International Journal of Machine Learning and Cybernetics (2023) 14:2945–2961
https://doi.org/10.1007/s13042-023-01811-y

ORIGINAL ARTICLE

RNON: image inpainting via repair network and optimization network

Yuantao Chen[1] · Runlong Xia[2,3] · Ke Zou[4] · Kai Yang[5]

Received: 30 August 2022 / Accepted: 23 February 2023 / Published online: 25 March 2023

Abstract

In the last few years, image inpainting methods based on deep learning models had shown obvious advantages compared with existing traditional methods. The former can better generate visually reasonable image structure and texture information. However, the existing premier convolutional neural networks methods usually causes the problems of excessive color difference and image texture loss and distortion phenomenon. The paper has proposed an effective image inpainting method using generative adversarial networks, which is composed of two mutually independent generative confrontation networks. Among them, the image repair network module aims to solve the problem of repairing the irregular missing areas of the image, and its generator is based on a partial convolutional network. The image optimization network module aims to solve the problem of local chromatic aberration in the repaired images, and its generator has based on deep residual networks. Through the synergy of the two network modules, the visual effect and image quality of the images has improved. The experimental results can show that the proposed method (RNON) performs better from comparisons of qualitative and quantitative evaluations with state-of-the-arts in image inpainting quality field.

Keywords Generative adversarial network · Image inpainting · Optimization network · Partial convolution · Residual network

1 Introduction

Image inpainting tasks, that is, filling the lost pixels area of the processing image, plays an important role in the research field of computer vision. It has been applied in many applied fields, such as image editing [1] and image rendering [2]. The core mission of image inpainting lied in how to fill in the missing areas to generate images with a visual effect closely to the real and reasonable semantic information. Image inpainting was originally a traditional problem of computer graphics, mainly based on mathematical and physical methods. For the defective part of the image, because the original appearance is often unclear, image inpainting has usually analyzed and executed according to human cognitive rules. This is essentially the problem with a serious morbid nature. The prior information it relied on needs to be stable. Once the prior information was slightly deviated and changed, it will lead to significant changes in the final solution of inpainting.

In recent years, as deep learning methods had achieved excellent results in the field of computer vision, more and more researchers had begun to turn their attention to deep learning. Image inpainting based on deep learning has become a research hotspot problem. The basic principle of early image inpainting is similar to the synthesis of textures information, which fills in missing images by matching and duplicating background patches. Barnes et al. [3] had used an iterative method to search for the most suitable patch to fill the missing area, which performed well in the task of

✉ Yuantao Chen
chenyt@hnuit.edu.cn

Runlong Xia
xiarunlong@vip.qq.com

Ke Zou
zouk@hnwjgk.cn

Kai Yang
yangkai@zoomlion.com

[1] School of Computer Science and Engineering, Hunan University of Information Technology, Changsha, Hunan, China

[2] Mountain Yuelu Breeding Innovation Center Limited, Changsha, China

[3] Is Also With Hunan Provincial Science and Technology Affairs Center, Changsha, China

[4] Hunan WUJO High-Tech Material Corporation Limited, Loudi, China

[5] Hunan ZOOMLION Intelligent Technology Corporation Limited, Changsha, China

background restoration, but because the restored background area was mostly smooth, it was difficult to capture high-level semantic information. Therefore, the proposed method can not repair image structures that had never appeared in the image or are more complex, such as human face images.

In the early researches, image inpainting methods mostly imitated the process of artists completing the repair work, that is, giving priority to the consistency of neighboring areas and the continuity of boundaries. Traditional image inpainting methods can be roughly divided into two categories: texture-based synthesis technology [9, 10] and search technology based on external databases [11, 43]. For example, Bertalmio et al. [9] and Ballester et al. [10] had relied on the use of mechanisms such as distance fields to spread the characteristic information of adjacent pixels to the object area, so that the missing information can be filled. However, these methods are only suitable for processing small missing areas and require small color and texture differences between the missing area and surrounding pixels, otherwise, it will cause excessive smoothing.

Recently, several image inpainting methods [4–6] are mostly based on deep convolutional networks and generative adversarial networks, usually combining U-Net [7] networks and generative adversarial networks to generate the most suitable image patches. For example, Iizuka et al. [5] had focused on the rectangular area around the center of the image and trained it together with the adversarial network to improve the consistency between the generated pixels and the original pixels, generate prediction patches with correct semantics, and finally Smoothly combine with the undamaged parts of the image. However, when the above deep convolutional neural network-based methods are used to repair irregular damage, there are often problems such as boundary artifacts, image distortion, or fuzzy textures that are inconsistent with surrounding areas. This is often due to the context information in the model and the missing areas in the original image caused by an invalid correlation.

The two parts of the generated confrontation network are weak. The generator will continuously optimize itself through the gradient information feedback by the discriminator to discriminate the data generated by itself, so that the generated fake data can deceive the discriminator, and eventually it will be in each co-evolution tends to Nash Equilibrium, and the repaired image generated at this time will also be infinitely close to the original real image. The generator itself is a very broad concept. Its network model can make convolutional networks, multi-layer perceptron, so the generation of confrontation networks has very few restrictions on the generator, and the discriminator can use any binary classifier, using supervision learn to ensure that the output is as true as possible under the given prior distribution.

In deep learning-based methods, the researchers had used deep networks to learn semantic information in images and use the prediction results generated by the network to fill in missing areas in the images. Among them, Pathak et al. [12] had proposed a pioneering deep learning-based image inpainting model named Context-Encoders, which implements image inpainting by combining the structure of the encoder with the generation of a confrontation network. The encoder has used to learn the overall features and infer the loss, and the generative adversarial network has used to distinguish the difference between the generated image and the real image. Yu et al. [6] had proposed a coarse-to-fine generative network repair framework and added an attention mechanism. The model enhanced the features of the deeper and middle regions in the convolutional network and obtained better repair results, but there is a distortion phenomenon. Liu et al. [8] had proposed stacking partial convolution and automatic updating of the mask. This method can generate complex and reasonable texture structure and tighter context semantic connection for the missing area of the image, but the repaired image has a certain degree of distortion and local chromatic aberration problems. Bhosale et al. [37–41] had proposed the applications of deep learning techniques in diagnosis of COVID-19 and deep convolutional neural network based COVID-19 classification from radiology X-ray images for IoT enabled devices. Table 1 is a summary of image inpainting methods based on deep learning.

The partial convolutional network can make the convolution results be determined only by the non-damaged area [8]. Through automatic mask update, only effective contextual information is left in the feature map, so that the image texture of the missing area can be filled, and it can keep the height consistent with the surrounding image texture. The proposed method solves the problem that the repaired image was blurred and the patch texture was inconsistent with the original image texture structure, but the repaired image still has the problem of mild boundary artifacts and local chromatic aberrations visible to the naked eye (Fig. 1).

To solve the problem of filling damaged images, this paper combines part of the convolutional neural networks [8] with the generative confrontation network and proposed an image repair network model constructed by two independent confrontation network modules. The network uses the characteristics of part of the convolutional neural network to obtain a texture structure close to the semantic information of the original image in the missing area, which improves the quality and authenticity of image inpainting and then enhances the deeper feature extraction through the confrontation network, which solves the problem of generating images. The problem of local chromatic aberration ensures that the local consistency of the repaired image is guaranteed, and the visual effect is further enhanced.

The paper has proposed an improved image inpainting method, which is composed of two independent generative confrontation network modules. Among them, the image repair network module is a combination of partial convolution and confrontation networks, which is used to repair irregular masks. At the same time, the overall texture structure and color of the image are more closely to the original image based on the feedback of the discriminator. To solve the problem of local chromatic aberration and mild boundary artifacts caused by the insufficiency of the repair network module, the paper has designed the image optimization network module, which combines the deep residual network with the confrontation network, and through the combination of adversarial loss, perceptual loss and content loss train the image optimization network module, so that the information of the non-missing area in the image is preserved, thereby maintaining the consistency of the texture structure of the image in the non-missing area, achieving the purpose of eliminating local chromatic aberration phenomenon and solving false borders.

For this paper, the main contributions are as follows:

(1) An effective two-stage image inpainting network model constructed by two independent network modules has proposed. The improved based on partial convolutional networks, and an image inpainting network module is proposed, which is composed of partial convolutional networks and generative adversarial networks. Aiming at the problem of local chromatic aberration in the output image of the image inpainting network module, an image optimization network module has proposed, which is the combination of deep residual network and confrontation network. The synergy of the two network modules can improve the repair accuracy and enhances the visual effect.

(2) During training network, the loss function of the image repair network module has composed of the repair loss of the unmasked area, the repair loss of the masked area, the perceptual loss, the style loss, the confrontation loss, and the total variation loss. The loss of the image optimization network module is composed of content loss, perceptual loss, and counter loss. The perceptual loss in the two modules is obtained by calculating the difference between the restored image and the real image in different layers of different pre-training networks to improve the ability to capture image features and enhance the effect of image inpainting and optimization.

