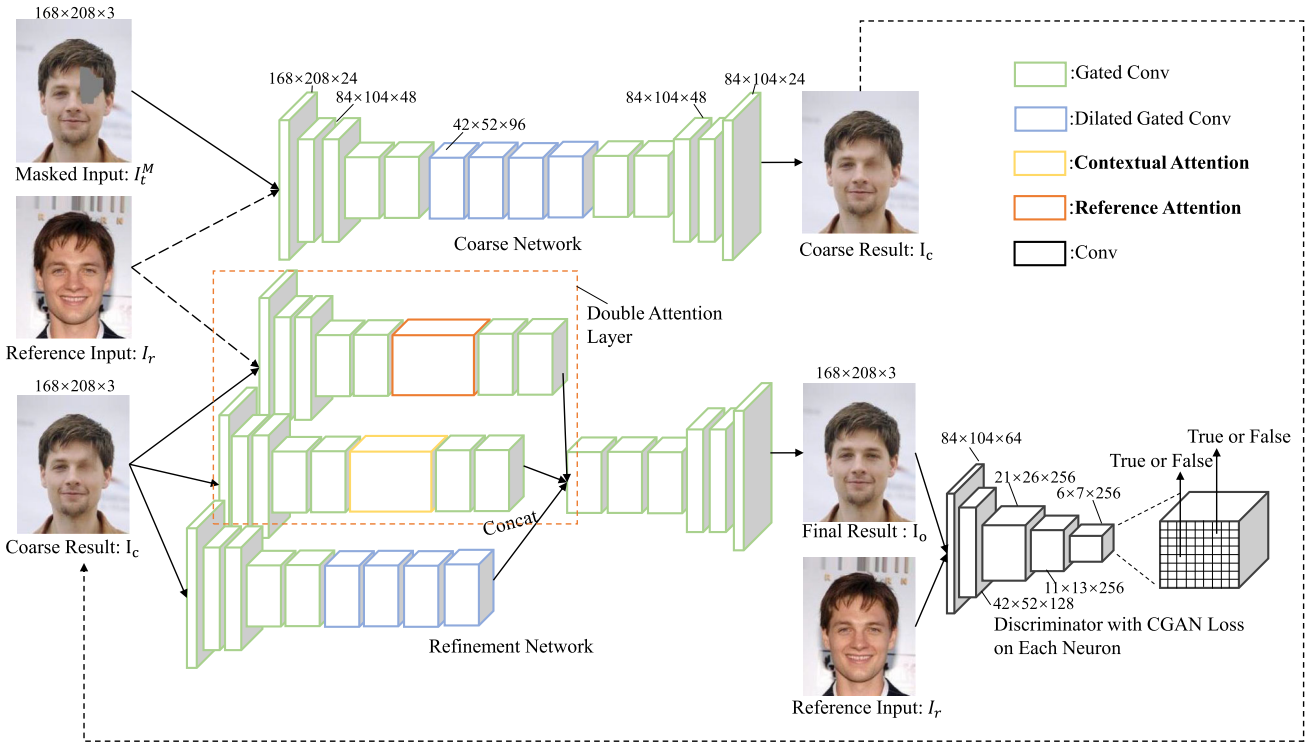

Fig. 1 Overview of our framework for reference-guided facial image inpainting. The reference input is utilized as a conditional input to constrain generated coarse results in the first stage and attended to provide more personal information by the reference attention layer in the second stage

three-dimensional feature of shape $\mathbb{R}^{w\times h\times c}$, where h, w, and c represent the height, width and number of channels, respectively. All feature elements are formulated to $w \times h \times c$ number of GANs focusing on different locations and different semantics of input image. In summary, our entire network consists of the above-mentioned two-stage generator and a discriminator, which are trained by our improved loss functions.

For training, given a target image X, we sample a mask M randomly. Input image I_t^M is corrupted from the target X and the mask M as $I_t^M = (1-M)\bigodot I_t$. Inpainting network G receives the image pair $(I_t^M,\ I_r)$, and then generates the repair result I_o. Training procedure is shown in Algorathm. 1, where i represents the number of loop, and $i = 1, 2, 3$ represent three times loop calculations. In each epoch of our training, the discriminator is updated three times with $(I_t,\ I_r)$, $(I_o,\ I_r)$ and Eq. (11) before the generator is updated.

Algorithm 1 The training procedure of the proposed reference-guided face inpinting network.

Require: The processed training dataset X of image pairs matched by identity label, each group of samples includes a target image I_t and a reference image I_r;
while the result of the generator has not converged **do**
  for $i = 1, 2, 3$ **do**
    Sample batch groups of image pairs I_t and I_r from dataset X;
    Generate random masks M(1 represents the masked regions) for I_t;
    Calculate I_t^M by $I_t \bigodot (1-M)$, and then get input pairs (I_t^M, I_r);
    Generate images $I_o = I_t^M + G\left(I_t^M, I_r, M\right)\bigodot M$;
    Update the discriminator with $(I_t,\ I_r)$ and $(I_o,\ I_r)$;
  end for
  Sample batch groups of image pairs I_t and I_r from dataset X;
  Generate random masks M for I_t;
  Get input pairs $\left(I_t^M, I_r\right)$;
  Generate images $I_o = I_t^M + G\left(I_t^M, I_r, M\right)\bigodot M$;
  Compute the training loss with Eq.9
  Update generator of the network with the training loss;
end while

3.3 Coarse inpainting

A conditional generative network based on Encoder-Decoder is designed to receive the masked input and the conditional reference input in the coarse inpainting stage. As shown in Fig. 1, five gated convolutions with different strides are stacked to highlight the masked regions and reference information in separate channels. Four dilated gated convolutions with stride 1 and kernel size 3 are stacked to capture the global features for each pixel of high-level. By this way, these features can be deconvoluted to generate better coarse inpainting results. The two convolutions are explained as follows.

