Multimedia Systems (2024) 30:101
https://doi.org/10.1007/s00530-024-01290-3

REGULAR PAPER

Image inpainting method based on AU-GAN

Chuangchuang Dong[1] · **Huaming Liu**[1] · **Xiuyou Wang**[1] · **Xuehui Bi**[1]

Received: 18 September 2023 / Accepted: 8 February 2024 / Published online: 29 March 2024

Abstract
Image inpainting refers to the process of filling in missing regions or removing objects, and has broad application prospects. The rapid development of deep learning has led to new technological breakthroughs in image repair technology, continuously improving the quality of image inpainting. However, when we inpaint large missing regions, the texture and structural features of the image cannot be comprehensively utilized. This leads to blurry images. To solve this problem, we propose an improved dual-stream U-Net algorithm that adds an attention mechanism to the two U-Net networks known as a dual AU-Net network to improve the texture details of the image. In addition, the location code (LC) of damaged regions is added to the network to guide network repair and accelerate the network convergence speed. Least squares GAN (LSGAN) loss is added to the generator's adversarial network to capture more content details and enhance training stability. The PSNR is 33.93 and the SSIM is 0.931 in the CelebA and Paris datasets. This method has been proven effective when compared to other methods.

Keywords AU-Net · Attention mechanism · Position coding · Image repair

1 Introduction

Image inpainting [1] is a technique that restores damaged pixel features within the image. This technique has numerous practical applications, including the restoration of cultural relics, calligraphy, paintings, and the removal of undesired objects or interference within images [2]. In the past two decades, numerous image inpainting methods have been proposed, which use prior data and the image data itself for inpainting. Ideally, image repair models should have the following three attributes: (1) Aesthetic and visual consistency in the image's structure and texture, (2) Cohesion in the image content as a whole, (3) Efficient and reliable network training.

Numerous scholars have proposed several image inpainting methods. Initially, scholars proposed traditional methods such as partial differential equations (PDE) [1], texture synthesis [3, 4], and sparse representation [5] for image restoration. However, the PDE method is only effective for inpainting small regions, while texture synthesis is suitable for restoring larger regions but can be limited by a lack of sample resources. The sparse representation method enables effective noise reduction but does not maintain image structure continuity. Additionally, these methods are generally unable to satisfy the third attribute and prove to be time-consuming.

Pathak et al. [6] were the first to propose the use of context encoders network for image inpainting, achieving remarkable results that sparked the interest of many scholars. With the growing success of deep learning in image processing, several inpainting methods have been proposed using neural networks. These inpainting techniques mainly comprise generative adversarial networks (GAN) [7–10], autoencoder networks [6, 11, 12], and transformer networks [13–15], with the capacity to achieve semantic inpainting results.

Generative adversarial networks (GAN) generate images through a continuous adversarial game between

Communicated by B. Bao.

✉ Huaming Liu
200806004@fynu.edu.cn

Chuangchuang Dong
320707643@qq.com

Xiuyou Wang
wangxiuyou@163.com

Xuehui Bi
bixuehui888@163.com

[1] School of Computer and Information Engineering, Fuyang Normal University, 100 Qinghe West Road, Fuyang 236037, Anhui Province, China

the generator and discriminator. Autoencoder networks extract image features [16, 17] for image inpainting, while transformer networks [15, 18] use attention mechanisms to capture global context information necessary for visual rationality repair. While these techniques successfully realize the second and third attributes of image repair, the first attribute may not be implemented as effectively.

While these three image inpainting modes can achieve good results, they have certain limitations. The GAN loss function is often faced with the problem of not effectively improving network training speed. Autoencoder networks [19] may not accurately restore texture details in images. The computationally intensive nature of transformer networks has been recognized as a drawback [20].

Is it possible to design a network that can solve the aforementioned problems simultaneously? To address this problem, we propose an inpainting network, called autoencoder U-shaped generative adversarial networks (AU-GAN), which utilizes GAN as the main architecture and incorporates a dual-stream AU-Net into the generator to generate high-quality images. We introduce attention mechanism, position coding, and LSGAN (least squares GAN) loss into the network to monitor the texture details of the image, and a parallel three-branch discriminator is designed to reduce the computational time. This method enhances the visual consistency and coherence of the images.

Texture and structure information are essential to the image. Focusing solely on one aspect can result in unsatisfactory outcomes [21]. A dual-stream coupled U-Net network can be utilized to extract texture and structure features of an image in two distinct encoding stages. The network possesses location coding and attention modules for the missing image regions. In the decoding stage, the texture feature supplements the structural feature reconstruction, and the structural feature supplements the image texture feature reconstruction. The two features are integrated to achieve a cohesive blend of image texture and structure information [22, 23]. The attention module assists the U-Net encoder in effectively locating the image's texture information [24] and highlighting the texture details [25]. Position coding of missing regions in U-Net allows the network to prioritize broken regions, leading to more efficient convergence. The discriminator applies a three-branch structure to recognize the repaired, edge, and grayscale images. In conclusion, our approach combines different loss functions, including LSGAN, perception, reconstruction, and confrontation, to monitor the repair outcome. This approach provides a global evaluation of image repair quality, improves training stability, and ensures reasonable visual consistency and coherence of the inpainted images.

To summarize, this article's primary contributions are:

- The AU-GAN method proposed by us combines the dual-stream AU-Net into the generator to improve the quality of the image. It utilizes skip connections and the texture information of the image to aid in the reconstruction of the image's structural information. Additionally, it utilizes skip connections and the structural information of the image to aid in the reconstruction of the image's texture information. This approach allows the texture and structural information of the image to be combined and used to guide the image repair process.
- To monitor the texture details of the image, we incorporated attention mechanism, position coding, and LSGAN loss into the network. Through the attention mechanism, the AU-Net network can better extract image texture, while the position coding module provides the network with directional information of the mask. LSGAN loss, based on the least squares method, penalizes samples based on their distance from the decision boundary, hence supervising the network repair results.
- To accelerate the convergence of the network, reduce time costs, and improve the visual consistency of the image, we adopt a discriminator with a parallel three-branch design. Among them, one branch discriminates the grayscale image of the repaired image, the other branch discriminates the edge image of the repaired image, and the third branch discriminates the repaired image. Finally, the outputs of the three branches will be merged for discrimination in the channel dimension.