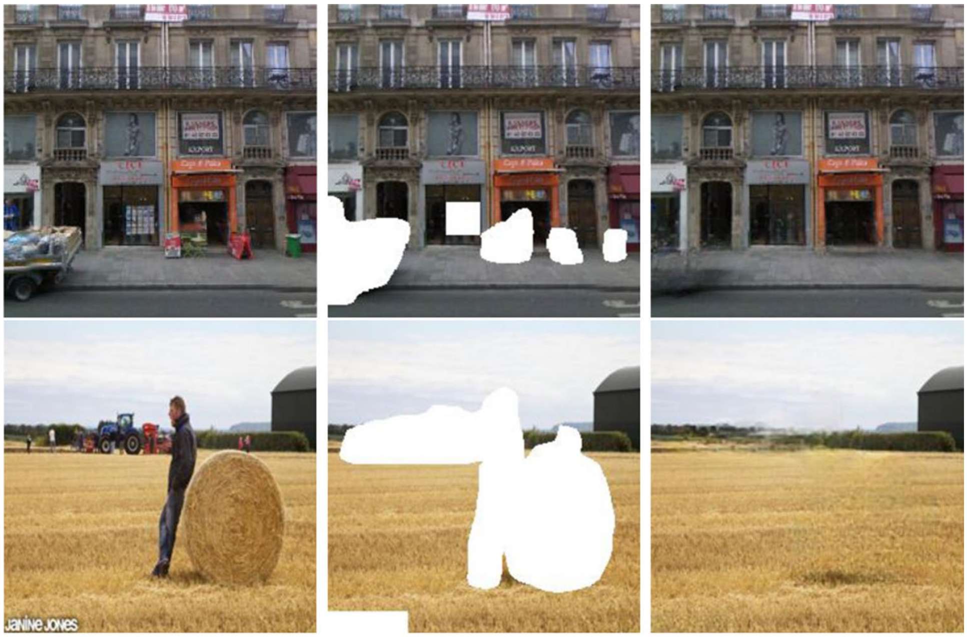
2 Approach

The image inpainting method proposed in the paper consists of two stages, the first stage is image inpainting, and the second stage is image inpainting optimization of results.

Table 1 A summary of image inpainting methods based on deep learning

Method	Main features	Existed problems	Training samples	Applications	Available code
Convolutional Neural Networks	The most widely studied and can handle high resolution images. The parameters are simple and the network structure is highly scalable. If the network is trained to generate the same resolution, it can easily repair any area of the image	Difficult texture inpainting	It is possible to converge on a dataset of a few thousand images of a specific type. Can also converge on a dataset of tens of thousands of images with diverse samples	Natural image inpainting	http://people.eecs.berkeley.edu/~pathak/context_encoder/ http://hi.cs.waseda.ac.jp/~iizuka/projects/completion/en/
Generative Adversarial Network	Clear and realistic images of the sample can be generated. Good restoration results can be achieved when a large amount of data is missing	The problem of contour discontinuity can occur, and sometimes training convergence is more difficult	Specific types of samples, low resolution samples	Clear inpainting of specific types of images	https://github.com/Avhirup/Semantic-Image-Completion
Recurrent Neural Network	A structurally coherent image of the inpainting results can be generated	Repair results are error-prone. Existing structure is not ideal for high resolution, large sample datasets	Specific types of samples, low resolution samples	Diversity inpainting for specific types of images	https://github.com/igul222/pixel_rnn

Fig. 1 Qualitative comparison of image inpainting results by ours results



(a) Original Image; (b) Input Image; (c) Ours Results

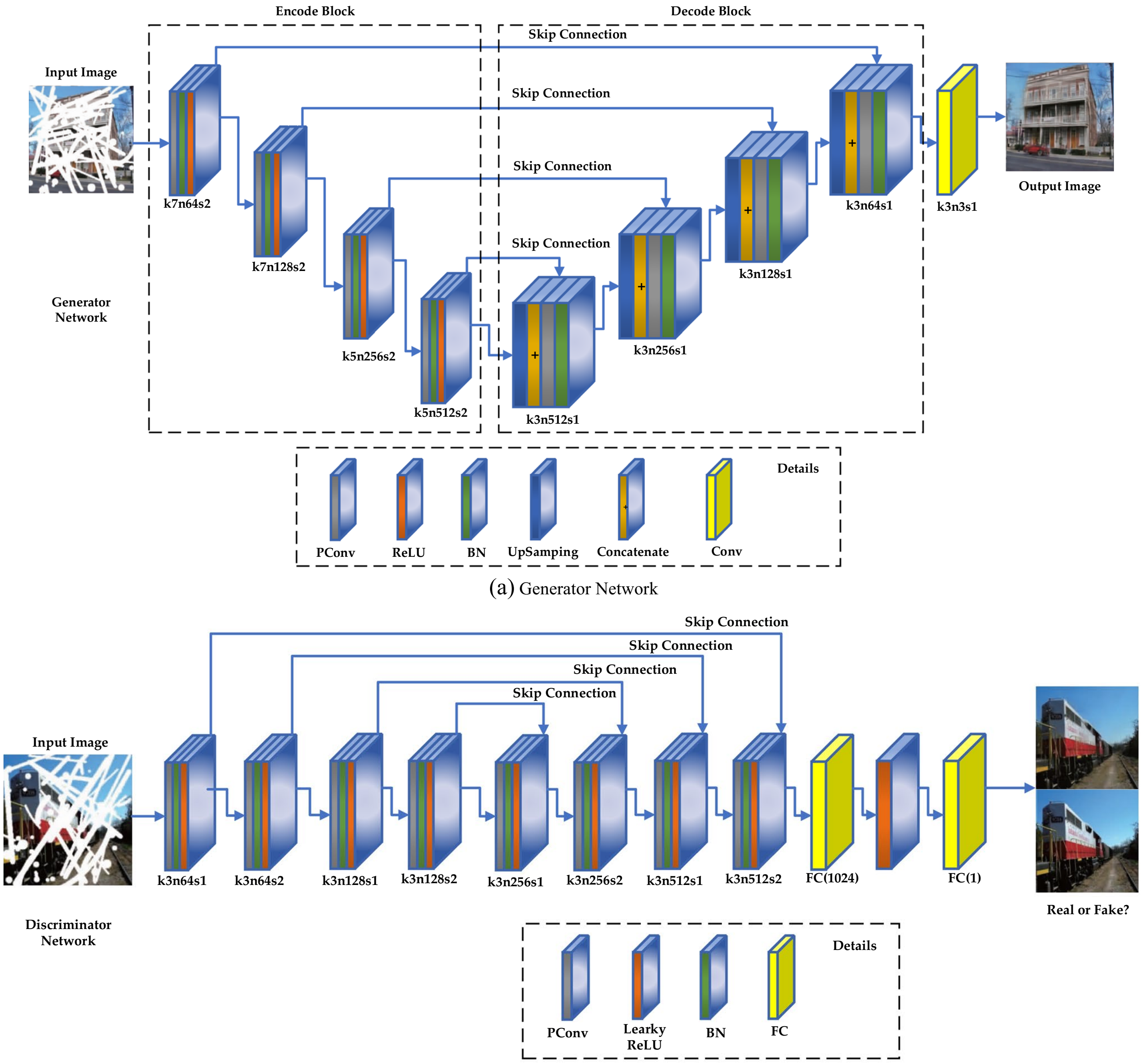
2.1 The image repair network module

The image inpainting network module proposed in the paper is shown in Fig. 2, including a generator network and a discriminator network according to Generative Adversarial Networks (GAN) structure. The generator network architecture has used the partial convolution proposed by Liu et al. [8] and Harley et al. [13] and the U-Net structure proposed by Olaf et al. [7], as shown in Fig. 2a.

The convolutional layer in the coding block of the generator network consists of three structures: a partial convolutional layer, a batch normalization layer, and a *ReLU* layer. Adding part of the convolutional layer can repair and learn the image while adding a batch normalization layer [14] to improve the fitting speed of network training and increase the generalization ability of the proposed network. As shown in Fig. 2a, the encoding stage includes a total of eight convolutional layers, of which the kernel sizes of some convolutional layers are 7, 5, 5, 3, 3, 3, 3, 3, and the number of channels is 64. 128, 256, 512, 512, 512, 512, 512, the step size is one. In the decoding stage, the image is expanded by up-sampling, because the experiments in the paper have proved that up-sampling can achieve better image inpainting effects than deconvolution. The decoding layer consists of four parts: up-sampling layer, connection layer, partial convolution layer, and batch normalization layer. The decoding stage includes a total of eight decoding layers, of which the expansion coefficient of the up-sampling layer is two, the kernel sizes of some convolutional layers are 3, 3, 3, 3, 3, 3, 3, 3, and the number of output channels is 512, 512, 512, 512, 256, 128, 64, 3, the step size is one, the *LeakyReLU* layer $\alpha = 0.2$. Except for the first and last partial convolutional layers, a batch normalization layer has used between each partial convolutional layer and the *ReLU*/*LeakyReLU* layer, as shown in Fig. 2a.