The gated convolution It introduces a feature gate and a mask gate to learn a dynamic feature selection mechanism for each channel and spatial location. According to context, mask, and reference, the mechanism not only selects the suitable feature but highlights the masked regions and reference information at the deep level to generate better results. The formula is as follows:

$$\begin{aligned} G_{x,y} &= W_g^T \cdot X, \\ F_{x,y} &= W_f^T \cdot X, \\ O_{x,y} &= \phi\left(F_{x,y}\right)\bigodot \sigma\left(G_{x,y}\right), \end{aligned} \tag{4}$$

where W_g and W_f represent the gating convolution filter and feature convolution filter to the input feature X. The σ stands for sigmoid function. Thus, the output value $\sigma\left(G_{x,y}\right)$ is between zero and one. ϕ can be any activation function. Our network chooses ELU function, which combines sigmoid and ReLU functions. It has soft saturation on the left side, no saturation on the right side and makes the experiment converge faster. The ELU function is formulated as:

$$f(x) = \begin{cases} x, & x \geq 0, \\ \alpha(e^x - 1) & x < 0. \end{cases} \tag{5}$$

The dilated convolution It can expand the receptive fields on each layer without increasing the number of weights by spreading the convolution kernel. Specifically, the dilated convolution operator for each pixel can be represented as:

$$O_{x,y} = \sum_{i=-k_h'}^{k_h'} \sum_{j=-k_w'}^{k_w'} W_{k_w'+i,k_h'+j} \cdot I_{x+\eta i, y+\eta j}, \tag{6}$$

where $k_h' = \frac{(k_h-1)}{2}$, $k_2' = \frac{(k_w-1)}{2}$. k_h and k_w are the kernel width and kernel height, respectively. η stands for the dilation factor, the above equation represents standard convolution with $\eta = 1$. W is the convolution filter. $I_{x,y}$ and $O_{x,y}$ are the input and the output, respectively. For simplicity, the bias is ignored.

As shown in Fig. 2, the dilated convolution can obtain more extensive spatial support for each pixel, compared with the standard convolution that may not obtain enough spatial support of the influencing region (such as the spatial support Ω_2 of pixel p_2 cannot contain any effective information outside of the hole). To overcome the problem, our model applies the dilated convolution, which can effectively perceive larger areas and more pixels outside the hole when computing each output pixel. It can be seen in the area Ω_4 of Fig. 2.

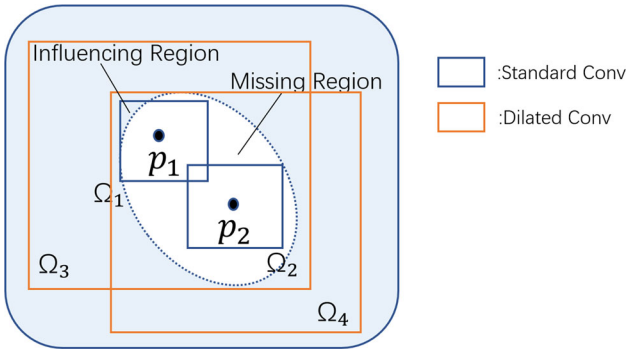

Fig. 2 Representation of spatial support. The dilated convolution can obtain more extensive spatial support than standard convolution. When using standard convolution, Ω_1 and Ω_2 represent the spatial support of the pixel p_1 and the pixel p_2. When using dilated convolution, Ω_3 and Ω_4 represent the spatial support of the pixel p_1 and the pixel p_2.

3.4 Refinement network with double attention

The dilated gated convolution layers process image features with local convolutional kernel layer by layer thus are not effective for restoring better texture details and more consistent semantic structures. To overcome the limitation, the attention mechanism is added to the refinement inpainting network to optimize the final result. The attention layer can learn where to borrow or copy feature information from known contextual and reference feature patches to generate missing patches. The attention mechanism is designed into the double attention layer as shown in Fig. 1 to pay attention to the contextual features and the reference features at the same time. In addition, in order to save more original input features information, a conventional dilated gated convolutional layer is added in parallel, thus forming a three-branch refinement repair encoder with the double attention layer. The entire refinement inpainting network composed of a three-branch Encoder and a Decoder learns the mapping from the coarse result I_c and the reference image I_r to the final output I_o, expressed as $G_2 : \{I_c, I_r\} \rightarrow I_o$. How they work will be introduced as follows.

3.4.1 Reference attention

Existing inpainting methods without a reference cannot obtain enough information to generate the semantic and structural details with the same identity as the input. In addition, these inpainting models usually lack generalization and diversity. Therefore, to refer to more information, the reference attention is designed in the stage of refinement inpainting. The reference attention layer can generate more diverse and more efficient image pixels via borrowing or copying from the reference image. In this part, the network matches the correlation between generated pixels from the holes and the reference pixels, and finds the patches that are most similar to the generated holes in the reference image. Before matching the correlation, the reference image I_r and the coarse result I_c are down-sampled into high-level feature maps. As shown in Fig. 1, two down samplings are performed by gated convolution with kernel size 3, stride 2, and ELU activation faction.

At the same time, the single-channel mask matrix is resized to the size of the high-level feature maps and sent to the attention layer. The feature map of the hole regions is extracted through calculation: $F_h = F_{in} \bigodot M$, where F_{in} and M represent the incomplete feature and the mask(the holes with value 1), and $\bigodot$ denotes element-wise multiplication. Then our model calculates the similarity between patches of holes and reference, and deconvolves to obtain the reconstructed holes feature map F_h^o, whose value of the area outside the hole is all set to 0. The remaining effective area after extracting the holes is $F_q = F_{in} \bigodot (1 - M)$, which is used to restore the final output feature map $F_r = F_h^o + F_q$. Due to the full use of the gated convolution of the single-channel mask, our model can respond to the mask in the middle layer, so that the extracted hole regions can obtain the expected result when performing the following similarity calculation and deconvolution.

