Multimedia Tools and Applications (2023) 82:47751–47771
https://doi.org/10.1007/s11042-023-15313-0

DGCA: high resolution image inpainting via DR-GAN and contextual attention

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

Received: 6 January 2023 / Revised: 25 February 2023 / Accepted: 6 April 2023 /
Published online: 5 May 2023

Abstract

The most image inpainting algorithms often have existed problems such as blurred image, texture distortion and semantic inaccuracy, and the image inpainting effect is limited for images with large missing regions and resolution level. To solve above problems, the paper proposes an improved two-stage image inpainting network based on parallel network and contextual attention. Firstly, the improved deep residual network is used to perform generative pixels filling on the missing area, and the first-stage adversarial network is used to complete the edges information. Then, the color features of the filling map are extracted, the edge map is fused and complemented, and the fusion map is used as the conditional label of the second-stage adversarial network. Finally, the image repairing result has obtained through the two-stage network with the contextual attention module. The experiments on public datasets can show that the proposed algorithm can obtain a more realistic repairing effect.

Keywords Image inpainting · Deep learning · Conditional generative adversarial network · Contextual attention · Parallel network

1 Introduction

Image inpainting, also known as image completion, is to automatically repair missing pixels in the image based on the known information of the original image. This is an important research field of computer vision and pattern recognition. With the development of digital image processing technology, the image inpainting had been widely used in image editing, object tracking, occluded object recognition, and intelligent aesthetics. Early image

✉ Yuantao Chen
chenyt@hnuit.edu.cn

[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] Hunan Provincial Science and Technology Affairs Center, Changsha, China

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

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

inpainting efforts focused on texture synthesis techniques [1, 2]. These traditional methods had used nearest neighbor search to duplicate the relevant image blocks and fill in the missing regions using image blocks from existing regions [23]. However, they performed poorly when there are no available duplicate textures in undamaged regions due to the inability to obtain high-level semantics from the images [18]. However, these methods usually can only copy pixels, stitching and transplantation, it is difficult to effectively obtain the global structure and semantic information of the original image, and it is also difficult to generate new image pixels for damaged area, so there are many limitations in the applications.

The development of deep learning and Convolutional Neural Networks (CNN), especially the Generative Adversarial Networks (GAN) has accelerated the development of image inpainting content [33, 37]. In the adversarial network, high-level semantic acquisition and low-level pixel synthesis are jointly trained to guide the network to reconstruct new content with real meaning in the missing area, and effectively make up for the shortcomings of traditional image inpainting algorithms. Due to the complexity and diversity of natural images, it is not enough to generate new pixels. It is also necessary to ensure that the inpainting results are visually realistic and vivid. At the same time, only using the classic single-order GAN architecture can also cause artifacts, blurring, and texture distortion.

In order to solve the above problems, researchers have made improvements and innovations on the basis of GAN. Iizuka et al. [11] used a global discriminator and a local discriminator to determine the semantic consistency of the generated image to improve the rationality of the content, but the local discriminator can only discriminate the rectangular mask, and it is difficult to deal with irregularities. Zheng et al. [39] proposed an image repairing method named Edge-Connect, which uses binary edge mapping as a network label to eliminate high-frequency textures and reduce artifacts. Wan et al. [26] proposed an image repairing method named Foreground-Aware, which uses a saliency detection algorithm to extract foreground objects in an image, and combines edge features as network labels. However, in addition to the two-stage generation network, this proposed method also needs to build three additional encoding and decoding networks, which increases the cost of network training, testing and practical application. At the same time, this proposed method will also cause a large amount of background information to be lost.

In edge detection, a lot of meaningful information are often been ignored, resulting in a lack of vividness in the generated content. Zhang et al. [35] divided the network into two modules, a rough content generator and a fine content generator, and used a parallel architecture to achieve repair. However, since the network uses the true map as the label in the initial stage, and the high frequency and non-critical details in the true map will have a great impact on the repairing result, this effect is usually negative. Pen et al. [22] proposed a structure and appearance flow image inpainting method, which obtained the image texture structure through the relative total variation metric, and then used the texture structure as the label input network for training, and achieved certain results. However, one of the prerequisites of this method is that structures with similar semantics in images also have similar details in a high probability, and when this assumption is not true, it will mislead the repair results.

Based on the above situations, this paper proposes a two-stage image inpainting network based on parallel confrontation and multi-condition fusion for high-resolution image inpainting. The network adopts a two-stage architecture, which is a label generation network and a fine content reconstruction network respectively. The label generation network is divided into two parallel modules, an edge generator and a rough content generator. At the same time, an edge pixel fusion operator is designed in the output link of the fine content reconstruction network to improve the pixel consistency of the edge repair. Experiments on the datasets of Places2 [31], CelebA [6], Facades [24] and Oxford Building [7] can show that the proposed network can obtain a better repair effect visually, and the evaluation index measurement can obtain a significant advantage.

The structure of this paper is as following: Sect. 1 is the introduction of related development with image inpainting. Section 2 depicts some related works and background by image pixel filling and image generation. Section 3 proposed the improved network named DGCA. Section 4 illustrates experimental results on datasets of image inpainting. Finally, the conclusion summarizes the proposed method and future research points.

2 Related works

2.1 Image pixel filling

Like most computer vision problems, the research on image inpainting problems predates machine learning and deep learning, such as interpolated pixel filling based on light intensity vector fields [10] and global image synthesis based on feature histograms [12]. However, these methods based on pixel expansion technology can usually only fill smaller areas in the image, such as ink dots, lines, or scratches. The method based on patch matching and texture synthesis technology can perform more repair tasks, such as filling larger defects or holes. Barnes et al. [1] proposed a random patch matching (Patch-Match) algorithm. On the basis of Patch-Match, by combining algorithms such as super pixel stitching [15], optimal patch search [8], and global consistency filling [30], real-time repairing [32] of the input image can be achieved.

However, this type of method mainly uses the low-level pixel features of the image, which is invalid for high-level semantics and complex structures, and cannot generate new content that does not exist in the input image. For this reason, Liu et al. [18] proposed a repair method based on a large external database drive. Assuming that the regions with similar contexts in the image also have the same content, the external database is searched for the most similar sample to the input image, and the part of the matching sample corresponding to the missing region of the input image is cut and transplanted to the input image. However, when there is no suitable sample on the called database, this kind of method will cause the repair result to be wrong. In addition, the external database that this kind of method needs to call is usually very large, which limits the actual development of related applications.

2.2 Attention module

In deep learning, the attention mechanism can reasonably allocate computing resources to the model according to the importance of input features. According to the dimension of the attention mechanism application, it can be divided into spatial attention [26] (*SA*) and channel attention [35] (*CA*). Because the contextual features of the spatial dimension can be used as the reference object of the damaged image area in the image inpainting process, the literature [28, 36, 37, 41] all apply the SA mechanism to the image inpainting network. Wang et al. [34] proposed a contextual attention layer, which uses SA to find the most similar background content for the missing image area. Wang et al. [28] proposed a multi-scale image context attention learning strategy based on the context attention layer, which enables the inpainting model to deal with rich background information more flexibly. After that, Zhang et al. [37] proposed a coherent semantic attention layer, which uses SA to improve the spatial semantic consistency within the repaired damaged image region. To make the image inpainting results consistent visually and semantically, Zhu et al. [41] proposed a pyramid context encoder network

(*PEN-Net*), which uses the pyramid network structure to learn from high-level semantic features. Attention information had been transferred to low-level features.