In Fig. 2, different colors are used to represent different types of layers. The size of the convolution kernel, the number of channels, and the step length of the layer are marked from left to right on the top of the layer. To enable the network to effectively integrate high-level semantic information and low-level spatial local information for image inpainting procedure, the generator network concatenates and fuses the feature maps output by each convolutional layer in the encoding block with the corresponding feature maps of the same size in the decoding block. The generator network finally uses a convolutional layer with a kernel size of one, an output channel of three, and $stride = 1$ to reduce the number of channels and use the sigmoid activation function to generate images.

The discriminator network structure refers to the discriminator network architecture proposed by Ledig et al. [15] and Goodfellow et al. [16], as shown in Fig. 2b. The discriminator convolution block consists of a convolution layer, a batch normalization layer, and a *LeakyReLU* layer. Among them, the convolution block is divided into two categories, one is the ***S1*** convolution block composed of a convolution layer with a step size of one, and the other is an ***S2*** convolution block composed of a convolution layer with a step size of two, the ***S1*** convolution block and ***S2*** convolution blocks are alternately combined into a discriminator convolution



(a) Generator Network

(b) Discriminator Network

Fig. 2 Generator network and Discriminator network in image repair network module

network, and finally the feature vector is output through the fully connected layer. The discriminator network includes a total of eight convolutional blocks. The kernel size of the convolutional layer is 3, 3, 3, 3, 3, 3, 3, 3, and the number of output channels is 64, 64, 128, 128, 256. 256, 512, 512, the step size is 1, 2, 1, 2, 1, 2, 1, 2 respectively. *LeakyReLU* layer $\alpha = 0.2$. Except for the first convolutional layer, a batch normalization layer is used between each convolutional layer and the *LeakyReLU* layer. Finally, there are two fully connected layers, with the number of channels being 1024 and 1, respectively, and the final output of the discrimination result.

2.2 Image optimization network module

Because the initial repaired image output by the image repair network module has a certain degree of local chromatic aberration, the paper has proposed an image optimization network module based on generative confrontation architecture. Its structure is shown in Fig. 3 and consists of generator and discriminator.

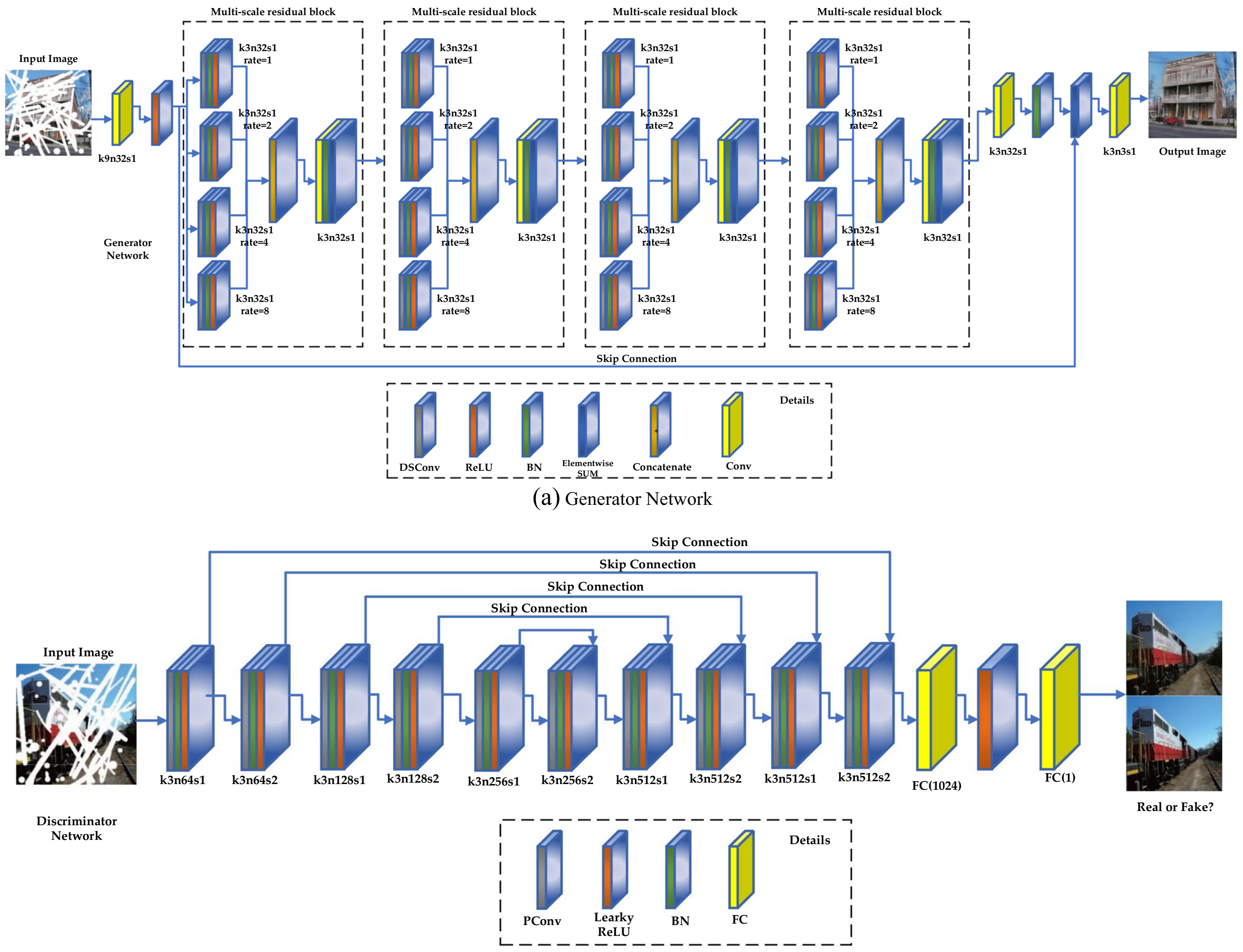
Inspired by the deep residual network proposed [15, 17] and the multi-scale expanded convolution fusion block proposed [18, 19], the generator in the image optimization network module proposed in the paper. The model structure is shown in Fig. 3a. The extraction of image features in the generator network is divided into two steps. The first stage is the preliminary extraction of the front-end convolutional layer. The kernel size is nine, the number of channels is 32, and the step size is one. The second stage is sixteen multi-scale expanded convolution residual blocks to learn the multi-scale depth features of the image. Each residual block is composed of two parts. The first part is composed of four dilated convolution blocks. The kernel size of each dilated scale convolution (*DSConv*) is three, the number of channels is 32, and the step size is equal. It is one, the expansion rate is 1, 2, 4, and 8, which enables it to extract features from four different sizes of receptive fields, which increases the feature extraction capabilities and multi-scale semantic learning capabilities. The second part is the residual connection block. This part concatenates the outputs of the four dilated convolution blocks in the first part, and then inputs them to a convolutional layer to sort the extracted features, and finally perform residual connection to relieve gradients dissipate problems and improve the ability of the network to express characteristics.

The structure of the discriminator is shown in Fig. 3b, which is similar to the network structure of the discriminator in the image inpainting network module. To allow the discriminator to better distinguish the subtle differences after image optimization, two convolution blocks are added based on Fig. 3b.

2.3 Loss function

2.3.1 Image repair network module

The purpose of image inpainting network module is to ensure that the color and spatial position of each reconstructed pixel can restore the original color and texture of the image to the



(a) Generator Network

(b) Discriminator Network

Fig. 3 Generator network and Discriminator network in image optimization network module

greatest extent. The total loss function L_1 of image inpainting network module is defined as shown in formula (1), which is composed of the restoration loss of the unmasked area, the restoration loss of the masked area, the perception loss, the style loss, the confrontation loss, and the total variation loss.

$$L_{total}^{inp} = 2L_{valid} + 12L_{hole} + 0.04L_{per} + 100\left(L_{style}^{1} + L_{style}^{2}\right) + 100L_{adv} + 0.3L_{\mathrm{var}} \tag{1}$$

The weight of each item on loss function is determined after analyzing the results of 100 independent experiments. The definition of the repair loss in the unmasked area is shown in formula (2), using the Manhattan distance between the repaired image and the unmasked area of the real image as the repair loss, where $\boldsymbol{I_{dam}}$ represents the damaged image and $\boldsymbol{M}$ represents the irregular binary mask (The corresponding area to be repaired in the mask is 0, and the others are 1), $\boldsymbol{I_{inp}}$ represents the repaired result image, and $\boldsymbol{I_{real}}$ represents the real undamaged image. The repair loss function of the masked area is shown in formula (3):

$$L_{valid} = \left\| M \times \left(I_{inp} - I_{dam}\right) \right\|_1 \tag{2}$$

$$L_{hole} = \left\| (1 - M) \times \left(I_{inp} - I_{dam}\right) \right\|_1 \tag{3}$$