The correlation between the reference and the hole region is calculated pixel-wisely in the reference attention layer. Specifically, as shown in Fig. 3, our network firstly extracts patches(33) from the reference feature map and reshapes them as convolution filters. To match the patches from hole regions $\{h_{patch}\}$ with reference patches $\{r_{patch'}\}$, our model normalizes the convolution filters and use them to perform convolution on the hole feature map. The

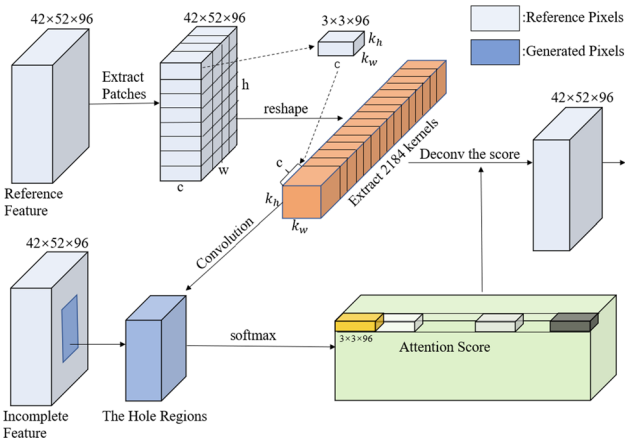

Fig. 3 Illustration of the reference attention layer. The convolution is utilized to compute the matching score of patches of hole regions with patches of reference feature(as convolution filters). Then the attention score can be obtained by softmax and be deconvolved for reconstruction

measurement formula with the normalized inner product is as follow:

$$S_{patch,patch'} = < \frac{h_{patch}}{\| h_{patch} \|}, \frac{r_{patch'}}{\| r_{patch'} \|} > , \quad (7)$$

where $S_{patch,patch'}$ represents similarity of patches between holes and the reference, $\frac{h_{patch}}{\|h_{patch}\|}$, $\frac{r_{patch'}}{\|r_{patch'}\|}$ represent normalized calculation and $<, >$ represents inner product calculation. Then our model weighs the similarity with scaled softmax to get an attention score for each pixel of hole regions. Furthermore, the reference patches extracted earlier are utilized as deconvolution filters, which deconvolves the attention score to reconstruct the feature map of the hole regions. The reconstructed feature map of the missing regions is merged with the effective regions remaining after the hole regions are extracted. Then the final feature map F_r with the same size as the output of the context attention layer is obtained. Finally, the final feature map is utilized to guide to generate the final inpainting result.

3.4.2 Contextual attention

The context attention layer is employed in conjunction with the reference attention layer to generate missing pixels by borrowing from its own surrounding pixels. In this part, the input feature map performs self-attention to obtain useful contextual information. Concretely, as shown in Fig. 4, our model takes the same method as above to extract patches and reshape them as convolutional filters. The difference is that the network extracts patches from the input feature map and matches the similarity between its own pixels. The network utilizes the normalized inner product to measure similarity:

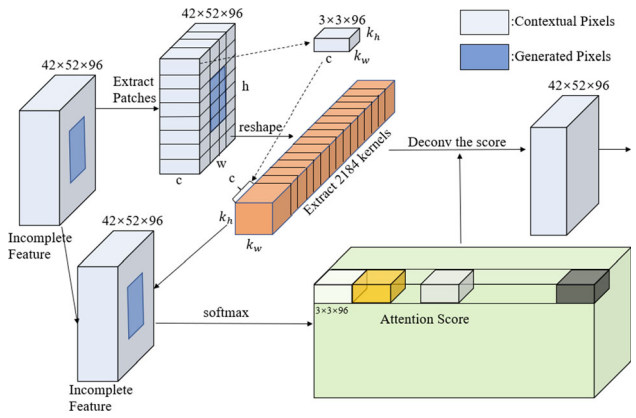

Fig. 4 Illustration of the contextual attention layer. The convolution is utilized to compute the matching score of patches with themselves. Then the attention score can be obtained by softmax and be deconvolved for reconstruction

$$S_{patch,patch'} = < \frac{f_{patch}}{\| f_{patch} \|}, \frac{f_{patch'}}{\| f_{patch'} \|} > , \quad (8)$$

where both f_{patch} and $f_{patch'}$ represent the patches from the same input feature map, $S_{patch,patch'}$ represents the similarity of patches between the input and itself. Then our model weighs the similarity with scaled softmax to get an attention score for each pixel. Notably, to avoid self-matching of patches in the masked region, the single-channel mask is applied to set those high scores obtained through self-matching of patches to zero after getting an attention score. Furthermore, the input patches extracted earlier are utilized as deconvolution filters, which deconvolves the attention score to reconstruct the second hole regions. In this way, the feature map F_c is obtained from the contextual attention layer. In addition, the feature map F_{dg} is obtained from the third dilated gated convolution layer. Finally, our network concatenates(concat) the feature maps F_r, F_c, and F_{dg} to reconstruct the final result with more details.

3.5 Improved loss functions

Our work designs an improved loss function by reproducing and making several improvements to the SN-PatchGAN loss, which consists of the pixel-wise reconstruction loss for reconstructing the missing region and conditional adversarial loss for generating more details with reference image, which can be expressed as:

$$\mathcal{L}_{all} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{adv}, \quad (9)$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{adv}$, respectively, represent the pixel-wise reconstruction loss and the conditional adversarial loss, λ_1, λ_2 stand for the weights of the reconstruction loss and the conditional adversarial loss, with default as 1.

Reconstruction Loss The network chooses $\mathcal{L}_1$ distance between the ground truth image and the restored image as our reconstruction loss function $\mathcal{L}_{rec}$:

$$\mathcal{L}_{rec} = \left|\left| I_{gt} - I_o \right|\right|_1, \quad (10)$$

where I_{gt} represents the ground truth image and I_o stands for the restored image by our network. We experimented with both $\mathcal{L}_1$ and $\mathcal{L}_2$ losses and found $\mathcal{L}_1$ loss performs better.