2.3 Image generation

CNN has been trained on the ImageNet dataset [32], which shows its excellent performance in acquiring high-level semantic features of images. The stacked deep convolutional network model with a large number of hidden layers trained through massive data can effectively obtain the nonlinear and complex mapping relationship between samples, which is consistent with the design idea of semantic inpainting based on image content. Pathak et al. [21] proposed a CNN-based encoding and decoding network structure. The encoder is a series of layer-by-layer down-sampling, while the decoder corresponds to the encoder. The network is trained with Euclidean distance and adversarial loss as constraints. Realize the generation of new pixels in the missing area. Fang et al. [6] used a multi-classification network to identify and classify the image texture, and then used the classification label as the content constraint item of the repair network, and synthesized the output result with fine texture through the multi-scale feature. Hu et al. [10] designed a set of repair network models containing two encoding and decoding structures, outputting rough content and fine content respectively, and outputting the previous term as the input condition of the latter term, and achieved a certain breakthrough in the task of repairing irregular masks.

In addition, the feature pyramid network [21], gated convolution network [27], Bidi-directional attention network [38] and coherent semantic attention network [40] all discuss the problem of image inpainting from different perspectives. Related experiments show that the above method generates new content with reasonable semantics in highly structured images (such as buildings, objects, people, landscapes, etc.). In addition, in tasks such as image style conversion [13, 24], image domain conversion [3, 25], image recoloring [4, 14] and image super-resolution reconstruction [34, 42], generative adversarial networks are also widely used.

3 Approach

The overall network architecture of the proposed algorithm in this paper is shown in Fig. 1.

The network adopts a two-stage architecture and consists of two parts: a label generation network G_L and a fine-grained content reconstruction network G_R. Specifically, the label generation network is divided into two parallel structures, a rough content generator G_C and an edge generator G_E.

3.1 Label generation network

3.1.1 Rough content generator

The rough content generator G_C is the structure of the label generation network and is used for generative pixel filling of missing images. Let I_{gt} be the true value image, and the missing image I_{mask} is as follows:

$$I_{mask} = I_{gt} \odot (1 - M) \tag{1}$$

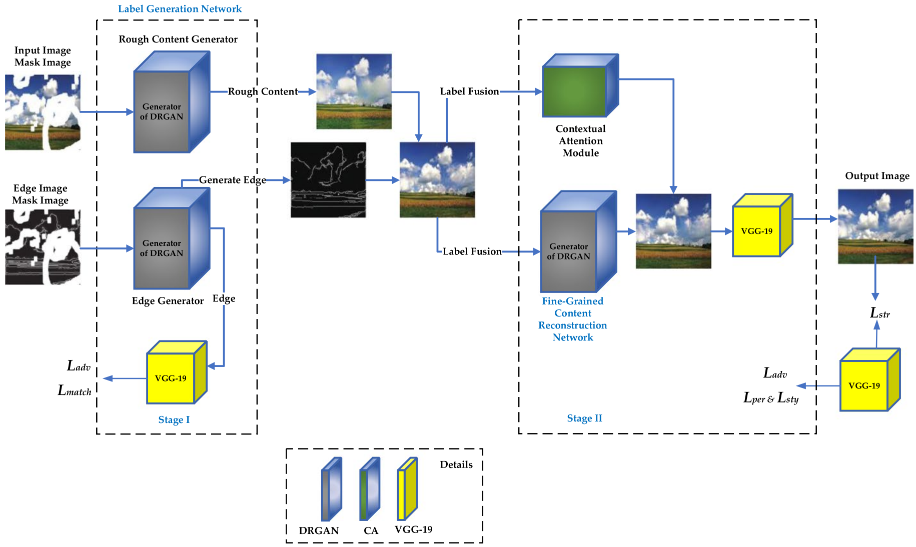


Fig. 1 The proposed network named as DGCA

Among them, M represents a binary mask (1 represents a missing area of the image, 0 represents an unmissed area), and $\odot$ represents a Hadamard product.

The specific structure of the rough content generator is similar to the generator part of the Deep Residual Generative Adversarial Network (DRGAN). In the down-sampling and the up-sampling, this paper sets the convolutional kernel to 3×3, and the step size is 2. At the same time, the design purpose of DRGAN is to perform super-resolution tasks, and in the repair task, due to the existence of large area missing, it is difficult for standard convolution to effectively extract the known information near some missing points. The known information is crucial for pixel reconstruction.

In order to make full use of known information, it is often necessary for the network to have a larger receptive field. Therefore, this paper improves on the basis of DRGAN and replaces the middle 8-layer standard convolution with a dilated convolution [11], as shown in Fig. 2. The dilated convolution kernel is 3×3, the dilated rate is 2, the step size is 1, the network uses *Leaky ReLU* as the activation function, and the generative pixel filling process of the missing image is expressed as follows:

$$I_{coarse} = G_C\left(I_{mask}, M\right) \tag{2}$$

Among them, $G_C(\cdot, \cdot)$ represents the calculation process of the rough content generator, and I_{coarse} represents the generated rough content. For the training of the generator, use Euclidean distance as the loss function:

After obtaining I_{coarse}, mark it as the content label part of the second stage network fusion label.

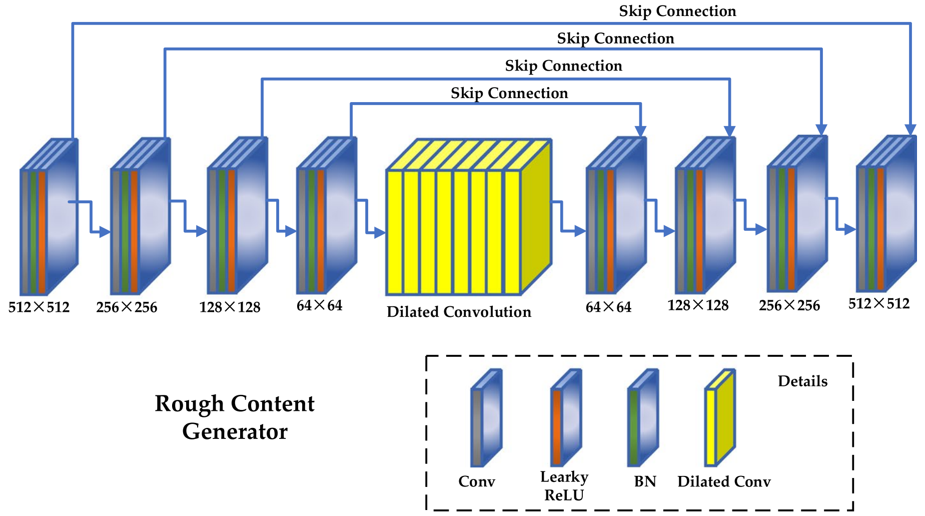


Fig. 2 The rough content generator

3.1.2 Edge generator

While performing generative pixel filling on the missing image, the edge generator G_E is used to complete the edge mapping of the missing image. The Canny edge detection is performed on the missing image I_{mask}, and the edge map to be completed is denoted as I^e_{mask}. In the edge detection process, in order to effectively eliminate the high-frequency noise of the image, a Gaussian filter with a kernel of 5×5 is performed, and the Canny threshold is set to 80 ~ 150. The edge generator adopts the DRGAN structure at the same time, the convolution kernel is 3×3, the middle eight convolution step length is 1, the remaining convolution step length is 2, and the activation function is *Leaky ReLU*. The completion process of missing edges is expressed as follows:

$$I^e_{generator} = G_E\left(I^e_{mask}, M\right) \tag{3}$$

Among them, $G_E(\cdot, \cdot)$ represents the operation process of the edge generator, and I^e_{gen} represents the complemented edge mapping.