The definition of perceptual loss is shown in formula (4). This paper uses different convolutional feature layers of multiple pre-trained networks to obtain the feature perception loss between the repaired result image and the real image and enhances the perception of the area to be repaired loss. In formula (4), $\boldsymbol{I_{com}}$ represents the real image of the non-missing area plus the predicted image of the missing area. As shown in formula (5), $\boldsymbol{m}$ represents the number of pre-training networks used, and $\boldsymbol{n}$ represents the number of convolutional feature layers used. The number ψ_j^i represents the ***jth*** layer convolution feature of the ***ith*** pre-training network and ω_i is the weight of the perceived loss of the ***ith*** pre-training network. After 50 independent experiments and comparisons, we finally used the *pool1*, *pool2*, and *pool3* layers of *VGG16* and *conv1*, *pool2*, and *pool3* layers of *DenseNet* in the experiments of *VGG16*. The output feature maps were used as the perception layer of the generation network. To calculate the perceived loss. The parameters of the pre-training network do not participate in training but are used to find the loss value. The perceptual loss obtained by the two pre-training networks is weighted and summed as the final perceptual loss. The weight setting used in the experiment in the paper is shown in formula (6).

$$L_{per}^{*} = \sum_{i=0}^{m-1}\left(\sum_{j=0}^{n-1}\left\|\Psi_j^i(I_{inp}) - \Psi_j^i(I_{real})\right\|_1 + \sum_{j=0}^{n-1}\left\|\Psi_j^i(I_{com}) - \Psi_j^i(I_{real})\right\|_1\right) \tag{4}$$

$$I_{com} = M \times I_{real} + (1 - M) \times I_{inp} \tag{5}$$

$$L_{per} = L_{per}^{VGG-16} + 30L_{per}^{DenseNet} \tag{6}$$

To make the repaired content close to the real image in style, this paper defines two style losses, A is the normalization factor acting on the ***jth*** layer of the ***ith*** pre-training network. The expression function is shown in formula (7) and formula (8):

$$L_{style}^{1} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left\|K_j^i\left(\Psi_j^i(I_{inp})^T\Psi_j^i(I_{inp}) - \Psi_j^i(I_{real})^T\Psi_j^i(I_{real})\right)\right\|_1 \tag{7}$$

$$L_{style}^{2} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left\|K_j^i\left(\Psi_j^i(I_{com})^T\Psi_j^i(I_{com}) - \Psi_j^i(I_{real})^T\Psi_j^i(I_{real})\right)\right\|_1 \tag{8}$$

The definition of the counter loss function is shown in formula (9):

$$L_{adv} = \frac{1}{N}\sum_{i=0}^{N-1}\left|D_{inp}\left(I_{inp}(x_i)\right) - D_{inp}\left(I_{real}(x_i)\right)\right| \tag{9}$$

$\boldsymbol{N}$ represents the total number of training samples in each batch and $I_{inp}(x_i)\,I_{real}(x_i)$ represents the ***ith*** repaired image of the batch and the real training sample image.

The total variation loss is a smooth penalty loss function for the area $\boldsymbol{P}$ proposed by Li et al. [20], $\boldsymbol{P}$ is the missing area of the damaged image, $\boldsymbol{i}$ and $\boldsymbol{j}$ represent the coordinate positions of the pixels in the missing area of the damaged image, and the total variation loss function is as follows (10) Shown:

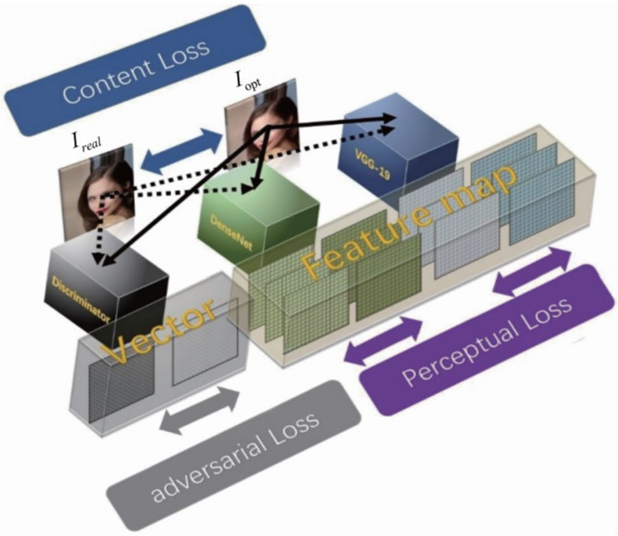


Fig. 4 Loss structure in image optimization network module

$$L_{\text{var}} = \sum_{(i,j)\in P,(i,j+1)\in P} \left\| I_{com}^{i,j+1} - I_{com}^{i,j} \right\|_1 + \sum_{(i,j)\in P,(i+1,j)\in P} \left\| I_{com}^{i+1,j} - I_{com}^{i,j} \right\|_1 \tag{10}$$

2.3.2 Image optimization network module

The purpose of the loss function of the image optimization network module is to maximize the preservation of the true and reasonable part of the image and to improve the area with local chromatic aberration. The image $\boldsymbol{I_{inp}}$ obtained by the image inpainting network module is used as the input image, and the $\boldsymbol{G_{opt}}$ of the image optimization network module outputs $\boldsymbol{I_{opt}}$. The composition of the total loss function of the image optimization network module is shown in Fig. 4, which is composed of content loss, perceptual loss, and counter loss after fifty independent experiments, the weight of each loss item is finally determined as shown in formula (11).

$$L_{total}^{opt} = 40L_{con} + L_{per} + 0.75L_{adv} \tag{11}$$

The content loss function is defined as the average absolute error with weight, as shown in formula (12). Where $\boldsymbol{N}$ represents the total number of training samples in each batch, $\boldsymbol{I_{opt}(x_i)}$ and $\boldsymbol{I_{real}(x_i)}$ represent the ***ith*** optimized image of the batch and the real training sample image, and $M\left(x_i\right)$ represents the ***ith*** binary mask of the batch.

$$L_{con} = \frac{1}{N}\sum_{i=0}^{N-1} \left| M\left(x_i\right) \times \left(I_{opt}\left(x_i\right) - I_{real}\left(x_i\right)\right) \right| \tag{12}$$

The definition of perceptual loss is similar to formula (4). The difference is that the pre-training network used is the *VGG-19* network and the *DenseNet* network pre-trained on ImageNet. The perceptual layer used is the *block3_conv4* layer of the *VGG-19* network and in the *pool2_conv* layer of the *DenseNet* network, the weights are set as shown in formula (13).

$$L_{per} = 1.25L_{per}^{VGG-19} + 50L_{per}^{DenseNet} \tag{13}$$

The generative loss function is shown in formula (14):

$$L_{adv} = -\frac{1}{N}\sum_{i=0}^{N-1} \left(D_{opt}\left(I_{opt}\left(x_i\right)\right)\log\left(D_{opt}\left(I_{real}\left(x_i\right)\right)\right) + \left(1 - D_{opt}\left(I_{opt}\left(x_i\right)\right)\right)\log\left(1 - D_{opt}\left(I_{real}\left(x_i\right)\right)\right)\right) \tag{14}$$

$\boldsymbol{N}$ represents the total number of training samples in each batch, and $\boldsymbol{I_{opt}(x_i)}$ and $\boldsymbol{I_{real}(x_i)}$ represent the ***ith*** optimized image sample and real image sample in each batch.

The training of the image inpainting network module and the image optimization network module is carried out sequentially. First, $\boldsymbol{I_{dam}}$ and $\boldsymbol{I_{real}}$ are input to the image inpainting network module for training, and then the output results of the image inpainting network module $\boldsymbol{I_{inp}}$ and $\boldsymbol{I_{real}}$ are input to the image optimization network module for training.

3 The experimental results and analysis

3.1 Experimental settings

The proposed model in the paper was trained and tested on the Places2 [21] and CelebA [22] datasets. Randomly select 200,000 images from each dataset as the training dataset and 4000 images as the testing dataset. The size of the image and the mask are both set to 256 × 256.

The CelebA dataset and COCO dataset [42] are used to conduct network model training and comparative evaluation of image inpainting effects. Among them, the CelebA dataset contains a total of more than 200,000 face images including more than 10,000 celebrities. It has the advantages of sufficient data volume, high quality, and rich image information annotations. Some samples are shown in Fig. 5. The COCO dataset contains 12 categories such as people, animals, and vehicles, which can be subdivided into 80 categories such as buses, airplanes, trains, cats, dogs, and birds. Some samples are shown in Fig. 6. The training dataset of the experiment is obtained by randomly selecting 80% of the dataset, and the remaining 20% is used as the test dataset. In fact, in the training process, the uncoded pictures that are used as references also need to be synchronized with resolution input from small to large in sequence (Table 2).

The initial weight setting method of the generator in the image repair network module selects the initialization method proposed by Zhang et al. [23], the mini-batch used in the experimental training is set to 4, a total of 100,000 iterations of training, and the optimization algorithm uses Adam [24], Decay = 1e − 7, and the learning rate is set to 2×10^{-4} and β = 0.05. The related parameters of Loss function are depicted in Section III.

The operating system platform is the Ubuntu 16.04, Intel Core i9-9900 k CPU, 32 GB memory, GeForce RTX 2080Ti GPU and 1 TB hard disk. The proposed model is trained using the GPU-based ***PyTorch*** 1.0.0 deep learning framework with Python 3.6.6, Cuda 10.0 and Cudnn 7.3.1. The Python IDE is ***PyCharm***, and the operating system is ***Ubuntu*** 16.04 LTS.