Adversarial Loss It aims to enhance the visual quality of the restored image through adversarial training. Our model adjusts the SN-PatchGAN loss for our conditional adversarial network training. The reference image is added as a conditional prior to the original adversarial loss. The generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_\mathcal{D}$ are as follows:

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathbb{P}_{I_t}}[D^{sn}(G(z|y))],$$
$$\mathcal{L}_D = \mathbb{E}_{x\sim\mathbb{P}_{I_{gt}}}[ReLU(\mathbf{1} - D^{sn}(x|y))] + \quad (11)$$
$$\mathbb{E}_{z\sim\mathbb{P}_{I_t}}[ReLU(\mathbf{1} + D^{sn}(G(z|y)))],$$

where $\mathbb{P}_{I_t}$, $\mathbb{P}_{I_{gt}}$ are the distributions of the original masked target input and the ground truth image, respectively. x, z represent real and fake data. D^{sn} represents spectral-normalized discriminator. y stands for reference input as the conditional prior, and $G(z|y)$ is the image inpainting network that takes the masked target input image I_{gt} and reference image I_t to generate the final output I_o.

4 Experiments

4.1 Datasets and evaluation metrics

Datasets Our method is trained and tested in CelebA dataset [46]. The dataset is divided into 68,754 groups of input data (a masked target image and a reference image) by the identity label. Our network randomly selects 57,996 groups for training and 10,758 groups for testing from the datasets, which totally contains 137,508 images. Our test set does not contain the training set. For each reference image, it matches each masked target input. The image pair and the inpainting result are shown in Fig. 5.

Evaluation Metrics To quantitatively evaluate our method on synthetic datasets against other inpainting methods, our work employs the Peak Signal-to-Noise Ratio(PSNR) and the Structural Similarity Measure(SSIM) as the evaluation metrics. PSNR can be mathematically defined as:

$$\mathrm{PSNR}(X, \hat{X}) = 20\log_{10}\left(\frac{2^n - 1}{\sqrt{MSE}}\right), \quad (12)$$

where X and $\hat{X}$ represent the original image and the complete image, respectively. n denotes the total number of pixels of $\hat{X}$, and MSE stands for the Mean Squared Error. SSIM can be formulated as:

$$\mathrm{SSIM}(X, \hat{X}) = \frac{(2\mu_X\mu_{\hat{X}} + C_1)(2\sigma_{X\hat{X}} + C_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + C_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + C_2)}, \quad (13)$$

where μ_X, $\mu_{\hat{X}}$ stand for the mean of X and $\hat{X}$. σ_X, $\sigma_{\hat{X}}$ represent the standard deviation of X and $\hat{X}$, respectively.

4.2 Quantitative evaluation

The PSNR and SSIM values of our generated image are quantitatively compared against the state-of-the-art inpainting methods on the CelebA dataset [46], as show in Table 1. The other methods including CA [9], EdgeConnect [32], LaFIn [39], SN-PatchGAN [11], and SPL [33].The data show that our proposed method performs favorably against other inpainting methods except SPL [33] in PSNR and SSIM. Specifically, EdgeConnect [32] and LaFIn [39], respectively, utilize an edge generator and a landmark predictor to guide the inpainting process, and they both perform well in inpainting tasks. However, the evaluation score of our method is high compared to EdgeConnect [32] and LaFIn [39] because our model can

Fig. 5 Our method's corresponding image groups, masked input, and inpainting results.

Table 1 Average PSNR and SSIM of different inpainting methods on CelebA dataset

Methods	PSNR	SSIM
CA [9]	29.0568	0.9325
LaFIn [39]	31.4414	0.9406
EdgeConnect [32]	30.3623	0.9289
SPL [33]	33.7557	0.9537
SN-PatchGAN [11]	32.1413	0.9435
Ours	33.2162	0.9505

obtain more reference information and attend more details by referring to another image. The PSNR and SSIM of CA [9] is obviously lower than our method. The restored effect of SN-PatchGAN [11] is close to our method, but due to the lack of personal information from single masked input, it is still slightly inferior to our method. In addition, our work tests different sizes of mask as shown in Fig. 6, ranging from 20 to 80% of the width of the original image, to evaluate the generalization ability of all the methods. Figures 7 and 8 show the different results of the PSNR and SSIM. The performance of these methods gradually drops, which is expected as the larger missing regions presents more uncertainties in pixel values. The line charts show that our method can also show better performance than other methods except SPL [33] when dealing with different sizes of occlusion.

As for SPL [33], the model adds semantic priors as a guide and exploits features of a multi-label classification model as the supervisions for leaning semantic priors. Therefore, the model achieves better results in the global evaluation of SSIM and PSNR. In comparison, our model has a slight disadvantage in terms of scores. This is mainly because our model, in order to recover more texture details and content features that are more similar to their own, pays attention to the reference information of another reference image while referring to the original incomplete image. Therefore, the reference image inevitably leads to our model showing a disadvantage in calculating the evaluation scores of PSNR and SSIM. Nevertheless, our model is still better than most other models. More importantly, compared with SPL, our method can visually recover missing information with more coherent content and more consistent semantics. As shown in Fig. 9, this result is encouraging. Due to the lack of processing of local information and insufficient reference information, the SPL causes problems such as blur, especially in the case of large occlusion.

4.3 Qualitative evaluation

The visual inpainting effect of our model is shown in Fig. 5. In the figure, it can be seen that the inpainting result of our model contains both context information and reference image information, and has the effect of consistent content and clear texture. Furthermore, when the angle of

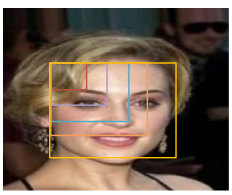

Fig. 6 The masked image with different proportions of the original image.

