Heritage Science

RESEARCH

Open Access

Ancient mural inpainting via structure information guided two-branch model

Xiaochao Deng[1] and Ying Yu[1*]

Abstract

Ancient murals are important cultural heritages for our exploration of ancient civilizations and are of great research value. Due to long-time exposure to the environment, ancient murals often suffer from damage (deterioration) such as cracks, scratches, corrosion, paint loss, and even large-region falling off. It is an urgent work to protect and restore these damaged ancient murals. Mural inpainting techniques refer to virtually filling the deteriorated regions by reconstructing the structure and texture elements of the mural images. Most existing mural inpainting approaches fail to fill loss contents that contain complex structures and diverse patterns since they neglect the importance of structure guidance. In this paper, we propose a structure-guided two-branch model based on the generative adversarial network (GAN) for ancient mural inpainting. In the proposed model, the mural inpainting process can be divided into two stages: structure reconstruction and content restoration. These two stages are conducted by using a structure reconstruction network (SRN) and a content restoration network (CRN), respectively. In the structure reconstruction stage, SRN employs the Gated Convolution and the Fast Fourier Convolution (FFC) residual block to reconstruct the missing structures of the damaged murals. In the content restoration stage, CRN uses the structures (generated by SRN) to guide the missing content restoration of the murals. We design a two-branch parallel encoder to improve the texture and color restoration quality for the missing regions of the murals. Moreover, we propose a cascade attention module that can capture long-term relevance information in the deep features. It helps to alleviate the texture-blur and color-bias problem. We conduct experiments on both simulated and real damaged murals, and compare our inpainting results with other four competitive approaches. Experimental results show that our proposed model outperforms other approaches in terms of texture clarity, color consistency and structural continuity of the restored mural images. In addition, the mural inpainting results of our model can achieve comparatively high quantitative evaluation metrics.

Keywords Ancient mural inpainting, Generative adversarial network, FFC residual block, Cascade attention module

Introduction

Ancient mural paintings are precious human cultural heritages, which record lots of historical, cultural, religious and artistic information, and vividly depict the social and religious stories of various ethnic groups in a certain historical period. Due to the exposure to the environment and the impact of human activities over hundreds or thousands of years, most of these ancient murals have suffered from various damages and degradations such as flaking, cracks, corrosion, paint loss, sootiness, aging, microorganism damage, scratches and many other forms of diseases [1]. These diseases may reduce the cultural and artistic values of ancient murals and even destroy the integrity of the mural contents. Therefore, the protective repair of ancient murals has become an urgent work for cultural heritage protection communities.

Ancient murals, which were created in different historical periods, vary greatly in color, style, and painting

*Correspondence:
Ying Yu
yuying.mail@163.com
[1] School of Information Science and Engineering, Yunnan University, Kunming 650500, China

techniques. Many murals have abundant contents including Buddha statues, architectures, decorative patterns, landscapes, dancers, silks, animals, etc. These makes the task of mural restoration more challenging. Existing physical protection measures and traditional manual inpainting work of ancient murals are very difficult and time-consuming. These protective operations may cause irreversible damage to the mural heritages. Benefiting from advances of computer technology, digital mural inpainting can virtually restore the visual appearance of ancient murals without intruding the original. The restored mural images can not only serve as references for the physical repairing, but also provide a permanent and replicable database for the mural cultural heritage.

Mural inpainting aims to fill the missing or damaged areas with realistic and fine-detailed contents by matching, copying, diffusing and other operations based on the information of the known areas. Traditional inpainting methods mainly includes the geometry-based methods and the patch-based methods. The geometry-based methods mainly use partial differential equations [2] to diffuse the structure information from the exterior to the interior of the missing hole. Jaidilert et al. [3] used different variational inpainting methods to inpaint the Thai murals. Chen et al. [4] improved the diffusion term of the curvature-driven diffusions algorithm, and introduced an adaptive control strategy and a smooth function. The geometry-based methods are generally suitable for repairing narrow and long cracks or scratches, but they do not perform well on large missing areas. The patch-based methods [5] fill the deteriorated regions by matching the most similar candidate patches from the known mural regions. Jiao et al. [6] proposed an improved block matching algorithm for Wutai Mountain murals. Cao et al. [7] proposed the adaptive sample block and local search algorithm to restore the flaking deterioration. Wang et al. [8] employed the line-drawings to maintain the structure coherence, and selected the target patches from a sparse representation model. The patch-based methods can produce satisfactory results for relatively large damaged areas. However, it cannot generate contents outside the undamaged mural areas, and sometimes results in block matching errors and inconsistent structures.

With the development of deep learning, a number of advanced natural image inpainting approaches built on deep convolutional neural networks [9] and generative adversarial networks [10] have achieved outstanding results on publicly available datasets. These approaches can adaptively capture the potential features in natural images through the learning process of massive data, and then generate the missing content of a damaged image. It has been proven that the deep learning-based approaches can produce more reliable results than traditional methods. In recent years, virtual mural restoration researchers started to utilize deep learning techniques to tackle with the mural inpainting problems. Wang et al. [11] proposed a Thanka mural inpainting method based on multi-scale adaptive partial convolution and stroke-like masks. Cao et al. [12] proposed a consistency enhanced generative adversarial network to restore Wutaishan murals. Lv et al. [13] proposed two generators connected image restoration networks to restore Dunhuang murals. Schmidt et al. [14] combined image super-resolution and deblurring techniques to restore the deteriorated cave paintings. Li et al. [15] proposed a generation-discriminator network model that mainly repaired damaged murals with dot-like defects. Yu et al. [16] adopted end-to-end networks with partial convolutional to repair the Dunhuang Grottoes Paintings, and designed two types of masks for simulating deteriorations. Li et al. [17] applied manual line-drawings for the missing region to guide the inpainting of damaged areas. Inspired by the image-making process from an artist's perspective, Ciortan et al. [18] proposed a multi-stage mural inpainting network based on "lines first, color palette after, color tones at last", and used four random-walk masks to imitate various degradations.

Existing deep learning-based approaches have achieved relatively good mural inpainting results. However, it is still a difficult and challenging task for some cases when recovering those heavily damaged regions with complex semantic structures and textures. Firstly, many mural images contain various damages and degradations, which will seriously affect the feature extraction of mural data in the training process. Secondly, the murals were created by multiple artists in different periods, and thus they have various painting styles. The CNN-based repair results will inevitably produce color disharmony between the restored regions and other mural areas. Thirdly, many ancient murals have large and complex damaged areas, and the original information was lost in history. This requires a mural inpainting method that can generate the missing contents that are consistent with the overall mural in artistic style, semantic perception and texture distribution.

It has been noticed that the structure information is very important for the reconstruction of the missing mural contents. In this paper we propose a structure-guided two-branch (SGTB) mural inpainting model to obtain high-quality mural restoration results. The proposed model is a two-stage generative adversarial network. The first-stage network generates the missing structures of the damaged murals. The second stage network leverages the generated structures to guide the restoration of the missing mural contents. The proposed two-stage model can achieve outstanding performance in the restoration of the structure, texture and color of the damaged murals. The main contributions of this paper are summarized as follows: (1) We build an ancient mural image dataset by collecting 3466 high-quality ancient mural images and expanding the number of these mural images to 10,398 by use of data augmentation techniques. (2) We employ the gated convolutions to extract low-level image features, and introduce the FFC residual blocks to capture the global context features of the mural images. (3) We propose a two-branch content restoration network. In the top-branch network, the encoder enlarges the size of the receptive field through layer-by-layer dilation gated convolutions and FFC residual blocks. In the bottom-branch network, the encoder focuses on the deep background features of interests. (4) We propose a cascaded attention module to refine the valid features of long-term information through channel-spatial interactions.

Proposed method

Ancient mural images usually contain complex structure, rich texture and color information. It is very difficult to restore the missing regions of the damaged mural images. It has been noted that most of the image information consists in the image structures. Therefore, the image structure information might play an important role in guiding the restoration process of the missing image contents. Motivated by this point, we propose a structure-guided two-branch (SGTB) model for inpainting the damaged ancient murals. Figure 1 shows the implementation process and network architecture of the proposed model. Figure 1a illustrates the implementation process, whereas Fig. 1b illustrates the details of network architecture. In the model, the mural inpainting process are divided into two stages: the structure reconstruction network (SRN) and the content restoration network (CRN). In the first stage, SRN predicts the structure map of a damaged mural image. By using the predicted structure map as guiding information, CRN performs the content restoration of the damaged murals in the second stage. Our two-stage networks decompose the restoration task of a damaged mural into two sub-tasks, i.e., the structure reconstruction and the content restoration. Each stage network is responsible for a specific inpainting task.

Given a ground truth mural I_{gt}, we first convert it into a grayscale mural I_{gray}. Then we combine the grayscale mural I_{gray} with a binary mask M to obtain

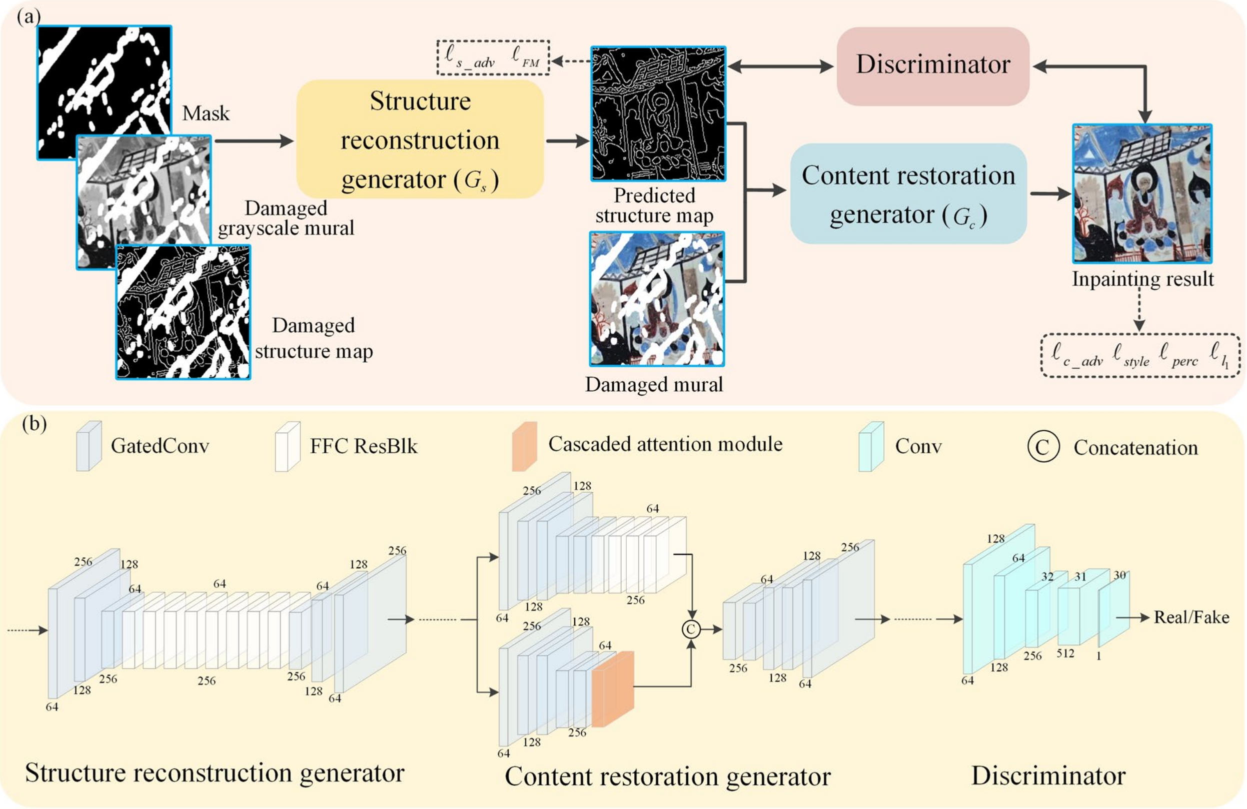
(a)
Mask
Damaged grayscale mural
Damaged structure map
Structure reconstruction generator (G_s)
ℓ_{s_adv} ℓ_{FM}
Predicted structure map
Damaged mural
Discriminator
Content restoration generator (G_c)
Inpainting result
ℓ_{c_adv} ℓ_{style} ℓ_{perc} ℓ_{l_1}
(b)
GatedConv
FFC ResBlk
Cascaded attention module
Conv
C Concatenation
Real/Fake
Structure reconstruction generator
Content restoration generator
Discriminator

Fig. 1 The implementation process and network architecture of the proposed model

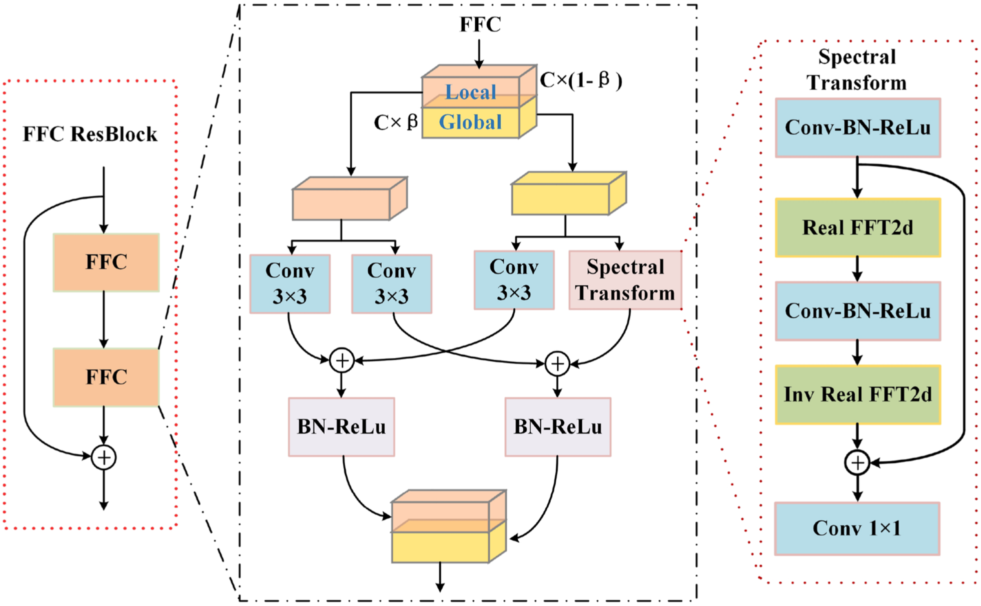

Fig. 2 FFC residual block

a masked grayscale mural by using the operation as $gray_{\mathrm{masked}} = I_{\mathrm{gray}} \odot (1 - M)$, where M indicates the damaged region that needs to be restored, and $\odot$ denotes the Hadamard product. We use the Canny edge detection operator [19] to extract the structure map S_{gray} of the grayscale mural I_{gray}, and then obtain the damaged structure map by using the operation as $S_{\mathrm{masked}} = S_{\mathrm{gray}} \odot (1 - M)$. First, we concatenate M, $gray_{\mathrm{masked}}$ and S_{masked} on the channel dimension, and feed it into the structure reconstruction generator (G_s). G_s focuses on predicting the structure of missing mural areas. The predicted structure map S_{out} is that:

$$S_{\mathrm{out}}=G_s\left(S_{\mathrm{masked}}, I_{\mathrm{gt}}, M\right) \tag{1}$$

Then, we concatenate the S_{out} with the damaged mural I_{masked} as the input of the two branches of the content restoration generator (G_c). G_c utilizes the predicted structure map and non-damaged regions to restore the missing contents of the damaged mural. The final inpainting results I_{out} is that:

$$I_{\mathrm{out}}=G_c(S_{\mathrm{out}}, I_{\mathrm{masked}}, M) \tag{2}$$

In each stage, the discriminator is responsible for distinguishing whether the structure maps and the mural images are authentic or generated by G_s and G_c. In addition, we propose a cascaded attention module to further alleviate the texture-blur and color-bias problem of the inpainting result. The proposed model can generate ancient mural images with continuous semantics, clear texture and lifelike colors under the guidance of the predicted structure map.

Structure reconstruction network (SRN)

Large missing areas will easily lead to disordered structure and poor consistency of the restored murals. As the structure of murals can assist the repair of the missing areas, we design a structure reconstruction network, also

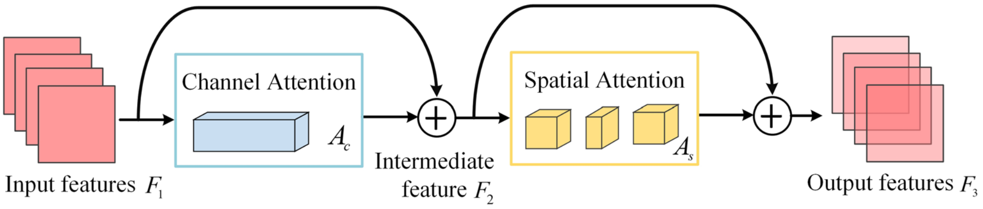

Fig. 3 The overview of cascaded attention module

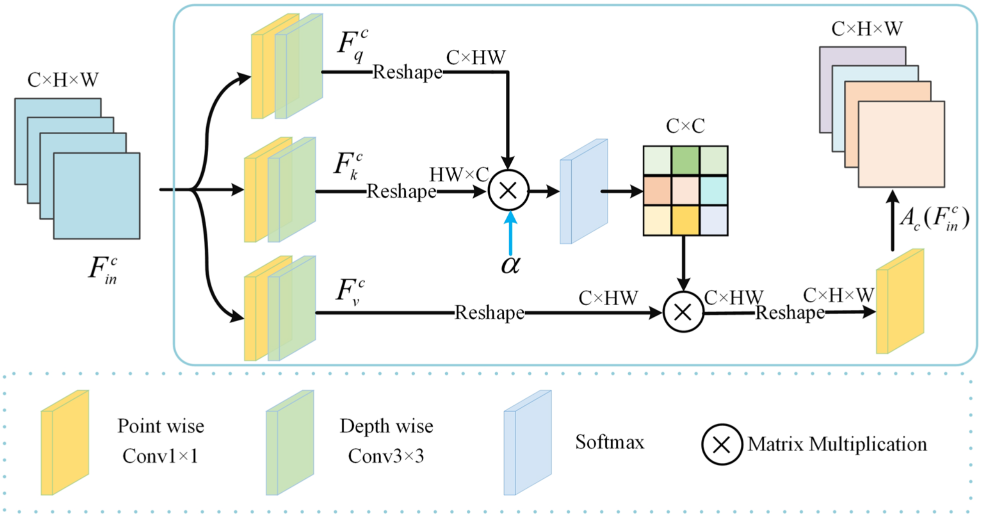

Fig. 4 Channel attention sub-module

referred to as the structure reconstruction generator in this work, to reconstruct (predict) the complete structure characteristics, which greatly improves the inpainting quality for the damaged mural with large holes.Taking the damaged structure map, the damaged grayscale mural and the mask as input, the structure reconstruction network generates a predicted structure map.

The structure reconstruction network consists of three down-sampling encoders, eight FFC residual blocks and three up-sampling decoders. The encoder contains a gated convolution (GatedConv), a normalization layer (Instance Normalization), and a ReLU activation function. Note that the structural information of a mural image is very sparse. In order to obtain a large receptive field, we set the down-sampling gated convolution with the kernel size of 7 × 7, 5 × 5, 3 × 3, respectively. We employ a normalization layer after each gated convolution layer to improve the training speed and the stability of the model, and we use a ReLU activation function to increase the fitting ability of the network.

Notice that there is less available structure information for the damaged murals with large missing areas. We consider utilizing the FFC blocks to capture the global context information in the early layer of the network as much as possible. The Fast Fourier Convolution (FFC) [20] is based on a channel-wise Fast Fourier Transform (FFT) [21], which splits all input channels into local and global branches. The local branch performs a local update of the feature map by using a vanilla convolution with a kernel size of 3 × 3. The global branch performs a Fourier transform of the feature map and updates the feature in the spectral domain. It can obtain the global context information of the mural structures. The specific implementation steps are as follows: (1) applies Real *FFT2d* to the input feature map, and concatenates real and imaginary parts across channel dimension:

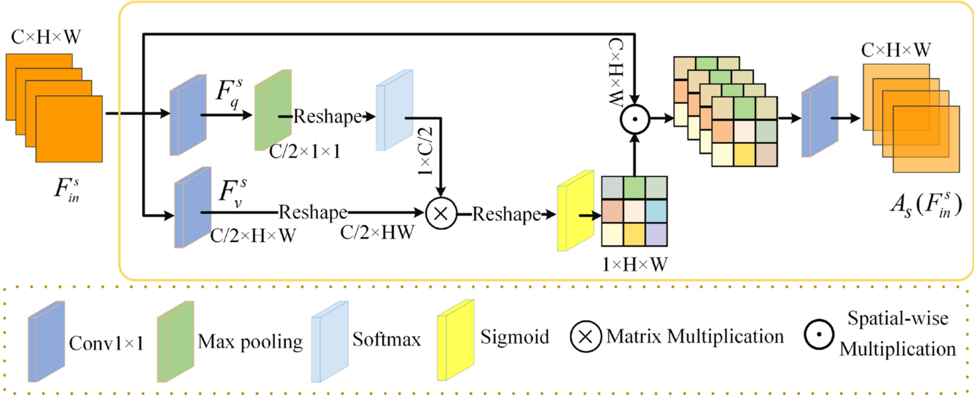

Fig. 5 Spatial attention sub-module

$$\mathbb{R}^{H\times W\times C} \xrightarrow{FFT2d} \mathbb{C}^{H\times \frac{W}{2}\times C} \xrightarrow{concat} \mathbb{R}^{H\times \frac{W}{2}\times 2C}$$

(2) applies convolution block in frequency domain:

$$\mathbb{R}^{H\times \frac{W}{2}\times 2C} \xrightarrow{Conv1\times 1\rightarrow BN\rightarrow ReLu} \mathbb{R}^{H\times \frac{W}{2}\times 2C}$$

(3) applies inverse Fourier transform to recover a spatial structure:

$$\mathbb{R}^{H\times \frac{W}{2}\times 2C} \xrightarrow{concat} \mathbb{C}^{H\times \frac{W}{2}\times C} \xrightarrow{iFFT2d} \mathbb{R}^{H\times W\times C}$$

Finally, the local branch of vanilla convolution and the global branch of FFT are fused to obtain a larger receptive field. We use the residual structure of FFC to avoid the gradient disappearance and explosion problems in the deep network. The FFC residual blocks is shown in Fig. 2. The decoder restores the feature information to the size of the original image through three up-sampling gated convolutions.

Content restoration network (CRN)

The goal of the second stage is to restore the texture details and color information of the mural based on the predicted structure map. The input to the content restoration network is the damaged mural image and the predicted structure map that is obtained in the first stage. In the content restoration network, we introduce two-branch parallel encoders and then merge them into a single decoder to achieve the inpainting of the murals. In the encoder of the top-branch, we employ the gated convolutions to extract the low-level feature of the murals, and then use the FFC residual blocks to obtain global feature information. Moreover, the cascaded attention module is employed in the encoder of the bottom-branch, which can flexibly capture long-term feature information of the murals. The decoder has a similar network structure to the encoder. ELU activation function is used in each layer to speed up the learning process in the deep neural network. The last up-sampling layer uses a Tanh activation function and converts the channel of the feature map to three channels. The details of the content restoration generator are shown in Fig. 1.

It should be noted that the mural inpainting needs long-term dependency [22] and multi-scale context information to generate the realistic mural image. However, the neural network can hardly capture the long-range relevance of the features. In order to obtain realistic mural inpainting results, we propose a cascaded attention module to refine the relevant feature information of the non-damaged regions. This module is capable of aggregating long-range pixel correlation and multi-scale context information among the spatial and the channel dimension. The overview of the cascaded attention module is illustrated in Fig. 3. We take the tensor $F_1 \in \mathbb{R}^{H\times W\times C}$ as the input feature map of the cascaded attention module. The intermediate feature map F_2 and the output feature map F_3 are defined as

$$F_2 = A_c(F_1) \oplus F_1 \tag{3}$$

$$F_3 = A_s(F_2) \oplus F_2 \tag{4}$$

where A_c and A_s are the channel and spatial attention maps, respectively. $\oplus$ denotes element-wise addition.

Table 1 Discriminative network structure

Layer	Kernel	Stride	Output(H × W × C)	Act-Func
Conv-0	4 × 4	2	128 × 128 × 64	Leaky-ReLu
Conv-1	4 × 4	2	64 × 64 × 128	Leaky-ReLu
Conv-2	4 × 4	2	32 × 32 × 256	Leaky-ReLu
Conv-3	4 × 4	1	31 × 31 × 512	Leaky-ReLu
Conv-4	4 × 4	1	30 × 30 × 1	Sigmoid

The channel attention sub-module can aggregate local and non-local pixel interactions. Specially, we generate the global attention map by calculating the cross-covariance of channels. The depth-wise convolution is employed to emphasize local context information. Figure 4 shows the details of the channel attention sub-module. We take the feature map F_{in}^{c} (extracted from the previous layer) as the input tensor, and compute the query $F_q^{\text{c}} = \text{C}_p^q\text{C}_d^q F_{\text{in}}^{\text{c}}$, the key $F_k^{\text{c}} = \text{C}_p^k\text{C}_d^k F_{\text{in}}^{\text{c}}$ and the value $F_v^{\text{c}} = \text{C}_p^v\text{C}_d^v F_{\text{in}}^{\text{c}}$. The $\text{C}_p^{(\cdot)}$ and $\text{C}_d^{(\cdot)}$ are 1 × 1 point-wise convolution and 3 × 3 depth-wise convolution, respectively. We reshape F_q^{c} and F_k^{c} to $\mathbb{R}^{C\times HW}$ and $\mathbb{R}^{HW\times C}$, respectively, and perform matrix multiplication on them. Then we use a Softmax function to generate a transposed-attention map A^c (in size of $\mathbb{R}^{C\times C}$). The parameter α is initially set as 0. In the training process, we gradually increase its value to enhance the relevance of generated information. Finally, A^c is multiplied with F_v^{c}, and pass a 1 × 1 point-wise convolution to form the channel attention feature map $A_c(F_{\text{in}}^{\text{c}})$. This computation process can be formulated as

$$A_c(F_{\text{in}}^{\text{c}}) = C_p\left(F_v^{\text{c}} \otimes \text{Softmax}(\alpha \cdot F_k^{\text{c}} \otimes F_v^{\text{c}})\right) \tag{5}$$

Figure 5 shows the details of the spatial attention sub-module. We use the standard 1 × 1 convolution layer on the input feature map $F_{\text{in}}^{\text{s}} \in \mathbb{R}^{C\times H\times W}$ to get $\left\{F_q^{\text{s}}, F_v^{\text{s}}\right\} \in \mathbb{R}^{C/2\times H\times W}$. Then, we employ the max-pooling and the Softmax operation on F_q^{s}. The output feature performs dot-product with F_v^{s}. Next, the attention map A^s is generated through a Sigmoid function. Finally, A^s performs a spatial-wise multiplication with F_{in}^{s} and pass a standard 1 × 1 convolution to form the spatial attention feature map $A_s\left(F_{\text{in}}^{\text{s}}\right)$:

$$A_s\left(F_{\text{in}}^{\text{s}}\right) = \text{conv}(F_{\text{in}}^{\text{s}} \odot \text{sigmoid}(A^S)) \tag{6}$$

$$A^s = \text{softmax}(\text{maxpool}(F_q^{\text{s}})) \otimes F_v^{\text{s}} \tag{7}$$

Discriminative network (DN)

During the training process, the discriminator judges whether the generated image is true or false, and feeds the judgement results to the generator for model optimization. Thus, the generator can involve to generate more natural and realistic images. Although the two generators (i.e., the structure reconstruction generator and the content restoration generator) are different, the purpose of both discriminators is to distinguish between generated images and ground truth. Thus, in two stages, we use the same Path GAN [23] as the underlying architecture of the discriminative network. For a 256 × 256px mural image, Path GAN can discriminate whether the 70 × 70 overlapping image patches are realistic. The specific structure of the proposed discriminative network is shown in Table 1.

Loss function

This paper aims to repair the damaged areas of ancient mural images. We perform the training of our proposed SGTB model by using a two-stage generative adversarial network (GAN). The GAN includes a generative network and a discriminative network. The generative network attempts to synthesize the mural contents that are reasonable and realistic. The discriminator attempts to distinguish whether an image is true or false. Through continuous training of the network, the synthesized mural images will gradually look realistic. GAN will achieve the optimal result when the following formula is satisfied:

$$\begin{aligned}\min_G \max_D V(D,G) &= \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] \\ &+\mathbb{E}_{z\sim P_{\text{out}}(z)}[\log(1-D(G(z)))]\end{aligned} \tag{8}$$

where x is the input data, and z denotes the noise. In this paper, the noise in the generative adversarial network comes from the mask. $P_{\text{data}}(x)$ is the probability distribution of the input mural images, whereas $P_{\text{out}}(z)$ represents the probability distribution of synthesized mural images. G is the generator whereas D is the discriminator.

In the first stage, the loss function is designed to guide the model to generate the structure information of the missing mural region. The loss function is composed of the adversarial loss $\ell_{\text{s\_adv}}$, and the feature-matching loss [24] ℓ_{FM}:

$$\ell_{\text{s\_G}} = \lambda_{\text{s\_adv}}\ell_{\text{s\_adv}} + \lambda_{\text{FM}}\ell_{\text{FM}} \tag{9}$$

where $\lambda_{\text{s\_adv}}$ and λ_{FM} are the weights of the adversarial loss and the feature-matching loss, respectively. The feature-matching loss compares the activation maps with those from the pre-trained VGG-19 network [25] in the intermediate layers of the discriminator, which is defined as

$$\ell_{\text{FM}} = \mathbb{E}\left[\sum_{i=1}^{m}\frac{1}{N_i}\left\|D^{(i)}(S_{\text{gt}}) - D^{(i)}(S_{\text{out}})\right\|_1\right] \tag{10}$$

where m is the number of the convolution layers of the discriminator. N_i is the number of the characteristic diagrams in the ith activation layer. $D^{(i)}$ is the feature map in the ith layer of the discriminator. S_{gt} is the structure map of the grayscale mural. S_{out} is the predicted structure map of the mural image.

In the second stage, the loss function is designed to guide the model to restore the missing content with clear textures and realistic colors. It needs to learn low-level pixel information and high-level semantic features of the mural images. For the content inpainting, the loss function consists of the l_1 loss (ℓ_{l_1}), adversarial loss ($\ell_{\text{c\_adv}}$), perceptual loss (ℓ_{perc}) [26], and style loss (ℓ_{style}) [27], which is defined as

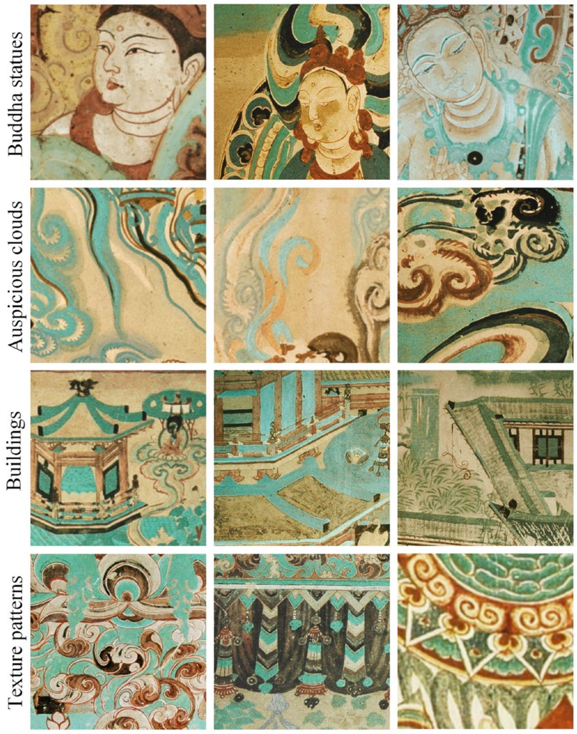

Fig. 6 Some examples of the ancient mural dataset

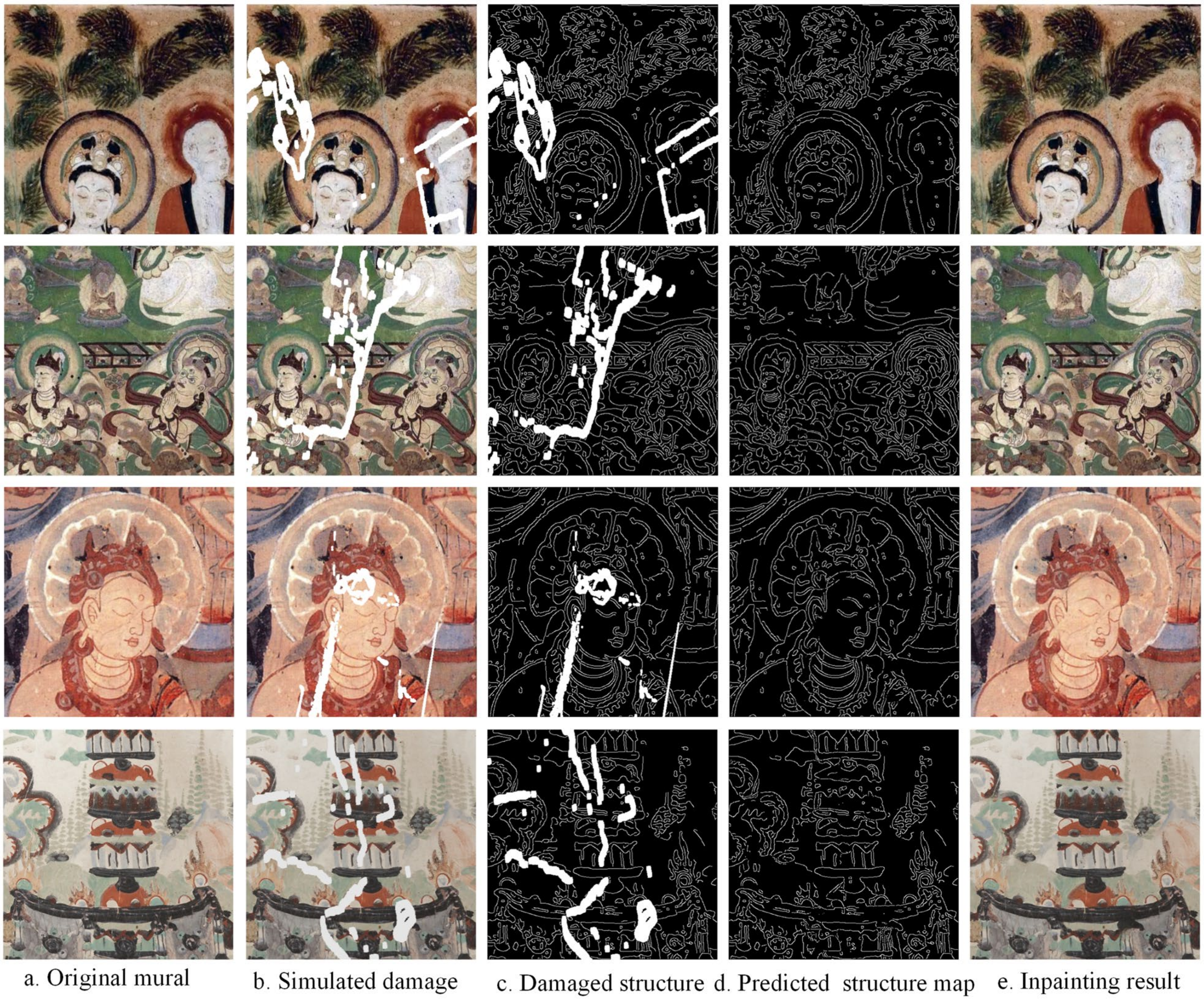

Fig. 7 Inpainting results of our proposed model on simulated damaged murals

$$\ell_{\mathrm{c\_G}} = \lambda_{l_1}\ell_{l_1} + \lambda_{\mathrm{c\_adv}}\ell_{\mathrm{c\_adv}} + \lambda_{\mathrm{perc}}\ell_{\mathrm{perc}} + \lambda_{\mathrm{style}}\ell_{\mathrm{style}} \tag{11}$$

The l_1 loss is used to measure the pixel-level difference between the real mural image and the synthetic image, which is calculated by

$$\ell_{l_1} = \|I_{\mathrm{out}} - I_{\mathrm{gt}}\|_1 \tag{12}$$

Mural images contain a lot of semantic-structure information and color-texture information. Besides l_1 loss and adversarial loss, we also introduce perceptual loss and style loss to improve the quality of the restored mural image. We compare the feature maps $\Psi_i(I_{\mathrm{gt}})$ of the real mural images I_{gt} from the ith pooling layer with the feature maps $\Psi_i(I_{\mathrm{out}})$ of the restored mural images I_{out}. The perceptual loss is calculated as

$$\ell_{\mathrm{perc}} = \mathbb{E}\left[\sum_{i=1}^{N} \left\|\Psi_i(I_{\mathrm{out}}) - \Psi_i(I_{\mathrm{gt}})\right\|_1\right] \tag{13}$$

where Ψ_i is the feature map of the i^{th} layer of the pre-trained VGG-19 networks. The style loss is used to calculate the L1 distance between the Gram matrix of the synthesized mural image and the real mural image. Assuming that the size of the ith layer feature map is $C_i \times H_i \times W_i$, $G_j^{\Psi}(\cdot)$ is a $C_j \times C_j$ Gram matrix that is constructed by feature map Ψ_j. The style loss is calculated as

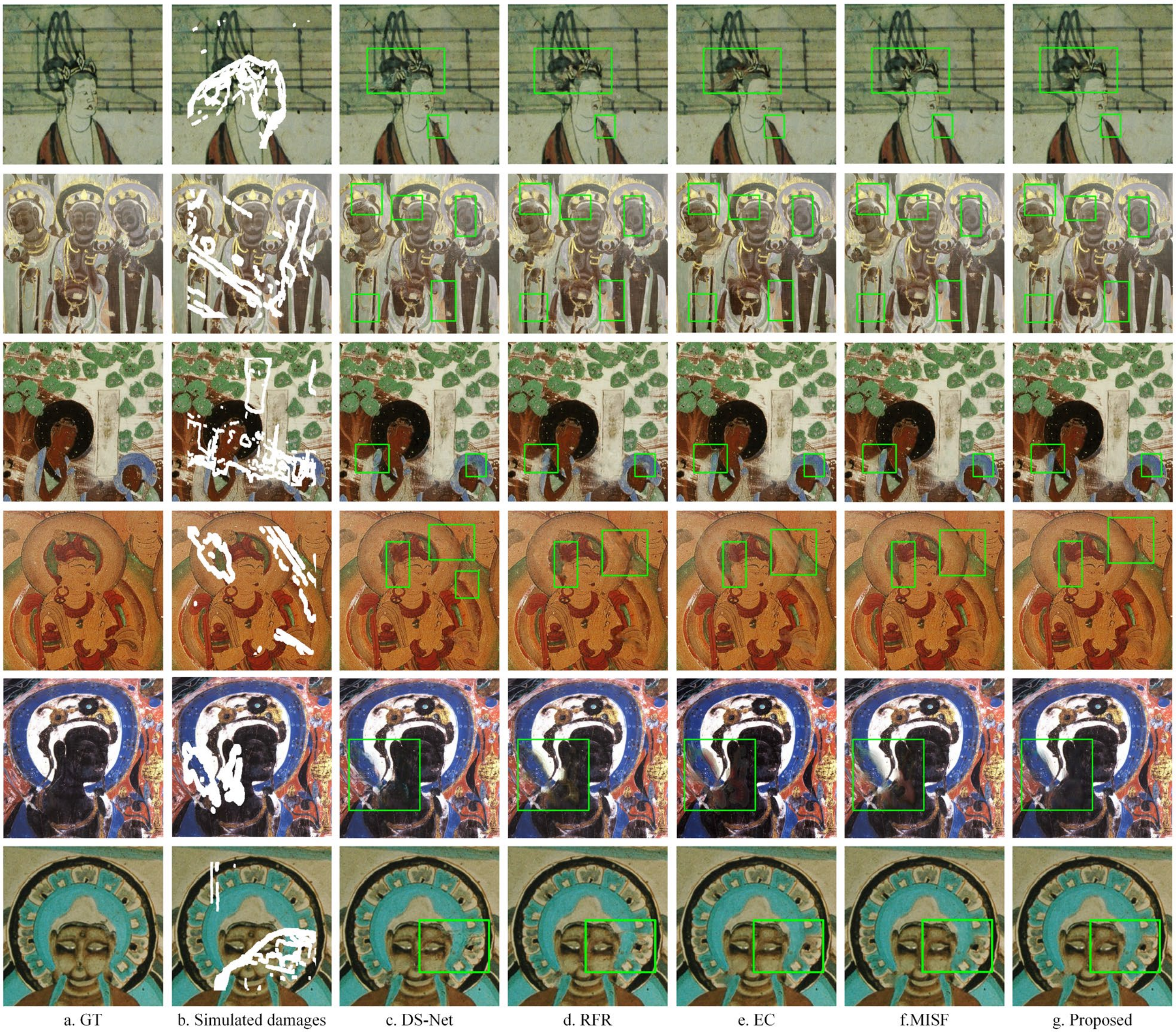

Fig. 8 Inpainting results of five approaches on simulated damaged murals

$$\ell_{\mathrm{style}} = \mathbb{E}\left[\sum_{j=1}^{N}\left\|G_j^{\Psi}(I_{\mathrm{out}}) - G_j^{\Psi}\left(I_{\mathrm{gt}}\right)\right\|_1\right] \tag{14}$$

In this work, we assign a different objective to the loss function at each stage. The ℓ_{l_1} aims to improve the quality of the restored mural image on pixel level. The ℓ_{adv} helps to improve the level of visual authenticity of the restored image. The ℓ_{style} tends to rectify the style consistency of the high-level structure. The ℓ_{FM} and ℓ_{perc} can help to keep the high-level semantic features of the whole mural. After several experimental verifications, good repair results were achieved when the weighting coefficients $\lambda_{\mathrm{s\_adv}}$, λ_{FM}, λ_{l_1}, $\lambda_{\mathrm{c\_adv}}$, λ_{perc} and λ_{style} for corresponding loss functions were set as 1, 10, 1, 0.1, 0.1, and 200 respectively.

Experimental results and analysis

To verify the inpainting performance of the proposed model, we conduct experiments on both irregular simulated damages and real damages of ancient murals. We compare our proposed model with four state-of-the-art approaches: DS-Net [28], RFR [29], EC [30] and MISF [31]. They were also trained on the same ancient mural dataset for comparison. We use peak signal-to-noise ratio (PSNR) [32], structural similarity (SSIM) [33] and learned perceptual image patch similarity (LPIPS) [34] to evaluate the inpainting results of murals. Since there is no ground-truth for the real damaged murals, we evaluate the quality of mural restoration by means of visual comparison. Moreover, we perform ablation experiments for each module and loss function of our model.

Table 2 Comparison of the PSNR and SSIM values of five approaches

Murals	DS-Net		RFR		EC		MISF		Proposed	
	PSNR/dB↑	SSIM↑	PSNR/dB↑	SSIM↑	PSNR/dB↑	SSIM↑	PSNR/dB↑	SSIM↑	PSNR/dB↑	SSIM↑
1	29.9886	0.9603	30.3743	0.9610	30.9715	0.9688	31.0584	0.9665	**31.1295**	**0.9670**
2	28.0542	0.9502	28.2087	0.9516	28.856	0.9580	28.9541	0.9590	**29.0945**	**0.9600**
3	29.4164	0.9603	29.9118	0.9637	29.9343	0.9631	29.9141	0.9629	**30.0545**	**0.9650**
4	32.0101	0.9456	32.7638	0.9548	31.0907	0.9409	32.7072	0.9504	**32.7638**	**0.9548**
5	27.0161	0.9510	26.0044	0.9441	25.4588	0.9480	26.3496	0.9526	**27.0314**	**0.9544**
6	32.2014	0.9715	32.3580	0.9730	32.3314	0.9726	32.6772	0.9736	**32.9086**	**0.9751**

The best result in each row is boldfaced

Table 3 Comparison of the LPIPS values of five approaches

Murals	DS-Net	RFR	EC	MISF	Proposed
	LPIPS↓	LPIPS↓	LPIPS↓	LPIPS↓	LPIPS↓
1	0.050	0.054	0.046	0.040	**0.036**
2	0.068	0.079	0.058	0.054	**0.052**
3	0.050	0.050	0.046	0.044	**0.041**
4	0.049	0.082	0.059	0.050	**0.040**
5	0.047	0.053	0.047	0.052	**0.042**
6	0.029	0.031	0.026	0.023	**0.020**

The best result in each row is boldfaced

Training settings

In our experiment, the hardware environment is configured as two NVIDIA GeForce RTX 2080Ti GPUs with 11 GB memory. We implement our model with PyTorch, running on an Ubuntu 18.01 system. All experiments are conducted in the same environment. Our model is trained by using 256 × 256 mural images, with batch size of 4. In the training process, we use the Adam algorithm as the optimizer in our model, and its hyperparameters β_1 and β_2 are set to 0.5 and 0.9. The generators and discriminator are trained with the learning rates of 1.0×10^{-4} and 1.0×10^{-5} respectively.

Dataset

Due to the limitations of equipment computing power, in this study, each mural image is cropped into several small sub-images with minimal overlap, and resized to 256 × 256px. we build a mural dataset that contains 3716 real mural images and replicas from the ancient mural album. The dataset contains images of ancient murals from different dynasties, styles and regions. Some examples of the mural dataset are shown in Fig. 6. The dataset is separated into training set and testing set. 3466 mural images of good quality are used for training, and other 151 intact murals and 99 real damaged murals are used for testing. In order to enhance the robustness of the model and alleviate the over-fitting problem, we use the data augmentation techniques [35] to expand the training samples. Through the augmentation techniques such as random mirror flipping, random 90-degree rotation and random crop, a total of 10,398 mural images are available for the network model training. We adopt the public mask dataset released in [36] for simulating irregular damages.

Experiment on simulated damage

In this subsection, we conduct experiment on the murals with simulated damages to demonstrate the inpainting performance of our model. We select several non-damaged mural images, and employ the random masks to imitate mural deterioration regions.

Figure 7 shows the inpainting results of our proposed model on four simulated damaged murals. Our model first predicts the structure map of the mural image (Fig. 7d). By use of the predicted structure map, the model recovers the texture of the mural and finally generates the final inpainting result (Fig. 7e). It can be seen that our model can predict complete structure information through the structure reconstruction network. Guided by the predicted structure information, the model can successfully perform the content restoration and generate a high-quality inpainting results.

Figure 8 gives the inpainting results of four comparative approaches on six Buddha murals with simulated damaged regions. It can be seen that DS-Net is capable of filling harmonious colors with the background, but cannot repair clear structures and produce some blurred textures for the damaged regions (e.g., 1st, 3rd, 4th, and 6th images). RFR tends to generate some blurry contents and color artifacts. For the 4th image, the Buddha's ears and clothes are not clearly recovered for the damaged regions. It also shows unrealistic texture and visual artifacts when repairing some large damaged regions (e.g., 5th image). EC is able to generate reasonable structure in the damaged regions due to its involvement of line drawings, but it will produce obvious artifacts and color distortion

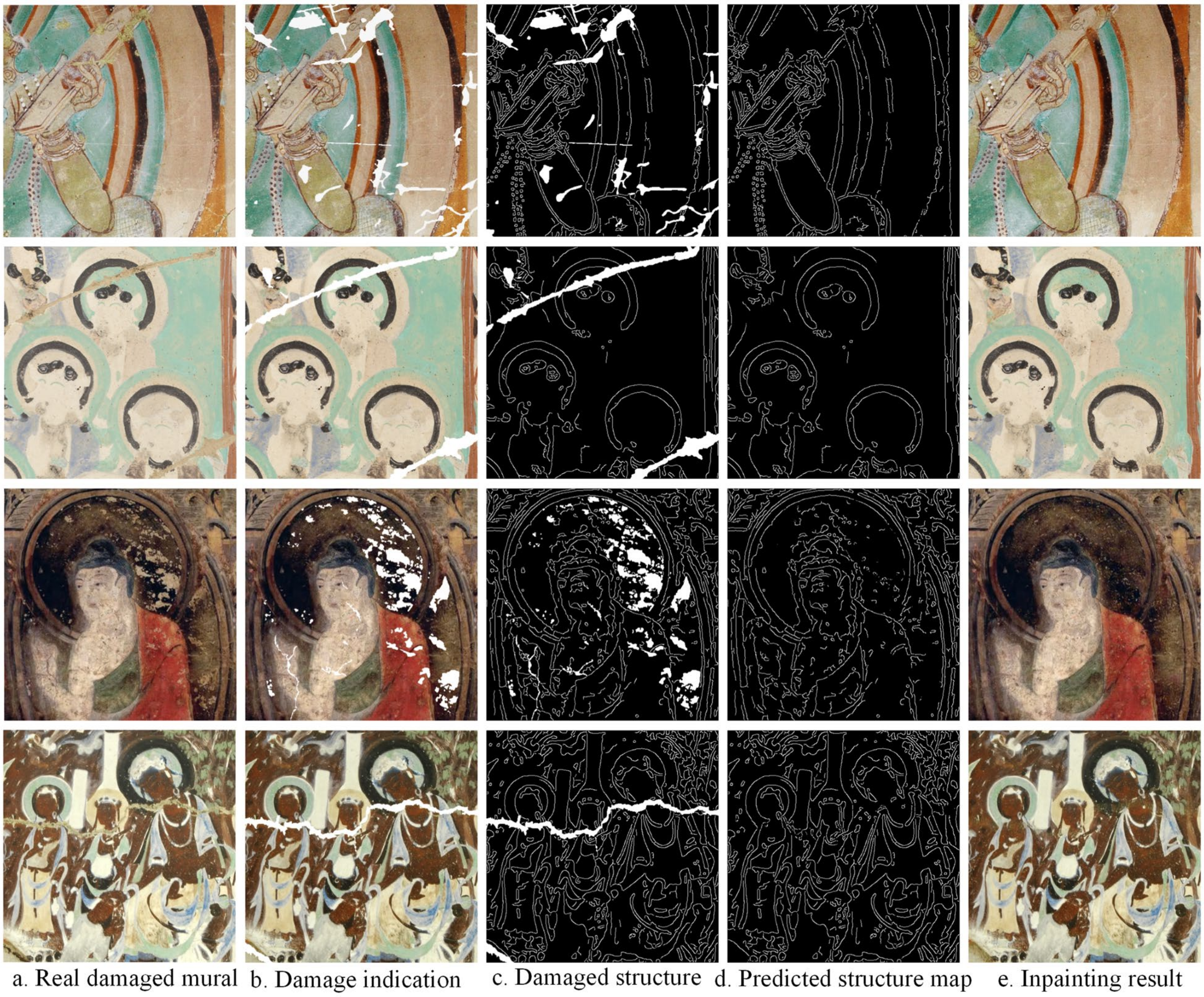

Fig. 9 Inpainting results of our proposed model on real damaged murals

in the inpainting results (e.g., 4th and 5th images). For MISF, obvious color-bias appears in some large missing regions (e.g., 5th image). Compared with other four approaches, our model not only predicts better structure information, but also generates clearer textures and fills more harmonious colors for the missing mural regions.

In order to further compare the restoration results in Fig. 8, we report the quantitative results in terms of PSNR, SSIM and LPIPS on our test mural images. PSNR is used to measure the quality of mural images, and a larger PSNR value indicates a better mural restoration. SSIM is a common metric in image processing to measure the structure similarity, and the closer its value is to 1, the higher the structure similarity between the restored murals and the ground-truth murals. LPIPS is used to evaluate the human perceptual disparity between two images, and the lower its value, the more similar the two images are. Tables 2 and 3 show the objective evaluation metrics for the restoration results of the simulated damaged mural images. It can be seen that our model is superior to other comparative approaches in quantitative metrics, which proves that our generated mural is closer to the ground-truth mural in terms of structure, pixel level and human perception.

Experiment on real damage

For the real damaged mural images, we label and mark the deteriorated regions of these historical relics manually to obtain the masks. In this experiment, we conduct our model on 99 real damaged mural images with various styles and disease types. Figure 9 gives the inpainting results of our model on four real damaged murals. As can be seen, our model can predict reasonable structure information, and ultimately generates clear textures, vivid colors and semantically continuous contents for the missing mural regions.

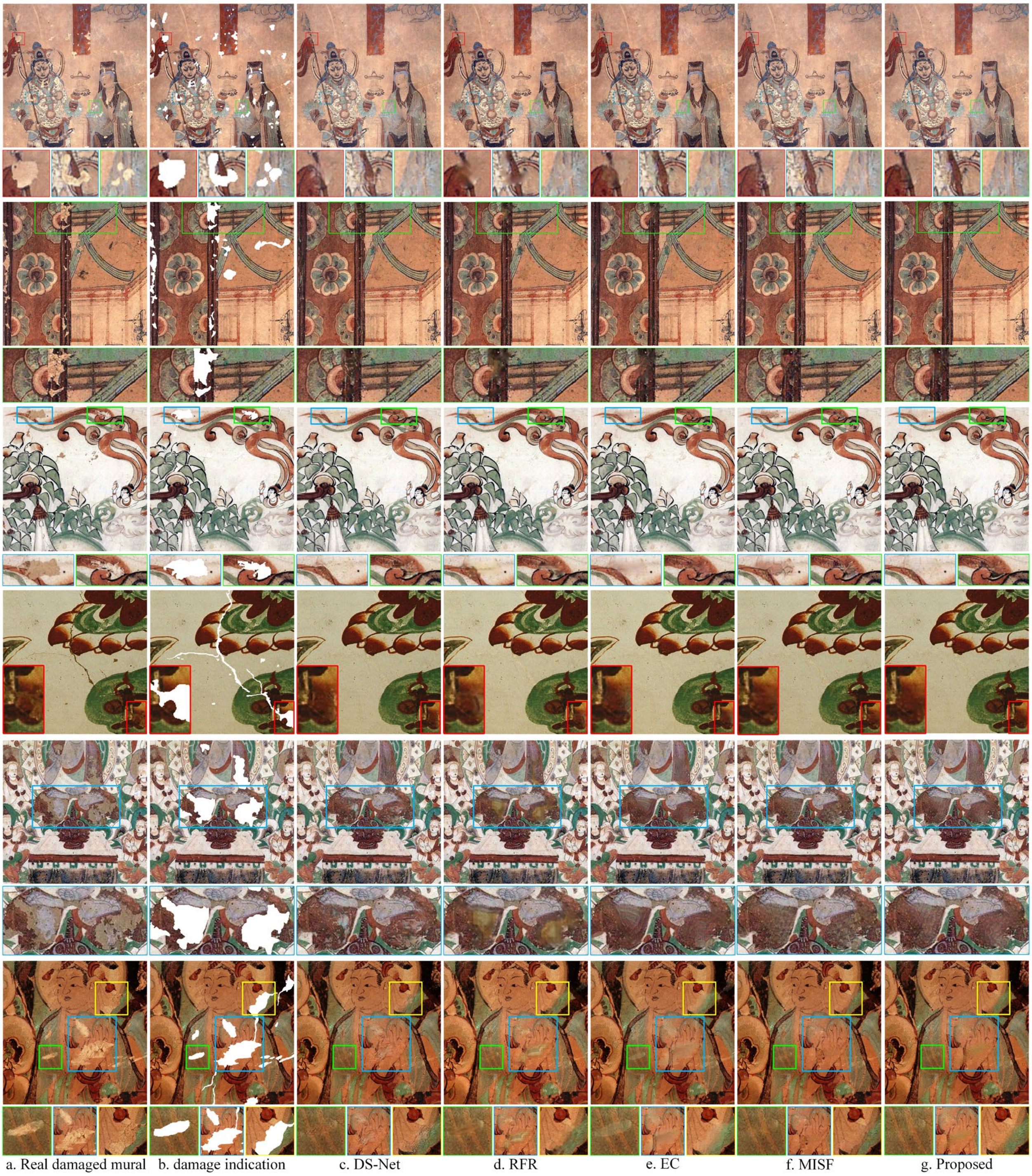

Fig. 10 Inpainting results of five approaches on real damaged murals

We also conduct comparative experiments on the real damaged murals. Figure 10 give the inpainting results of five approaches on 6 samples of the real damaged murals. We zoom in on the region marked by the color box, and the results of which are located below the corresponding mural image. As can be seen from the 1st image, DS-Net and RFR are unable to repair the missing regions with complex structures, and the repaired regions are somewhat blurry and over-smooth. EC and MISF produce unnatural color in the repaired region as compared

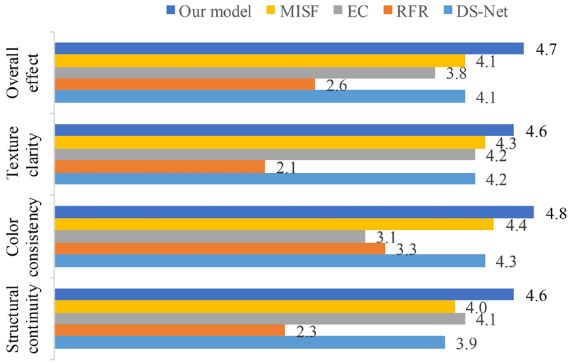

Fig. 11 Comparison of the evaluation scores of five approaches

with its surroundings. In the 3th image, DS-Net and EC produce consistent structural information, but they will bring some pixel diffuse such as the repaired silks in the left marker box. RFR and MISF generate unwanted colors and textures for the damaged region within the blue box. In the 4th image, all approaches can generate reasonable contents for the cracks, but our model generate better results for the bottom-right region of color-degradation. The 5th and 6th images have the large area falling-off disease. DS-Net and MISF tend to generate unrealistic pixel blocks. RFR produces obvious color distortion in the deteriorated region. Although EC is capable of restoring the missing structures, it will produce some white watermarks and unrealistic textures for the repaired areas. As compared with other approaches, our proposed model achieves the best visual quality of the inpainting results for these real damaged murals.

Since there is no ground truth for real damaged murals, we cannot evaluate the inpainting results quantitatively. To make the visual comparison more convincing, we invite 20 volunteers to score the structural continuity, color consistency, texture clarity and overall effect of the restoration results of the four comparative approaches. We select 10 restored mural images as the test subjects. We rank the user ratings into five levels with corresponding scores of 5, 4, 3, 2, 1. The higher the score, the better the evaluation. Figure 11 shows the average scores of 20 volunteers on the restoration results of the five comparative approaches. It can be seen that our model achieves the highest scores in the test of visual comparison.

Ablation study

In this subsection, we conduct ablation experiments on each of the proposed components and loss functions to verify the effect of them on the repair results.

Ablation study of the proposed components

To begin with, we study the effects of the proposed main components. We take the two-stage Edge-Connect as the baseline network and abbreviate it as $G_1 + G_2$. G_1 and G_2 represent the first and second stages of the baseline, respectively. We replace the first stage network of the baseline with our proposed Structure Reconstruction Network (SRN), and keep the other network unchanged. This combined module is referred to as $(G_s + G_2)$. We also remove the FFC residual block (FFC ResBlk) and the cascaded attention module (CAM) of the proposed Content Restoration Network (CRN) from our model, respectively. In this paper, we abbreviate this operation as G_s + CAM and G_s + FFC ResBlk, respectively. Then, we train our model with the proposed components for ablation experiments and test it on 151 simulated damaged murals.

Figure 12 shows the visual comparison of the ablation experimental results of each component. Row 1 shows three original mural images. Row 2 shows the corresponding masked mural images that are partially cropped to magnified for better observation. Row 3 shows the results of the baseline network. Rows 4 to 6 show the ablation results of three kinds of combined module, respectively. Row 7 shows the inpainting results that are generated by our complete model $(G_s + G_c)$. In Fig. 12, we use yellow arrows to mark the repair areas with significant differences. As shown in Row 3, the two-stage baseline model inevitably produces pixel diffusion and color deviation. Particularly, we can observe an unnaturally distorted structure for the repaired roof in the middle image. Note that the $G_s + G_2$ model improves the structural quality of the mural and alleviates the pixel diffusion and color deviation in the deteriorated region. This proves that the structure information of a mural image can be utilized to improve the quality of texture restoration. It can be also observed in Rows 5 and 6 that each component of the CRN is conducive to texture and color restoration. It can be seen that our complete model $(G_s + G_c)$ has the best visual quality in this ablation study, which can generate high-fidelity mural contents with continuous structures, vivid colors and semantically plausible textures.

We also provide an objective evaluation of the above ablation experiment on 151 simulated damaged mural images. Table 4 gives the performance of the inpainting results that are generated by five test combination modules in terms of PSNR, SSIM and LPIPS. Compared with the baseline network, the restoration results for each proposed component show obvious improvement

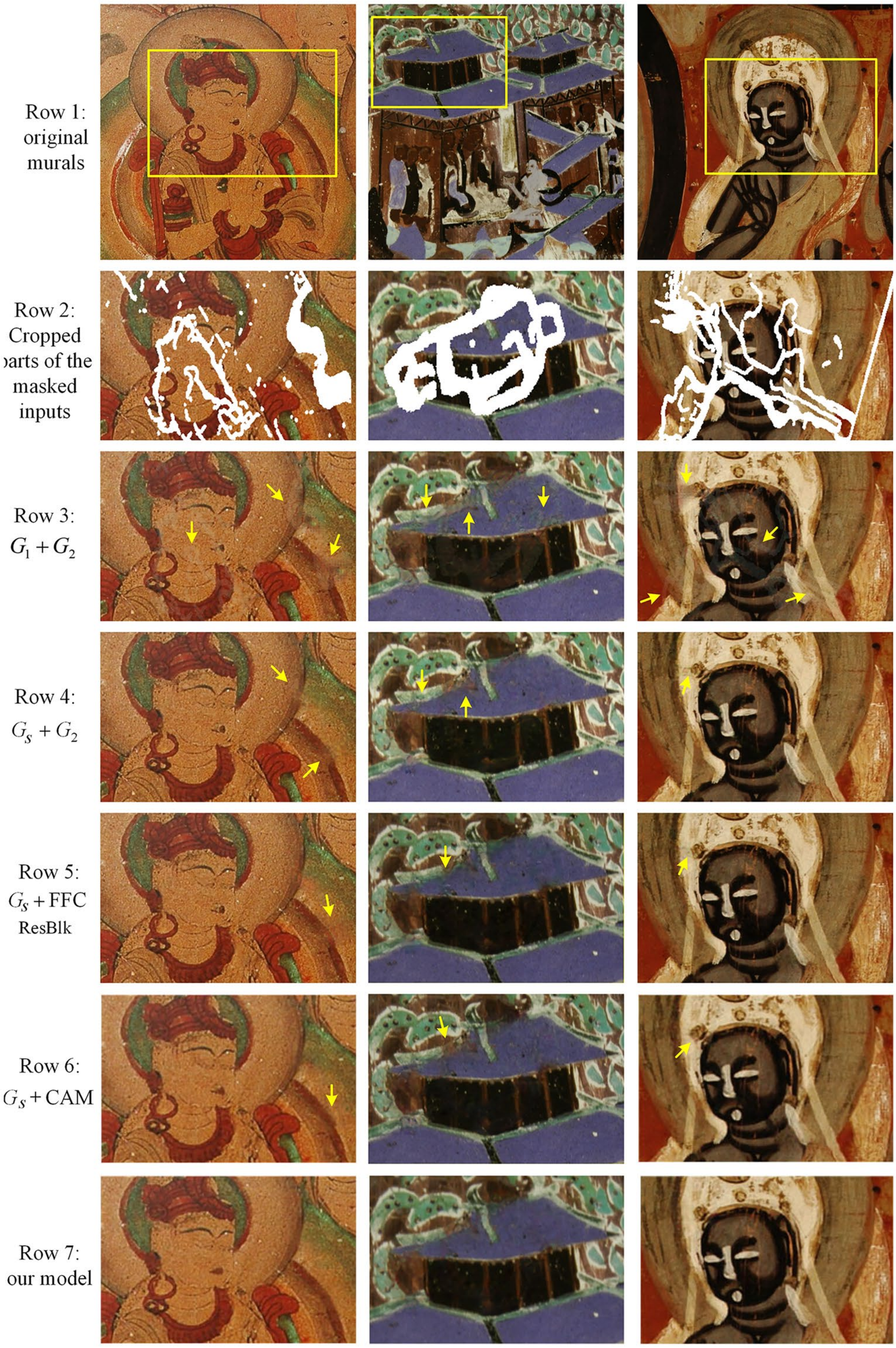

Fig. 12 Inpainting results of five approaches on real damaged murals

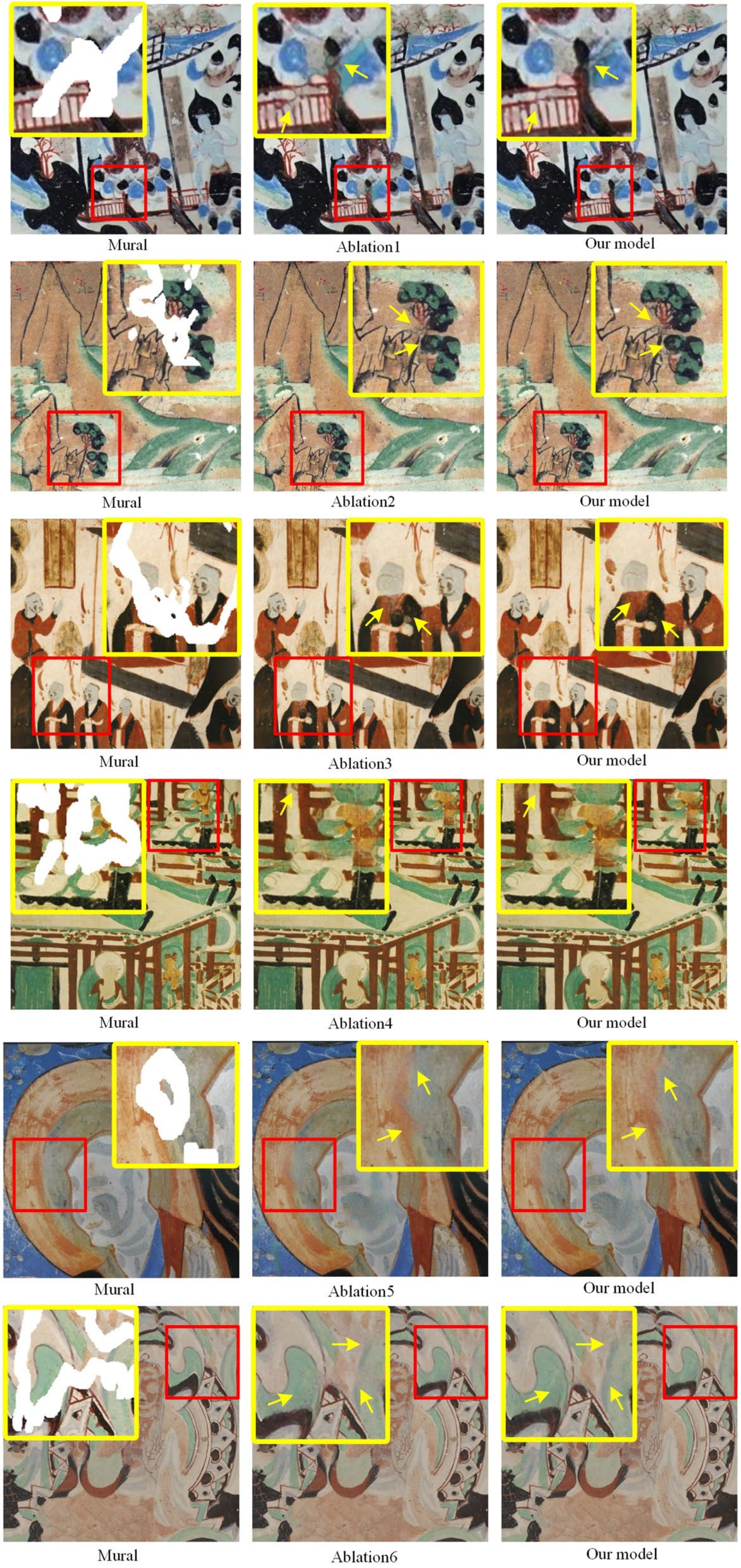

Fig. 13 Inpainting results of five approaches on real damaged murals

Table 4 Metrics for each component of our model

Module combination	PSNR/dB↑	SSIM↑	LPIPS↓
Baseline ($G_1 + G_2$)	28.2139	0.9655	0.0459
$G_s + G_2$	28.5658	0.9681	0.0397
G_s + FFC ResBlk	31.3839	0.9756	0.0348
G_s + CAM	31.2826	0.9753	0.0365
Our model	**31.4522**	**0.9760**	**0.0324**

The best result in each column is boldfaced

Table 5 Metrics for each component of our model

Loss strategy	Stage1 loss	Stage2 loss
Proposed	ℓ_{stage1}	ℓ_{stage2}
Ablation1	$\ell_{\text{stage1}} - \ell_{\text{s\_adv}}$	ℓ_{stage2}
Ablation2	$\ell_{\text{stage1}} - \ell_{\text{FM}}$	ℓ_{stage2}
Ablation3	ℓ_{stage1}	$\ell_{\text{stage2}} - \ell_{l1}$
Ablation4	ℓ_{stage1}	$\ell_{\text{stage2}} - \ell_{\text{c\_adv}}$
Ablation5	ℓ_{stage1}	$\ell_{\text{stage2}} - \ell_{\text{perc}}$
Ablation6	ℓ_{stage1}	$\ell_{\text{stage2}} - \ell_{\text{style}}$

"–" Indicates the removed item

in both PSNR and SSIM metrics. The LPIPS metrics for each component are lower than the baseline network. This demonstrates that each proposed component in our model plays an important role in the restoration of the damaged murals.

Ablation study of the loss functions

In this test, we conduct an ablation study on the loss functions to analyze the effect of them. We remove the loss functions from the two-stage network of our proposed model one by one, and obtain six different ablation strategies (Ablation1, 2, 3, 4, 5, 6) that are illustrated in Table 5. The symbol "–" denotes the "remove" item.

Figure 13 shows some visual comparisons of these six ablation strategies and our proposed model. In each group, our proposed model is compared to an ablation strategy that has removed a certain loss function. As indicated by the yellow arrow marking area, each loss function has a specific effect on improving quality of the restored murals. When removing a certain loss function from the model, the repair results appears some degradation as compared with the proposed model.

Conclusion

In this work, we proposed a structure guided two-branch (SGTB) model to virtually restore the deteriorated regions of the ancient murals. The two restoration stages of the model are reframed by using the generative adversarial network. In the structure reconstruction stage, FFC residual blocks are introduced to extract the global features of the murals, and thus the model can accurately predict a complete mural structure map. In the content restoration stage, the two-branch parallel encoders are designed to improve the restoration quality of the mural textures and colors. In addition, the proposed model employs a cascaded attention module to focus on long-term relevance of the feature information to refine the texture and color restoration of the damaged murals. Our proposed model is performed on both simulated and real deteriorated murals. The experimental results demonstrate that our model can effectively repair the ancient murals with various deteriorated regions. As compared with four existing approaches, our model can obtain better mural inpainting quality when evaluated by use of visual comparison and objective metrics.

It is worth stating that the deep learning-based image inpainting requires large amount of training data. Most available ancient Chinese murals have varying degrees of diseases such as erosion, falling off, crack, scratches, etc. It is very difficult for us to collect enough high-quality training data. Although we expand the training data set by using the data augmentation techniques such as rotation, cropping, flipping, etc., the data augmentation will bring information redundancy. This will probably affect the generalization ability of the deep-learning model. Although our proposed model achieves better performance in restoring the deteriorated ancient murals than existing approaches, it still suffered from the limitations of the lack of high-quality training data. In our future work, we will collect more ancient mural images through field visits. Moreover, we will consider utilizing some intelligent algorithms to build a synthetic mural training dataset by using super-resolution or style transfer based on deep neural networks. As we all know, the size of the disease range has an impact on the inpainting effect. In the future, we will consider adopting different scale restoration for the diseased areas of murals.

Abbreviations

FFC ResBlk	Fast Fourier convolution residual block
CAM	Cascade attention module
VGG	Visual geometry group network
Act-Func	Activation function

Acknowledgements
None.

Author contributions
All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 62166048, Grant No. 61263048), by the Applied Basic Research Project of Yunnan Province (Grant No. 2018FB102), and by the Postgraduate Research and Innovation Foundation of Yunnan University (2021Y258).

Availability of data and materials

The datasets used and/or analyzed in the current study are available from the corresponding author by reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 22 March 2023 Accepted: 8 June 2023
Published online: 20 June 2023

References

1. Yue YQ. Condition surveys of deterioration and research of wall paintings in Maijishan cave-temple. Study Nat Cult Herit. 2019;4(2):127–31 **(in Chinese with an English abstract)**.
2. Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting. Proceedings of the 27th annual conference on Computer graphics and interactive techniques. 2000: 417-424.
3. Jaidilert S, Farooque G. Crack detection and images inpainting method for Thai mural painting images. 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2018: 143–148.
4. Chen Y, Ai YP, Guo HG. Inpainting algorithm for Dunhuang Mural based on improved curvature-driven diffusion model. J Comput-Aided Design Comput Graph. 2020;32(05):787–96 **(in Chinese with an English abstract)**.
5. Criminisi A, Perez P, Toyama K. Object removal by exemplar-based inpainting. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. IEEE, 2003, 2: II-II.
6. Jiao LJ, Wang WJ, Li BJ, et al. Wutai mountain mural inpainting based on improved block matching algorithm. J Comput-Aided Design Comput Graph. 2019;31(01):118–25 **(in Chinese with an English abstract)**.
7. Cao J, Li Y, Zhang Q, et al. Restoration of an ancient temple mural by a local search algorithm of an adaptive sample block. Herit Sci. 2019;7(1):1–14. https://doi.org/10.1186/s40494-019-0281-y.
8. Wang H, Li Q, Zou Q. Inpainting of Dunhuang murals by sparsely modeling the texture similarity and structure continuity. J Comput Cult Herit (JOCCH). 2019;12(3):1–21.
9. Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: feature learning by inpainting. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536–2544.
10. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44.
11. Wang N, Wang W, Hu W, et al. Thanka mural inpainting based on multi-scale adaptive partial convolution and stroke-like mask. IEEE Trans Image Process. 2021;30:3720–33. https://doi.org/10.1109/TIP.2021.3064268.
12. Cao J, Zhang Z, Zhao A, et al. Ancient mural restoration based on a modified generative adversarial network. Herit Sci. 2020;8(1):1–14. https://doi.org/10.1186/s40494-020-0355-x.
13. Lv C, Li Z, Shen Y, et al. SeparaFill: two generators connected mural image restoration based on generative adversarial network with skip connect. Herit Sci. 2022;10(1):1–13. https://doi.org/10.1186/s40494-022-00771-w.
14. Schmidt A, Madhu P, Maier A, et al. ARIN: adaptive resampling and instance normalization for robust blind inpainting of Dunhuang Cave Paintings. 2022 Eleventh international conference on image processing theory, tools and applications (IPTA). IEEE, 2022: 1–6. https://doi.org/10.1109/IPTA54936.2022.9784144.
15. Li J, Wang H, Deng Z, et al. Restoration of non-structural damaged murals in Shenzhen Bao'an based on a generator-discriminator network. Herit Sci. 2021;9(1):1–14. https://doi.org/10.1186/s40494-020-00478-w.
16. Yu T, Lin C, Zhang S, et al. End-to-end partial convolutions neural networks for Dunhuang grottoes wall-painting restoration. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.
17. Li L, Zou Q, Zhang F, et al. Line drawing guided progressive inpainting of mural damages. arXiv preprint arXiv:2211.06649, 2022.
18. Ciortan IM, George S, Hardeberg JY. Colour-balanced edge-guided digital inpainting: applications on artworks. Sensors. 2021;21(6):2091.
19. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;6:679–98.
20. Chi L, Jiang B, Mu Y. Fast Fourier convolution. Adv Neural Inf Process Syst. 2020;33:4479–88.
21. Brigham EO, Morrow RE. The fast Fourier transform. IEEE Spectrum. 1967;4(12):63–70.
22. Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5505–5514.
23. Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125–1134.
24. Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8798–8807.
25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
26. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. European conference on computer vision. Cham: Springer; 2016. p. 694–711.
27. Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2414–2423.
28. Wang N, Zhang Y, Zhang L. Dynamic selection network for image inpainting. IEEE Trans Image Process. 2021;30:1784–98.
29. Li J, Wang N, Zhang L, et al. Recurrent feature reasoning for image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7760–7768.
30. Nazeri K, Ng E, Joseph T, et al. Edgeconnect: generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212, 2019.
31. Li X, Guo Q, Lin D, et al. MISF: multi-level interactive Siamese filtering for high-fidelity image inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1869-1878.
32. Gupta P, Srivastava P, Bhardwaj S, et al. A modified PSNR metric based on HVS for quality assessment of color images. 2011 International Conference on Communication and Industrial Application. IEEE, 2011: 1–4.
33. Hore A, Ziou D, Image quality metrics: PSNR vs. SSIM. 20th international conference on pattern recognition. IEEE. 2010;2010:2366–9.
34. Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 586–595.
35. Ma DG, Tang P, Zhao LJ, et al. Review of data augmentation for image in deep learning. J Image Graph. 2021;26(03):487–502 **(in Chinese with an English abstract)**.
36. Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions. Proceedings of the European conference on computer vision (ECCV). 2018: 85–100.

Publisher's Note