Fig. 5 Some samples of CelebA dataset [22]

Fig. 6 Some samples of COCO dataset [42]

Table 2 Datasets Description

Dataset	Year	Content Category	Image Type	Size
Places2 [21]	2017	Miscellaneous	RGB	10Mil, 256×256
CelebA [22]	2018	Face Images	RGB	200 K, variant
COCO [42]	2014	Miscellaneous	RGB	2.5Mil, 256×256

3.2 Comparisons to state-of-the-arts

This section analyzes and compares the proposed method in the paper with three deep learning image inpainting methods proposed in recent years. In the experimental results, Context Encoder (CE) represents the method proposed by Pathak et al. [12], and Globally and Locally (GL) represents the method proposed by Iizuka et al. [5], Partial Convolutions (PConv) represents the method proposed by Liu et al. [8]. Both CE and GL are representative research work in the field of image inpainting of regular masks, and PConv can represent the cutting-edge method in image inpainting of irregular masks.

3.2.1 Qualitative evaluations

Figure 7 shows the method proposed in the paper and three advanced methods: CE [12], GL [5], and PConv [8] qualitatively compare the results of the CelebA and Places2 datasets. From the results, the CE repaired image cannot be produced. For complex textures, although the pixels in the filled area have a certain semantic similarity with the surrounding pixels, the resulting image is blurry and too smooth. As shown in the second row and second column of Fig. 7, although the eyes of the image have a rough outline, but it is very fuzzy and has obvious local chromatic aberration. The image repaired by GL achieves better results in contextual semantics and further refines the texture, but it cannot guarantee the consistency of the local and the whole. There are still problems with obvious regional boundaries and serious loss of texture details, as shown in the fifth row of Fig. 7. The repair results in the third column are shown.

The texture details of the restored house are seriously missing. The image repaired by *PConv* further refines the texture structure, obtains reasonable texture details and roughly correct context semantics, but cannot guarantee local consistency, and there is a phenomenon of local chromatic aberration. As can be seen from the repair results in Fig. 7, this paper proposes a method to repair in the post image, the repaired area has no false boundaries and can produce a reasonable texture structure and correct context semantics. It can significantly improve the local chromatic aberration problem, greatly enhance the visual effect and image quality, and reflect the irregularity of the method proposed in the paper. The superiority of image inpainting in the case of masks is shown in the third row and fifth column of Fig. 7. When restoring the side face image of a person, the results of the method proposed in the paper are repaired, and detailed results are generated in both the face and the background area of the person. Texture and local chromatic aberration have also been significantly improved, and the repaired image is more natural and realistic. As shown in Fig. 8, there are comparisons of the image inpainting visual results on the Places dataset. As shown in Fig. 9, there are comparisons of the image inpainting visual results on the CelebA dataset. As shown in Fig. 10, there are comparisons of the image inpainting visual results on the COCO dataset.

3.2.2 Quantitative evaluations

This paper adopts the structural similarity (SSIM) proposed by Zhou et al. [25] and the Peak Signal-to-Noise Ratio (PSNR) proposed by Quail et al. [26] as an image inpainting performance metric. The peak signal-to-noise ratio is to evaluate the quality of the image by the error of the corresponding pixels of the two images. The larger the value, the better the inpainting result of the image, as shown in the following equations:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|x(i,j) - y(i,j)\|^2 \tag{15}$$

$$PSNR{=}10 \cdot \log_{10}\left(\frac{MAX_x^2}{MSE}\right) \tag{16}$$

wherein, m, n represent the size of the image size, MSE represents the mean square error between the images. MAX_x^2 represents the maximum value taken in the image.

The SSIM indicator measures the structural similarity of the image from the brightness, contrast and structural information of the image, thereby evaluating the degree of distortion of the image. The larger the value, the smaller the distortion, and the closer the repaired image is to the original image, as shown in the following equations:

$$SSIM = \frac{\left(2\sigma_x\sigma_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\sigma_x^2 + \sigma_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{17}$$

wherein, σ_x and σ_y are the mean values of the original image and the repaired image, respectively, σ_x^2 and σ_y^2 are the variances of the original image and the repaired image, respectively, σ_{xy} is the covariance, and C_1 and C_2 are constants used to maintain stability. As shown in Table 3, this section compares the SSIM and PSNR values of the restoration results

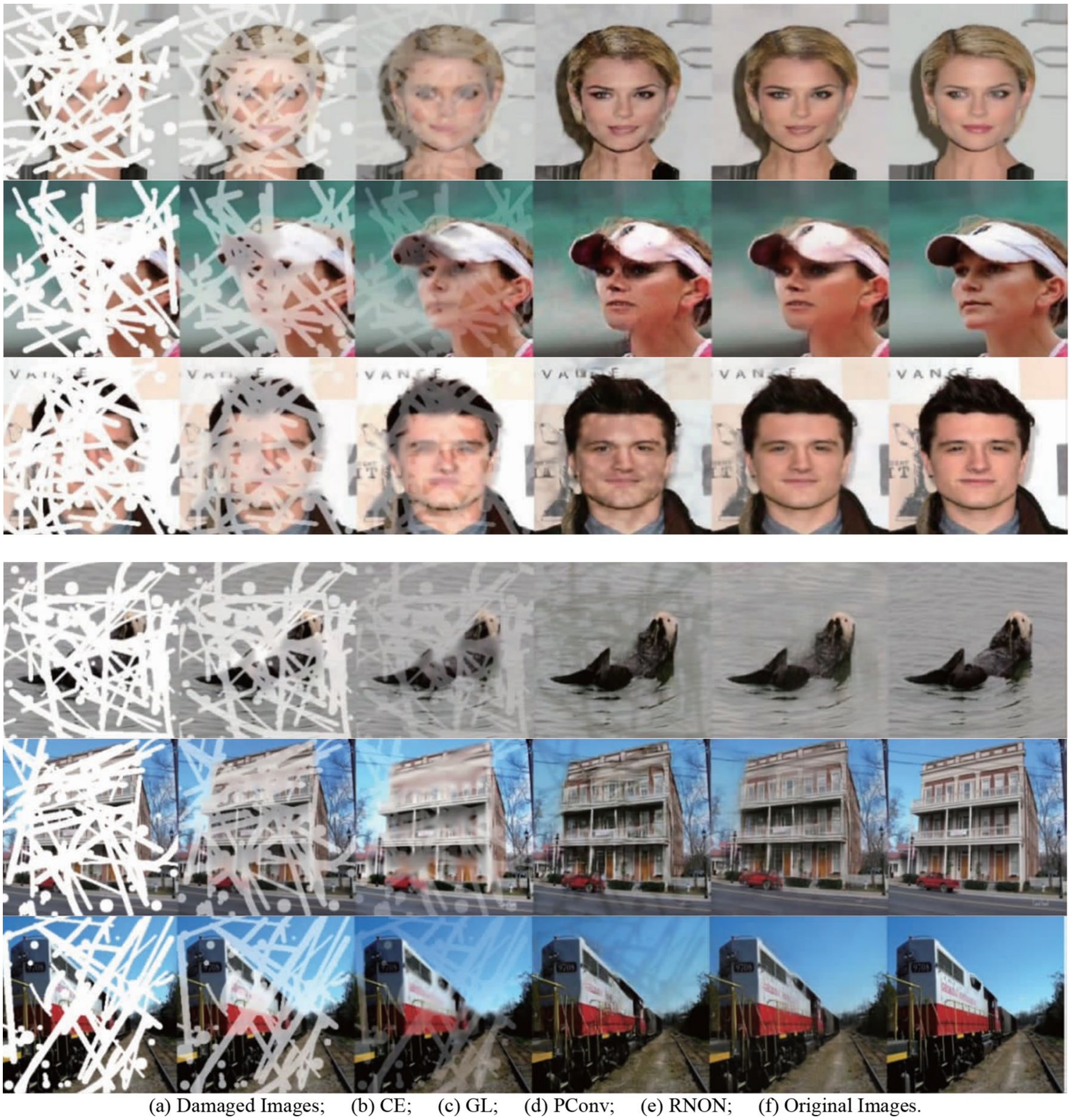

(a) Damaged Images; (b) CE; (c) GL; (d) PConv; (e) RNON; (f) Original Images.

Fig. 7 Comparisons of the repair results on the CelebA and Places2 datasets, from left to right, are **a** damaged images, **b** CE, **c** GL, **d** PConv, **e** RNON, and **f** original images

of *CE* [12], *GL* [5] and *PConv* [8] on the CelebA dataset and Places2 dataset. From the results in Table 3, it can be seen that on the two datasets, the SSIM and PSNR values of this method have obvious advantages over *CE* and *GL*.