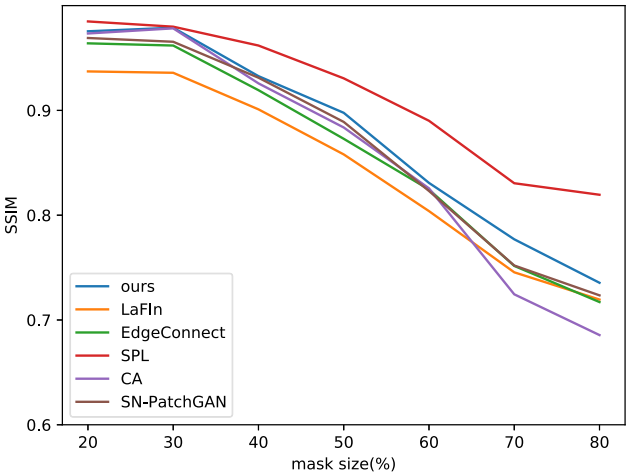

Fig. 7 Evaluations of SSIM on different square mask sizes of all the methods in the CelebA test dataset [46].

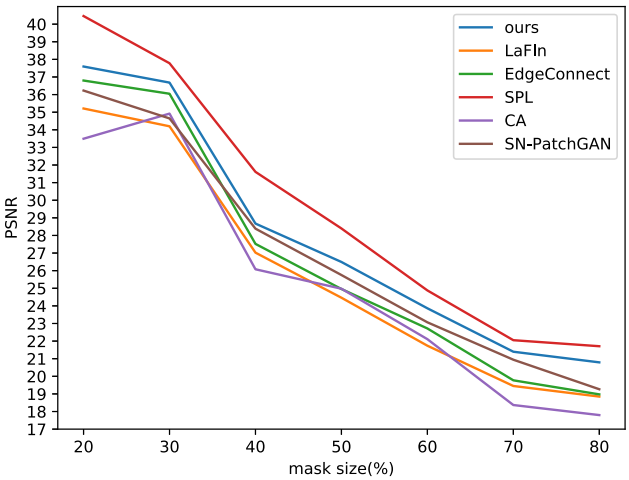

Fig. 8 Evaluations of PSNR on different square mask sizes of all the methods in the CelebA test dataset [46].

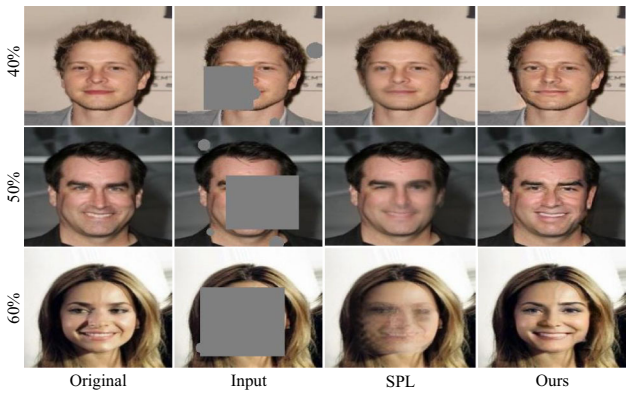

Fig. 9 The comparison between our method and the SPL [33] method in the case of different sizes of occlusion.

the reference image and the target image are different, or the reference image with glasses and the target image without glasses, or the reference image without glasses and the target image with glasses, our repair results still achieve the desired effect. It shows that our model has certain robustness in dealing with situations where the reference image and the target image are very different.

In addition, the visual quality of our complete images is further compared with the generated images of other reference inpainting methods on the CelebA dataset [46]. As shown in Fig. 10, the first three rows are, respectively, the reference image, the original image, and the masked input image, each of the last four rows represents the result of each method. In addition, each column represents the result of a different method on the same example. From the sixth row in Fig. 10, it can be seen that there are apparent artifacts in the inpainting results of the SPL [33]. Especially in the case of large occlusion, the restored results are obviously distorted. In contrast, the result of our method is visually more coherent and semantically more consistent. From the fourth and fifth rows in Fig. 10, the results of EdgeConnect [32] and LaFIn [39] are relatively good, but

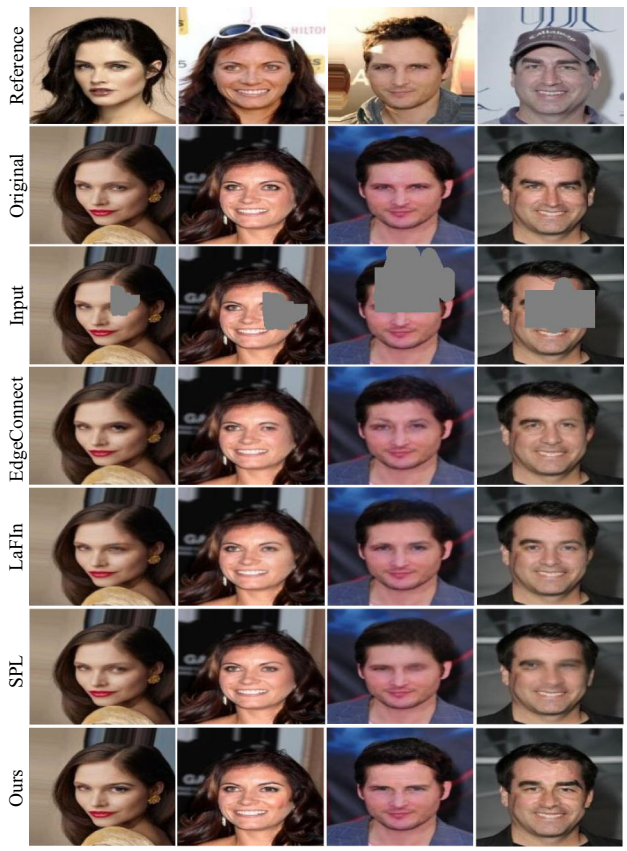

Fig. 10 Four examples of qualitative evaluation on CelebA dataset [46]. SPL [33] is based on the semantic priors to restore the image. EdgeConnect [32] and LaFIn [39], respectively, take an edge generator and a landmark predictor as a guidence

they have low similarity with people's own images or even no longer similar to themselves. Our methods with reference images can generate images sharper in vision and more consistent texture than other methods, which can be clearly seen in the last column of Fig. 10. More importantly, the image generated is more similar to people's own images in vision by paying more attention to the reference image in the stage of refinement inpainting. For example, in column 1, the eye corners of the image repaired by our method are more similar to her own reference image. In column 4, the structural features of the eyes and eyebrows are clearly derived from his own reference image and so on. The repair results of other methods distort their own personal characteristics. Both quantitative and qualitative evaluations demonstrate the superiority of our proposed method on the CelebA [46] dataset.

