Applied Intelligence (2023) 53:1162–1179
https://doi.org/10.1007/s10489-022-03387-6

Parallel adaptive guidance network for image inpainting

Jinyang Jiang[1] · Xiucheng Dong[1] · Tao Li[1] · Fan Zhang[1] · Hongjiang Qian[1] · Guifang Chen[1]

Accepted: 13 February 2022 / Published online: 26 April 2022

Abstract

Motivated by human behavior, dividing inpainting tasks into structure reconstruction and texture generation helps to simplify restoration process and avoid distorted structures and blurry textures. However, most of tasks are ineffective for dealing with large continuous holes. In this paper, we devise a parallel adaptive guidance network(PAGN), which repairs structures and enriches textures through parallel branches, and several intermediate-level representations in different branches guide each other via the vertical skip connection and the guidance filter, ensuring that each branch only leverages the desirable features of another and outputs high-quality contents. Considering that the larger the missing regions are, less information is available. We promote the joint-contextual attention mechanism(Joint-CAM), which explores the connection between unknown and known patches by measuring their similarity at the same scale and at different scales, to utilize the existing messages fully. Since strong feature representation is essential for generating visually realistic and semantically reasonable contents in the missing regions, we further design attention-based multiscale perceptual res2blcok(AMPR) in the bottleneck that extracts features of various sizes at granular levels and obtains relatively precise object locations. Experiments on the public datasets CelebA-HQ, Places2, and Paris show that our proposed model is superior to state-of-the-art models, especially for filling large holes.

Keywords Image inpainting · Parallel adaptive guidance network · Joint-contextual attention mechanism · Multiscale receptive fields

1 Introduction

Image inpainting is an important yet challenging computer vision task. Its goal is to predict appropriate pixels of the missing areas. It serves a wide range of applications, such as photoediting, decapping, and removing unwanted objects from photos. Well-repaired areas should have reasonable semantic structures and visually realistic textures.

Earlier traditional algorithms [2, 4, 6] fill holes by dealing with suitable patches or pixels from known regions. However, these methods cannot understand meaningful image semantic priors, and the repaired areas might exhibit incorrect semantic features.

In recent years, with the rapid development of deep learning, some methods [8, 16, 24, 26, 37, 39] have learned image data distributions by adopting convolutional neural networks(CNNs). Although these approaches generate semantically reasonable contents in the missing areas, the completed regions often suffer from either distorted structures or texture artifacts due to incompatibility with the human restoration behavior of the structure and texture of missing regions separately. To solve this problem, a multistage network [1, 28, 29, 33] repairs edge or structure information firstly, and use it to guide full image recovery. A single-stage network [35] utilizes the properties of the convolutional encoder-decoder itself, which extracts structures and textures at the deep layers and shallow layers of the encoder respectively and repairs them. However, as the holes become larger, most of separate repair networks easily generate discontinuous structures and unsatisfactory textures. This mainly lies in two reasons. One is failure to depict accurate edges or structures for guiding the texture generation, due to the constraint of a small amount of known information. The other is not offering the specific effective strategies to handle large holes. The progressive

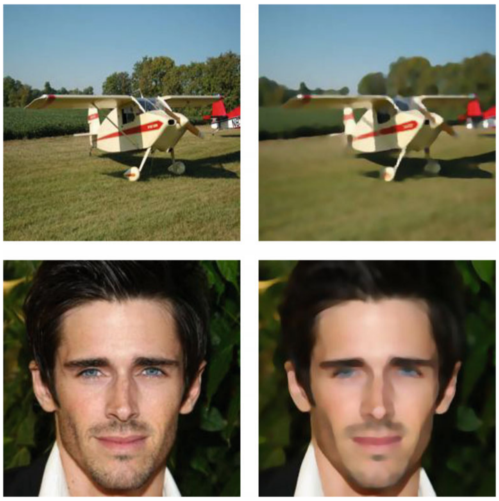
✉ Xiucheng Dong
dxc136@163.com

Jinyang Jiang
1085220997@qq.com

[1] School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, China

repair model [31, 32] fills large holes gradually by repeating the same modules, as they consider the reconstruction of structures and textures as a whole, making it more likely to produce inferior content that affects the restoration of the next stage.

To overcome the above limitations, we design multiple tactics to aid in fixing large holes on the basis of separate restoration. Specifically, we propose a novel framework called parallel adaptive guidance network (PAGN), in which the structure reconstruction branch and texture enrichment branch restore the structured image (the image after edge-preserving smoothing) and full image, respectively. As shown in Fig. 1, the structured image consists of large-scale objects such as edges and flat areas and excludes small-scale objects (e.g., details). The complete image includes semantically reasonable structures and rich textures. Unlike unidirectional structure-guided repair [28, 29, 33], our parallel branches guide each other at multiple intermediate layers via a skip connection to achieve "win-win cooperation", in which the structure reconstruction branch provides strong structure priors for the repair process of complete image, and the well-designed texture enrichment branch carrying reasonable structural information contributes to the recovery of structure maps. In addition, in contrast to the previously repaired pixels directly guiding the recovery of the remaining pixels [31, 32], we adopt the guidance filter to avoid useless information from one branch directly passing into another. In this way, the model can adaptively utilize contributing information in the mutual guidance process and produce more visually pleasing results, especially on large hole inpainting tasks. To the best of my knowledge, we are the first to use a mutual guidance structure to allow each branch to take advantage of additional information that is only beneficial to its own restoration results.



(a) Complete image (b) Structured image

Fig. 1 The difference between a complete image with details and a structured image

To make the recovered content more consistent with the background (undamaged regions), some methods [9, 22, 24, 32] build long-term connections between distant pixels by modifying contextual attention [25]. However, all of these existing attention designs fill missing areas by searching patches with high similarity within the same scale. When the image is badly damaged and only scarce information is available, exploiting limited known information through various perspectives is particularly essential. Thus, we devise a joint-contextual attention mechanism(Joint-CAM), which explores feature dependencies not only at the same scale but also different scales. This can utilize finite contextual information maximally to infer the missing contents and ensure structural continuity to some extent.

Furthermore, precise feature representation helps to understand the semantic content of an image, which is good for the inpainting task. Existing methods [3, 30, 34] utilizes utilize layerwise multiscale structures to extract different-scale features. However, these learned representations are relatively coarse due to the lack of intralevel feature fusion. To solve this issue, we design a novel multibranch module in the end of the encoder, namely attention-based multiscale perceptual res2block(AMPR). In addition to the multiple parallel convolution layers with different receptive fields, AMPR contains intralayer residual connections and attention mechanism. Thus, AMPR not only extracts multiscale features of the whole image at a granular level and retains more accurate spatial location information, but also fuses features from various branches effectively through an attention mechanism.

In summary, the main contributions of our work can be described as follows:

1. A novel framework called the parallel adaptive guidance network (PAGN) for image inpainting not only specializes in repairing structures and enriching textures through parallel branches separately but also facilitates features from one branch to adaptively accept useful information from another via a skip connection with a guidance filter.
2. We introduce strategies for repairing large holes, including the joint-contextual attention mechanisms(Joint-CAM), which utilize limited known information maximally, and an attention-based multiscale perceptual res2block(AMPR), which effectively recovers missing objects of various sizes. These tactics help to generate the results with clear textures and continuous structures
3. Our method is more effective than state-of-the-art models for dealing with large holes, and achieves high-quality results on facial and natural datasets.

2 Related work

2.1 Image inpainting

Currently, image inpainting is divided into two main parts: traditional methods or deep learning methods. Traditional approach [2, 4, 6] spreads the neighborhood pixels or patches from background to the target hole. Although these methods can repair better textures, they have many limitations: 1. They must assume that the content of the missing regions can be found in the input image or external image libraries. 2. The reconstructed areas often exhibit incorrect structures due to a lack of understanding of the high-level semantics of images. 3. Broken images are repaired with regular-shaped masks.

With the rapid development of deep learning, convolutional neural networks (CNNs) have shown outstanding performance in image inpainting tasks because they effectively capture local features and high-level abstract features of an image. Pathak et al. [7] proposed a context encoder and was the first work incorporating generative adversarial loss into an encoder-decoder architecture to repair broken pixels. However, the repaired results always contain texture artifacts due to the restriction of the channelwise fully connected layer. Subsequently, Iizuka et al. [8] improved the image quality by employing cascaded dilated convolution, and use the local and global discriminators together to ensure both the consistency of the repaired region and the entire image. Chen et al. [39] let the local discriminator identify the similar patches in different images but in the same type, which improves the discriminative ability of the network. Yu et al. [25] proposed a coarse-to-fine network, which produces the rough prediction first through a coarse network, and further optimizes the coarse intermediate results into more high-quality images through a refinement network. Chen et al. [27] added context-awareness loss to make the repaired regions more realistic by constraining the similarity of local features. Liu et al. [16] and Yu et al. [26] replace the partial convolution and gated convolution with an ordinary convolution, respectively, to avoid color incongruity and edge response.

These methods often fail to reconstruct continuous structures or fine details as they recover the holes without plausible strong constraints. Nazeri et al. [29] depicted the lines of the missing areas first and added the colors and textures based on these restricted lines. Ren et al. [33] split the whole inpainting into two steps: first, they repaired the missing structures and then provided the completed structures to the texture generator to direct the synthesis the vivid textures through appearance flow. Shao et al. [1] utilized fusion images of edge maps and blurred images which provide color information as labels to guide the reconstruction of the refined image. Liu et al. [35] captured structure and texture features using the deep and shallow layers of the encoder respectively, and filled the holes of different-type features via separate multiscale blocks. Guo et al. [31] and Li et al. [32] considered a progressive inpainting policy, in which dilated pixels gradually form known regions to the hole center by using repeating blocks or modules. Zhu et al. [40] utilized multiple decoders to refine the reconstructed results.

2.2 Contextual attention inpainting

To keep the generated textures realistic and consistent with the surrounding features, Yu et al. [25] proposed a contextual attention layer that borrows similar feature patches from context to fill missing regions. Chen et al. [39] preprocessed the images by using a similar block around the damaged area to update the damaged block. Zeng et al. [24] designed a pyramid-context encoder, which progressively applies a contextual attention mechanism from latent feature maps to the original image, to ensure both semantic and visual coherence in the repaired regions. Liu et al. [9] found that the repaired results will show discontinuous pixels if they focus only on feature dependency inside and outside the holes. Thus, they explore relationships between pixels in the holes as well. Li et al. [32] considered the consistency between the attention scores from different recurrences and devised the knowledge attention layer for recurrent architecture. However, these methods ignore the strong correlations between feature patches at different scales. Exploring it will obtain more accurate matching patches in missing regions, especially when background information fades considerably.

2.3 Multiscale design

Influenced by the way neurons in the human brain are connecte, multiscale structure is adopted to capture features of different sizes in many computer vision tasks. Inception [12] and atrous spatial pyramid pooling (ASPP) [21] are the most common multiscale designs are implemented in various networks. In the object detection field, Liu et al. [42] proposed a receptive field block, which absorbs the advantages of ASPP and Inception, to enhance feature robustness and improve detection accuracy. In the super resolution field, Li et al. [38] combined ASPP and channel attention at the bottleneck. In the 3D reconstruction field, Ding et al. [44] estimated a more accurate depth map by using continuous multiple ASPP blocks, which is vital for better 3D reconstruction. In the image deraining filed, Wang et al. [45] utilizeed multiscale kernels and multiresolution feature maps to capture rain streaks with different sizes and scales.

In the image inpainting field, multiscale feature representation is essential for understanding the semantic information of images. Wang et al. [30] devised the multicolumn network, which contains three parallel encoder-decoder branches with different filter sizes and spatial resolutions, to extract different levels of features. Chen et al. [34] designed two parallel encoders with different receptive fields to obtain global semantic features and local detail features respectively.

3 Parallel adaptive guidance network

Given a defective image and the corresponding binary mask, our goal is to output a well-repaired image with visually realistic content. Xiong et al. [28], Nazeri et al. [29], Ren et al. [33] showed that repairing the structures and textures of an image separately would reduces texture artifacts and oversmoothed boundaries. Therefore, we design a parallel adaptive guidance network(PAGN) as shown in Fig. 2, where the structure reconstruction branch aims to reconstruct the structures of a damaged image, and the texture enrichment branch simultaneously enriches the textures and recover the complete image. These features from different branches guide each other's restoration using skip connections and guidance filters. Moreover, the utilizing of the joint-contextual attention mechanism(Joint-CAM) in the texture enrichment branch helps to output completed images with realistic details and continuous structures, and designed attention-based multiscale perceptual Res2Blocks(AMPR) in the bottleneck helps to capture the features with different receptive fields in a way closer to the human eye.

We first introduce the idea of mutual adaptive guidance in Section 3.1. Then, we describe the Attention-based Multiscale Perceptual Res2block and Joint Contextual Attention Mechanism in Sections 3.2 and 3.3 respectively. Finally, we provide the corresponding loss functions of our model in Section 3.4.

3.1 Mutual adaptive guidance

Different elements of an image, structure and texture are interrelated. In contrast to the traditional idea, which only takes the guidance map (structure map, edge map, etc.) as one of the inputs to provide extra information, we make two improvements. First, one-way guidance is replaced with mutual guidance in favor of the restoration of each element. Specifically, in the decoder, recovered structures can be integrated into the texture enrichment branch to provide strong priors. Once the correct structures are completed, the inpainting task can be treated as a detail-enrich problem. In the encoder, extracted full features involving rich texture and reasonable structure information can also help the repair of accurate structures in turn. Second, each branch incorporates multiple intermediate-level features from another branch so that the guidance information is considered at multiple layers, avoiding only affecting previous layers of the deep network.

As the mask area increases, the feature maps used in guidance inevitably contain wrong or invalid information

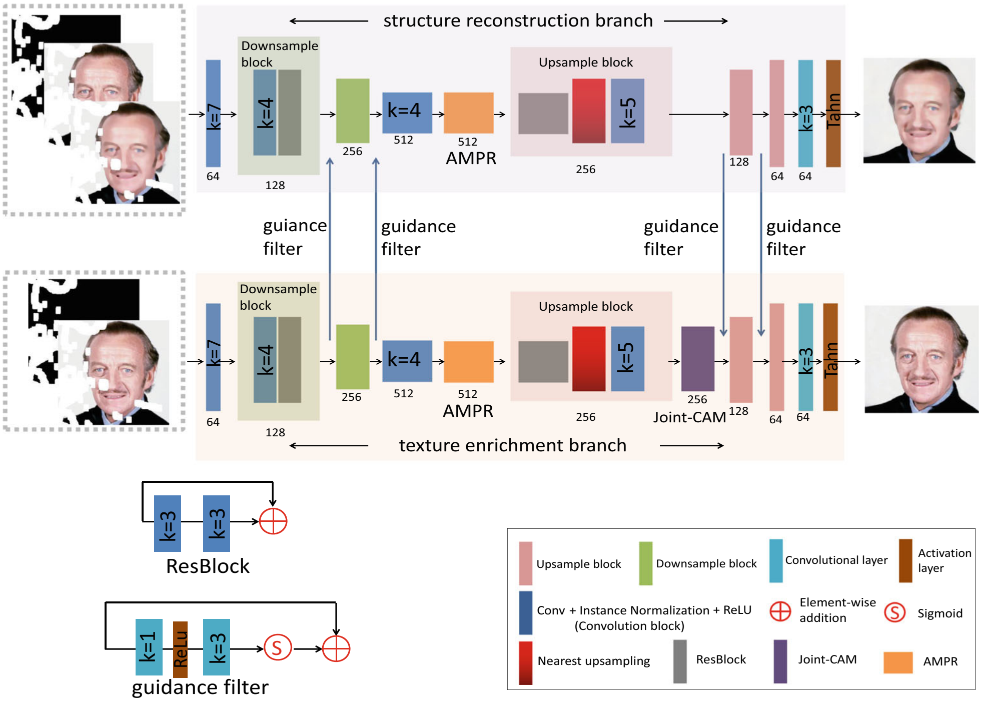
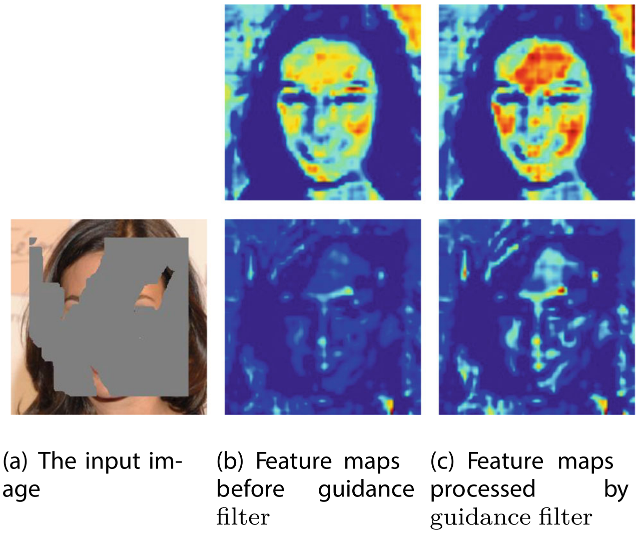


Fig. 2 The overall pipeline of the PAGN. It consists of structure reconstruction branch and texture enrichment branch. Each branch adopts a typical encoder-decoder structure. The two branches guide each other by skip connection and guidance filter. In the encoder, downsample blocks, AMPR, etc. are used to understand the semantic information. In the decoder, upsample blocks, Joint-CAM, etc. are used to reconstruct the image. The downsample block consists of convolution block and Resblock, and the upsample block consists of ResBlock, nearest upsampling and convolution block with normalization and activation layer. The number k represents the kernel size of the convolutional layer

during the restoration process. To prevent less informative features from one branch directly going to another via skip connection, an extra guidance filter is added to highlight contributing features adaptively and suppress poor features. The guidance filter consists of two convolution layers with different kernel sizes and a sigmoid function, in which a 1∗1 convolution is utilized to integrate information and compress channels of input feature maps, and another 3∗3 convolution and a sigmoid function are used to yield an attention map. This attention map is then used to recalibrate the feature map to be directed to another branch. The feature contrast diagram of the input and output feature of the guide filter is shown in Fig. 3, and the processed feature maps highlight the important objectives such as mask regions and key points of the face.(feature maps are upsampled to $256 * 256$ for observation).

3.2 AMPR In the encoder

Robust feature representations facilitate inpainting networks to yield semantically accurate contents and clear details. We think that robust feature representations are reflected in two aspects: features with multiscale receptive fields and precise spatial location. The former usually contains global and local information, which helps network understand the missing semantic content and enrich texture details. The latter is vital for visual systems. For up to these purposes, we propose a new multiscale structure called attention-based multiscale perceptual res2block(AMPR) at the bottleneck, which consists of multiscale perceptual block, intralayer residual connection similar to [19] and convolutional block attention module(CBAM) [36]. The multiscale perceptual block (Section 3.2.1) aims to capture scale-diverse features effectively. Intralayer residual connection (Section 3.2.2) enhances information exchange between different branches, contributing to understanding the contents of missing regions at a granular level and retaining relatively precise spatial location information. The convolutional block attention module(CBAM) helps to reduce redundant features when fusing multiscale features. The design inspiration of this block originates from the receptive fields block(RFB) [42] as shown in Fig. 4a, but has three special modifications as shown in Fig. 4b, where d represents the dilation rate, and k represents the kernel size. Next, we introduce the corresponding modifications in detail.


(a) The input image

(b) Feature maps before guidance filter

(c) Feature maps processed by guidance filter

Fig. 3 Visual contrast diagram of the input and output of the guidance filter

3.2.1 Multiscale perceptual block

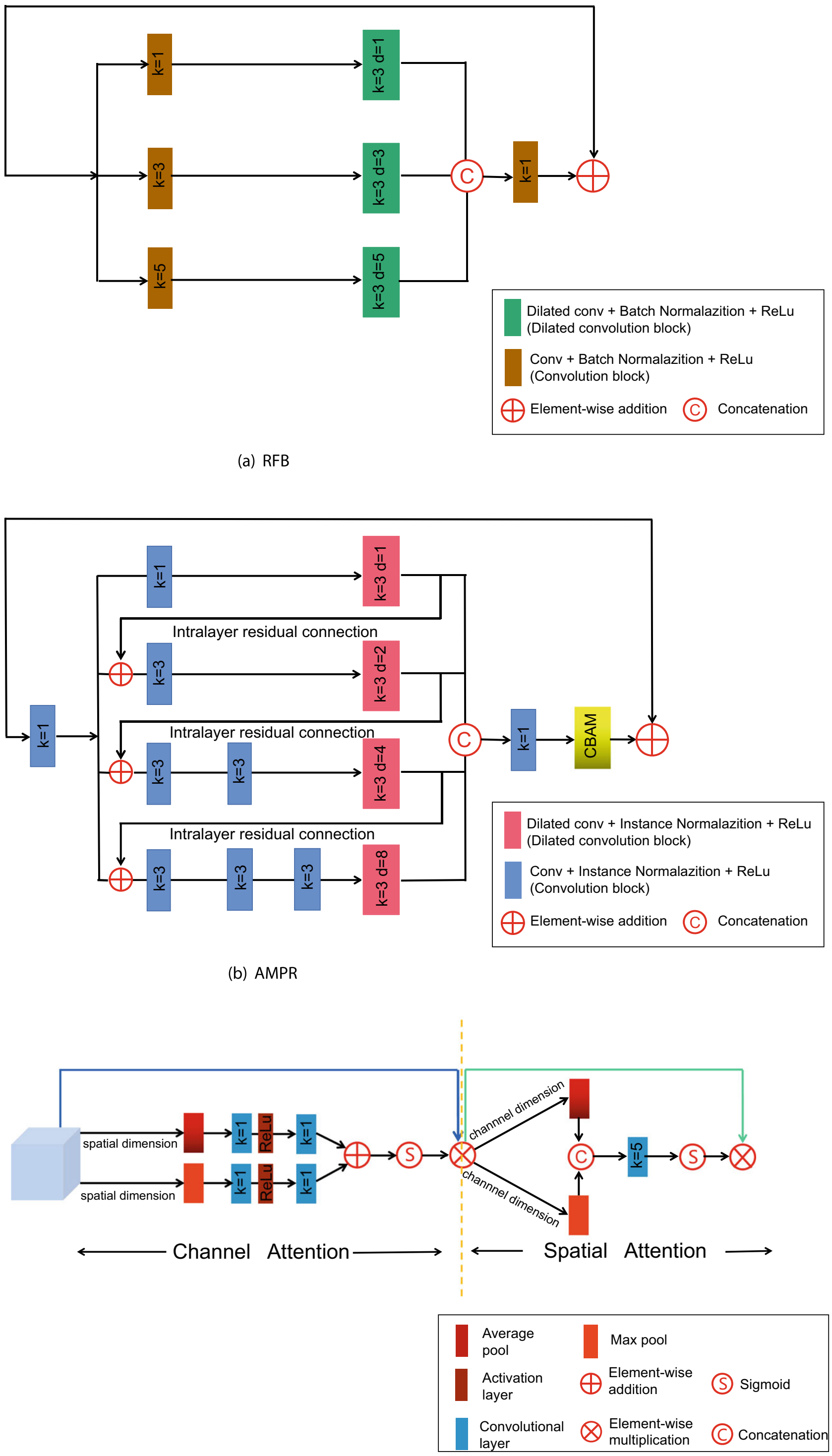
Inspired by receptive fields block [42], both the size and eccentricity of the receptive fields play important roles in human vision. Thus, the multiscale perceptual block can be divided into two components as well: convolution layers with different kernels and dilated convolution layers [18] with individual eccentricity. To reduce network parameters, a $1 * 1$ convolution is employed to compress the channels of the feature map before going to multibranch structure. As shown in Fig. 4b, in four parallel branches with various filter groups, for the convolution layer part, the kernel size is 1, 3, 5 and 7, and the eccentricity is fixed at 1. For the dilated convolution layer part, the eccentricity is 1, 2, 4 and 8, and the kernel size is fixed at 3.

Since replacing a large-scale convolution kernel with multiple small-scale convolution kernels can both reduce the parameters and increase the depth of the network with the same receptive field, we use the corresponding $3 * 3$ convolution kernels instead of a large-scale convolutional kernel of 5,7.

3.2.2 Intralayer residual connection

Intralayer feature fusion has been successfully applied in many computer vision tasks, but is less explored in image inpainting tasks. Our proposed intralayer residual connection allows adequate information fusion between different-branch features. This operation not only helps to capture multiscale features at a finer level, but also preserves accurate spatial information, which leads to key point (nose, eyes, etc.) localization. We denote $C_i()$ as the i-th branch in the multiscale perceptual block and F_i is the output of $C_i()$. The specific implementation is to add the input feature Xi to the output of the previous branch F_{i-1}, and then feed into $C_i()$. As shown in Fig. 4b, we use the element-wise sum instead of a concatenation operation to fuse features inside each branch, aiming to avoid redundant feature information as the convolution layers continuously

Fig. 4 The comparison between (a) RFB and (b) AMPR. The CBAM module of AMPR is shown in (c). The parameters k and d represent the kernel size and the dilated rates of the dilated convolution(or convolution), respectively. In convolution, the default value of d is 1



(a) RFB

(b) AMPR

(c) CBAM

increase. The output of each branch can be computed as follows:

$$F_i = \begin{cases} C_i(X_i) & i = 1 \\ C_i(X_i + F_{i-1}) & i = 2, 3, 4 \end{cases} \quad (1)$$

3.2.3 Convolutional block attention module

Actually, each feature from different sources is treated equally if we stack them directly in channel dim, which is not consistent with human vision. Thus, attention mechanism is utilized to tell us which features need special attention and elevate the redundant information. Here, we choose the popular attention module - convolutional block attention module(CBAM) [43] - to fuse features effectively. As shown in Fig. 4c, the CBAM includes channel attention and spatial attention. Within channel attention, we concern about what is meaningful information. The 3-dimensional input (simple stacking of feature maps from multiple branches) is compressed into two different 2-dimensional features that only have the channel information by global maximum pooling and global average pooling along the spatial dimmention. Maximum pooling is used to search for unique semantic features and average pooling is used to count the semantic information. Then the shared neural network and sigmoid function are explored for internal correlations between channels to produce a channel attention mask.

Supplementation to channel attention, spatial attention is used to concern where the important information is. We perform maximum pooling and average pooling on features of passing channel attention along the channel dimension, which represent two different kinds of spatial information. Then, the two pooled feature maps are stacked on the channel dimension to generate the spatial attention mask by convolution and sigmoid activation.

3.3 Joint-contextual attention mechanism

With the help of contextual attention [25], we can copy distant patches from surrounding areas to synthesize better quality textures. However, existing attention modules [9, 24, 25] for inpainting task reconstruct the missing patch by using the similar patches at the same scale, which fails to make the most of the helpful information. With the missing areas increase, the scale restriction make the contextual attention prone to generate the distorted structures. In the task of super resolution reconstruction, the paper [43] utilize the cross-scale non-local module to explore the correlations between the low-resolution patch and high-resolution patches, which ensures the structural consistency. Inspired by it, we introduce the joint-contextual attention mechanism(Joint-CAM)to enlarge the search scope of similar patches. Joint-CAM contains in-scale contextual attention(is-ca), cross-scale contextual attention(cs-ca) and the residual connection. The is-ca considers similarity between same-scale feature patches, which help to generate the clearer textures. The cs-ca explores the dependencies between cross-scale feature patches, which facilitates the recovery of reasonable structures and details. The residual connection can help the Joint-CAM target feature patches that need to be filled with patches of known regions. These two contextual attention are described in detail below.

3.3.1 In-scale contextual attention

We take the center hole as an example. For the in-scale contextual attention(is-ca), we usually measure the similarity for all the patch pairs inside and outside the holes(foreground and background) using the cosine similarity:

$$sim_{x,y,\bar{x},\bar{y}} = \langle \frac{f_{x,y}}{\| f_{x,y} \|}, \frac{f_{\bar{x},\bar{y}}}{\| f_{\bar{x},\bar{y}} \|} \rangle \tag{2}$$

Where $sim_{x,y,\bar{x},\bar{y}}$ represents the similarity between the foreground patch $f_{x,y}$ at location (x, y) and the background patch $f_{\bar{x},\bar{y}}$ at the location $(\bar{x}, \bar{y})$ in the same feature maps. The in-scale attention score of each background patch $f_{\bar{x},\bar{y}}$ is calculated by softmax function with a scale. Finally, each foreground patch is reconstructed by aggregating weighted background patches. In practice, the above steps usually can be simplified by the convolution, a channelwise softmax function, and the transposed convolution. The paper [25] described the simplified process in detail.If you are interested, you can review it. The green and black arrows in Fig. 5 show the process of is-ca clearly as well.

3.3.2 Cross-scale contextual attention

Cross-scale feature similarity was proposed for image super-resolution by [43], and this idea is then extended to our restoration tasks. Cross-scale contextual attention(cs-ca) models long-range dependency without same-scale restriction. The similarity between cross-scale features is obtained by measuring the correlations between low-resolution patches $(k * k)$ and higher-resolution patches $(sk * sk)$ in the same feature map. However, applying cosine similarity directly is infeasible since the spatial dimensions of different resolutions are different. Thus, we first downsample the feature map, and the low-resolution patches $(k * k)$ in the downsampled maps have the same receptive fields as the higher-resolution $(sk * sk)$ patches in the original maps. Then, the cross-scale attention scores are derived by calculating the cosine similarity between the patches from the downsampled map and the those with the resolutions from the original map, which can be achieved along the red and black arrows in Fig. 5. Specifically, we assume that the input feature map is f$(H * W)$. First, f is downsampled to g $(H/s * W/s\ )$. Then, the cross-scale attention scores between patches$(k * k)$ in f and those in g are calculated by convolution and a softmax function. Finally, the corresponding patches$(sk * sk)$ in f are used as deconvolution filters to reconstruct the missing patches in f and generate high-frequency details. Notably, unlike the single image super resolution task [38], the stride of transposed convolution(deconvolution) is set to 1 so that the feature maps are not zoomed upon when reconstructing

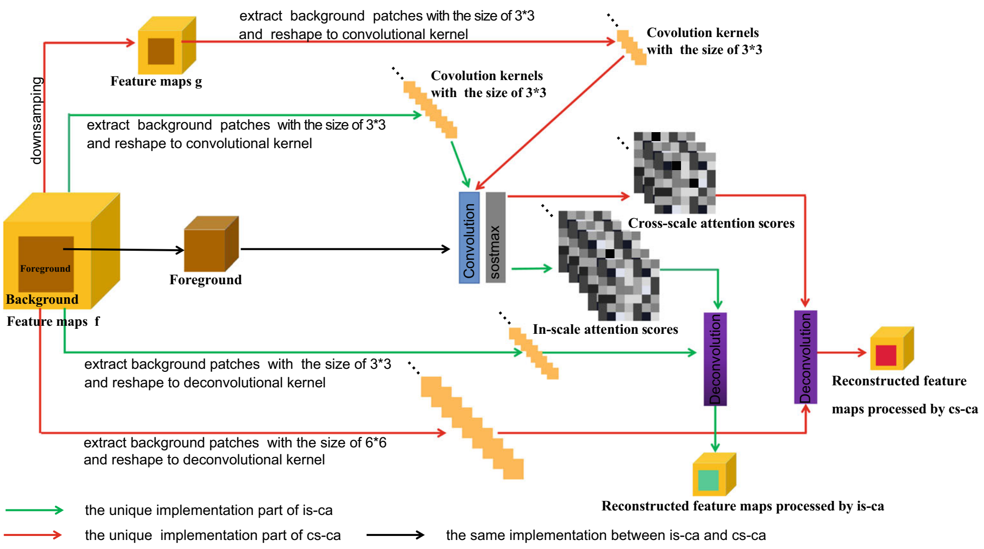

extract background patches with the size of 3*3
and reshape to convolutional kernel
Covolution kernels
with the size of 3*3
Feature maps g
downsampling
Covolution kernels
with the size of 3*3
extract background patches with the size of 3*3
and reshape to convolutional kernel
Foreground
Background
Feature maps f
Foreground
Convolution
sostmax
Cross-scale attention scores
In-scale attention scores
Deconvolution
Deconvolution
Reconstructed feature
maps processed by cs-ca
extract background patches with the size of 3*3
and reshape to deconvolutional kernel
extract background patches with the size of 6*6
and reshape to deconvolutional kernel
Reconstructed feature maps processed by is-ca
the unique implementation part of is-ca
the unique implementation part of cs-ca
the same implementation between is-ca and cs-ca

Fig. 5 The implementation of in-scale contextual attention(is-ca) and cross-scale contextual attention(cs-ca). The green and red lines represent different parts between two types of contextual attention, and the black lines represent similar parts

the missing patches. In this paper, we choose k to be 3 and s to be 2, and the downsampling operation is bilinear interpolation.

When merging the two independent feature maps generated by the is-ca F_{in_scale} and cs-ca F_{cross_scale} into the unified feature maps, we use residual convolution ResConv() to learn residual features R_{attn} between different sources instead of adding or concatenating them directly, which allows the network to focus on only the distinct information while bypassing the same information, to reduce redundant features in the merged maps. In addition, building a skip connection between the input feature F_{input} and output of Joint-CAM F_{attn} allows the network focus more on hole patches that have similar patches in known regions. This improves the discriminative ability of the network. The merging process is shown in (3)-(5).

$$R_{attn} = F_{in\_scale} - F_{cross\_scale} \tag{3}$$

$$F_{attn} = ResConv(R_{attn}) + F_{in\_scale} \tag{4}$$

$$F = F_{attn} + F_{input} \tag{5}$$

3.4 Loss function

For the structure reconstruction branch, we jointly use generative adversarial loss and L1 loss. For the texture enrichment branch, we add extra perceptual loss as well as style loss. In our paper, the weight setting of the above losses is the same as that of StructureFlow(SF) [33]. In addition, we use depth-supervised perceptual loss in each deconvolution layers of two branches to refine the predictions at each scale. The refined estimation of the missing regions ensures that the joint contextual attention mechanism performs well.

3.4.1 Depth-supervised perceptual loss

Compared with pixel-by-pixel loss, the perceptual loss is shown to be more consistent with the human visual system and generate more details. Thus, we use the depth-supervised perceptual loss to progressively optimize the predictions at each deconvolution layer of parallel branches. Taking the texture enrichment branch as an example, we first use activation layers $\{relu3_1\}$and $\{relu2\_1\}$ of VGG19 [17] to extract two-resolution feature maps of real images, and then calculate the loss L_{deep} between the extracted real features and the features predicted by our corresponding deconvolution layer. The depth-supervised perceptual loss in the structure reconstruction branch L^S_{deep} is obtained in the same way:

$$\begin{aligned} L_{deep} &= \sum_{i=1}^{P} \|\Phi_i(I_{gt}) - F_i\|_1 \\ L^S_{deep} &= \sum_{i=1}^{P} \|\Phi_i(S_{gt}) - S_i\|_1 \end{aligned} \tag{6}$$

Here, I_{gt} and S_{gt} represent real images and real structured images, respectively. Φ_i represents the i-th selected activation layer of the VGG19 [17] network. F_i represents the feature maps with the same size as $\Phi_i(I_{gt})$, predicted by the deconvolution layer of texture enrichment branch. S_i represents the structure feature maps with the same size as $\Phi_i(S_{gt})$, predicted by the deconvolution layer of the structure reconstruction branch.

3.4.2 Structure reconstruction branch loss

The pixel-to-pixel loss of the structure reconstruction branch L^s_{l1} is defined as the L1 loss between the reconstructed structured image $\hat{S}_{re}$ and the real structured image S_{gt}, which constrains the major content in the missing regions, as shown in (7).

$$[]L^s_{l1} = \|S_{gt} - \hat{S}_{re}\|_1 \tag{7}$$

In fact, the image inpainting task is an ill-posed problem with multiple feasible restoration results in the missing regions. To make the repaired results look more realistic and contain more details, we use the generative adversarial loss L^s_{adv} [13]:

$$L^s_{adv} = E[log(1 - D_s(G_s(I_{in}, S_{in}, M)))] + E[log(D_s(S_{gt}))] \tag{8}$$

Here, I_{in} denotes the broken input image, and S_{in} denotes the edge-preserving smoothed structure of the corrupted image. S_{gt} represents the real structured image. M represents the binary mask, in which pixel value 0 represents the background and pixel value 1 represents the missing region. G_s is our structure reconstruction branch. D_s is the structure discriminator, which discriminates whether the restored structured image is the same as the real one or not. If the identification results is true, the output is 1, otherwise, the output is 0. The best output is 0.5, i.e., restored image is so realistic that fools the discriminator. In this paper, we adopt PatchGAN [20] as our discriminator to discriminate the authenticity of all image patches instead of the whole image.

Eventually, the structure reconstruction branch is trained together using (9), and λ^s_{l1}, λ^s_{adv} and λ^s_{deep} are 4,1,0.01, respectively.

$$\min_G \max_D L^s(G, D) = \lambda^s_{l1} L^s_{l1} + \lambda^s_{adv} L^s_{adv} + \lambda^s_{deep} L^s_{deep} \tag{9}$$

3.4.3 Texture enrichment branch loss

The pixel-to-pixel loss of the texture enrichment branch L^t_{l1} is defined as the L1 loss between the reconstructed image $\hat{I}_{re}$ and the real image I_{gt}, as shown in (10):

$$L^t_{l1} = \|I_{gt} - \hat{I}_{re}\|_1 \tag{10}$$

Additionally, the generative adversarial loss is added in the texture enrichment branch, as shown in (11), where G_t is the texture enrichment branch that generates the final restored result image containing rich textures. The structure of discriminator D_t is the same as D_s, which determines whether the given input is real or fake.

$$L^t_{adv} = E[log(1 - D_t(G_t(I_{in}, M)))] + E[log(D_t(I_{gt}))] \tag{11}$$

To ensure that the repaired image matches the human vision system, the perceptual loss L^t_{per} [15] is shown in (12). $\Phi_i()$ represents the activation layer of the VGG19 [17] network, and $\{relu1\_1\}$, $\{relu2\_1\}$, $\{relu3\_1\}$, $\{relu4\_1\}$ and $\{relu5\_1\}$ are selected in this paper. As shown in (13), the style loss L^t_{style} [15] is also additionally included in

Table 1 Quantitative evaluation of the CelebA-HQ dataset and Places2 dataset

		0.2-0.3			0.3-0.4			0.4-0.5			0.5-0.6		
	MODEL	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS
CelebA-HQ	EC(ICCVW 2019) [29]	0.94	27.49	0.0330	0.92	25.65	0.0822	0.89	24.04	0.0730	0.79	21.07	0.1245
	CSA(ICCV 2019) [9]	0.95	28.07	0.0295	0.92	26.27	0.0455	0.91	24.59	0.0710	0.82	21.55	0.1346
	RFR(CVPR 2020) [32]	0.96	28.14	0.0310	0.93	26.08	0.0482	0.90	24.52	0.0681	0.82	21.85	0.1121
	PAGN	0.95	28.06	0.0329	0.94	26.35	0.0488	0.91	24.85	0.0672	0.84	22.20	0.1077
Places2	EC(ICCVW 2019) [29]	0.92	25.40	0.0641	0.86	22.43	0.1202	0.80	20.94	0.1681	0.64	18.42	0.2663
	SF(ICCV 2019) [33]	0.93	26.59	0.0665	0.87	23.06	0.1092	0.82	21.20	0.1423	0.68	18.70	0.2263
	MEDEF(ECCV 2020) [35]	0.94	27.05	0.0529	0.87	23.08	0.0976	0.82	21.50	0.1333	0.67	19.01	0.2158
	PAGN	0.94	27.32	0.0507	0.89	23.80	0.0958	0.84	22.08	0.1331	0.72	19.48	0.2280

Table 2 Quantitative evaluation of the Paris dataset under the large holes, and the mask ratio is 40%-60%

	MODEL	SSIM	PSNR	LPIPS
Paris	EC(ICCVW 2019) [29]	0.75	21.00	0.1884
	CSA(ICCV 2019) [9]	0.77	21.73	0.2133
	MEDEF(ECCV 2020) [35]	0.76	21.50	0.2180
	RFR(CVPR 2020) [32]	0.78	21.80	0.1835
	PAGN	0.81	22.27	0.1796

the texture enrichment branch, to reduce the checkerboard artifacts caused by the perceptual loss and to maintain the consistent texture style with the real image. G^{ϕ}_{j} is a style matrix of size $C * C$. Because the result of model trained with a small style loss weight has many fish scale artifacts, the weight of style loss is much bigger than other losses.

$$L^{t}_{per} = E[\sum_{i}^{P}\frac{1}{N_i}\|\Phi_i(I_{gt}) - \Phi_i(\hat{I}_{re})\|_1] \tag{12}$$

$$L^{t}_{style} = E_j[\|G^{\phi}_{j}(I_{gt}) - G^{\phi}_{j}(\hat{I}_{re})\|_1] \tag{13}$$

Finally, the texture enrichment branch is trained together using (14), where λ^t_{adv} ,λ^t_{l1} , λ^t_{per}, λ^t_{sty}, λ^t_{deep} are 1, 5, 0.01, 180, 0.1 respectively in the experiment.

$$\begin{aligned}\min_{G}\max_{D} L^{t}(G, D) = {} & \lambda^{t}_{l1}L^{t}_{l1} + \lambda^{t}_{adv}L^{t}_{adv} + \lambda^{t}_{deep}L_{deep} \\ & +\lambda^{t}_{sty}L^{t}_{style} + \lambda^{t}_{per}L^{t}_{per}\end{aligned} \tag{14}$$

The total losses are shown in (15). Inspired by PEPSI [11] that slightly reduces the weights of one path loss to focus on the other path. At the beginning of training, we want to provide a strong structure prior to the texture enrichment branch, and the penalty of the structure reconstruction branch is strong. As the training processes, we focus on image detail restoration gradually, and the penalty of the structure reconstruction branch slowly weakens. I represents current iterations, and I_{max} represents the maximum number of iterations.

$$L_{total} = \min_{G}\max_{D} L^{t}(G, D) + (1 - \frac{I}{I_{max}})\min_{G}\max_{D} L^{s}(G, D) \tag{15}$$

4 Experiment

4.1 Implementation details

We train our method on the Place2 [10], CelebA-HQ [46] and Paris [23] datasets. Place2 contains more than 10 million images and covers 365 natural scenes, and we follow the original training and validation splits. The CelebA-HQ dataset consists of 30,000 highly structured face images with the resolution of 1024∗1024, where 3000 images are randomly selected into the test set. For Paris, it is 6,000 images of Paris street buildings, where 50 images belong to the test set. For mask, we use challenging irregular mask datasets provided by [16]. All the irregular masks and images used for training and testing are resized to 256∗256.

During the training, we use the Adam optimizer [41] with β1=0 and β2=0.999. The batch size and the learning rate are set to 8 and $1 \times 10e^{-}4$, respectively. Our proposed method is implemented in PyTorch, and conducted on a single NVIDIA 3090 GPU. In addition, the end-to-end training strategy is adopted, so the structure reconstruction branch and texture enrichment branch are trained simultaneously. The training process of the CelebA-HQ model, Paris model and Placces2 model took 3 days, 2.5 days and 10 days respectively.

To obtain a smoothed structure of the ground truth, a rolling guidance filter (RGF) [14] is utilized to process real images, which leaves critical structures and edges and removes texture details. As a real time edge-preserved method, its parameters σs and σr are used to control spatial and range scale of smooth window. As the σs and σr larger, more details are smoothed. Here, we select σs as 3 and σr as 0.05.

Table 3 NR scores for various mask ratios on CelebA-HQ and Paris

	MODEL	20–40%	40–60%
NR	EC(ICCVW 2019) [29]	0.22	0.10
	CSA(ICCV 2019) [9]	0.38	0.18
	RFR(CVPR 2020) [32]	0.41	0.24
	PAGN	0.67	0.58

Table 4 FR scores for various mask ratios on CelebA-HQ and Paris

	MODEL	20–40%				40–60%			
		Rank1	Rank2	Rank3	Rank4	Rank1	Rank2	Rank3	Rank4
FR	EC(ICCVW 2019) [29]	0	0.22	0.27	0.58	0	0.05	0.50	0.55
	CSA(ICCV 2019) [9]	0	0.27	0.54	0.30	0	0.11	0.47	0.52
	RFR(CVPR 2020) [32]	0.30	0.23	0.30	0.19	0.05	0.88	0.13	0
	PAGN	0.69	0.30	0	0	0.94	0.05	0	0

No.1 means that the completed image is considered by volunteers to be the best and No.4 means the worst image among the images inpainted by different methods

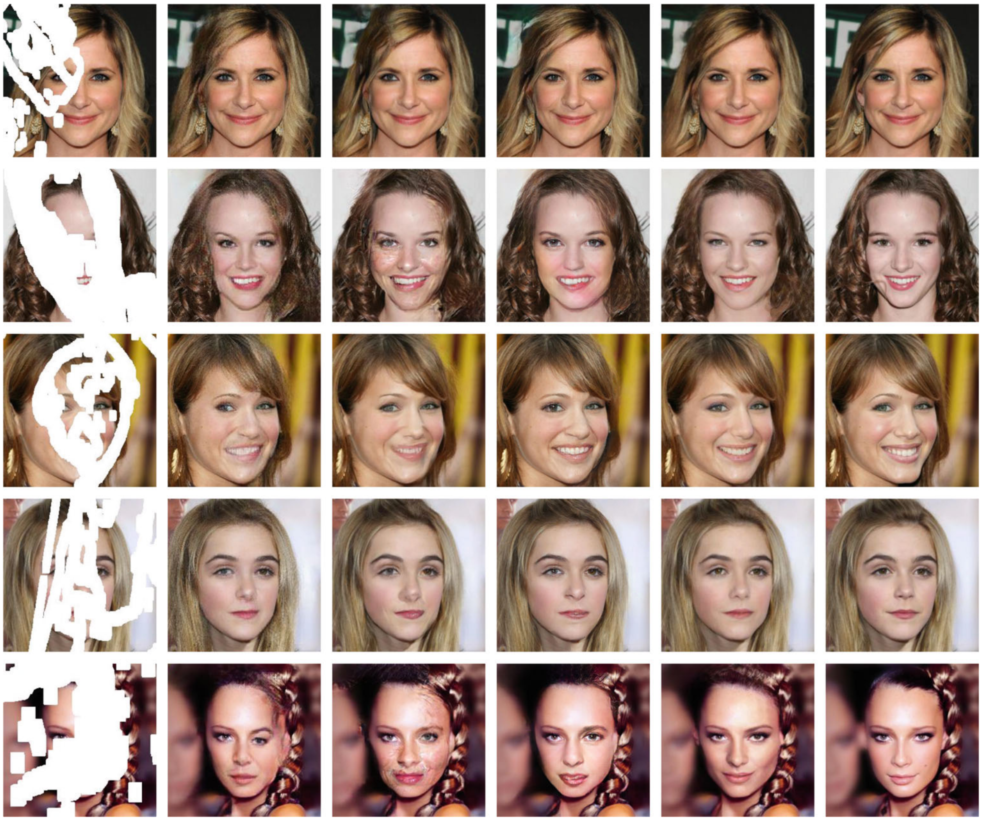
4.2 Quantitative comparisons

4.2.1 objective evaluation

The image inpainting task lacks reasonable objective evaluation metrics that can accurately reflect the image performance. However, when our method is compared with other methods, objective evaluation metrics are essential because they are relatively fair and independent of human will. Thus, the commonly used metrics - PSNR, SSIM and LPIPS- are adopted for image quantitative comparisons. PSNR measures the difference between the reference image and the predicted result based on the pixel-level errors, and the larger the value is, the less distorted the image is. SSIM evaluates the image in terms of luminance, contrast and structure, and its value range is [0,1]. The closer the value is to 1, the closer the image is to the reference image. LPIPS [5] based on AlexNet, measures the distance between the deep features of restored image and those of real image. The caculation of LPIPS is shown (16).

$$D(X_{gt}, X_{re}) = \sum_{i} \frac{1}{H_i \times W_i} \sum_{h,w} \|W_i \odot (f_{gt}^{i} - f_{re}^{i})\|_2 \qquad (16)$$

The specific algorithm is as follows. Firstly, we use the $relu1 \sim relu5$ layers from AlexNet to extract the feature stack f_{gt}^{i} and f_{re}^{i} from the ground truth and generated



(a) Input (b) EC [29] (c) CSA [9] (d) RFR [32] (e) PAGN (f) Ground truth

Fig. 6 Qualitative results on the CelebA-HQ dataset

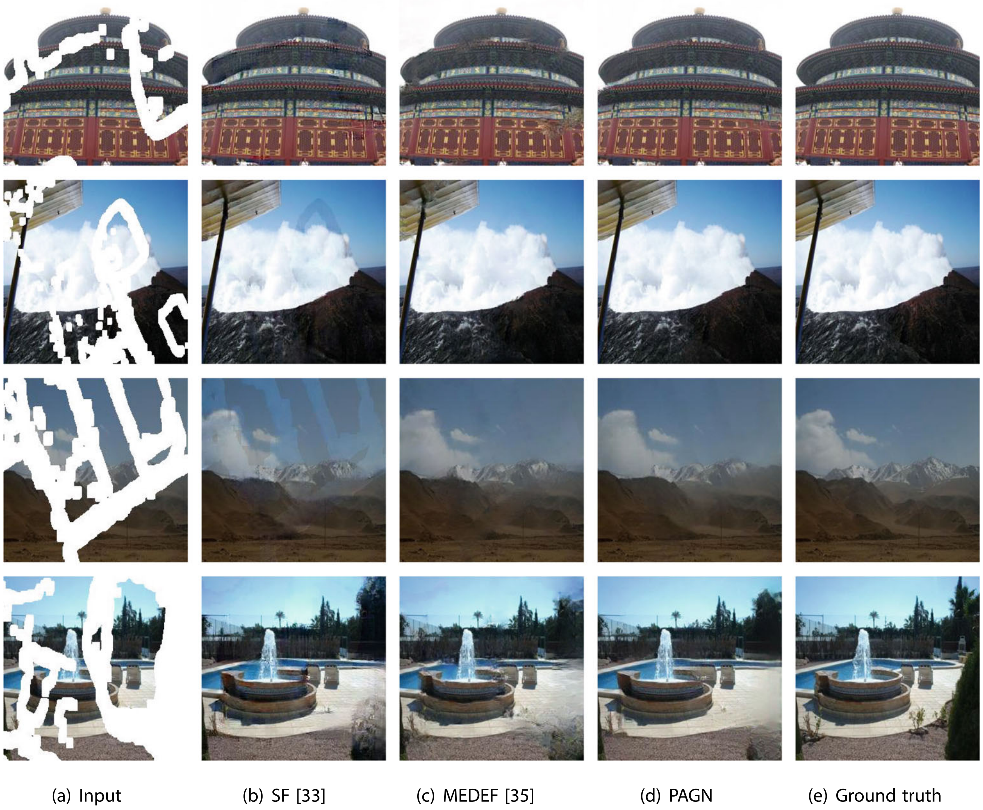

(a) Input (b) SF [33] (c) MEDEF [35] (d) PAGN (e) Ground truth

Fig. 7 Qualitative results on the Places2 dataset

image, respectively. And we assign f_{gt}^{i}, $f_{re}^{i} \in R^{H_i \times W_i \times C_i}$ for activation layer i (i=1∼5). Secondly, we scale the channels of each feature by vector W_i and compute the l_2 distance to get the difference map. The W_i consists of a $1*1$ convolution layer where the output channel is set 1. Finally, all stacked difference maps are averaged in the spatial dimension and accumulated in the channel dimension.

We compare our approach with several state-of-the-art models on the CelebA-HQ [11], Place2 [10] and Paris [23] datasets. These models are: EdgeConnect (ICCVW 2019) [29] and CSA (ICCV 2019) [9], SF (ICCV 2019) [33], RFR (CVPR 2020) [32] and MEDEF (ECCV 2020) [35]. During the training process, for the Paris dataset, three pretrained models have been adopted(EC (ICCVW 2019) [29], MEDEF (ECCV 2020) [35] and RFR (CVPR 2020) [32]) through their official websites. And the CSA model is trained by us by using the official code and default parameters. For the Places2 dataset, the pretrained model of EC (ICCVW 2019) [29], SF (ICCV 2019) [33] and MEDEF (ECCV 2020) [35] have been adopted through their official websites. And for CelebA-HQ dataset, since the official pretrained CelebA model of comparison methods [9, 29, 32] cannot be generalized to the CelebA-HQ dataset, we retrained them on the CelebA-HQ dataset by using the default parameters and official code. And we choose the best results as the final output. Additionally, to evaluate fairly, we conduct experiments on same irregular holes provided by Liu [16]. These masks are classified based on different hole-to-image area ratios(e.g.,0-10%,10-20%,etc.). During testing, the comparison methods and our methods are evaluated by the same test dataset and irregular masks.The results are shown in Tables 1 and 2.

4.2.2 Subjective evaluation

The objective evaluation may be inconsistent with the subjective perception of people. In this section, subjective evaluation, as an essential complement to objective evaluation indicators, is utilized to assess our proposed method.

According to whether the ground truth is required, we divided the user study into two types of experiments: no reference(NR) and full reference(FR). No reference means that users do not know what the real image looks like and where the mask is, and full reference is the opposite. Specifically, we invited 20 participants, both engaged and not engaged in image processing direction. In the first set of experiments, the participants are asked to whether the image is real in a bunch of random images that contain repaired images and ground truth. The results are summarized in Table 3, and the numbers in the table represent the probability that the completed images are considered as the real image. We can see that images inpainted by our method

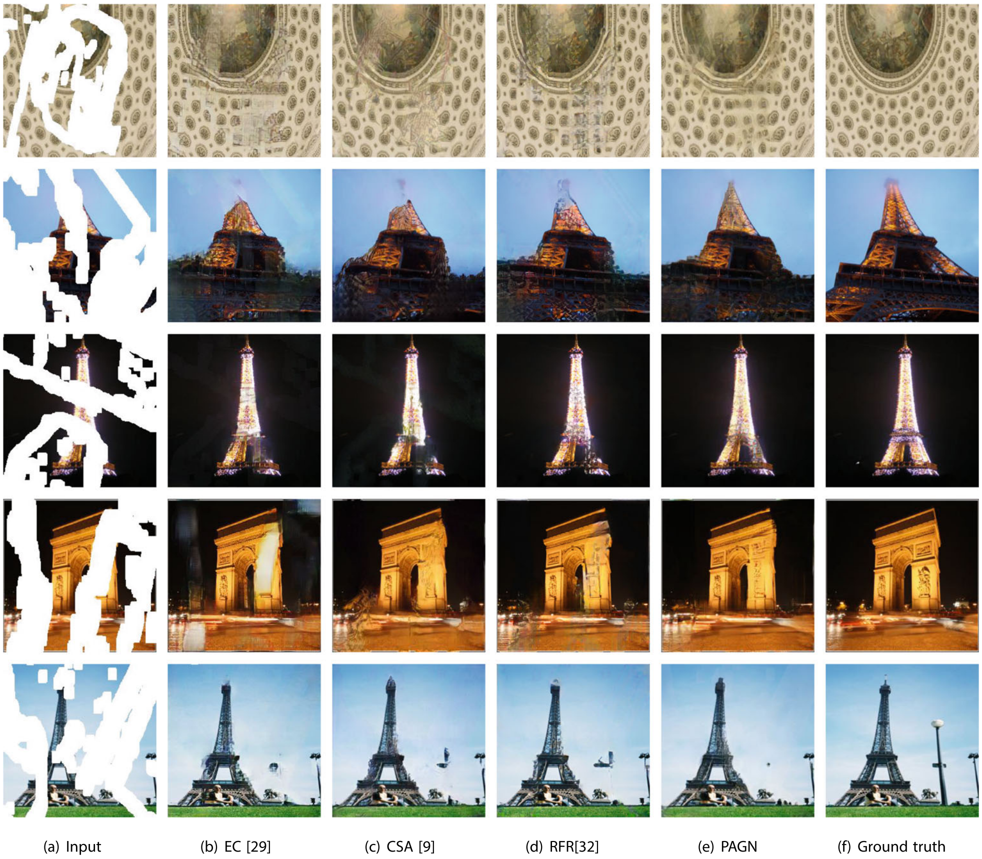


Fig. 8 Qualitative results on the Paris dataset

are always the most realistic under different mask ratios. In the second set of experiments, participants needed to rate the quality of displayed images, and these images are repaired from one broken image by our method and other comparison methods. Table 4 shows the results that sorted by participants, and the number in the table represents the probability that the image restored by a certain model is of that rank(No.1, No.2...). From Table 4, we can conclude that images completed by our method have the highest probability of being thought to be the best in human vision, indicating that our models can generate more natural images. Notably, images for subjective evaluation come from a portion of CelebA-HQ and Paris, and all images are shown to participants for no time limitation.

4.3 Qualitative comparisons

The qualitative comparison can test image quality in a intuitive way. Figures 6, 7 and 8 show the results of our method that compared with other state-of-the-art methods on the CelebA-HQ, Places2 and Paris datasets, respectively. Compared with other methods, the images repaired by our method have less noticeable discontinuous structures and blurry textures and are the most genuine in terms of the human visual system. In Fig. 6, we find that EC easily generate irrational structures, and CSA fails to generate clear textures. We guess that when the missing hole is large, EC is difficult to repair the edges of missing areas accurately, and CSA has insufficient information to understand the connection between pixels in the missing region. For RFR, although the use of recurrent feature reasoning is suitable for repairing large holes, the restored images still have some unnatural content.

Combining Fig. 6 with the Table 1, we observe the following two points. First, the higher PSNR value of an image does not mean better vision for human eye. For example, in the CSA method, repaired images look visually poor, but their PSNR value is not low. Therefore, there is a gap between objective indicator and subjective

Table 5 The effectiveness of PAGN

	Two-stage network	One-way guidance	w/o guidance filter	Our PAGN
PSNR	21.88	22.05	22.05	22.20
SSIM	0.82	0.83	0.83	0.84
LPIPS	0.1299	0.1114	0.1113	0.1077

These statistics are based on irregular masks with a size of 50%-60% of the entire image

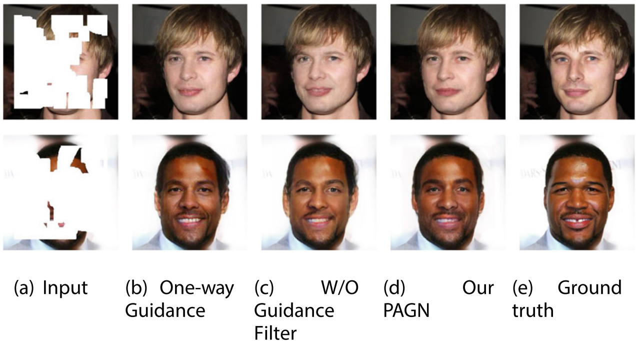

(a) Input (b) One-way Guidance (c) W/O Guidance Filter (d) Our PAGN (e) Ground truth

Fig. 9 Effect of the guidance filter on inpainting results, and the mask ratio is 50%-60%

perception. Second, as the missing area increases, the advantages of objective metrics tested by our model are gradually emerge. As the mask ratio increases and the information in known areas is reduced, other methods generate more apparently distorted pixels than our method, so both objective indicators and subjective vision are inferior to our model.

4.4 Ablation study

In this study, the quantitative metrics are calculated from validation images on the CelebA-HQ dataset [11].

4.4.1 The effectiveness of mutual adaptive guidance

To demonstrate the effectiveness of the idea of mutual adaptive guidance, we compare the PAGN with several variants: a two-stage network, a network with one-way guidance, a network without guidance filter. The two-stage network divides image inpainting into two stages without end-to-end training. The first stage is used to recover damaged structures, and the second stage would recover the textures based on recovered structures.The network with one-way guidance means that the guidance direction in the encoder is the same as that in the decoder. We use ablation experiments which are seen in Table 5 and Fig. 9 to show the inpainting performance of the full network and corresponding variants.

Table 6 The effectiveness of Joint-CAM,and the mask ratio is 50%-60%

	w/o cs-ca	Joint-CAM(different-level feature map as cross-scale map)	Our Joint-CAM
PSNR	21.94	22.02	22.20
SSIM	0.81	0.82	0.84
LPIPS	0.1141	0.1109	0.1077

The objective results are given in Table 5, demonstrating that the proposed PAGN performs better than other variants in terms of PSNR, SSIM, LPIPS. Figure 9 shows the visual results. By observing the mouth in the first row, the network without guidance filter generates the wrong semantic structures, and the network with one-way guidance produces the inconsistent color. The results in the second row show that our PAGN can produce less blurring around the hair and ear.

4.4.2 Effect of joint-contextual attention mechanism

To investigate the effectiveness of cross-scale contextual attention(cs-ca), we perform two groups of experiments to make a comparison. The first group only use in-scale contextual attention(is-ca) [25] in the texture enrichment branch. The visual comparison is shown in Fig. 10. By observing the left eye and mouth of the first black man on the top row and the ear of the second man on the bottom row, we can see that the model with Joint-CAM

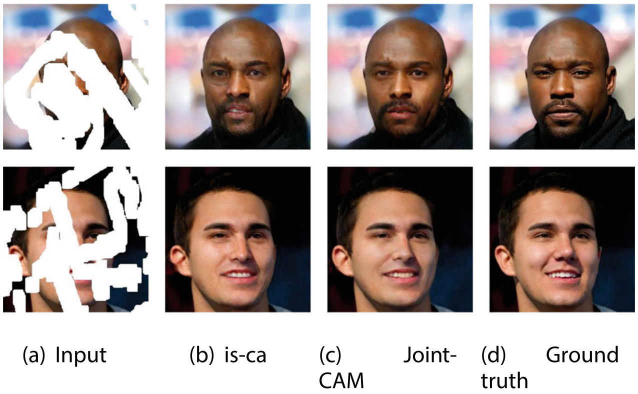
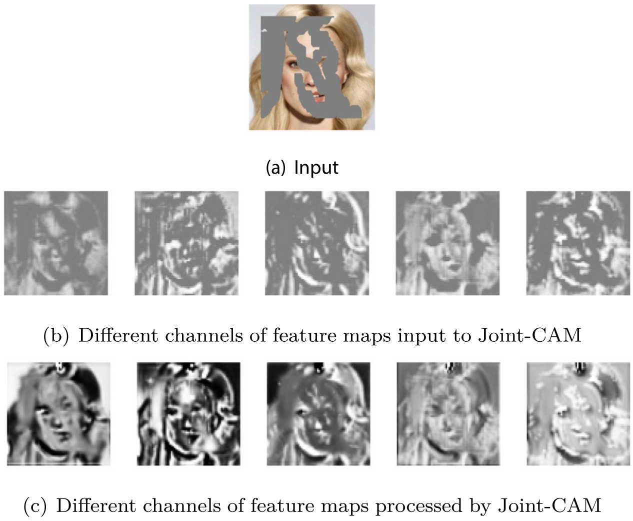

(a) Input (b) is-ca (c) Joint-CAM (d) Ground truth

Fig. 10 Results of different contextual attention


(a) Input

(b) Different channels of feature maps input to Joint-CAM

(c) Different channels of feature maps processed by Joint-CAM

Fig. 11 The visual contrast diagram of the Joint-CAM

Table 7 The effectiveness of AMPR and the mask ratio is 50%–60%

	w/o AMPR	w/o intralayer residual connection	w/o CBAM	AMPR
PSNR	21.48	21.95	22.12	22.20
SSIM	0.80	0.83	0.83	0.84
LPIPS	0.1244	0.1141	0.1124	0.1077

helps to recover accurate structures and realistic textures with fewer artifacts. In the next group, we adopt the previous high-level feature map instead of the downsampling map to explore the relationship between different-scale features. The comparison objective metrics are shown in Table 6.

In addition, we further demonstrate the effectiveness of our proposed Joint-CAM in the feature space. Since the gray-scale features maps are easier for observation, we display the input and output feature maps of Joint-CAM module in gray-scale. As shown in Fig. 11, we find that the feature maps processed by Joint- CAM have a more reasonable structures, clearer textures and brighter color.

4.4.3 Effective of attention-based multiscale perceptive Res2Block

AMPR is used to capture different-scale features and improve model generalization. We conducted experiments to evaluate the importance of different components of AMPR. Table 7 shows that both PSNR and SSIM are the highest with the addition of all components. The visual quality as shown in Fig. 12, as the mask ratio increases, the areas repaired by the network without AMPR have more obvious texture artifacts. Noteworthy, a convolutional layers are applied in the last of the encoder in the case of without AMPR.

To further validate the AMPR, we show the feature visualization comparison as shown in Fig. 13. Specifically, we up-sample the size of feature maps with different sources from 32×32 to 64 × 64 and display them in colors(COLORMAP_JET) for better observation. One is generated by the last layer of encoder which is consist of a convolutional layer. And the other is generated by the last layer of encoder which is consist of AMPR. We can find that the feature maps generated by AMPR have larger receptive fields than those of convolutional layer. Other than that, the localization of key points and missing regions is more accurately by applying the AMPR module.

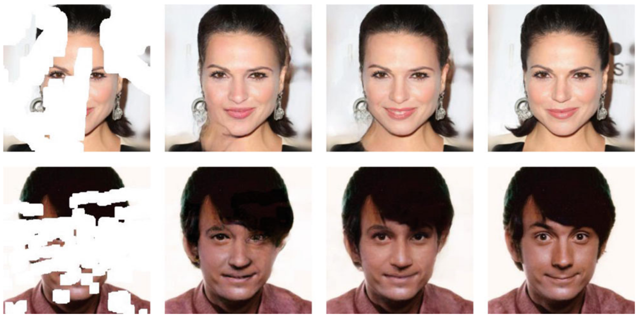
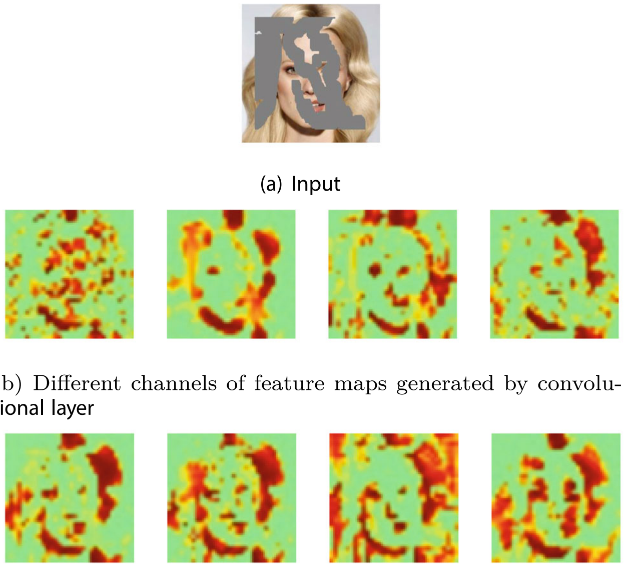

(a) Input (b) W/O AMPR (c) Add AMPR (d) Ground truth

Fig. 12 Visual comparison results with or without AMPR, and the mask ratio is 30%-40%. Among the images without AMPR, blurry textures exist in the neck, eyes.


(a) Input

(b) Different channels of feature maps generated by convolutional layer

(c) Different channels of feature maps generated by AMPR

Fig. 13 The visual contrast diagram of the AMPR

5 Conclusion and future work

In this paper, we introduce the parallel adaptive guidance network(PAGN), which repairs broken structures and textures separately in a parallel manner within one stage. Structural and texture features mutually guide through skip connections with guidance filters, which allows to pay attention on the communication of useful features. Furthermore, in the texture enrichment branch, we apply the joint-contextual attention mechanism(Joint-CAM), which leverages limited context from multiple perspectives, making it easier to yield details and accurate structures in missing regions. Finally, to give our inpainting network has robust feature representation, a novel multiscale structure called attention-based multiscale perceptual res2block(AMPR) is adopted into the bottleneck, to extract different-level features at a finer level. Experiments on the public datasets verified the effectiveness of our proposed models, which are especially suitable for repairing large holes.

In future work, we aim to combine inpainting tasks with super resolution to efficiently repair high-resolution images with complex textures and rich colors.

Acknowledgments Supported by National Natural Science Foundation of China (61901392 and 11872069), the Applied Basic Research Projects of Science and Technology Department of Sichuan Province(2021YJ0109), Central Government Funds of Guiding Local Scientific and Technological Development for Sichuan Province(2021ZYD0034), National,Ministry of Education “Chunhui Plan” Scientific Research Project (Z2017076), Chengdu Science and Technology Program(2016-YF04-00044-JH) and Xihua University Graduate Innovation Fund (YCJJ2021139 and YCJJ2021066).

References

1. Shao H, Wang Y, Fu Y, Yin Z (2020) Generative image inpainting via edge structure and color aware fusion. Signal Process Image Commun 87(115929)
2. Criminisi A, Perez P, Toyama K (2004) Region filling and object removal by exemplar -based image inpainting. IEEE TIP 13(9):1200–1212
3. Wang N, Wang W, Hu W, Fenster A, Li S (2021) Thanka Mural Inpainting Based on Multi-Scale Adaptive Partial Convolution and Stroke-Like Mask. IEEE TIP 30:3720–3733
4. Darabi S, Shechtman E, Barnes C, Goldman DB, Sen P (2012) Image melding: Combining inconsistent images using patch-based synthesis. ACM TOG 31(4):82
5. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Proceedings of the 2018 CVPR, pp 586–595
6. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) Patchmatch: a randomized correspondence algorithm for structural image editing. ACM ToG 28(3):24
7. Pathak D, Krahenbühl P, Donahue J, Darrell T, Efros AA (2016) Context Encoders: Feature Learning by Inpainting. In: Proceedings of the 2016 CVPR, pp 2536–2544
8. Iizuka S, Simo-Serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graph (TOG) 36(4):107
9. Liu H, Jiang B, Xiao Y (2019) Coherent Semantic Attention for Image Inpainting. In: Proceedings of the 2019 ICCV, pp 4169–4178
10. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: a 10 million image database for scene recognition. IEEE TPAM 40(6):1452–1464
11. Sagong MC, Shin YG, Kim SW, Park S, Ko SJ (2019) PEPSI : Fast Image Inpainting With Parallel Decoding Network. In: Proceedings of the 2019 CVPR
12. Christian S, Vincent V, Sergey I, Jonathon S, Zbigniew W (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 CVPR, pp 2818–2826
13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D (2014) Generative adversarial nets. In: Proceedings of the 2014 NeurIPS, pp 2672–2680
14. Zhang Q, Shen X, Xu L, Jia J (2014) Rolling Guidance Filter. In: Proceedings of the 2014 ECCV, pp 815–830
15. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the 2016 ECCV, pp 694–711
16. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B (2018) Image Inpainting for Irregular Holes Using Partial Convolutions. In: Proceedings of the 2018 ECCV, pp 85–100
17. Simonyan K, Zisserman A (2014), Very deep convolutional networks for Large-Scale image recognition. In: Proceedings of the 2014 ICLR
18. Yu F, Koltun V (2016) Multi-Scale Context aggregation by dilated convolutions. In: Proceedings of the 2016 ICLR
19. Gao H, Chen M, Zhao K, Zhang Y, Yang H, Torr P (2019) Res2net: A New Multi-Scale Backbone Architecture. IEEE TPAM 43(2):652–662
20. Isola P, Zhu J, Zhou T, Efros AA (2017) Image-to-Image Translation with Conditional Adversarial Networks. In: Proceedings of the 2017 CVPR, pp 5967–5976
21. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the 2018 ECCV, pp 801–818
22. Liu J, Jung C (2020) Facial image inpainting using attention-based multi-level generative network. Neurocomputing 437:95–106
23. Philbin J, Zisserman A The Paris Dataset, https://www.robots.ox.ac.uk/~vgg/data/parisbuildings/
24. Zeng Y, Fu J, Chao H, Guo B (2019) Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. In: Proceedings of the 2019 CVPR, pp 1486–1494
25. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T (2018) Generative Image Inpainting with Contextual Attention. In: Proceedings of the 2018 CVPR, pp 5505–5514
26. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T (2019) Free-Form Image Inpainting With Gated Convolution. In: Proceedings of the 2019 ICCV, pp 4470–4479
27. Chen Y, Liu L, Tao J, Xia R, Zhang Q, Yang K, Xiong J, Chen K (2021) The improved image inpainting algorithm via encoder and similarity constraint. Vis Comput 37:1691–1705
28. Xiong W, Yu J, Lin Z, Jiang J, Lu X, Barnes C, Luo J (2019) Foreground-Aware Image Inpainting(2019),in:Proceedings of the 2019 CVPR, pp 5833–5841
29. Nazeri K, Ng E, Joseph T, Qureshi F, Ebrahimi M (2019) EdgeConnect: Generative Image Inpainting With Adversarial Edge Learning[J]. In: Proceedings of the 2019 ICCVW
30. Wang Y, Tao X, Qi X, Shen X, Jia J (2018) Image inpainting via generative multi-column convolutional neural networks. Adv Neural Inf Process Syst:331–340
31. Guo Z, Chen Z, Yu T, Chen J, Liu S (2019) Progressive Image Inpainting with Full-Resolution Residual Network. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 2496–2504
32. Li J, Wang N, Zhang L, Du B, Tao D (2020) Recurrent Feature Reasoning for Image Inpainting. In: Proceedings of the 2020 CVPR, pp 7757–7765
33. Ren Y, Yu X, Zhang R (2019) StructureFlow: Image Inpainting via Structure-aware Appearance Flow. In: Proceedings of the 2019 ICCV, pp 181–190
34. Chen M, Liu Z, Ye L, Wang Y (2020) Attentional coarse-and-fine generative adversarial networks for image inpainting. Neurocomputing 405:259–269

35. Liu H, Jiang B, Song Y, Huang W, Yang C (2020) Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations. In: Proceedings of the 2020 ECCV, pp 725–741
36. Woo S, Park J, Lee JY CBAM: Convolutional block attention module (2018). In: Proceedings of the 2018 ECCV, pp 3–19
37. Zheng C, Cham TJ, Cai J (2021) Pluralistic Free-Form Image Completion. Int J Comput Vis
38. Li T, Dong X, Lin H (2020) Guided Depth Map Super-Resolution Using Recumbent Y Network. IEEE Access:122695–122708
39. Chen Y, Zhang H, Liu L, Chen X, Zhang Q, Yang K, Xia R, Xie J (2021) Research on image Inpainting algorithm of improved GAN based on two-discriminations networks. Appl Intell 51:3460–3474
40. Zhu M, He D, Li X, Li C, Li F, Liu X, Ding E, Zhang Z Image inpainting by end-to-end cascaded refinement with mask awareness. IEEE Trans Image Process:4855–4866
41. Kingma DP, Adam J. B. a. (2015) A Method for stochastic optimization. In: Proceedings of the 2015 ICLR
42. Liu S, Huang D, Wang Y (2018) Receptive Field Block Net for Accurate and Fast Object Detection. In: Proceedings of the 2018 ECCV, pp 404–419
43. Mei Y, Fan Y, Zhou Y, Huang L, Huang T, Shi H (2020) Image Super-Resolution With Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining. In: Proceedings of the 2020 CVPR, pp. 5689–5698.
44. Ding Y, Lin L, Wang L, Zhang M, Li D (2020) Digging into the multi-scale structure for a more refined depth map and 3D reconstruction. Neural Comput Appl 32:11217–11228
45. Wang C, Wu Y, Cai Y, Yao G, Wang ZH (2020) Single image deraining via deep pyramid network with spatial contextual information aggregation. Appl Intell 50:1437–1447
46. Karras T, Aila T, Laine S, Zhang M, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196

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

Jinyang Jiang received the B.S. degree in information and communication engineering from Chengdu College of Electronic Science and Technology, Chengdu, China, in 2019. She is currently pursuing the master degree in Xihua University, Chengdu, China. Her research interests include image processing and computer vision.

Xiucheng Dong received the B.S. and M.S. degrees from Chongqing University, China, in 1985 and 1990, respectively. He is currently a Professor with the School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, China. He is also the Dean of the Electrical Engineering and Electronic Information, Xihua University. His research interests include intelligent control,modeling of nonlinear systems, machine vision,and virtual reality.

Tao Li received the B.S. degree in electronics and information engineering and the Ph.D. degree in communication and information system from Sichuan University, Chengdu, China, in 2005 and 2017, respectively. She joined the School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, in 2017. Her research interests include image/video compression and restoration, image/video super-resolution, and computer vision.

Fan Zhang received the B.S. degree in information engineering and the M.S. degree in signal and information processing from Xihua University, Chengdu, China, in 2010 and 2013, respectively. He is currently a Ph.D candidate of Sichuan University. His research interests include 3D reconstruction, deep generative model and the theory of deep learning.

Hongjiang Qian received the B.S. degree in electrical engineering from Xihua University, Chengdu, China, in 2019. he is currently pursuing the master degree in the same university, Chengdu, China. His research interests include machine vision and intelligent control.

Guifang Chen received the B.S. degree in electrical engineering from Xihua University, Chengdu, China, in 2019. She is currently pursuing the master degree in the same university, Chengdu, China. Her research interests include machine learning and fault diagnosis.