Compared with *PConv*, for the metric PSNR, on the CelebA and Places2 datasets, the method in the paper has achieved better results under the different damage levels represented by the four different mask rates shown in Table 2. On the Places2 dataset, when the mask rate is between (0.01,0.25] and (0.40,0.50), the difference between the proposed method in the paper and *PConv* is very small, only 0.42 and 0.04. In other cases, it is better than that. Nearly 0.6 and 0.5. For the metric SSIM, on the two datasets, when there is a greater degree of damage, that is, when the mask rate is larger, the method in the paper achieves a higher SSIM value than *PConv*, only when the mask When the rate is (0.01, 0.25], the score is equivalent to *PConv*, with a difference of about 0.03. The possible reason is that the *PConv* method uses a local repair mechanism. In the case of a low mask rate, the area of the missing area is small, and the local chromatic aberration of the image generated by *PConv*. The problem is not obvious, so good results have been achieved.

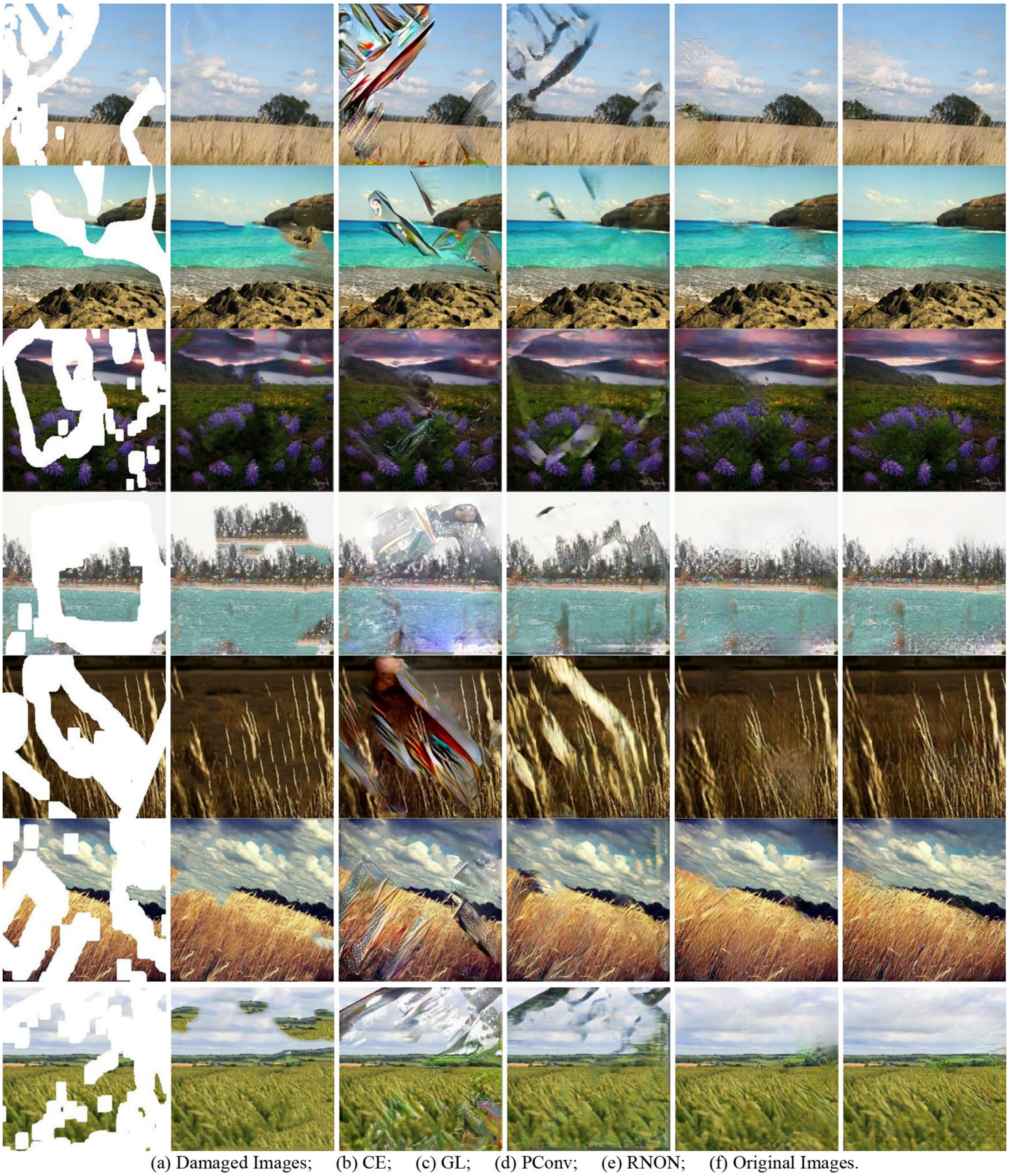
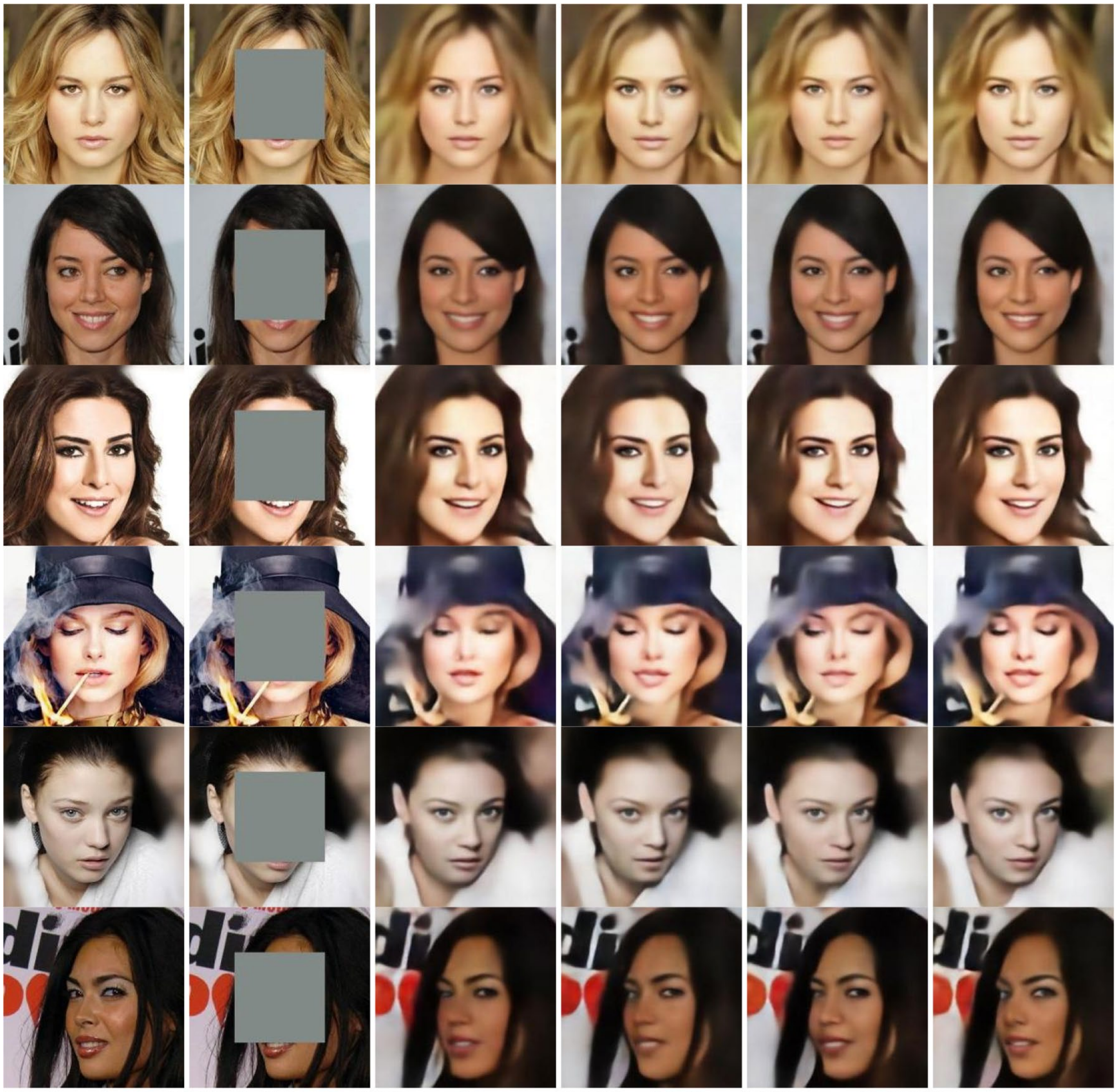


Fig. 8 Comparisons of the repair results on the Places2 datasets, from left to right, are **a** damaged images, **b** CE, **c** GL, **d** PConv, **e** RNON, and **f** original images

The mechanism of the method in the paper is a local repair mechanism combined with a global repair mechanism. In the case of a large missing area, the local chromatic aberration phenomenon can be greatly reduced and a better visual effect can be achieved. Based on the two metrics on the two datasets, the performance of the proposed method in the paper is better than that of *CE*, *GL*, and *PConv*.

3.3 Failure cases

The visual comparison of different activation functions or 3×3 filter for mask updating has shown in Fig. 11. Figure 12 shows some failure cases of RNON model. The proposed model struggles to recover the high-frequency details while the damaged areas are too large or the



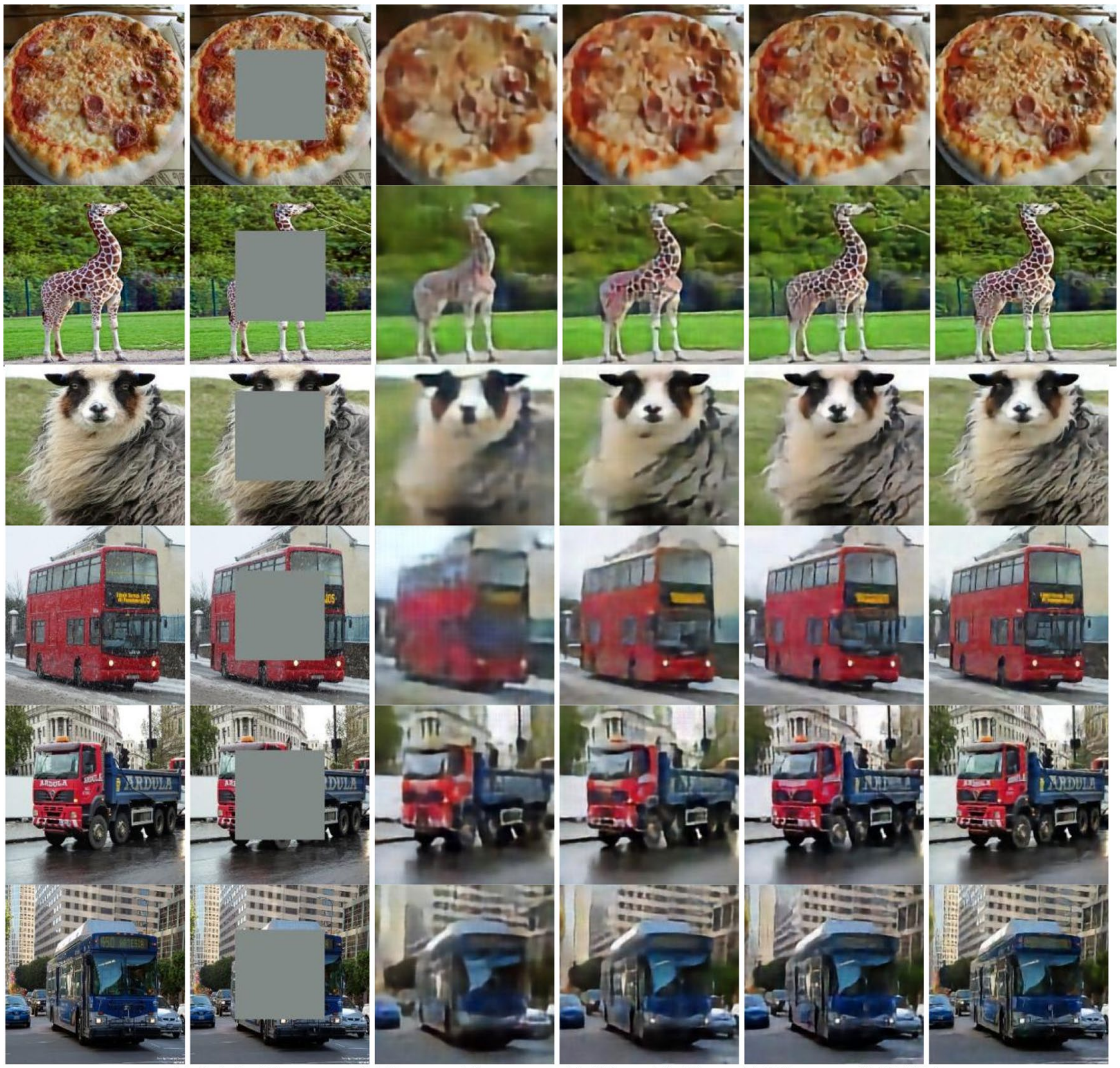
(a) Original Images; (b) Damaged Images; (c) CE; (d) GL; (e) PConv; (f) RNON.

Fig. 9 Comparisons of the repair results on the CelebA dataset, from left to right, are **a** original images, **b** damaged images, **c** CE, **d** GL, **e** PConv, and **f** RNON

background are too complex. In some cases, the mask covers a large portion of a specific object, like a car, it is still difficult for proposed model to recover the original shape.

4 Conclusion

Aiming at the computer vision task of image inpainting, this paper proposes a model that combines partial convolution and generative adversarial networks. This model can fill in missing regions of any size, shape, and position in the image. Through experimental comparison with other cutting-edge technologies, the proposed model in the paper can get a more natural restoration image with a more natural visual effect. Compared with GL and CE, the method in the paper can generate a more semantically reasonable texture structure and context connection, thereby solving the problem of image distortion and texture blur. Compared with the partial convolutional network, the method designed in the paper can generate more integrated repaired images, which makes up for the local chromatic aberration caused by the limitations of partial convolution. However, the image optimization network module is fully optimized for the entire image, so it is in the process of eliminating local chromatic aberrations. At the same time, the overall color of the image will also have a small deviation, which will reduce the scores of the PSNR and SSIM indicators in the evaluation effect. However, in terms of



(a) Original Images; (b) Damaged Images; (c) CE; (d) GL; (e) PConv; (f) RNON.

Fig. 10 Comparisons of the repair results on the COCO dataset, from left to right, are **a** original images, **b** damaged images, **c** CE, **d** GL, **e** PConv, and **f** RNON

Table 3 PSNR/SSIM for Celeba and places2 datasets

Evaluations	Methods	Mask Rate (CelebA/Places2)			
		(0.01, 0.25)	(0.25, 0.40)	(0.40, 0.50)	(0.50, 0.60)
SSIM	CE [12]	0.8983/0.8841	0.7908/0.7929	0.5267/0.5277	0.4469/0.4321
	GL [5]	0.9136/0.9087	0.8079/**0.8084**	0.5438/0.5434	0.4403/0.4635
	PConv [8]	**0.9253**/**0.9291**	0.7993/0.7908	0.5957/0.5958	0.4897/0.4769
	RNON	0.9232/0.9254	**0.8235**/0.8016	**0.6032**/**0.6097**	**0.4926**/**0.5029**
PSNR	CE [12]	25.98/23.86	19.87/18.36	18.03/17.53	16.88/15.96
	GL [5]	26.76/26.23	20.21/20.66	19.13/19.80	17.96/16.69
	PConv [8]	28.20/**28.73**	23.94/23.86	21.57/**22.53**	19.21/18.77
	RNON	**28.75**/28.31	**24.30**/**24.46**	**22.52**/22.49	**19.99**/**19.23**

Bold-underline value indicates that the best experimental result

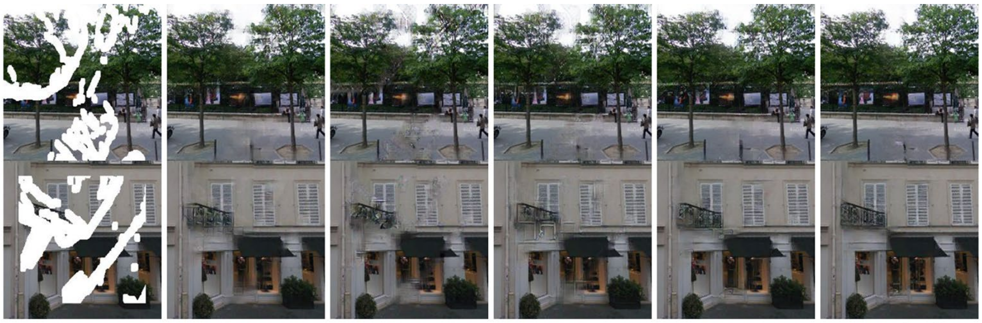

(a) Input Images; (b) Sigmoid as activation function; (c) LeakyReLU with slope of 0.2 as activation fuction; (d) ReLU; (e) 3×3 filter for mask updating; (f) Proposed model. All images are scaled to 256×256 .

Fig. 11 Visual comparison of different activation functions or 3×3 filters on the bidirectional attention maps, from left to right, are **a** Input Images; **b** Sigmoid as activation function; **c** LeakyReLU with slope of 0.2 as activation fuction; **d** ReLU; **e** 3×3 filter for mask updating; **f** Proposed model. All images are scaled to 256×256

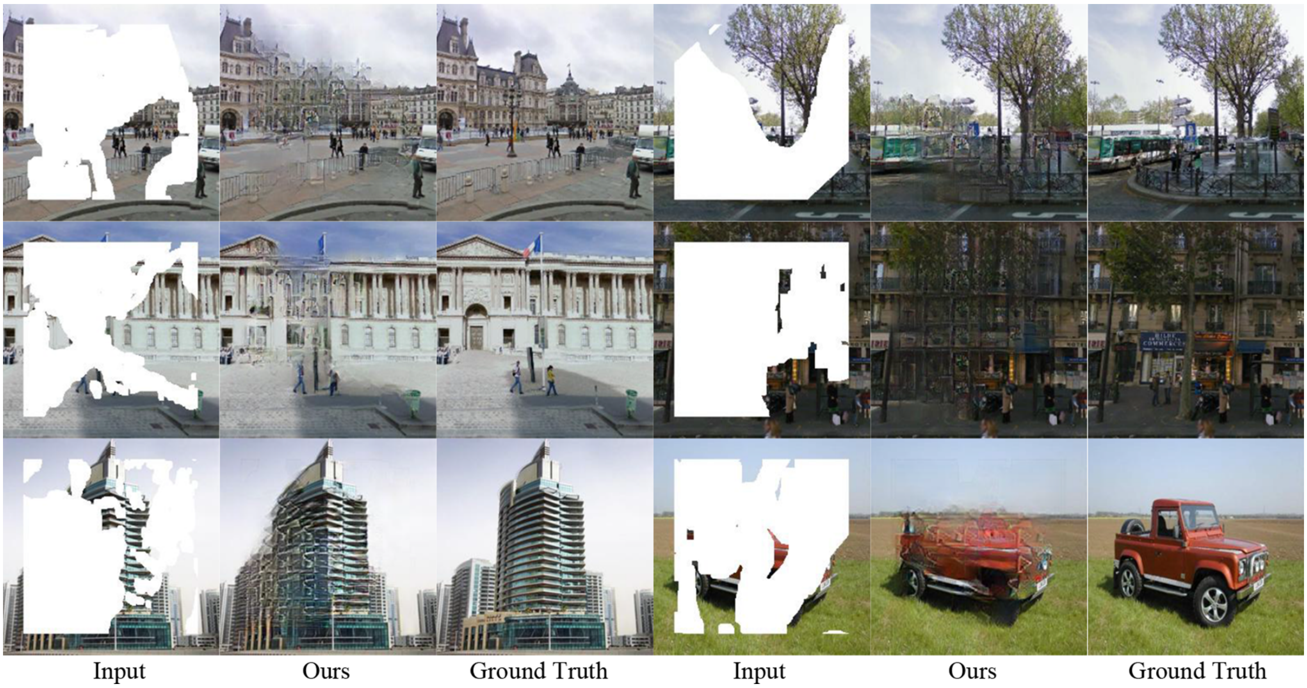


Fig. 12 Failure cases of proposed model. Each group is ordered as input image, our result and ground truth. All images are scaled to 256×256

the overall quality and visual effects of the image, optimizing the network can still largely improved the performance of image inpainting. Therefore, the repaired image obtained by the method in the paper not only significantly improves the authenticity of the semantic level, but also greatly reduces the sense of violation in the visual effect.

Aiming at the phenomenon of the overall chromatic aberration shifting in the repaired image obtained in the paper, further work will propose a new solution to this problem. In addition, the paper will also try to apply model migration to video restoration work.

Acknowledgements The authors would like to thank CelebA, Places2 and COCO datasets, which allowed us to train and evaluate the proposed model. The authors would also like to thank all the reviewers for their insightful comments.

Author contributions YC: methodology, validation, writing—original draft, data curation. RX: software, data curation. KY: writing—review and editing, formal analysis. KZ: validation, resources, writing—review and editing.

Funding This work is supported by A Project Supported by Scientific Research Fund of Hunan Provincial Education Department under Grant 22A0701, and 2022 Institute Scientific Research Project of Hunan University of Information Technology under Grant XXY022ZD01.

Data availability No additional data are available.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Newson A, Almansa A, Fradet M, Gousseau Y, Perez P (2014) Video inpainting of complex scenes. Siam J Imaging Sci 7(4):1993–2019
2. Levin A, Zomet AA, Peleg S and Weiss Y (2004) Seamless image stitching in the gradient domain. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 377–389
3. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) Patch match: a randomized correspondence algorithm for structural image editing. ACM Trans Graphics 28(3):2–11
4. Zheng C, Cham T and Cai J (2019) Pluralistic image completion. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1438–1447
5. Iizuka S, Simo-Serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graphics 36(4):1–14
6. Yu JH, Lin Z, Yang YM, Shen XH, Lu X and Huang TS (2018) Generative image inpainting with contextual attention. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5505–5514
7. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. Siam J Image Sci 18(4):55–67
8. Liu GL, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B (2018) Image inpainting for irregular holes using partial convolutions. ACM Trans Graphics 9(3):37–51
9. Bertalmio M, Sapiro G and Caselles V (2000) Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pp 417–424
10. Ballester C, Bertlmio 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
11. Hays J, Efros A (2007) Scene completion using millions of photographs. ACM Trans Graphics 26(3):4–12
12. Pathak D, Krahenbuhl P, Donahue J, Darrell T and Efros AA (2016) Context encoders: feature learning by inpainting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2536–2544
13. Harley AW, Derpanis KG and Kokkinos I (2017) Segmentation-aware convolutional networks using local attention masks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017, pp 22–29
14. Ioffe S and Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning (ICML), pp 448–456
15. Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken AP, Tejani A, Totz J, Wang ZH and Shi WZ (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 105–114
16. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC and Bengio Y (2014) Generative adversarial nets. In Proc. Adv. Neural Info. Process. Syst., pp 2672–2680
17. He KM, Zhang XY, Ren SQ and Sun J (2016) Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778
18. Zeng YH, Fu JH, Chao HY and Guo BN (2019) Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1486–1498.
19. Yu F and Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In: Proceedings of 4th International Conference on Learning Representation (ICLR), pp 23–32
20. Johnson J, Alahi A and Li FF (2016) Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 694–711
21. Zhou BL, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464
22. Liu ZW, Luo P, Wang XG and Tang XO (2015) Deep learning face attributes in the wild. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 3730–3738
23. Liu XD, Gao JF, He XD, Deng L, Duh K and Wang YY (2015) Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In: Proceedings of 2015 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp 912–921
24. Kingm D and Ba J (2014) Adam: a method for stochastic optimization. In: Proceedings of 4th International Conference on Learning Representation (ICLR), pp 58–64
25. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
26. Korhonen J and Junyong Y (2012) Peak signal-to-noise ratio. In: Proceedings of International Workshop on Quality of Multimedia Experience Electronics Letters (QoMEX), pp 37–38
27. Guo Q, Gao SS, Zhang XF, Yin YL, Zhang CM (2018) Patch-based image inpainting via two-stage low rank approximation. IEEE Trans Vis Comput Graphics 24(6):2023–2026
28. Xie CH, Liu SH, Li C, Cheng MM, Zuo WM, Liu X, Wen SL, Ding E (2019) Image inpainting with learnable bidirectional attention maps. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 8857–8866
29. Yu JH, Lin Z, Yang JM, Shen XH, Lu X, Huang TS (2019) Free-form image inpainting with gated convolution. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 4470–4479
30. Zeng YH, Fu JL, Chao HY, Guo BN (2019) Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1486–1494
31. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 7132–7141
32. Liu WW, Xu D, Tsang IW, Zhang WJ (2019) Metric learning for multi-output tasks. IEEE Trans Pattern Anal Mach Intell 41(2):408–422

33. Liu HY, Jiang B, Xiao Y, Yang C (2019) Coherent semantic attention for image inpainting. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 4169–4178
34. Ding D, Ram S, Rodriguez JJ (2019) Image inpainting using nonlocal texture matching and nonlinear filtering. IEEE Trans Image Process 28(4):1705–1709
35. Wang N, Li JY, Zhang LF, Du B (2019) Musical: multi-scale image contextual attention learning for inpainting. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp 3748–3754
36. Doersch C, Singh S, Gupta A, Sivic J, Efros AA (2015) What makes paris look like paris?. Communications of the ACM, pp 103–110
37. Bhosale YH, Patnaik KS (2022) Application of deep learning techniques in diagnosis of COVID-19 (coronavirus): a systematic review. Neural Process Lett. https://doi.org/10.1007/s11063-022-11023-0
38. Bhosale YH, Patnaik KS (2022) IoT deployable lightweight deep learning application for COVID-19 detection with lung diseases using raspberryPi. Proc Int Conf IoT Blockchain Technol (ICIBT). https://doi.org/10.1109/ICIBT52874.2022.9807725
39. Bhosale YH, Zanwar S, Ahmed Z, Nakrani M, Bhuyar D and Shinde U (2022) Deep convolutional neural network based Covid-19 classification from radiology X-ray images for IoT enabled devices. In: Proceedings of 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS). https://doi.org/10.1109/ICACCS54159.2022.9785113
40. Bhosale YH, Singh P, Patnaik KS (2022) COVID-19 and associated lung disease classification using deep learning. In: Proceedings of International Conference on Innovative Computing and Communications, pp 283–295
41. Bhosale YH, Patnaik KS (2022) PulDi-COVID: chronic obstructive pulmonary (lung) diseases with COVID-19 classification using ensemble deep convolutional neural network from chest X-ray images to minimize severity and mortality rates. Biomed Signal Process Control. https://doi.org/10.1016/j.bspc.2022.104445
42. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick CL (2014) Multi-scale context aggregation by dilated convolutions. In: Proceedings of European Conference on Computer Vision (ECCV), pp 740–755
43. Banan A, Nasiri A, Garavand AT (2020) Deep learning-based appearance features extraction for automated carp species identification. Aquacult Eng 89:102053