5 Ablation study

The above part has quantitatively and qualitatively shown that our proposed method performs favorably against other inpainting methods. In this section, the paper further discusses our proposed reference-guided inpainting by ablation study.

Quantitative evaluation In this part, our work conducts experiments and quantitative evaluations on the two datasets of CelebA [46] and LFW [47]. As shown in Tables 2 and 3, the part compares the PSNR and SSIM values of the image generated by our methods, including the first stage of coarse inpainting without a reference image, the first stage of coarse inpainting with a reference image, the final results without a reference image, and the final results with reference image. The group of A,B and the group of C,D from Table 2 and Table 3 illustrate that our inpainting method with reference image is better than the method without reference image in both the coarse result and the final results. It proves that our introduced reference image is effective for improving the inpainting effect. At the same time, when the reference image is missing, our repair results are also higher than some other typical methods, which shows that our model has a certain degree of robustness to deal with the case of reference image missing. This is because our model retains the dilated gated convolutional layer in the three branches and the role played by our contextual attention layer. In addition, the group of A,C and the group of B,D from the Table show that the indicators of PSNR and SSIM have all improved. It proves that our designed second stage of refinement inpainting and reference attention are essential to generate better results. The experimental results in the celeba dataset [46] show that when there is no reference image as conditional input, the result has not improved very

Table 2 Average PSNR and SSIM of our method in different situations on CelebA [46] dataset

Methods	PSNR	SSIM
(A) The First Stage of Coarse Inpainting without Reference	29.4183	0.9188
(B) The First Stage of Coarse Inpainting with Reference	30.6402	0.9315
(C) Final Result without Reference	31.7201	0.9438
(D) Final Result with Reference	33.2162	0.9505

Stage 1 result represents the coarse inpainting result by our method

Table 3 Average PSNR and SSIM of our method in different situations on LFW dataset [47]

Methods	PSNR	SSIM
(A) The First Stage of Coarse Inpainting without Reference	26.8153	0.8907
(B) The First Stage of Coarse Inpainting with Reference	27.8464	0.9187
(C) Final Result without Reference	27.3297	0.9201
(D) Final Result with Reference	28.8656	0.9283

Stage 1 result represents the coarse inpainting result by our method

significantly even if the second stage of refinement inpainting stage is added. In contrast, the final repair effect with the reference guide is significantly improved compared to the first stage repair effect. It demonstrates that our designed reference attention layer needs the reference image, and can utilize the image to generate better textural and structural content. Due to the small number of images in the LFW dataset [47], the convergence is not high, so the above phenomenon is not obvious enough, but the data results can also show the superiority of the repair effect after adding the reference image. The quantitative experiment illustrates the effectiveness of our proposed reference-guide method in the two stage inpainting works.

In addition, the paper conducts an ablation study on the three-branch network layer of the refinement inpainting stage. The quantitative experimental results are shown in Table 4. The results of Table 4 show that when all attention layers are removed, the results of PSNR and SSIM decrease significantly. It proves the significance of the attention layer introduced by us for inpainting. The results B,D and C,D of Table 4 show that when the attention layer is added to our model, the PSNR and SSIM will be improved. And when two attention layers exist at the same time, the best performance of the model can be achieved. That is because the network of the context attention layer and the reference attention layer can respectively match more similar patch blocks from the original input and the reference input to guide the network learning. Finally, when we remove the dilated gated convolutional layer, our model will be unable to retain more effective information, resulting in a decrease in model performance.

Qualitative evaluation As shown in Fig. 11, this part firstly shows the comparison of the final results of two different situations. The one case is the images restored with the reference image, which can be seen in the fourth row. And another case is the images restored without the reference image, which can be seen in the last row. The first column shows that the eyes of the generated image by inpainting method without a reference is bigger than by inpainting method with a reference. And it can be seen that the generated image with reference is more similar to the original image and the reference image. This is because the reference attention layer of the inpainting stage refers to more similar textural details from the reference image. In addition, the second column, the third column, and the last column of Fig. 11 illustrate that the hairs, the eyebrows, and the nose of the generated image by inpainting method with a reference are more similar to the original input and

Table 4 Average PSNR and SSIM of our ablation study of three-batch Encoders on CelebA [46] dataset

Methods	PSNR	SSIM
(A) Our Method without Conventional Convolution Branch	32.1620	0.9361
(B) Our Method without Contextual Attention Branch	32.8396	0.9428
(C) Our Method without Reference Attention Branch	32.7761	0.9411
(D) Our Method without Any Attention	30.3011	0.9335
(E) Our Method	33.2162	0.9505

Fig. 11 Qualitative evaluation of our method with and without reference-guided image

the reference image. As shown in the green box and red box of Fig. 11, the similarity between generated parts and original parts is reduced when the reference-guided image is removed. It also proves that the reference image can provide more information to the model.

Furthermore, the qualitative experimental results of the three-branch network layer are shown in Fig. 12. When the dilated gated convolutional layer is removed, the repair result is obviously blurred as shown in Fig. 12a. It proves

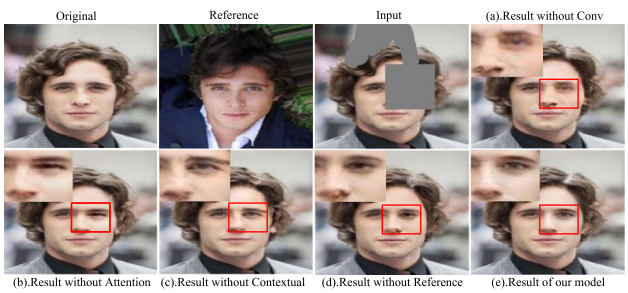

Fig. 12 The ablation study on CelebA dataset of each layer of the three-branch Encoders, where **b** represents the results after removing all the attention layers. **a**, **c** and **d**, respectively, represent the results after removing each layer in the three-branch network.

that the convolution layer has the effect of perceiving the original input features. When there is no attention layer, our model restores a distorted result as shown in Fig. 12b. When the context attention layer or reference attention layer is introduced, the model can overcome some distortion and blurring by referring to useful information as shown in Fig. 12c and d. But these repair effects are all not as good as our three-branch repair network with double attention layer. That is because our method is aware of contextual images and reference images structures and can adaptively borrow information from surrounding areas and reference areas to help the synthesis and generation.