This paper is organized as follows: Sect. 1 provides background information on inpainting and trends in image inpainting methods. Sect. 2 reviews prior work on inpainting and introduces the primary image inpainting method. Sect. 3 introduces the proposed method, including the network architecture, generator, location code, discriminator, and loss function. Sect. 4 describes the experimental analysis conducted. Finally, Sect. 5 concludes the paper and presents future work.

2 Related work

2.1 Image inpainting based on traditional methods

Traditional image inpainting can be divided into three categories: partial differential equation (PDE) method [26], (2) texture synthesis method [27, 28], and (3) sparse representation method [29].

Partial differential equations The PDE methodology [1, 30] utilizes boundary information to regulate the diffusion direction and rate, diffusing gradually from the missing boundary toward its interior. Bertalmio et al. were the first

to apply this method in image inpainting [1] in 2000. The central concept behind the PDE technology is to distribute pixel information to the missing area following the isophotes direction while utilizing the propagation mechanism to restore the image. It has generated remarkable outcomes in small regions.

Texture synthesis Drori et al. were the first to pioneer the use of texture patches as a whole to eliminate the randomness of pixel filling [31] in 2003. In 2004, Criminisi et al. proposed an exemplar [3] based on priority order for image inpainting. This technique computes the priority function using the product of confidence and data items to determine the priority order of the missing boundaries, finds the best matching patch through global search, and fills it directly. These methods can generate filling errors that influence the subsequent fillings. Additionally, the priority calculation can be illogical, resulting in incorrect filling order and unsatisfactory repair outcomes. Furthermore, in cases where sample resources are unavailable, repair is often impossible. Barnes et al. introduced the PatchMatch technique [32] for image inpainting, which utilizes a random search to improve repair efficiency, but it has some limitations such as sensitivity to sample size. Another method relies on external databases [4] and proves effective in scenarios where sample images with adequate visual similarity can be obtained. However, if the restored image is missing from the sample database, it results in an incorrect filling, leading to an inadequate final result. Although the sample resources may exist, they can still be disrupted in complex situations such as scale, rotation, and illumination, and may lead to unsatisfactory repair results [33].

Sparse representation The main concept behind sparse representation inpainting [34] is that filled regions and unmissing regions have similar and sparse characteristics. To obtain sparse coefficients, image information can be sparsely represented, and reconstruction algorithms can then be employed to recover the image signal, leading to complete image restoration [35]. This method averts the dissimilarity of a single texture block and inserts multiple texture blocks to fill in similar regions. Nonetheless, missing samples make it impossible to fix them. Sparse representation methods cannot incorporate semantic information; instead, they fill in the missing regions by applying known information in the image. If sample resources are unavailable for the repair process, filling in missing regions will become problematic. Hays et al. resorted to using similar samples from an image library [4] to fill in the missing regions. If no corresponding image blocks can be found in the library, it generates repair errors, and a lengthy search process is initiated.

2.2 Image inpainting based on deep learning

Deep Learning method Recently, image inpainting has been undergoing developments with the application of deep learning technology. Deep learning has a higher capacity for both feature learning and expression compared to conventional algorithms. Additionally, it can capture more features, refresh various task indicators on a regular basis concerning computer vision, and has made significant strides in image inpainting. We can categorize these methods into three global categories: image inpainting technology based on GAN [9, 17, 36, 37], autoencoder [38–40], and transformer network [15, 18, 41].

The primary concept of image inpainting technology based on GAN involves continuous image generation by the generator while the discriminator ensures the validity of the image inpainting results. Subsequently, both the generator and discriminator engage in an adversarial game with continuous optimization. [42] incorporated photo style into the GAN network and introduced a new normalization and regularization method for the GAN generator to tackle issues such as speckles, thereby improving the overall image quality. However, this method tends to create blurry images. [43] proposed a network structure based on wavelet decomposition. The method decomposes the image into various bands with apparent missing areas, and then uses the discrete wavelet transform to retain the spatial information. Lastly, a new normalization method is designed to fuse multi-frequency features to improve image repair, but texture details in the restored image are not perfect. To enhance the detail of face images during image repair, Zhou et al. proposed to use a GAN network [44] with dual spatial attention modules and multiple discriminators. Despite achieving commendable results, the method is not suitable for repair of significantly sized defects.

The image inpainting network structure based on encoding and decoding is widely employed [45]. A multi-scale network [46] with an encoding and decoding structure enhances the network's general comprehension of the image by extracting edges and lines of the image as prior information, to improve the quality of image inpainting. Images with significant gaps, however, show mediocre repair results because the line and edge effects of the image are roughly the same. Thus, [47] employed the semantic information of the images and texture as prior information to guide in the repair process. They also used a semantic intelligent communication module to refine the image texture to prevent texture confusion in the image. Nonetheless, the approach may cause distortion while fixing large regions. Positing that repairing an image based on local texture and semantic information can often cause texture blur, Zhou et al. proposed a method that extracts key feature points [48] according to the homography of the

source image, clusters them, and then obtains the repaired image through the color space conversion module and feature fusion module. However, this method does not utilize the texture and structure information of the image, resulting in artifacts. To address this, a hierarchical network [49] was proposed that generates multiple coarse results with varying structures in the coarse stage. In the fine stage, a structure attention module refines each coarse result via texture enhancement and synthesizes the discrete structure features to combine the image texture and output the repaired image. The result of this method may contain artifacts. Guo et al. proposed a dual-stream network structure [50] for image inpainting. The structure and texture features of images are extracted and combined through two U-Net networks that interact, thereby effectively utilizing image information for image inpainting and achieving superior results. However, the method cannot integrate the image's texture and structural information more effectively for inpainting large missing regions, often resulting in a blurred inpainted image. What sets it apart from our proposed inpainting method is: (1) We incorporate attention mechanisms, which can enhance the visual coherence of image restoration. (2) We introduce position encoding to enable the network to perform efficient and precise repairs on damaged areas.

The transformer architecture has garnered significant attention since its introduction. Its core strength lies in the attention idea, which enables optimal exploitation of context information. Zheng et al. creatively approached the task of image repair as a directionless prediction task [51] from sequence to sequence, leveraging the attention mechanism of transformers. The authors proposed a novel attention perception layer that effectively exploits high-frequency features, while balancing the attention between missing and visible areas. On a similar note, Yu et al. observed that the convolutional neural network's modeling effect [52] on remote context information is subpar, leading to distortion in the repair results. To overcome this problem, the authors integrated an attention module into the network and added the perception mechanism and content awareness layer to the two U-Net network structures. Consequently, the authors successfully improved the inpainting network's modeling ability. To conduct effective image inpainting, a multistage repair network [53] was developed, comprising of an autoencoder and a single-scale network. The network was further equipped with an attention mechanism at each stage, which aids in optimizing the feature flows from one network to the next. The integration of a cross-stage information exchange facilitated minimal data loss, leading to better image recovery. However, the authors did not adequately integrate the attention mechanism with the texture and structure information of the image. This limitation warrants further improvement to enhance the quality of the repaired image.

Through an analysis of relevant studies, it can be found that the GAN network can use the generator and discriminator to continuously improve the quality of the image. Attention mechanism can increase the network's focus on missing areas, while position coding can provide directional information for the mask to the network. Therefore, it is necessary to combine the strengths of GAN network, attention mechanism and position coding, and rationally add them to the network, further improving the quality of image inpainting.

3 Method

3.1 Network architecture

This paper proposes an image inpainting method that integrates autoencoder U-shaped generative adversarial networks (AU-GAN) into the GAN framework. The proposed generator configuration consists of a two-stream AU-Net network structure, which utilizes the contextual transformer for spatially disoriented gated fusion (CTSDG) module [50] for bidirectional gated feature fusion (Bi-GFF) and the contextual feature aggregation (CFA) module. To improve network training, the discriminator adopts a tripartite structure for discriminating the grayscale map of the inpainted result, the inpainted image, and the edge map of the inpainted image, sharing the weights of the discriminator of the three branches. This paper incorporates position encoding for the mask in the first convolutional layer of the generator. The structure consists of seven consecutive convolutional modules, where the encoding stage is followed by an attention module to capture global contextual information. The proposed network diagram is presented in Fig. 1.

Following the attention module is the decoding module. Our approach utilizes a layer-by-layer incremental and jump-connection approach. This approach jump-connects not only the levels in the decoder and the encoder corresponding to the scale but also the dual-stream AU-Net network employed to extract the structural and texture information of the image separately. The extracted structural features guide the texture reconstruction in the decoding stage, whereas the extracted texture features guide the reconstruction of the structural features of the image in the other AU-Net. Finally, the generated images are produced using Bi-GFF and CFA modules.

In AU-Net's coding layer, the first layer does not use batch normalization, whereas the remaining six layers adopt batch normalization and average pooling to lower the network's complexity and parameter count. The activation function for each layer is LeakyReLU with $\alpha = 0.2$. The added position encoding in the first convolutional layer allows the mask position information to be represented using sine and cosine functions with distinct frequencies. Four kernels layered masks are subsequently utilized to capture the position

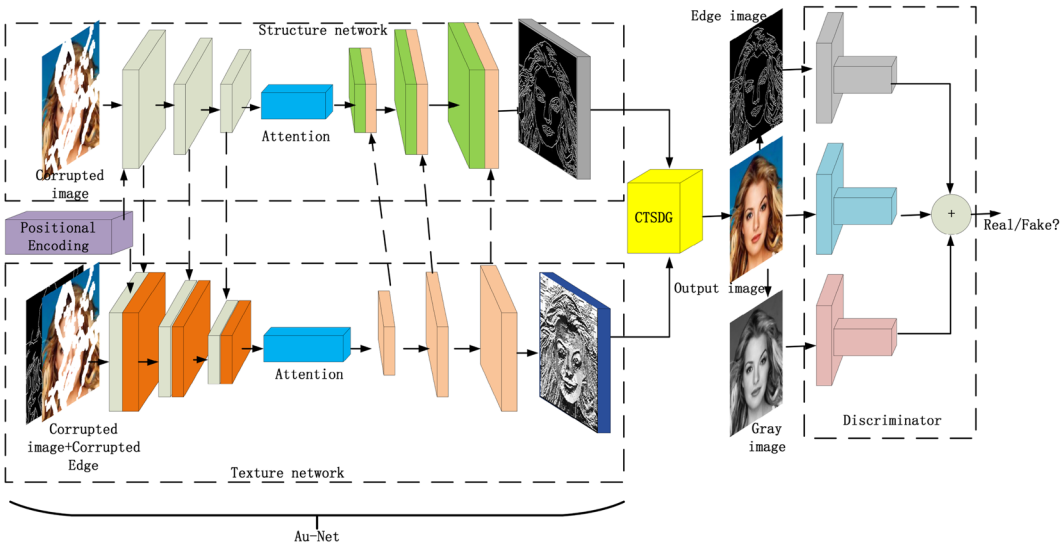

Fig. 1 Image inpainting network structure diagram. The network generates an adversarial network with the generator composed of AU-Net and CTSDG modules. Within AU-Net, the U-Net network is the primary framework, with position encoding denoting the position encoding of the mask, attention indicating the attention module, and CTSDG representing the feature fusion and context aggregation module. Furthermore, the discriminator features a tripartite structure, with grayscale representing the gray image of the image and detected edge indicating the edge image

information related to the mask. The attention module estimates the attention score of each patch by calculating the cosine similarity between feature values. Next, the texture features are reconstructed based on the attention scores to accomplish the task of inpainting the details of image texture information.

The discriminator comprises a tripartite structure to evaluate the gray image, edge image, and generated image. Each branch for the discriminator has five convolution layers with Leaky ReLU as the activation function using the $\alpha = 0.2$.

To guarantee the naturalness and plausibility of the inpainted image, we utilize an amalgamation of multiple loss functions, which include style loss, confrontation loss, reconstruction loss, and perception loss. In this paper, we utilize LSGAN loss, which is based on the least squares approach, to improve the network's overall training stability.

The proposed network design and parameter settings of each module guarantee the network structure's soundness, enhance the correlation between the known area and the missing area, and improve the quality of the final restored image.

3.2 Generator

The generator of the network is seen as the structure of two intertwined AU-Nets. The damaged image is fed into the network, and the position encoding of the mask is added to the first convolutional layer, providing the network with mask information. Both AU-Nets contain an attention module. This module searches for texture information in the background of the image to fill in missing areas and provide more detailed texture information for feature reconstruction.

Furthermore, the two AU-Net networks contain skip connections to provide additional information for the decoding stage, enabling the application of complex feature information. To enhance the interaction between separately extracted texture and structure features, both types of features are merged for decoding. This enables mutual promotion of texture and structure feature decoding.

3.2.1 Location code

The location of the mask can be passed to the neural network through zero padding in the convolutional neural network. However, this approach has limitations. The zero padding can only provide the network with spatial location information of the image, such as the orientation of the mask. When the image has large missing regions, the zero padding becomes less effective in providing information to aid the network. Instead, it can lead to problems like ghosting in the inpainted image, thereby decreasing image quality. In image inpainting, the main goal is to restore the missing regions. Providing location information about the non-missing regions is unnecessary. Therefore, providing precise location information of the missing regions is necessary, and by adding location coding, this problem can be effectively solved. Position coding represents the position relationship between the missing and non-missing regions in an image. In an image, the distance and direction of the mask are denoted by P_s and P_r, respectively. For

a 256 × 256 mask where masked and unmasked regions are denoted by 0 and 1, respectively, a 3 × 3 kernel with values of 1 is used to calculate the mask distance D_{dis} at each position in the masked region. P_s is then obtained by clipping and applying sinusoidal position coding, as shown in Fig 2b. Then, the distance is clipped and mapped by the sinusoidal positional encoding (SPE) [54] to get $P_s \in \mathbb{R}^{256\times256\times d}$

$$\begin{aligned} P_{s,2i} &= \sin(\text{clip}(D_{\text{dis},}0, D_{\max})/10000^{\frac{i}{d}}), \\ P_{s,2i+1} &= \cos(\text{clip}(D_{\text{dis},}0, D_{\max})/10000^{\frac{i}{d}}), \end{aligned} \tag{1}$$

where i and d denote the index of the channel, the total number of channels in P_s, respectively, and we set $D_{\max} = 128, d = 64$.

Yet, the sinusoidal location algorithm described above can only provide spatial location information for the missing region and can calculate the location code for a predetermined size. So P_s use the nearest neighbor interpolation to constantly adjust the size, in order to train the position coding that can scale to any size. For the mask direction, a 4 channel vector D_{dir} can be obtained from four different cores, which can represent the nearest direction of the mask position and the unmasked position. Its value depends on which kernel can finish covering the mask area first, as shown in Fig. 2c. Then, D_{dir} is projected into W_{dir} with learnable embedding parameters using formula 2, and finally mask direction P_r is obtained. P_r is defined as:

$$P_r = D_{\text{dir}} \times W_{\text{dir}} \in \mathbb{R}^{256\times256\times d}, \tag{2}$$

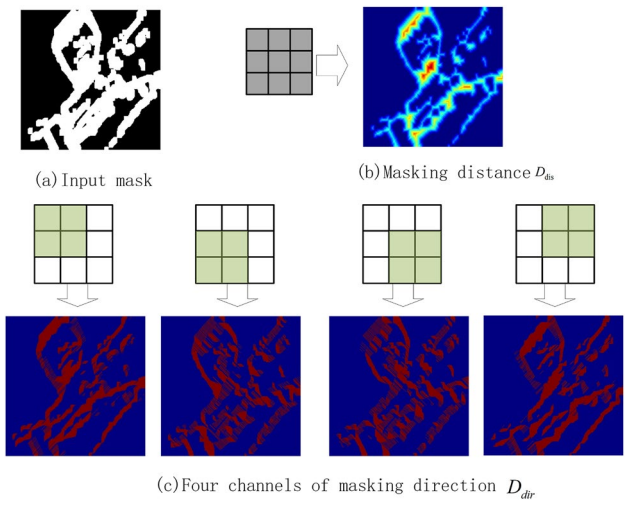

Fig. 2 The illustration of our masking relative position encoding. **a** Input mask, **b** masking distance D_{dis} and the all-one 3 × 3 kernel, **c** masking directions D_{dir} and their kernels

3.2.2 Attention module

During deconvolution coding of the image features, network computation cost becomes excessive. Therefore, an attention module is added after the encoding of AU-Net for this purpose. The attention module collects texture features to reconstruct texture details.

To be specific, given a feature map F, we first extract the patches of 3×3 pixels and calculate their cosine similarities as:

Within the attention module, where the given feature map F are considered, the calculation of cosine similarity between corresponding pairs of feature pixels is computed by

$$S^{i,j}_{\text{contextual}} = \left\langle \frac{f_i}{\|f_i\|_2}, \frac{f_j}{\|f_j\|_2} \right\rangle, \tag{3}$$

where f_i and f_j correspond to the $i-th$ and $j-th$ patch of the feature map, respectively.

We then obtain the attention score of each patch:

$$\hat{S}^{i,j}_{\text{contextual}} = \frac{\exp\left(S^{i,j}_{\text{contextual}}\right)}{\sum_{j=1}^{N} \exp\left(S^{i,j}_{\text{contextual}}\right)}. \tag{4}$$

After obtaining the attention score of each patch in the feature map, we combine the score of each patch with its corresponding patch to reconstruct the texture features of the image in this way by

$$\tilde{f}_i = \sum_{j=1}^{N} f_j \cdot \hat{S}^{i,j}_{\text{contextual}}, \tag{5}$$

where $\tilde{f}_i$ is the i-th patch of the reconstructed feature map.

3.3 Discriminator

In designing a GAN for image inpainting, the discriminator plays a critical role in image quality. Therefore, the discriminator is designed as a three-branch structure to distinguish the inpainted image from the real image, grayscale image, and edge image. In particular, three convolution layers are used to apply four steps of convolution operation. The convolution kernels have a size of three, and sigmoidal activation functions are used. Using the Canny operator, the edge image of the real image is detected, and the gray image is detected using the weighted average method. Thereby, the outputs of the three branches are combined to identify the images and continuously improve the quality of restoration performed by the generator.

3.4 Loss function

The model employs a mixed loss function comprising reconstruction, confrontation, perception, style, and LSGAN loss. This combination ensures the consistency of inpainted contents based on their weight.

The loss function formula denotes G and D as the generator and discriminator of GAN, respectively. Additionally, I_{gt}, E_{gt}, and M_{gt} represent the real, edge, and gray images, respectively. The binary mask is denoted as M, where $I_{in} = I_{in} \odot M_{in}$ represents the input of the network. Additionally, $E_{in} = E_{in} \odot M_{in}$ represents the input of the edge graph, and $Y_{in} = Y_{gt} \odot M_{in}$ denotes the input of the grayscale graph. The generator produces two outputs: I_{out} and E_{out}, which are represented by $I_{out}, E_{out} = G(I_{in}, E_{in}, Y_{in}, M_{in})$.

Reconstruction loss The L_1 distance between I_{in} and I_{out} is set to minimize the difference between them. This is shown in formula 6.

$$\mathcal{L}_{rec} = \mathbb{E}[I_{out}, I_{gt}]. \tag{6}$$

Perceptual loss To account for the lack of sensitivity to high-level semantics in the reconstruction loss, the overall loss is augmented with a perceptual loss. The L_1 distance between I_{gt} and I_{out} is calculated using the VGG19 network model in the feature space, as depicted in formula 7.

$$\mathcal{L}_{perc} = \mathbb{E}\left[\sum_i \left\|\phi_i(\boldsymbol{I}_{out}) - \phi_i(\boldsymbol{I}_{gt})\right\|_1\right], \tag{7}$$

where ϕ_i denotes the feature of the i-th pooling layer as the input of VGG19 network.

Style loss The addition of style loss to the overall loss ensures that the repaired content is consistent and coherent throughout the entire image. Style loss is calculated by measuring the L_1 distance between features, as demonstrated in formula 8.

$$\mathcal{L}_{style} = \mathbb{E}\left[\sum_i \left\|(\psi_i(\boldsymbol{I}_{out}) - \psi_i(\boldsymbol{I}_{gt}))\right\|_1\right], \tag{8}$$

where ψ_i denotes the GRAM matrix constructed by the feature.

Adversarial loss Adversarial loss is essential in producing visually impressive generated images. The traditional adversarial loss, based on maximal and minimal values, does not provide gradient feedback on fake samples, making it ineffective for supervision. To counteract this, the LSGAN loss uses least squares to calculate the distances from the decision boundary, penalizing the samples to transmit gradients. This results in improved network stability and enhanced supervision for image inpainting. Formula 9 shows how the LSGAN loss is added to the network.

$$\mathcal{L}_{adv} = \begin{cases} \min\limits_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{(I_{in})\sim p_{data}(I_{in})}\left[(D(I_{in}) - 1)^2\right] + \\ \frac{1}{2}\mathbb{E}_{I_{in}\sim p_{I_{out}}(I_{out})}\left[(D(G(I_{out})))^2\right], \\ \min\limits_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{I_{in}\sim p_{I_{out}}(I_{out})}\left[(D(G(I_{out})) - 1)^2\right]. \end{cases} \tag{9}$$

Intermediate losses Intermediate losses must be added to F_s and F_t to improve the utilization of both structural and texture features. Formula 10 presents the respective intermediate losses.

$$\begin{aligned} \mathcal{L}_{inter} &= \mathcal{L}_{structure} + \mathcal{L}_{texture} \\ &= \mathrm{BCE}(\boldsymbol{E}_{gt}, \mathcal{P}_s(\boldsymbol{F}_s)) + \ell_1(\boldsymbol{I}_{gt}, \mathcal{P}_t(\boldsymbol{F}_t)), \end{aligned} \tag{10}$$

where F_s and F_t represent structural features and texture features, respectively, $\mathcal{P}_s$ and $\mathcal{P}_t$ represent the projection functions implemented by residual blocks and convolution layers.

Overall loss To obtain clearer and more natural images, we combine the various loss functions introduced above to provide better supervision for image generation. The network's overall loss is

$$\begin{aligned} \mathcal{L}_{joint} &= \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} \\ &\quad + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{inter}\mathcal{L}_{inter}, \end{aligned} \tag{11}$$

The total loss function composed of various loss functions can effectively supervise the network to restore the image, and gradually improve the restored image in terms of image style, texture, structure, and semantic consistency according to different weights, where λ_{rec}, λ_{perc}, λ_{style}, λ_{adv}, and λ_{inter} are the trade-off parameters

4 Experimental analysis

The CelebA and Paris datasets were used to perform experiments. Results were objectively evaluated, along with necessary ablation experiments.

4.1 Experimental setup

The experiment uses the CelebA and Paris datasets in addition to the irregular dataset for the masks. The experiment is conducted with masks of varying sizes. The resolution size is 256×256 pixels. We assigned weight parameters to the loss function with values of $\lambda_{rec} = 10$ for reconstruction

loss, $\lambda_{perc} = 0.1$ for perceptual loss, $\lambda_{style} = 250$ for style loss, $\lambda_{adv} = 0.1$ for adversarial loss, and $\lambda_{inter} = 1$ for the intermediate loss.

We implemented the network architecture using PyTorch framework and trained it on a NVIDIA GeForce GTX 1080 Ti GPU (12GB) with a batch size of 6. We utilized the Adam optimizer for optimization. The training is split into two stages, using a learning rate of 10^{-4} to train the model for 300,000 iterations in the first stage, and fine-tuned using a learning rate of 5×10^{-5} for another 300,000 iterations in the second stage. The discriminator training rate was set at 1/10 of that of the generator. Models were trained on CelebA and Paris streetscape datasets for about four days, followed by a fine-tuning of one day.

4.2 Qualitative comparison

To verify the effectiveness and innovativeness of our method, we compared it with other state-of-the-art models of the same category. We used peak signal-to-noise ratio (PSNR) and Structure SIMilarity (SSIM) as metrics for qualitative comparison of results, where the higher PSNR and SSIM values indicate a more satisfactory inpainted result. The PSNR is computed using formula 12.

$$\begin{aligned} \text{PSNR} &= 10 \times \log\left(\frac{(2^n-1)^2}{\text{MSE}}\right) \\ \text{MSE} &= \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|A(i,j)-B(i,j)\|^2 \end{aligned} \tag{12}$$

The term MSE refers to the mean square error between the original image X and the processed image Y. The variable n represents the bit depth of the pixel value, and in the case of grayscale images, $n = 8$.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

here, σ_x, σ_y, σ_{xy},μ_x, and μ_y represent the standard deviations, cross-covariance, and local means for image X and image Y. The variables C_1 and C_2 denote constants.

The findings are presented in Fig. 3. In Fig. 3a, the deep learning method for inpainting was employed. Specifically, the PConv technique utilized gated convolution for image post-processing [39], but the resulting image structure information was incomplete. DeepFillv2 was another method that employed gated convolution for generative image inpainting [55]. However, this approach was prone to producing blurred images. In contrast, the MED [24] technique employed a mutual encoding and decoding CNN network to leverage texture and structure information for image inpainting. Unfortunately, texture information in the repaired images created by the MED method was not precise enough. The RFR [23] method was a progressive image repair network that repaired the image gradually, starting from the damaged edge, with the repaired results as prior information. However, this method also resulted in blurred images and unclear image structure information. Another repair network that integrated information on image texture and structure was the CTSDG [50] method, but its repair output was not natural enough.

While the algorithm's score for the repaired image displays some objectivity, it does not provide a complete appraisal of the image inpainting quality. To address this issue, we conducted a subjective evaluation by soliciting

Fig. 3 Qualitative comparison on CelebA and Paris StreetView: **a** input corrupted images, **b** PConv[39] **c** DeepFillv2 [56], **d** MED[24], **e** RFR[23], **f** CTSDG[50], **g** Ours, and **h** ground-truth images

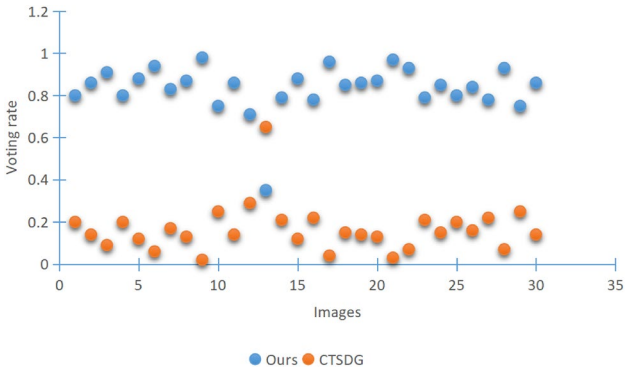

Fig. 4 Comparison of the inpainting results of 30 images using our method and the CTSDG method, respectively

artificial assessments of the inpainted outcomes. Specifically, we randomly selected 30 images from the test set and performed inpainting with both our method and CTSDG method. We then asked 20 evaluators to subjectively rate the 30 pairs of images. The images with the majority of the votes represent the ones with better inpainting quality. The results are shown in Fig. 4, indicating that, for the most part, our method outperformed the CTSDG method.

To demonstrate the effectiveness of the innovative approaches outlined in this method, we conducted ablation experiments for each approach, with the results portrayed in Fig. 5. Specifically, the impact of the position module, LSGAN loss, and attention modules is evaluated in Fig. 5b, c, and d, respectively. The findings indicate that the network requires the attention module to effectively restore the texture details of the image. Without positional encoding, the inpainted image is blurry, and the training process is prolonged. Moreover, without LSGAN loss, the overall visual effect of the repaired image is less coherent and inconsistent.

Figure 5e depicts the network with both location coding and LSGAN loss, which resulted in clearer inpainting outcomes and reduced training time. Similarly, Fig. 5f demonstrates that the inclusion of the attention module and position coding together in the network improved the texture details of the inpainted image and accelerated network training. The network with both attention module and LSGAN loss was developed for an enhanced texture information and a visually consistent and coherent effect, as shown in Fig. 5g. In Fig. 5h, the position coding, LSGAN loss, and attention module were incorporated into the network which led to the improvement in the network's ability to enhance image details' repair, while also speeding up network training and achieving a more visually consistent and coherent image.

In Fig. 6, we present the effect of position coding on the network's training speed. Figure 6a depicts the loss convergence diagram of the network without the incorporation of position coding. After 450,000 training iterations, the loss reaches a plateau. In contrast, Fig. 6b reveals that after adding position coding to the network, the loss converges to a fixed value approximately 150,000 iterations earlier, at around 300,000 training iterations.

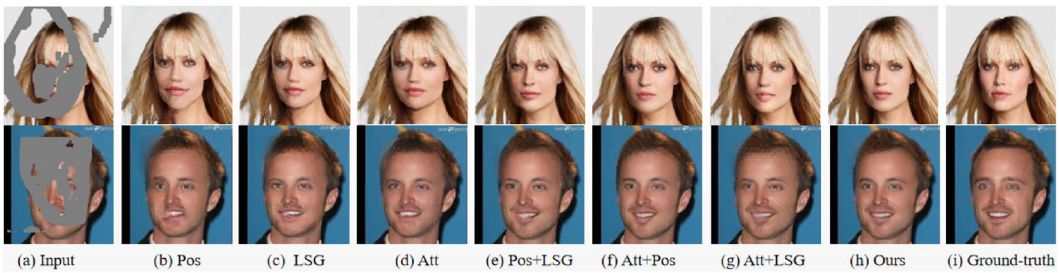

Fig. 5 Comparison of results of ablation experiment

Fig. 6 Comparison of convergence speed of network training

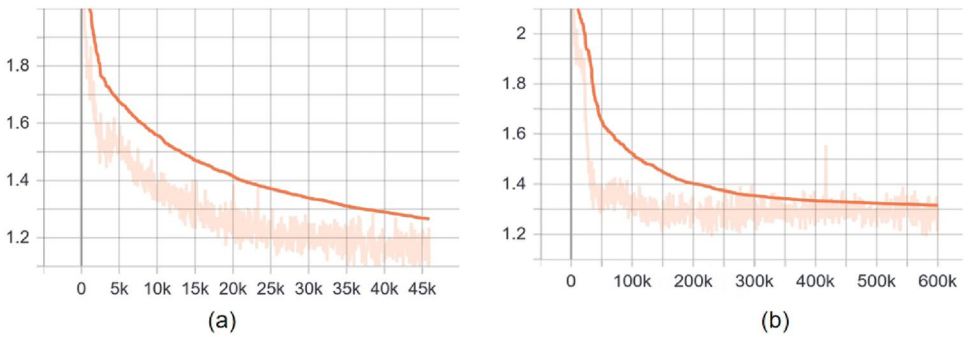

Table 1 Ablation experiment of CelebA data set

Components			Metrics		
Att	Pos	LSG	PSNR↑	SSIM↑	LPIPS↓
Mask ratio			0–20 20–40 40–60%	0–20 20–40 40–60%	0–20 20–40 40–60%
✓			29.17 23.23 18.02	0.724 0.549 0.418	0.071 0.145 0.246
		✓	30.15 24.45 18.85	0.784 0.608 0.456	0.067 0.135 0.235
✓			31.11 25.13 19.24	0.805 0.655 0.504	0.056 0.103 0.199
	✓	✓	31.58 26.12 20.45	0.875 0.701 0.545	0.035 0.097 0.185
✓	✓		33.01 27.43 22.70	0.902 0.788 0.609	0.030 0.087 0.179
✓		✓	33.54 27.91 23.05	0.915 0.781 0.618	0.028 0.081 0.170
✓	✓	✓	**33.93 28.10 23.54**	**0.931 0.793 0.623**	**0.024 0.071 0.165**

The bold markings in the table represent the best repair results

Table 2 Quantitative comparison of experimental results

Metrics	PSNR↑	SSIM↑	LPIPS↓
	0–20 20–40 40–60%	0–20 20–40 40–60%	0–20 20–40 40–60%
PConv [39]	31.89 26.48 21.32	0.899 0.750 0.588	0.046 0.107 0.214
DeepFillv2 [56]	32.48 26.93 21.70	0.906 0.757 0.569	0.040 0.107 0.214
MED [24]	32.68 27.01 21.86	0.907 0.763 0.575	0.037 0.081 0.179
RFR [23]	33.03 27.13 22.69	0.916 0.780 0.603	0.031 0.090 0.185
CTSDG [50]	33.49 27.43 22.70	0.920 0.788 0.609	0.028 0.081 0.179
Ours	**33.93 28.10 23.54**	**0.931 0.793 0.623**	**0.024 0.071 0.165**

The bold markings in the table represent the best repair results

The results of the ablation experiment are shown in table 1. Through comparative analysis, it was found that attention mechanisms, positional encoding, and LSGAN loss all have a positive effect on image restoration. Three different damaged images (0–20%, 20–40%, and 40–60%) were repaired, and the PSNR, SSIM, and LPIPS indicators were used. Table 1 indicates the modules added to the network with a checkmark (✓). The attention mechanism had the greatest impact on improving image inpainting, followed by the enhancement of LSGAN loss, while the positional encoding had the least improvement ability. Looking at the pairwise combination strategies, it can be seen that the positional encoding and LSGAN loss have the smallest improvement, while the attention mechanism and LSGAN loss have the largest improvement. Using all three strategies, namely attention mechanisms, positional encoding, and LSGAN loss, resulted in the optimal inpainting of the images, which also confirms the effectiveness of the proposed innovative idea in this paper.

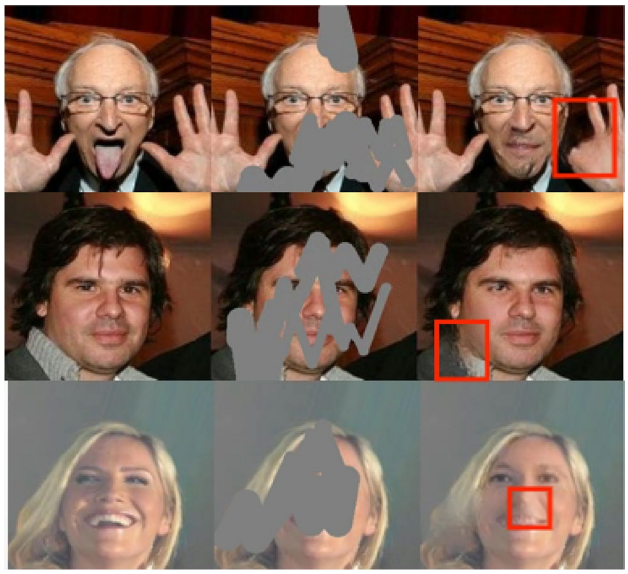

Fig. 7 Poor repair results

4.3 Quantitative comparison

We performed experimental comparisons on the CelebA dataset with different mask ratios. The mask sizes were categorized into three groups: 0–20%, 20–40%, and 40–60%. Our method showed better performance compared to other approaches for all three mask types, as shown in Table 2.

Although this model performs well in image repair, there are still some challenges to address. For instance, in Fig. 7, the first group of images lacks clarity in their structural

characteristics, and the shape of the hand in these images is not present in the training set, which leads to the omission of the finger in the repaired images. In the second group, the gray and white collar becomes entangled in the repair process, causing the model to struggle in its attempt to infer the clothing structure and rendering a poor quality of repair. Finally, the third group of images appears too vague to capture a consistent style, which limits the model's ability to restore image details and texture.

5 Conclusion

This paper proposes a dual-stream image inpainting network with an AU-Net module that utilizes an attention mechanism to extract image structure and texture features resulting in higher quality reconstructed images. The addition of mask position coding enables the network to pay greater attention to missing regions, improving the accuracy and efficiency of the inpainting process. The proposed method also incorporates loss of LSGAN along with perception, reconstruction, and style to comprehensively monitor image quality and ensure more accurate and detailed image generation. This method was compared with others using the CelebA and Paris Street View Datasets, our method outperformed the other models being compared.

Image inpainting is a complex problem due to the varied types of images and damaged or missing information. Various solutions exist such as infusing information around the damaged regions or texture synthesis inpainting. However, image understanding is crucial for effective inpainting, and attentional mechanisms and transformer networks have shown promising results. The diffusion model and graph convolutional networks can further improve inpainting by better understanding target objects in images. Future research aims to incorporate more semantic understanding into the image inpainting process for clear and natural results.

Acknowledgements This work was supported by several grants including: the Major Natural Science Research Project of Higher Education Institutions in Anhui Province (Grant No. KJ2020ZD46), the High-level Talents Research Start-up Project of Fuyang Normal University (Grant No. 2020KYQD0032), and the Fuyang Normal University Project (Grant No. Rcxm202001, No. rcxm202106, No. 2021FSKJ01ZD), as well as the Fuyang City School Cooperation Project (Grant No. SXHZ202103) and the Industrial Chain Research and Innovation Team of Fuyang Normal University Fuyang (Grant No.CYLTD202213).

Author contribution Dong, C. and Liu, H.L. proposed the idea, designed and performed the simulations, and wrote the paper. Wang, X. and Bi, X. analyzed the data. All authors have read and agreed to the published version of the manuscript.

Data availability The data that support the findings of this study are openly available in CelebA at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.Paris Dataset can be downloaded from http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/. Irregular mask Dataset are openly available in Training Set at https://www.dropbox.com/s/qp8cxqttta4zi70/irregular_mask.zip?dl=0.

Declarations

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

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 417–424 (2000)
2. Guillemot, C., Le Meur, O.: Image inpainting: overview and recent advances. IEEE Signal Process. Mag. **31**(1), 127–144 (2014)
3. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
4. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graph. **26**(3), 4 (2007)
5. Gao, F., Wang, J., Yu, Q., Zhang, J.: Sparse representation image repair algorithm for distinguishing structure and texture. Comput. Eng. **42**(3), 242–248 (2016)
6. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
7. Zhang, H., Hu, Z., Luo, C., Zuo, W., Wang, M.: Semantic image inpainting with progressive generative networks. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1939–1947 (2018)
8. Sun, Q., Zeng, X.: Image restoration based on generation countermeasure network. Comput. Sci. **45**(12), 229–234 (2018)
9. Sagong, M.-C., Shin, Y.-G., Kim, S.-W., Park, S., Ko, S.-J.: Pepsi : fast image inpainting with parallel decoding network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
11. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6721–6729 (2017)
12. Liao, L., Hu, R., Xiao, J., Wang, Z.: Edge-aware context encoder for image inpainting. In: IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), pp. 3156–3160 (2018)
13. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)
14. Dong, Q., Cao, C., Fu, Y.: Incremental transformer structure enhanced image inpainting with masking positional encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11358–11368 (2022)
15. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10758–10768 (2022)
16. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edge-connect: structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3265–3274 (2019)
17. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**(4), 1–14 (2017)
18. Liu, Q., Tan, Z., Chen, D., Chu, Q., Dai, X., Chen, Y., Liu, M., Yuan, L., Yu, N.: Reduce information loss in transformers for pluralistic image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11347–11357 (2022)
19. Liu, K., Wang, X., Xie, Y., Hu, J.: Edge guided Gan: boundary information guided depth image restoration. Chin. J. Image Graph. **26**(01), 186–197 (2021)
20. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4692–4701 (2021)
21. Yu, T., Guo, Z., Jin, X., Wu, S., Chen, Z., Li, W., Zhang, Z., Liu, S.: Region normalization for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12733–12740 (2020)
22. Yang, J., Qi, Z., Shi, Y.: Learning to incorporate structure knowledge for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12605–12612 (2020)
23. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7760–7768 (2020)
24. Liu, H., Jiang, B., Song, Y., Huang, W., Yang, C.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: European Conference on Computer Vision, pp. 725–741 (2020)
25. Zhao, L., Shen, L., Hong, R.: A review of research progress in image inpainting. Comput. Sci. **48**(3), 14–26 (2021)
26. Ma, S., Jiang, C.: Blind image restoration model based on fourth order partial differential equation. Chin. J. Image Graph. **1**, 26–30 (2010)
27. Grossauer, H.: A combined pde and texture synthesis approach to inpainting. In: European Conference on Computer Vision, pp. 214–224 (2004)
28. Rane, S.D., Sapiro, G., Bertalmio, M.: Structure and texture filling-in of missing image blocks in wireless transmission and compression applications. IEEE Trans. Image Process. **12**(3), 296–303 (2003)
29. Shi, P., Lian, Q., Shang, Q.: Image inpainting algorithm based on three-layer sparse representation. Comput. Eng. **36**(13), 189–191 (2010)
30. Benseghir, M., Nouri, F.Z., Tauber, P.C.: A new partial differential equation for image inpainting. Boletim da Sociedade Paranaense de Matemática **39**(3), 137–155 (2021)
31. Drori, I., Cohen-Or, D., Yeshurun, H.: Fragment-based image completion. In: ACM SIGGRAPH 2003 Papers. SIGGRAPH '03, pp. 303–312. Association for Computing Machinery, New York (2003)
32. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24–12411 (2009)
33. Xu, R., Guo, M., Wang, J., Li, X., Zhou, B., Loy, C.C.: Texture memory-augmented deep patch-based image inpainting. IEEE Trans. Image Process. **30**, 9112–9124 (2021)
34. Mo, J., Zhou, Y.: The research of image inpainting algorithm using self-adaptive group structure and sparse representation. Cluster Comput. **22**(3), 7593–7601 (2019)
35. Lou, X., Tang, X., Zhang, Y.: Sparse representation image restoration of similar matching block group. Chin. J. Image Graph. **24**(07), 1055–1066 (2019)
36. Yeh, R.A., Chen, C., Yian Lim, T., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5485–5493 (2017)
37. Zhu, M., Zheng, G.: Overview of improved occlusive face recognition algorithms. Comput. Sci. Appl. **12**, 1569 (2022)
38. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: image inpainting via deep feature rearrangement. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–17 (2018)
39. Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 85–100 (2018)
40. Wang, W., Zhang, J., Niu, L., Ling, H., Yang, X., Zhang, L.: Parallel multi-resolution fusion network for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14559–14568 (2021)
41. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H.: Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728–5739 (2022)
42. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
43. Yu, Y., Zhan, F., Lu, S., Pan, J., Ma, F., Xie, X., Miao, C.: Wavefill: A wavelet-based generation network for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14114–14123 (2021)
44. Zhou, T., Ding, C., Lin, S., Wang, X., Tao, D.: Learning oracle attention for high-fidelity face completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7680–7689 (2020)
45. Wang, Y., Chen, Y.C., Tao, X., Jia, J.: Vcnet: a robust approach to blind image inpainting. In: European Conference on Computer Vision, pp. 752–768 (2020)
46. Cao, C., Fu, Y.: Learning a sketch tensor space for image inpainting of man-made scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14509–14518 (2021)
47. Liao, L., Xiao, J., Wang, Z., Lin, C.-W., Satoh, S.: Image inpainting guided by coherence priors of semantics and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6539–6548 (2021)
48. Zhou, Y., Barnes, C., Shechtman, E., Amirghodsi, S.: Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, pp. 2266–2276 (2021)
49. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10775–10784 (2021)
50. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14134–14143 (2021)
51. Zheng, C., Cham, T.-J., Cai, J., Phung, D.: Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11512–11522 (2022)
52. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514 (2018)
53. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14821–14831 (2021)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł, Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 5998–6008 (2017)
55. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4471–4480 (2019)
56. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4471–4480 (2019)