The discriminator D_E of the edge generator uses the *VGG-19* architecture that has been pre-trained on the ImageNet dataset, and the activation function is *Leaky ReLU*. The input of the discriminator is the edge map I^e_{gt} obtained by the *Canny* detection of the true image and the edge map I^e_{gen} obtained by the complement of the missing image. The objective function of the discriminator is composed of three parts: adversarial loss, distance loss and matching loss. The adversarial loss is expressed as follows:

$$L^E_{adv} = E\left[\lg\left(1 - D_E\left(G_E\left(I^e_{mask}, M\right)\right)\right)\right] + E\left[\lg D_E\left(I^e_{gt}\right)\right] \tag{4}$$

In the adversarial training process, the discriminator judges the true value edge I^e_{gt} as true; for the generated edge I^e_{gen}, the discriminator judges the false. The generator and the discriminator each update the parameters to minimize the confrontation loss. This process can be seen as finding a Nash equilibrium solution in a zero-sum game. At the same time, a distance loss is expressed as follows:

$$L_{l_2}^{E} = \left\| I_{gt}^{e} - I_{gen}^{e} \right\|_2 \tag{5}$$

In addition, the matching loss is similar to the functional form proposed by Johnson et al. [15], which stabilizes the training by comparing the features of the activation function layer in the discriminator. The matching loss is expressed as follows:

$$L_{match}^{E} = E\left[\sum_i \frac{1}{N_i} \left\| D_E^{(i)}\left(I_{gt}^{e}\right) - D_E^{(i)}\left(I_{gen}^{e}\right) \right\|_1 \right] \tag{6}$$

Among them, i represents the number of discriminator convolutional layers, $D_E^{(i)}$ represents the i^{th} activation function layer of the discriminator (for this network, $D_E^{(i)}$ is the ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1 layers in VGG-19), and N_i represents the discriminator The number of elements in the i^{th} activation function layer.

The joint loss function of the edge generator is expressed as follows:

$$\min_{G_E} \max_{D_E} L^E = \lambda_{adv}^{E} L_{adv}^{E} + \lambda_{l_2}^{E} L_{l_2}^{E} + \lambda_{match}^{E} L_{match}^{E} \tag{7}$$

Among them, λ_{adv}^{E}, $\lambda_{l_2}^{E}$, and λ_{match}^{E} are regularization parameters. This article sets $\lambda_{adv}^{E} = 1$, $\lambda_{l_2}^{E} = 5$, and $\lambda_{match}^{E} = 10$ respectively.

After the complementary edge I_{gen}^{e} is obtained through the above process, the part corresponding to the missing area of the missing edge and the part that is not missing from the missing edge are spliced. The purpose of this operation is to make full use of the known information of the image to guide the subsequent network to complete the repair. The splicing process is expressed as follows:

$$I_{label}^{e} = I_{mask}^{e} \odot (1 - M) + I_{gen}^{e} \odot M \tag{8}$$

Among them, I_{label}^{e} represents a binary edge map obtained by stitching (1 represents an edge, 0 represents a background area), and this paper uses this mapping as the edge label part of the second stage network fusion label (Fig. 3).

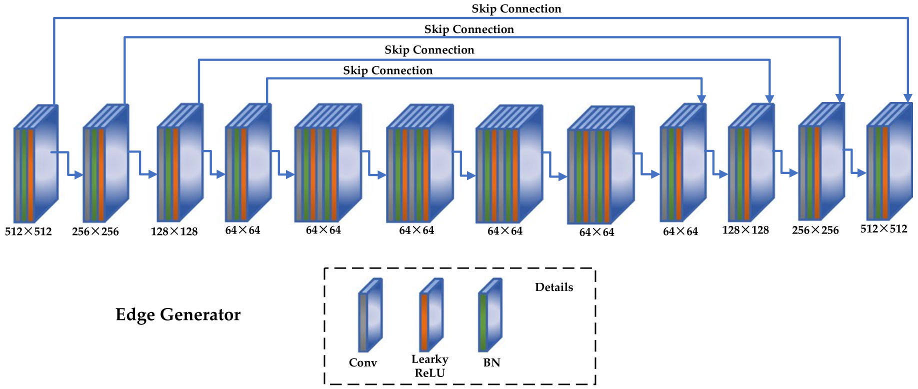


Fig. 3 The edge generator

3.2 Fine-grained content reconstruction network

The fine-grained content reconstruction network G_R uses the fused image of the rough content map I_{coarse} and the edge map I^e_{label} with the same resolution as the label. Since I_{coarse} is a color image and I^e_{label} is a binary image, the label fusion process is expressed as follows:

$$I_{label} = I_{coarse} \odot \left(1 - I^e_{label}\right) \tag{9}$$

Among them, I_{label} represents the condition label of the fine content reconstruction network.

The fine-grained content reconstruction network adopts the DRGAN architecture with the context attention module, and the specific operation process of the context attention module is shown in Fig. 4. First, for the feature map input in the preamble, the pixel blocks belonging to the generated part and the original part are extracted respectively. Then use the original part as a convolution filter to process the generated part, and calculate the similarity score between the generated part and the original part. Finally, the deconvolution operation is performed with the score as the weight, and a new feature map is reconstructed and output. The reconstruction process of the fine content reconstruction network is expressed as follows:

$$I_{gen} = G_R\left(I_{label}\right) \tag{10}$$

Among them, $G_R(\cdot)$ represents the calculation process of the generator of the fine content reconstruction network, and I_{gen} represents the global result of the generation.

The discriminator D_R of the fine-grained content reconstruction network uses the VGG-19 architecture that has been pre-trained on ImageNet, and takes the true value image I_{gt} and the global generated image I_{gen} as input. For the objective function, in this paper, the forms of adversarial loss, perceptual loss, style loss, and structure loss, the adversarial loss is expressed as follows:

$$L^R_{adv} = E\left[\lg\left(1 - D_R\left(G_R\left(I_{label}\right)\right)\right)\right] + E\left[\lg D_R\left(I_{gt}\right)\right] \tag{11}$$

The L^R_{adv} here is similar to L^E_{adv} in the edge generator, that is, it is judged to be true for I_{gt}, and it is judged to be false for I_{gen}. At the same time, the perceptual loss L^E_{per} is similar to the matching loss in the edge generator:

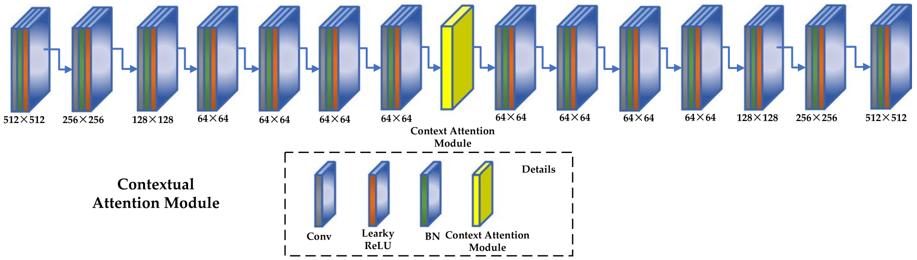


Fig. 4 The contextual attention module

$$L_{per}^{R} = E\left[\sum_{i} \frac{1}{N_i} \left\| D_R^{(i)}\left(I_{gt}\right) - D_R^{(i)}\left(I_{gen}\right) \right\|_1\right] \tag{12}$$

In addition, each activation function layer in the discriminator is also used to calculate the style loss. The style loss is the covariance between the features of the activation function layer. Given the feature map of refueling $H_k \times W_k \times C_k$, the style loss is expressed as follows:

$$L_{sty}^{R} = E\left[\left\| T_k\left(I_{gt}\right) - T_k\left(I_{gen}\right) \right\|_1\right] \tag{13}$$

Among them, T_k is the Gram Matrix of $C_k \times C_k$ constructed according to the activation function layer $D_R^{(k)}$. Korhonen et al. [17] have proved that similar structures can effectively eliminate artifacts in the generated results. At the same time, since both I_{gt} and I_{gen} are color images, this paper uses structure loss instead of distance loss, and the structure loss is expressed as follows:

$$L_{str}^{R} = 1 - \frac{\left[2\mu_1\mu_2 + (0.01\varepsilon)^2\right]\left[2\sigma_{12} + (0.03\varepsilon)^2\right]}{\left[\mu_1^2 + \mu_2^2 + (0.01\varepsilon)^2\right]\left[\sigma_1^2 + \sigma_2^2 + (0.03\varepsilon)^2\right]} \tag{14}$$

Among them, μ_1 represents the pixel average value of I_{gt}, μ_2 represents the pixel average value of I_{gen}, σ_{12} represents the covariance of I_{gt} and I_{gen}, σ_1^2 represents the pixel variance of I_{gt}, σ_2^2 represents the pixel variance of I_{gen}, and ε represents the dynamic range of the pixel (Fig. 5).

The joint loss function of the fine-grained content reconstruction network is expressed as follows:

$$\min_{G_R}\max_{D_R} L^R = \lambda_{adv}^{R} L_{adv}^{R} + \lambda_{per}^{R} L_{per}^{R} + \lambda_{sty}^{R} L_{sty}^{R} + \lambda_{str}^{R} L_{str}^{R} \tag{15}$$

Among them, λ_{adv}^{R}, λ_{per}^{R}, λ_{sty}^{R}, and λ_{str}^{R} are regularization parameters. This article sets $\lambda_a^R = \lambda_p^R = 0.5$, $\lambda_{str}^{R} = 1$, and $\lambda_{sty}^{R} = 200$ respectively.

After the global generated image I_{gen} is obtained, the edge pixel fusion operator processing I_{gen} is designed to obtain the final output result:

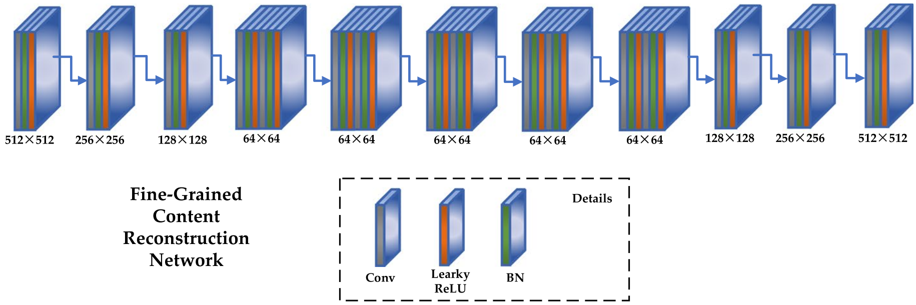


Fig. 5 The fine-grained content reconstruction network

$$\underset{reu}{\tilde{I}} = \frac{\exp\left(\frac{\varphi}{\Phi}\right) - 1}{\exp\left(\frac{\varphi}{\Phi}\right)} I_{mask} + \frac{1}{\exp\left(\frac{\varphi}{\Phi}\right)} I_{gen} \tag{16}$$

Among them, Φ is the width of the processing area of the edge pixel fusion operator (this article is set to 5% of the graphic size), and φ is the shortest distance from a point on the processing area to the mask area. This operator can effectively eliminate the artifacts of the generated edge and increase the consistency of the edge pixels between the repaired area and the original area.

4 Experimental analysis and results

4.1 Experimental settings

In order to verify the scientific and effectiveness of the algorithm in this paper, the network was trained, tested and verified on the datasets of Places2 [31], CelebA [6], Facades [24], and Oxford Building [7].

The Places2 dataset contains 365 independent scene categories, using 1,800,000 images in the high-resolution standard dataset and 6,200,000 images in the high-resolution challenge set to train the network. The CelebA dataset contains 202,599 face images of people, which are used for portrait inpainting. The Oxford Building dataset contains 5,062 structured building images with an image resolution of up to 1024 × 1024, which is used for high-resolution verification. The Facades dataset contains 606 architectural style images, all of which have been tilt-corrected and used for ablation research (Table 1).

The mask used in this paper to simulate the missing image comes from the irregular mask dataset provided by Wang et al. [29]. The dataset contains 12,000 irregular mask images, which are divided into six categories according to the ratio of the mask to the image area: [0.01, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4), (0.4,0.5], (0.5,0.6]. This paper uses the fusion image I^{e}_{label} of the rough content I_{coarse} generated by the complemented edge mapping I_{label} as the final reconstruction network label. The training of the network is divided into three parts: the edge generator, the rough content generation Trainer and fine content reconstruction network are trained separately. The network is built under the TensorFlow framework, with an initial learning rate of 2×10^{-4}, and Adam as the optimizer. The batch size is set to 8, the size of input image is based on datasets, and training is terminated. The condition is 20 iterations. In the stages of testing, verification and practical application, the algorithm in this paper can achieve end-to-end image inpainting.

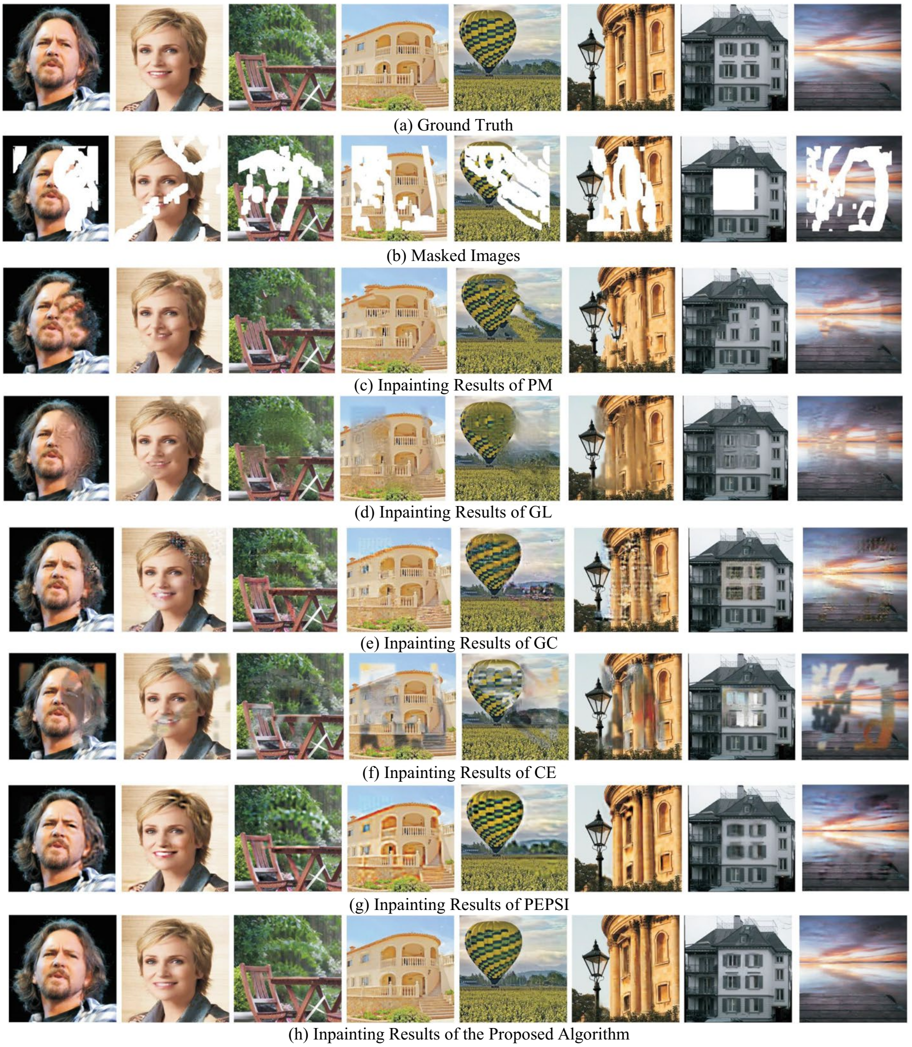
Table 1 Details of image inpainting datasets

Dataset	Type	Image Size	Scale	Website
Places2 [31]	Scene	256 × 256	10 M	http://places2.csail.mit.edu
CelebA [6]	Face	178 × 218	202 K	http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
Facades [24]	Scene	250 × 250 ~ 1024 × 1024	0.6 K	http://cmp.felk.cvut.cz/~tylecr1/facade/
Oxford Building [7]	Scene	1024 × 1024	4.9 K	http://oeb.griffith.ox.ac.uk/

4.2 Comparative experiments

The comparison algorithm is as follows: Patch-Match (PM) [1], Context Encoder Network (CE) [33], Globally and Locally Consistent Network (GL) [20], Contextual Attention Network (CA) [22], Gated Convolutional Network (GC) [27], Edge-Connect (EC), Parallel Decoding Network (PEPSI) [35].

The qualitative comparison experiment results of each algorithm are shown in Fig. 6. It can be seen from the Fig. 6 that Patch-Match based on random patch matching technology cannot accurately obtain image semantic information, nor can it generate new pixels that do not exist in the missing image. Although it can suppress checkerboard artifacts, it will result in inaccurate semantics. GL is not enough to restore color and structure. There are artifacts in the GC effect, and the ability to balance the texture is


(a) Ground Truth

(b) Masked Images

(c) Inpainting Results of PM

(d) Inpainting Results of GL

(e) Inpainting Results of GC

(f) Inpainting Results of CE

(g) Inpainting Results of PEPSI

(h) Inpainting Results of the Proposed Algorithm

Fig. 6 Comparisons of image inpainting results of different algorithms

not good. This is because these two algorithms only use a single-stage adversarial network and fail to adopt a layer-by-layer repair plan from edge to content. Edge-Connect can effectively obtain and restore a certain image edge structure, because the algorithm itself uses binary edge mapping as the network label. However, the label of Edge-Connect is only a binary map, which fails to reflect the color information part of the image, and the effect of color inpainting and reconstruction is not ideal. PEPSI can restore the global structure of the image and restore the color information of the image more accurately. However, the algorithm mainly uses context, rough content and texture as the basis for inpainting, and there are still some shortcomings in fine reconstruction.

The algorithm in this paper adopts a second stage structure that takes into account both the edge structure and global color information. It can generate new pixels that Patch-Match can't generate, suppress GL and GC artifacts, and repair more realistic colors than Edge-Connect repairs and more realistic colors than PEPSI. Sharp edges, so a better repair effect can be obtained visually.

Since there is still a lack of quantitative visual evaluation indicators for image inpainting tasks, in order to measure the inpainting results as accurately as possible, the following indicators are used: Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), Mean Absolute Error (MAE), and Mean Square Error (MSE). PSNR measures the quality of repaired images. SSIM compares the similarity between the true value map and the result map in terms of brightness, contrast, and structure. MAE measures the deviation of the corresponding position between the truth map and the result map. MSE measures the degree of difference between the truth map and the result map. The higher the value of PSNR, the higher the quality of the repair result map, the higher the value of SSIM, the lower the values of MAE and MSE, the closer the repair result map is to the true value map.

The index evaluation results of the statistical comparison algorithm for 10,000 random images on the Places2 dataset are shown in Tables 2 and 3.

In the PSNR indicator in Table 1, "Global" refers to the comparison between the generated global map and the truth map, and "Local" refers to the comparison between the final output result map and the truth map after the global map and the missing map are spliced. It can be seen from Tables 1 and 2 that the algorithm in this paper is only slightly lower than PEPSI in the case of repairing irregular masks, and generally maintains a relatively high level. For MAE and MSE, the algorithm in this paper also has strong advantages.

Table 2 Evaluation index results of different algorithms (rectangular mask)

Methods	PSNR		SSIM	MAE	MSE
	Global	Local			
PM	-	-	-	16.1	3.9
CE	17.7	23.7	87.2	-	-
GL	19.4	25.0	89.6	9.3	2.2
CA	19.0	24.9	89.8	8.6	2.1
GC	18.7	24.7	89.5	-	-
PEPSI	19.5	25.6	90.1	8.6	**2.0**
Proposed	**22.9**	**27.2**	**91.0**	**8.3**	2.4

The bold font is the best result in every column

Table 3 Evaluation index results of different algorithms (unregular mask)

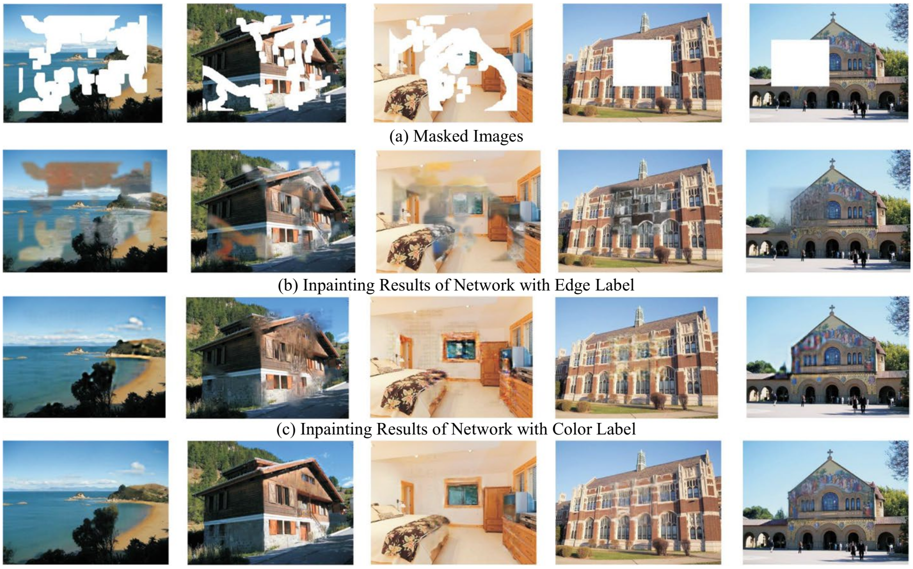
Methods	PSNR		SSIM	MAE	MSE
	Global	Local			
PM	-	-	-	11.3	2.4
CE	9.7	16.3	79.4	-	-
GL	15.1	21.5	84.3	21.6	7.1
CA	12.4	18.9	79.8	17.2	4.7
GC	21.2	26.4	91.0	-	-
PEPSI	22.0	28.6	**92.9**	9.1	**1.6**
Proposed	**24.5**	**30.5**	91.8	**8.3**	2.5

The bold font is the best result in every column

4.3 Ablation experiments

This section analyzes the influence of each part of the fusion tag on the final repair result from the color information and edge structure. The specific results are shown in Fig. 7.

First, assume that color information is very important for image inpainting tasks. For this reason, the rough content fill map is used as the color condition to be incorporated into the label of the fine content reconstruction network. In order to verify this hypothesis, the rough content generator is deleted, and only the edge structure map repaired by the edge generator is used as the label of the fine content reconstruction network. The repair result is shown in Fig. 7(b). It can be seen from Fig. 7 that if the color information is missing from the label, the algorithm will be difficult to effectively reconstruct the new color, that is, to demonstrate the design significance of the rough content generator. Then turn your attention to the edge structure information. It is assumed that the edge mapping can effectively


(a) Masked Images

(b) Inpainting Results of Network with Edge Label

(c) Inpainting Results of Network with Color Label

(d) Inpainting Results of Proposed Algorithm

Fig. 7 Comparisons of image inpainting results of color label and edge label

Table 4 Evaluation index comparison of ablation studies on color content label and edge structure label

Label	CelebA				Facades			
	PSNR	SSIM	MAE	MSE	PSNR	SSIM	MAE	MSE
Missing Color Label	17.76	79.76	16.8	4.9	21.54	81.26	16.6	4.8
Missing Edge Label	19.09	81.51	15.6	4.5	23.64	82.18	15.6	4.4
Full Label Model	**28.96**	**90.32**	**8.8**	**2.6**	**29.41**	**90.67**	**8.7**	**2.5**

The bold font is the best result in every column

Table 5 Evaluation index comparison of ablation studies on different loss function regularization parameters

Parameter	PSNR	SSIM	MAE	MSE
$\lambda_{match}^{E} = 1$	28.62	87.96	17.6	7.5
$\lambda_{match}^{E} = 5$	26.61	84.87	20.2	7.8
$\lambda_{match}^{E} = 20$	25.55	83.34	20.3	7.4
$\lambda_{sty}^{R} = 50$	25.12	83.35	20.1	7.6
$\lambda_{sty}^{R} = 150$	25.83	84.26	21.2	7.1
$\lambda_{sty}^{R} = 250$	23.79	82.19	20.4	8.0
Proposed	**29.27**	**90.55**	**9.2**	**3.5**

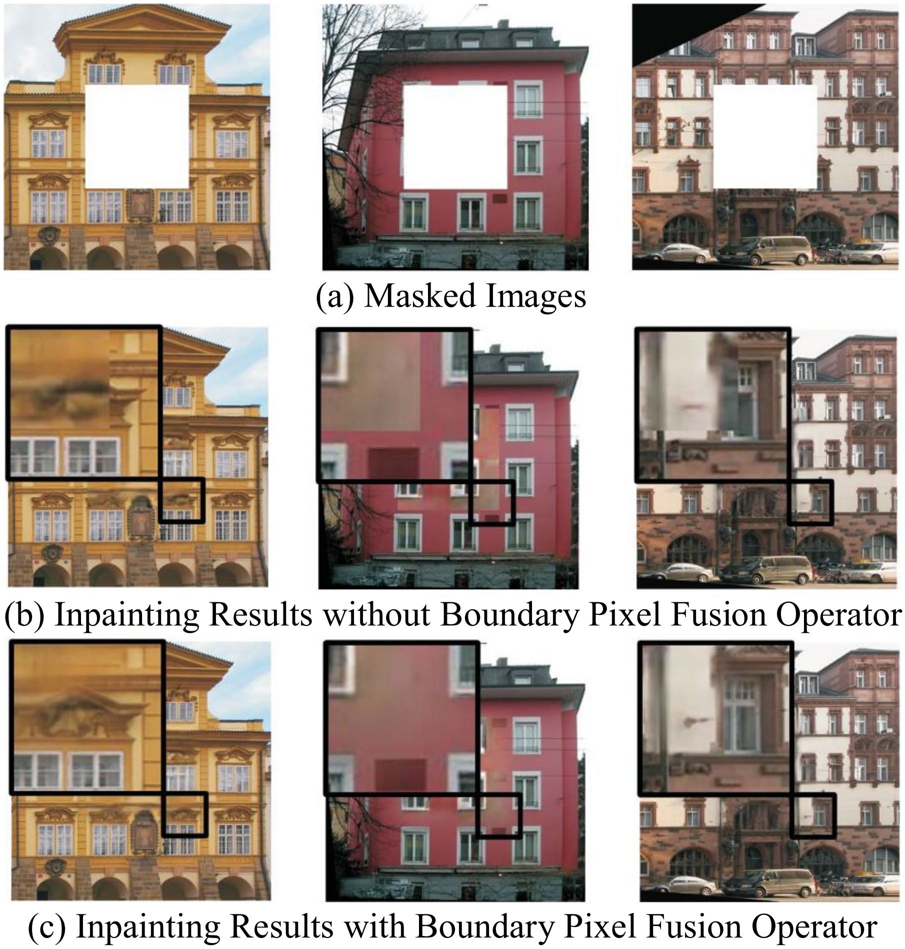
The bold font is the best result in every column

represent the objective structure of the image. In order to verify this hypothesis, the edge generator is deleted, and only the truth map is used as a label to train the fine content reconstruction network. The visualized result after repair is shown in Fig. 7(c). It can be seen from Fig. 7 that if the edge information is not used, the fineness of the content of the restored image is significantly reduced.

At the same time, 500 images were selected on the CelebA and Facades datasets, and the evaluation results of the algorithms trained with different tags on the PSNR and SSIM were counted. The results are shown in Table 4. It can be seen from Table 4 that, regardless of image quality or structural similarity, the algorithm trained by fusion label is more effective than the algorithm trained by single edge label and color label.

However, how to accurately obtain color and edge information has become the key to the research problem. This paper found that if the edge recovery is too little, the result will be a checkerboard artifact; if the edge recovery is too much, the result will be high-frequency noise. The same applies to the use of color labels: if the rough content is too smooth, the resulting color will not be vivid enough; if the rough content is too fine, the result will produce high-frequency noise. Therefore, adjust the regularization parameters of the loss function and select the parameters that can obtain the best repair effect. Randomly select 1000 images on the Places2 dataset for statistics. Table 5 shows the effect of regularization parameters on the results of choosing different edge generators for matching loss and fine content reconstruction network style loss. It can be seen from Table 5 that when $\lambda_{match}^{E} = 10$ and $\lambda_{sty}^{R} = 200$ are set, the algorithm in this paper can obtain the best effect.

In the network output link, an edge pixel fusion operator is designed to perform pixel processing on the missing edges of the repaired image, improve the edge pixel consistency of the pixel image, and eliminate artifacts. Figure 8 is the processing effect of the edge pixel fusion operator in this paper. It can be seen by partially zooming in the image that the operator in this paper can effectively improve the pixel consistency of the edge repair.



(a) Masked Images

(b) Inpainting Results without Boundary Pixel Fusion Operator

(c) Inpainting Results with Boundary Pixel Fusion Operator

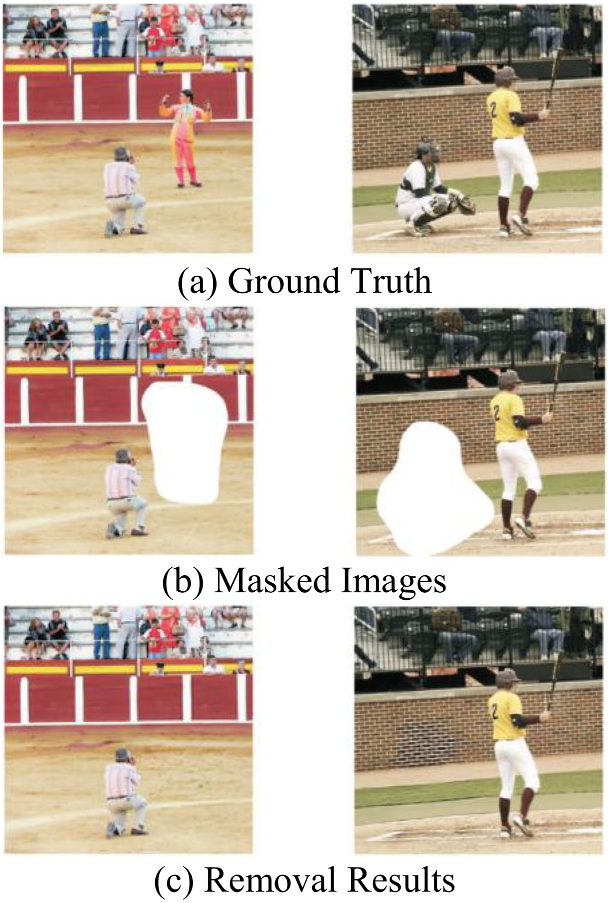
Fig. 8 Comparison of inpainting results with and without boundary pixel fusion operator

4.4 Interactive image editing

The proposed algorithm in this paper can also be used for interactive image editing tasks while performing image inpainting tasks. Masks are set for some characteristic targets in the image, and these targets are removed by pixel generation, as shown in Fig. 9. It can be seen from Fig. 9 that this method can erase some characters in the Fig. 9 and generate new pixel content with reasonable semantics based on context information.

In addition, the algorithm in this paper changes the generated results by manually marking the information on the mask area, as shown in Fig. 10. It can be seen from Fig. 10 that the letters in the first row of warning signs and the window frame style in the second row are modified. These two interactive experiments for image editing tasks can also reflect the superiority and reliability of this algorithm from the side.

Fig. 9 Illustration of the proposed algorithm on target removal task


(a) Ground Truth

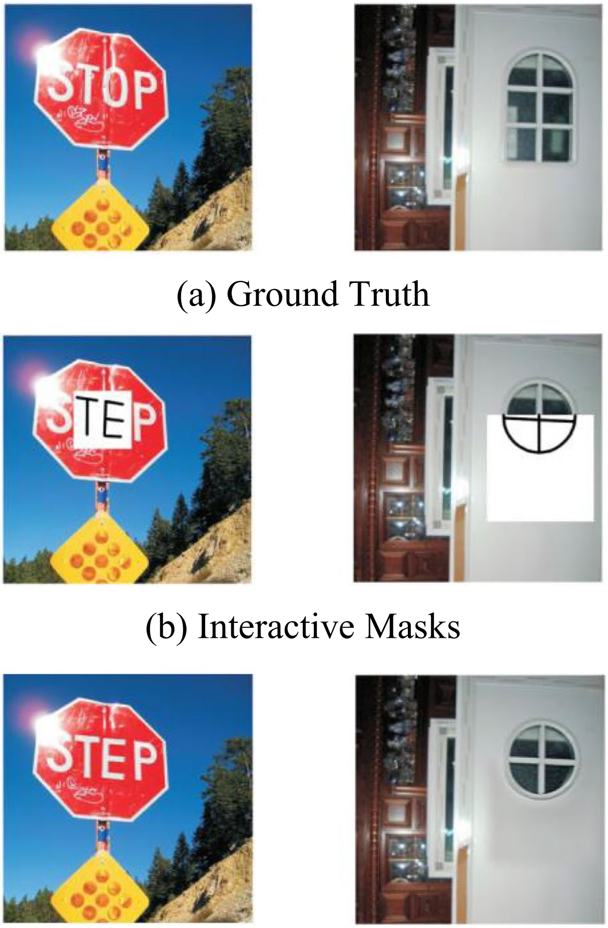
(b) Masked Images

(c) Removal Results

5 Conclusion

This paper proposes a generative high-resolution image inpainting algorithm based on parallel confrontation and multi-condition fusion. The image repairing network has divided into two parts, and the missing area has gradually repaired from coarse level to fine level, from edge information to content information. At the same time, the architecture, label and loss function of Generative Adversarial Network have improved, and edge pixel fusion operator has been proposed. The experimental results on four standard datasets can show that compared with the popular repairing methods in recent three

Fig. 10 Illustration of the proposed algorithm on image editing



(a) Ground Truth

(b) Interactive Masks

(c) Interactive Results

years, the proposed model in the paper can obtain more effective texture detail information, and the repairing effect for complex structures and textures is more realistic and more accurate. In addition, the proposed algorithm not only performs better in image semantic tasks, but also achieves satisfactory results in tasks such as object removal and image editing mission. However, limited by the existing technologies and hardware conditions, when faced with an image with an overly complex scene or an overly stylized image, there may still be artifacts or inaccurate semantics. This is the direction of continued research in the future.

Acknowledgements This work is supported by A Project Supported by Scientific Research Fund of Hunan Provincial Education Department under Grant 22A0701, 2022 Institute Scientific Research Project of Hunan University of Information Technology under Grant XXY022ZD01, College Students' Innovative Entrepreneurial Training Plan Program of Hunan University of Information Technology under Grant X202213836002, China University Innovation Funding—Beslin Smart Education Project under Grant 2022BL055.

Data availability All data generated or analysed during this study are included in this published article and its supplementary information files. The authors would like to thank Places2, CelebA, Façade and Oxford Building datasets, which allowed us to train and evaluate the proposed model.

Declarations

Conflicts 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. Barnes C, Shechtman E, Finkelstein A, Goldman D (2009) PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans Graph 28(3):24
2. 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
3. Chen T, Zhang X, Hamann B, Wang D, Zhang H (2022) A multi-level feature integration network for image inpainting. Multimed Tools Appl 81:38781–38802
4. Ding D, Ram S, Rodriguez J (2019) Image inpainting using nonlocal texture matching and nonlinear filtering. IEEE Trans Image Process 28(4):1705–1709
5. Doersch C, Singh S, Gupta A, Sivic J, Efros A (2012) What makes Paris look like Paris? ACM Trans Graphics 31(4):101
6. Fang Y, Li Y, Tu X, Tan T, Wang X (2020) Face completion with hybrid dilated convolution. Signal Process Image Commun 80:115664
7. Gao S, Cheng M, Zhao K, Zhang X, Yang M, Torr P (2019) Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662
8. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, pp 2672–2680
9. Guo Q, Gao S, Zhang X, Yin Y, Zhang C (2018) Patch-based image inpainting via two-stage low rank approximation. IEEE Trans Visual Comput Graphics 24(6):2023–2026
10. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141
11. Iizuka S, Simo-Serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graphics 36(4):107
12. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning, pp 448–456
13. Jiang B, Huang W, Yang C, Huang Y (2022) Image inpainting based on cross-hierarchy global and local aware network. Multimed Tools Appl.https://doi.org/10.1007/s11042-022-14245-5
14. Jiang Y, Yang F, Bian Z, Lu C, Xia S (2022) Mask removal: Face inpainting via attributes. Multimed Tools Appl 81:29785–29797
15. Johnson J, Alahi A, Li F (2016) Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of European Conference on Computer Vision, pp 694–711
16. Kingm D, Ba J (2014) Adam: A method for stochastic optimization. In Proceedings of 4th International Conference on Learning Representation, pp 58–64
17. Korhonen J, Junyong Y (2012) Peak signal-to-noise ratio. In Proceedings of International Workshop on Quality of Multimedia Experience Electronics Letters, pp 37–38
18. Liu G, Reda F, Shih K, Wang T, Tao A, Catanzaro B (2018) Image inpainting for irregular holes using partial convolutions, In: Proceedings of European Conference Computer Vision, pp 89–105
19. Liu W, Xu D, Tsang I, Zhang W (2019) Metric learning for multi-output tasks. IEEE Trans Pattern Anal Mach Intell 41(2):408–422
20. Pan J, Sun D, Zhang J, Tang J, Yang J, Tai Y, Yang M (2022) Dual convolutional neural networks for low-level vision. Int J Comput Vision 130:1440–1458
21. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros A (2016) Context encoders: feature learning by inpainting. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2536–2544
22. Pen H, Wang Q, Wang Z (2021) Boundary precedence image inpainting method based on self-organizing maps. Knowl-Based Syst 216:106722

23. Qin Z, Zeng Q, Zong Y, Xu F (2021) Image inpainting based on deep learning: A review. Displays 69:102028
24. Quan W, Zhang R, Zhang Y, Li Z, Wang J, Yan D (2022) Image inpainting with local and global refinement. IEEE Trans Image Process 31:2405–2420
25. Song Y, Yang C, Lin Z, Liu X, Huang Q, Li H, Kuo C (2018) Contextual-based image inpainting: infer, match and translate. In: Proceedings of European Conference on Computer Vision, pp 3–18
26. Wan R, Shi B, Li H, Duan L, Kot A (2021) Face image reflection removal. Int J Comput Vision 129:385–399
27. Wang Y, Tao X, Qi X, Shen X, Jia J (2018) Image inpainting via generative multi-column convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp 329–338
28. Wang N, Li J, Zhang L, Du B (2019) MUSICAL: multi-scale image contextual attention networks. In: Proceedings of International Joint Conference on Artificial Intelligence, pp 3748–3754
29. Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
30. Wang N, Zhang Y, Zhang L (2021) Dynamic selection network for image inpainting. IEEE Trans Image Process 30:1784–1798
31. Xie C, Liu S, Li C, Cheng M, Zuo W, Liu X, Wen S, Ding E (2019) Image inpainting with learnable bidirectional attention maps. In: Proceedings of IEEE International Conference on Computer Vision, pp 8858–8867
32. Xu R, Guo M, Wang J, Li X, Zhou B, Loy C (2021) Texture memory-augmented deep patch-based image inpainting. IEEE Trans Image Process 30:9112–9124
33. Zeng Y, Gong Y, Zhang J (2021) Feature learning and patch matching for diverse image inpainting. Pattern Recogn 119:108036
34. Zhang L, Chang M, Chen R (2023) Image inpainting based on sparse representation using self-similar joint sparse coding. Multimed Tools Appl.https://doi.org/10.1007/s11042-023-14337-w
35. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of European Conference on Computer Vision, pp 294–310
36. Zhang Y, Ding F, Kwong S, Zhu G (2021) Feature pyramid network for diffusion-based image inpainting detection. Inf Sci 572:29–42
37. Zhang X, Wang X, Shi C, Yan Z, Li X, Kong B, Lyu S, Zhu B, Lv J, Yin Y, Song Q, Wu X, Mumtaz I (2022) DE-GAN: domain embedded GAN for high quality face image inpainting. Pattern Recogn 124:108415
38. Zheng C, Cham T, Cai J (2019) Pluralistic image completion. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1438–1447
39. Zheng C, Cham T, Cai J (2021) Pluralistic free-form image completion. Int J Comput Vision 129:2786–2805
40. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 50(6):1452–1464
41. Zhu M, He D, Li X, Li C, Li F, Liu X, Ding E, Zhang Z (2021) Image inpainting by end-to-end cascaded refinement with mask awareness. IEEE Trans Image Process 30:4855–4866
42. Zhuo L, Tan S, Li B, Huang J (2022) ISP-GAN: inception sub-pixel deconvolution-based lightweight GANs for colorization. Multimed Tools Appl 81:24977–24994

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yuantao Chen received the B.S. degree in Computer Science and Technology from Jianghan Petroleum Institute. He received the M.S. degree in Geodetection and Information Technology from Yangtze University. He received the Ph.D. degree in Control Science and Engineering from Nanjing University of Science and Technology, Nanjing, China, in 2014. He is an associate professor at School of Computer Science and Engineering, Hunan University of Information Technology. His research interests include computer vision, deep learning, etc.

Runlong Xia received the B.S. degree in Electronic Commerce from Hunan Normal University in 2010. He is a researcher with Mountain Yuelu Breeding Innovation Center Limited and Hunan Provincial Science and Technology Affairs Center. His research interests include news communication and public opinion analysis, etc.

Kai Yang received the Master degree in Mechanical Engineering from Jilin University in 2014. He is an engineer with Hunan ZOOMLION Intelligent Technology Corporation Limited. His research interests include mechanical engineering, intelligent control technology, etc.

Ke Zou received the B.S. degree in Computer Science and Technology from Northeastern University in 2005. He is an director and researcher at Informatized Office, Hunan WUJO High-Tech Material Corporation Limited. His research interests include big data processing, information processing, etc.