Interestingly, additional discovery is shown in Fig. 13. When a false image(not belong to the identity of the original input and different in gender) is the reference input, the model will learn the structure and texture of this false image in the refinement inpainting stage. Although the repair results in the first stage are not much different, it causes gender distortion in the final repair result. Nevertheless, the consistency of texture content and visually reasonable results can still be guaranteed, which shows that our model is robust to processing image pairs of different labels. According to previous work experience and our

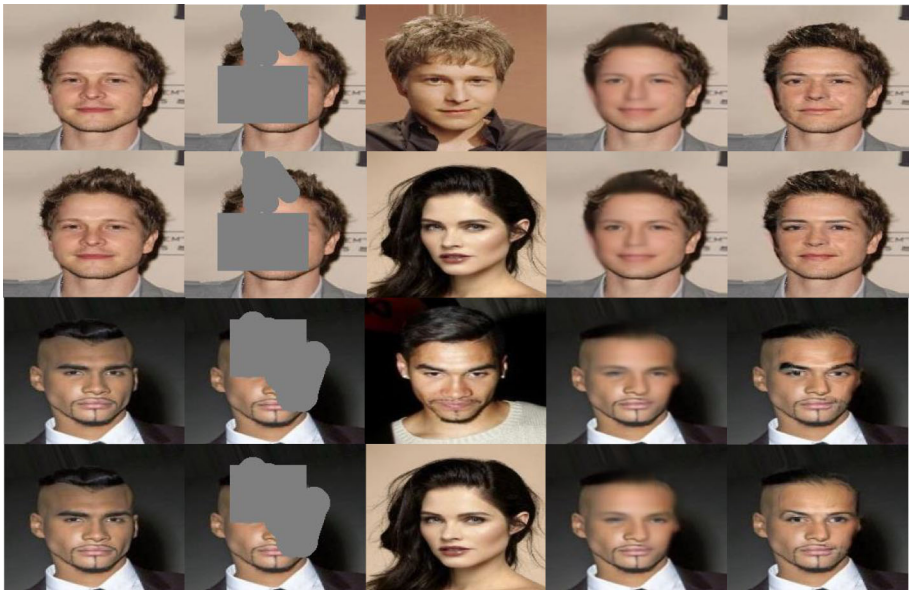

Fig. 13 Qualitative evaluation of our method with a ture reference (belong to the identity of the original input) and a false reference(not belong to the identity of the original input)

analysis, this is closely related to the structural characteristics of the GAN network. This phenomenon fully illustrates the guiding roles of our reference-guided mechanism. And our reference attention can learn the guidance of reference images to reconstruct the missing regions. The above quantitative and qualitative ablation evaluations demonstrate the superiority and effectiveness of our proposed method for face inpainting.

6 Conclusion

In this paper, we propose a reference-guided inpainting method for irregularly masked face images. By utilizing the CGAN framework to take the reference image as a conditional input and introducing a reference attention module to attend on the reference, our method generated a final result with better visual effects and more similar to people's own photographs. Both quantitative and qualitative evaluations on the CelebA dataset and LFW dataset demonstrate the effectiveness of the proposed method. However, only two valuable celebA dataset and LFW dataset with the identity label are found, which can be divided into tens of thousands or thousands of image pairs by identity. In future work, we will continue to explore how to select the reference image with the most similar label from a large number of images and explore high-resolution reference-guided facial image inpainting.

Acknowledgements This work was supported by Hebei University High-level Scientific Research Foundation for the introduction of talent (No.521100221029).

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Trans Graph 28(3):24
2. Hays J, Efros AA (2007) Scene completion using millions of photographs. ACM Trans Graph (ToG) 26(3):4-es. https://doi.org/10.1145/1400181.1400202
3. Efros AA, Freeman WT (2001) Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp 341-346 https://doi.org/10.1145/383259.383296
4. Efros AA, Leung TK (1999) Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision, IEEE vol 2, pp. 1033–1038. https://doi.org/10.1109/iccv.1999.790383
5. Wilczkowiak M, Brostow GJ, Tordoff B, Cipolla R (2005) Hole filling through photomontage. In: BMVC 2005-Proceedings of the British Machine Vision Conference 2005 https://doi.org/10.5244/c.19.52
6. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105. https://doi.org/10.1145/3065386
7. Wang H, Peng J, Jiang G, Fu X (2021) Learning multiple semantic knowledge for cross-domain unsupervised vehicle re-identification. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1-6 https://doi.org/10.1109/icme51207.2021.9428440
8. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544 https://doi.org/10.1109/cvpr.2016.278

9. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5505–5514 https://doi.org/10.1109/cvpr.2018.00577
10. Yang C, Lu X, Lin Z, Shechtman E, Wang O, Li H (2017) High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6721–6729 https://doi.org/10.1109/cvpr.2017.434
11. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2019) Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4471–4480 https://doi.org/10.1109/iccv.2019.00457
12. Iizuka S, Simo-Serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graph (ToG) 36(4):1–14. https://doi.org/10.1145/3072959.3073659
13. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B (2018) Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 85–100 https://doi.org/10.1007/978-3-030-01252-6_6
14. Li Y, Liu S, Yang J, Yang MH (2017) Generative face completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3911–3919 https://doi.org/10.1109/cvpr.2017.624
15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, WardeFarley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27
16. Shim G, Park J, Kweon IS (2020) Robust reference-based super-resolution with similarity-aware deformable convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8425–8434 https://doi.org/10.1109/cvpr42600.2020.00845
17. Yang F, Yang H, Fu J, Lu H, Guo B (2020) Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5791–5800 https://doi.org/10.1109/cvpr42600.2020.00583
18. Zhang Z, Wang Z, Lin Z, Qi H (2019) Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7982–7991 https://doi.org/10.1109/cvpr.2019.00817
19. Zheng H, Guo M, Wang H, Liu Y, Fang L (2017) Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system. In: Proceedings of the IEEE international conference on computer vision workshops, pp 2481–2486 https://doi.org/10.1109/iccvw.2017.292
20. Zheng H, Ji M, Wang H, Liu Y, Fang L (2018) Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European conference on computer vision (ECCV), pp 88–104 https://doi.org/10.1007/978-3-030-01231-1_6
21. Zheng H, Ji M, Han L, Xu Z, Wang H, Liu Y, Fang L (2017) Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In:BMVC, vol 1, p 2 https://doi.org/10.5244/c.31.138
22. Ballester C, Bertalmio M, Caselles V, Sapiro G, Verdera J (2001) Filling-in by joint interpolation of vector fields and gray levels. IEEE Trans Image Process 10(8):1200–1211. https://doi.org/10.1109/83.935036
23. Bertalmio M, Sapiro G, Caselles V, Ballester C (2000) Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pp 417–424
24. Levin A, Zomet A, Weiss Y (2003) Learning how to inpaint from global image statistics. ICCV 1:305–312. https://doi.org/10.1109/iccv.2003.1238360
25. Simakov D, Caspi Y, Shechtman E, Irani M (2008) Summarizing visual data using bidirectional similarity. In: 2008 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8 https://doi.org/10.1109/cvpr.2008.4587842
26. Wang H, Peng J, Zhao Y, Fu X (2020) Multi-path deep CNNs for fine-grained car recognition. IEEE Trans Veh Technol 69(10):10484–10493. https://doi.org/10.1109/tvt.2020.3009162
27. Ding Y, Ma Z, Wen S, Xie J, Chang D, Si Z, Liang WuMH (2021) AP-CNN: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Trans Image Process 30:2826–2836. https://doi.org/10.1109/tip.2021.3055617
28. Chang D, Ding Y, Xie J, Bhunia AK, Li X, Ma Z, Wu M, Guo J, Song YZ (2020) The devil is in the channels: mutual-channel loss for fine-grained image classification. IEEE Trans Image Process 29:4683–4695. https://doi.org/10.1109/tip.2020.2973812
29. Deng Q, Li Q, Cao J, Liu Y, Sun Z (2020) Controllable multi-attribute editing of high-resolution face images. IEEE Trans Inf Forensics Secur 16:1410–1423. https://doi.org/10.1109/tifs.2020.3033184
30. Yeh RA, Chen C, Yian Lim T, Schwing AG, Hasegawa-Johnson M, Do MN (2017) Semantic image inpainting with deep generative models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5485–5493 https://doi.org/10.1109/cvpr.2017.728
31. Que Y, Li S, Lee HJ (2020) Attentive composite residual network for robust rain removal from single images. IEEE Trans Multimedia. https://doi.org/10.1109/tmm.2020.3019680
32. Nazeri, Ng, Joseph, Qureshi, and Ebrahimi. Nazeri K, Ng E, Joseph T, Qureshi FZ, Ebrahimi M (2019) Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv:190100212
33. Zhang W, Zhu J, Tai Y, Wang Y, Chu W, Ni B, Wang C, Yang X (2021) Context-aware image inpainting with learned semantic priors. arXiv:210607220
34. He R, Cao J, Song L, Sun Z, Tan T (2019) Adversarial cross-spectral face completion for NIR-VIS face recognition. IEEE Trans Pattern Anal Mach Intell 42(5):1025–1037. https://doi.org/10.1109/tpami.2019.2961900
35. Zhao Y, Price B, Cohen S, Gurari D (2019) Guided image inpainting: Replacing an image region by pulling content from another image. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1514–1523 https://doi.org/10.1109/wacv.2019.00166
36. Whyte O, Sivic J, Zisserman A (2009) Get out of my picture! internet-based inpainting. In: BMVC, vol 2, p 5 https://doi.org/10.5244/c.23.116
37. Hays J, Efros AA (2007) Scene completion using millions of photographs. ACM Trans Graph (ToG) 26(3):4-es. https://doi.org/10.1145/1276377.1276382
38. Huang JB, Kang SB, Ahuja N, Kopf J (2014) Image completion using planar structure guidance. ACM Trans Graph (ToG) 33(4):1–10. https://doi.org/10.1145/2601097.2601205
39. Yang Y, Guo X (2020) Generative landmark guided face inpainting. In: Chinese conference on pattern recognition and computer vision (PRCV), Springer, pp 14-26 https://doi.org/10.1007/978-3-030-60633-6_2
40. Li J, Li Z, Cao J, Song X, He R (2021) Faceinpainter: high fidelity face adaptation to heterogeneous domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5089–5098
41. Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., Catanzaro, B. (2018) High-resolution image synthesis and semantic

manipulation with conditional GANs. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp 8798–8807). https://doi.org/10.1109/cvpr.2018.00917
42. Zhang R, Zhu JY, Isola P, Geng X, Lin AS, Yu T, Efros AA (2017) Real-time user-guided image colorization with learned deep priors. arXiv:170502999
43. Sangkloy P, Lu J, Fang C, Yu F, Hays J (2017) Scribbler: Controlling deep image synthesis with sketch and color. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5400–5409 https://doi.org/10.1109/cvpr.2017.723
44. Li Y, Luo Y, Lu J (2021) Reference-guided deep deblurring via a selective attention network. Appl Intell. https://doi.org/10.1007/s10489-021-02585-y
45. Zhou Y, Barnes C, Shechtman E, Amirghodsi S (2021) Transfill: reference-guided image inpainting by merging multiple color and spatial transformations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2266–2276
46. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3730–3738 https://doi.org/10.1109/iccv.2015.425
47. Huang GB, Mattar M, Berg T, Learned-Miller E (2008, October) Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition