npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01710-1

Mural inpainting via two-stage generative adversarial network

Qiongshuai Lyu[1] ✉, Na Zhao[2] ✉, Junke Song[1], Yu Yang[1] & Yuehong Gong[1]

The digital restoration of Dunhuang murals is of extremely high value for the research and dissemination of mural culture and art. However, a large number of murals have been damaged to varying degrees. In this paper, we propose a two-stage coarse-to-fine digital mural restoration framework to solve the large area of irregular shape damage on the mural. The first stage is used to achieve coarse-grained semantic reconstruction, and the second stage is used to achieve fine-grained feature reconstruction. To improve the repair quality, we also designed a new building block (STMA) that integrates the Swin transformer module (SwinT) and the multi-scale dilated convolution attention module (MSDA). Meanwhile, the proposed loss function is to further empower the proposed model to repair damaged murals. Extensive comparative experiments show that the proposed model can effectively restore the missing content of the mural and exceed the comparative methods in both quantitative and qualitative evaluation.

Dunhuang murals, with their unique artistic style, exquisite painting skills, and profound cultural connotation, have become one of the most precious heritages in the history of human civilization. Some mural images are shown in Fig. 1. In the long history, due to climate change, sand erosion, man-made destruction, and other factors, a large number of murals have been damaged to varying degrees, such as paint shedding, fading, peeling, cracking, and so on[1]. These factors have seriously restricted the cultural exchange between the East and the West in murals, and have limited the exploration of the historical changes, artistic characteristics, and cultural connotations of Dunhuang murals. Therefore, it is of great significance to explore an effective mural restoration method with the concept of "minimal intervention principle" for the study of mural culture and the development and breakthrough of other painting art categories.

Manual restoration methods play an important role in the conservation of murals. Through observation, we found that the manual restoration of murals can be simply described as a two-step process. The first step is to outline the contours and lines of the missing areas of the murals by highly skilled painters, and the second step is to color the missing areas according to the lines and the surrounding colors. The first step requires the painter to draw the structural information of the missing mural based on personal experience, which requires the painter to be very familiar with the painting style of the mural. The second step requires coloring the missing areas according to the style of the mural based on structural information. These two steps are relatively high technical requirements for mural restorers. Moreover, the manual restoration method is irreversible and tends to restore directly on the mural, which is prone to cause secondary damage to the mural[2]. Inspired by medicine, the protection staff of cultural relics in many countries and areas apply the protection concept of "minimal intervention" to the restoration task of murals and tend to use digital image restoration methods to complete the restoration of murals.

With the continuous development of computer vision technology, many scholars have proposed image restoration technologies using different strategies in recent years, making digital image restoration technology has made great progress in solving the problem of repairing natural images[3–7]. Although these methods achieve good performance in solving natural image restoration problems, they are not suitable for the task of mural restoration. The main reasons are: (1) The styles of murals and natural images are very different. Although transfer learning can integrate the prior information in natural images into the restoration task of murals, the unique style displayed by murals is obviously different from natural images. This limits the effectiveness of transfer learning methods. (2) The lines in the mural are uneven, the colors are rich and diverse, and the patterns are complex and changeable. Moreover, the existing feedforward neural network method based on simple convolution cannot capture the rich high-frequency texture information and low-frequency color patch information contained in the mural, which affects the recovery performance of the model.

To alleviate these problems, we propose a new mural restoration framework based on generative adversarial networks (GAN) to solve the large area of irregular shape damage on the mural. The proposed framework adopts the strategy of first contour structure and then texture

[1]School of Software, Pingdingshan University, Pingdingshan, China. [2]School of Journalism and Communication, Pingdingshan University, Pingdingshan, China. ✉e-mail: qiongshuailyu@pdsu.edu.cn; 5229@pdsu.edu.cn

Fig. 1 | Some mural images.

coloring, which is inspired by the method of manually inpainting murals. This strategy is significantly different from the inpainting strategy for natural images. The framework contains two generative networks to model the two steps in the manual mural restoration method. The first is a coarse-grained network, which is mainly used to reconstruct the semantic information of the lines and contours of the damaged mural. The coarse-grained network simulates the modeling of mural contour and structure restoration during manual restoration. The second is a fine-grained network, which is mainly used to reconstruct fine-grained features to repair texture details and color correction of damaged murals. The fine-grained network simulates the modeling of mural texture details and colorization restoration during manual inpainting. Moreover, the Swin transformer module (SwinT)[8,9] can model long-range dependency with the shifted window scheme. The capture of multi-scale features helps to extract information at different granularities[10,11]. Since different image reconstruction methods have complementary image prior modeling capabilities[12,13], we propose a new building block (STMA) that fuses the SwinT module with the multi-scale dilated convolution attention module. To better improve the performance of the network, we also design a new joint loss function to impose guidance on the two-stage training process of the model from multiple dimensions. Experiments show that our proposed mural repair method has better performance. In summary, the main contributions of this paper are summarized as follows: (1) We propose a two-stage GAN mural inpainting framework to solve the large area of irregular shape damage on the mural, which models the manual mural inpainting process and can achieve coarse-to-fine mural inpainting. (2) A novel building block (STMA) is proposed to effectively extract rich contextual information. It combines a Swin transformer module and a multi-scale dilated convolution attention module. (3) The joint loss function designed based on multiple dimensions can not only stabilize the training process of the model but also improve the mural restoration performance of the model.

Methods

Preliminaries: generative adversarial network

Inspired by game theory, Goodfellow et al.[14] first proposed GAN. It contains a generative network G and a discriminative network D. The two networks learn the distribution of data through adversarial training. In the game process between the generative network G and the discriminative network D, the generative network G generates realistic samples as much as possible, and the discriminative network D tries to judge the authenticity of the sample. To realize the process of the game, the following loss function is used when training the GAN,

$$\min_G \max_D V(D,G) = E_{x\sim P_{data(x)}}\left[\log D(x)\right] + E_{z\sim P_{(z)}}\left[\log(1-D(G(z)))\right] \tag{1}$$

where $x \sim P_{data(x)}$, $P_{data(x)}$ represents the real data distribution. z represents the random noise $z \sim P_{(z)}$, $P_{(z)}$ represents random distribution, such as Gaussian distribution.

At present, GAN have been widely used in image inpainting tasks. To generate multiple image inpainting results, a probabilistic diverse GAN (PD-GAN)[15] is proposed. PD-GAN is built upon a vanilla GAN that generates images based on random noise. Nazeri et al.[16] proposed EdgeConnect by decomposing image restoration tasks into structure prediction and image completion. This method alleviates the problem of image completion when significant portions of the image are missing. By combining the advantages of convolutional neural networks (CNN) and transformers, Wan et al.[17] proposed a high-fidelity pluralistic image completion method based on GAN. This method utilizes a transformer to restore coarse textures and enhances local texture details using CNN under the guidance of mask images.

Related work on mural inpainting

As a research hotspot in the field of image restoration, the digital restoration of murals has attracted more and more attention. Several scholars have proposed mural restoration schemes employing different strategies, including the partial differential equation method[18–20], texture synthesis method[21,22], sparse representation method[23–25], and deep learning method[26–31]. Chen et al.[19] proposed an adaptive inpainting method for Dunhuang murals based on the improved curvature-driven model, which solved the defects of the CDD algorithm and shortened the inpainting time. To solve the problem of repairing murals that have fallen off and eroded by diseases, Cao et al.[21] proposed a virtual restoration method of murals based on an adaptive local search of sample patches, which makes the composition characteristics of repaired murals more reasonable. Chen et al.[25] proposed a mural inpainting method based on Gabor transform and group sparse representation, which alleviates the problems of blurred structure and discontinuous lines in the inpainted murals. Although these traditional methods have achieved good mural inpainting performance, due to the excellent performance of deep learning methods in the field of image reconstruction in recent years, many mural inpainting methods based on deep learning have also emerged. To effectively repair ancient

murals, Cao et al.[26] proposed a consistency-enhanced GAN to repair missing murals. To utilize the auxiliary information provided by the line drawing, Li et al.[27] proposed a line drawing-guided mural restoration method, which decomposed the mural restoration process into two stages: structure reconstruction and color correction. In the structural reconstruction stage, line drawings are used to ensure the authenticity of the large-scale restoration content and the stability of the structure. In the color correction stage, the missing pixels are corrected and the local color is adjusted. The two-stage restoration strategy effectively improves the quality of mural restoration. Considering the problems of insufficient feature extraction and loss of detail reconstruction, Chen et al.[28] proposed a GAN mural restoration method based on multi-scale features and attention fusion. From the perspective of attention and loss function, Li et al.[29] proposed an approach for inpainting damaged areas of Dunhuang murals based on a recurrent feature reasoning network. To exploit multi-scale information and mask features, Wang et al.[30] proposed a Thanka mural inpainting method based on multi-scale adaptive partial convolution and stroke-like masks.

STMA-GAN Network

Inspired by the manual restoration mural scheme, we proposed the STMA-GAN model, as shown in Fig. 2. STMA-GAN is built up on a GAN. By modeling the manual restoration process, STMA-GAN is designed as a model that contains two generators and a discriminator. Note that the same discriminator structure is used during the two-stage training process. Among the two generators, one is a coarse-grained network G^C, which is used to model the first step of the manual restoration process to reconstruct the lines and contours of the missing content of the mural, and the other is a fine-grained network G^F, which is used to refine the texture and details of the image reconstructed by the coarse-grained network. Let x_L and x_M be the line drawing and mask image. x_L and x_M are the inputs to the coarse-grained network. x_C is the output of the coarse-grained network. It can be represented as:

$$x_C = G^C(x_L, x_M) \tag{2}$$

The coarse-grained network consists of an encoder, a stack of STMA blocks, and a decoder. The encoder consists of four convolution-batch channel normalization-ReLU layers (Conv-BCN-ReLU), BCN[32] can adaptively combine the information of channel and batch dimension to improve the generalization of the model. The role of the encoder is to encode the input x_L and x_M to extract the key features $F^C_{e_i}$,

$$F^C_{e_i} = \begin{cases} G^C_{e_{i-1}}(x_L, x_M) & i = 1 \\ G^C_{e_{i-1}}\left(F^C_{e_{i-1}}\right) & i \in \{2,3,4\} \end{cases} \tag{3}$$

where $G^C_{e_i}$ is the i-th layer function operation of the encoder. Four stacked STMA blocks form the backbone of the coarse-grained network. This way of stacking building blocks enables the network to extract richer image contextual information. It can be expressed as:

$$F^C_{s_j} = \begin{cases} G^C_{s_{j-1}}(F^C_{e_4}) & j = 1 \\ G^C_{s_{j-1}}(F^C_{s_{j-1}}) & j \in \{2,3,4\} \end{cases} \tag{4}$$

where $G^C_{s_j}$ is the j-th function operation of the STMA block. Then, four upsampling-convolution-BCN-ReLU (Up-Conv-BCN-ReLU) layers are used to decode the features from STMA blocks to the initial inpainted image as the output result of the coarse-grained network. Because the deep network structure will make the features disappear in the transmission process from the shallow layer to the deep layer, which weakens the performance of the model to reconstruct the image[33,34]. To alleviate this issue, skip connections are applied between the encoder and the decoder.

$$F^C_{d_k} = \begin{cases} G^C_{d_{k-1}}(F^C_{s_4} + F^C_{e_4}) & k = 1 \\ G^C_{d_{k-1}}\left(F^C_{d_{k-1}}\right) + F^C_{e_{4-k+1}} & k \in \{2,3,4\} \end{cases} \tag{5}$$

where $G^C_{d_k}$ is the k-th layer function operation of the decoder. Finally, the output of the coarse-grained network is reconstructed by a convolutional layer G^C_{cov},

$$x_C = G^C_{\text{cov}}(F^C_{d_4}) \tag{6}$$

Similar to the coarse-grained network G^C, the architecture of the fine-grained network G^F adopts the same design idea. Let x_m be the binary mask image, the fine-grained network receives x_m and x_C as input and outputs the

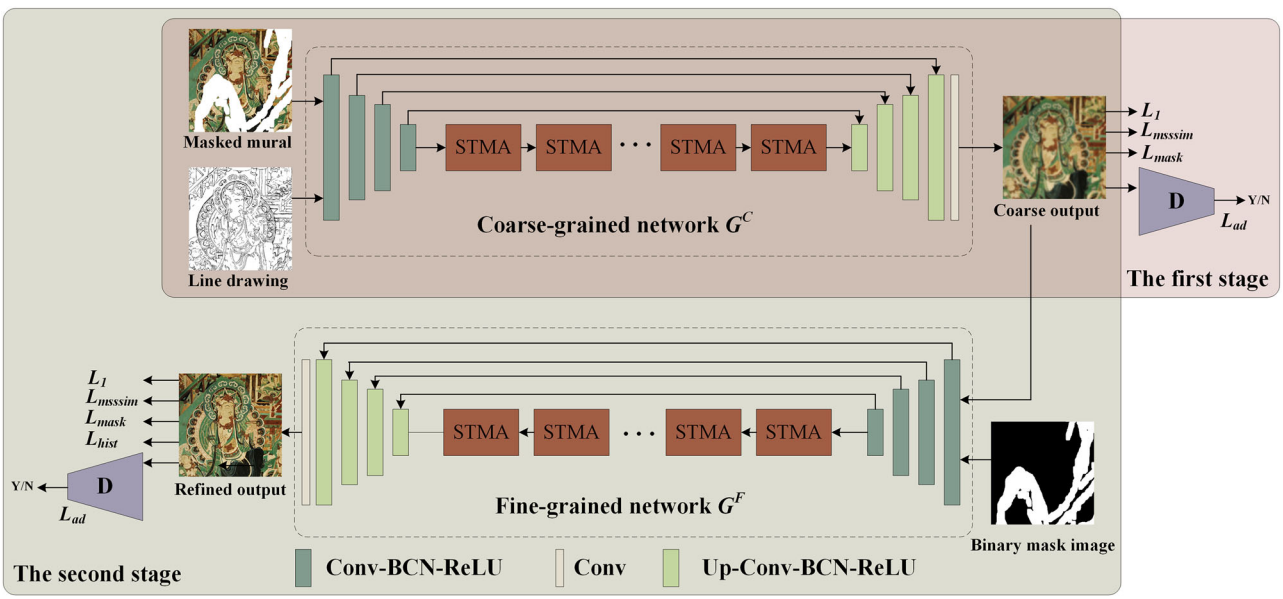

Fig. 2 | The overall pipeline of our proposed STMA-GAN.

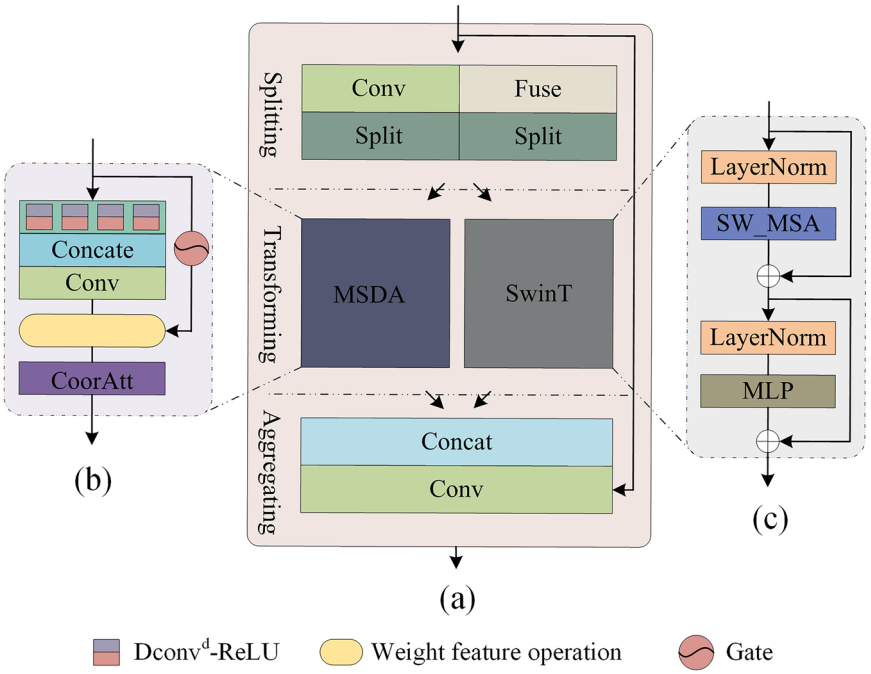

Fig. 3 | The structure of the STMA block. a STMA. **b** MSDA. **c** SwinT.

result $\hat{y}$ as the final repaired mural by the model. It can be represented as:

$$\hat{y} = G^F(x_m, x_C) \tag{7}$$

However, to achieve the refinement of mural texture and detail restoration, a two-staged training strategy and different loss functions are used during training, see the loss function subsection for details.

The discriminator, its role is to determine whether the repair results are good or bad. To stabilize the training of the discriminator, spectral normalization[35] is introduced into the discriminator. Spectral normalization can constrain the spectral norm of each layer of the discriminator so that the discriminator satisfies Lipschitz continuity[36], which can reduce the blurring effect of the mural repair results and enhance realism[5]. We use a stack of five convolution-spectral normalization-Leaky ReLU (Conv-SN-LReLU) layers to form the backbone of the discriminator, a single sigmoid function is used for the output of the discriminator.

STMA block

STMA block fuses the Swin transformer module (SwinT) and the multi-scale dilated convolution attention module (MSDA) by the strategy of the split-transformation-merge[37].

Splitting: Inspired by ref.[13], the feature tensor X_{in} is fed into a 1×1 convolution and split into two sub-feature tensors equally. It can be expressed as:

$$x_1, x_2 = Split(Conv_{1\times1}(X_{in})) \tag{8}$$

Following the weight fusion strategy, we obtain the fusion weights by

$$w_1, w_2 = Split(Fuse(X_{in})) \tag{9}$$

where $Fuse(.) = Softmax(MLP(GAP(.)))$, GAP represents the global average pooling, MLP represents the multi-layer perceptron. We use the weights w_1 and w_2 to fuse x_1 and x_2 via $X_1 = w_1 x_1$ and $X_2 = w_2 x_2$.

Transforming: The two sub-feature tensors X_1 and X_2 are fed into the SwinT module and MSDA module to perform different transformations, respectively.

$$X_s, X_m = SwinT(X_1), MSDA(X_2) \tag{10}$$

Specifically, the SwinT module is a Transformer block, which includes a multi-head self-attention mechanism based on a shifted window (SW_MSA) and a multi-layer perceptron (MLP). SW_MSA can use the interactive information between local attention and shift window to better model the long-distance dependency, and it is also beneficial to MLP for further feature transformation[38]. A LayerNorm layer is added before SW_MSA and MLP, and the residual connection is used to facilitate the transfer of information, see Fig. 3. The SwinT module is computed as:

$$\begin{aligned} X_1' &= SW\_MSA(LN(X_1)) + X_1 \\ X_s &= MLP(LN(X_1')) + X_1' \end{aligned} \tag{11}$$

where X_s is the output of the SwinT module. The MSDA module is a multi-scale dilated convolution attention module, which contains convolution operations with different dilation rates. Dilated convolution can systematically aggregate multi-scale contextual information without losing resolution[4,39,40]. Inspired by this, we use convolutions with different dilation rates to implement multi-scale feature transformation on the input tensor to capture rich contextual information. Convolution operations with larger dilation rates have larger receptive fields, which help extract more semantic information. Convolution operations with smaller dilation rates have smaller receptive fields, which is beneficial in paying more attention to local patterns. The input feature tensor X_2 is evenly divided into four sub-feature tensors $x_{sub}^i, i \in \{1, 2, 3, 4\}$ along the channel dimension. Each sub-feature tensor x_{sub}^i is fed separately to the convolution with different dilation rates $d \in \{1, 2, 4, 8\}$. We combine convolution operations with different dilation rates in a parallel manner and then concatenated contextual features of different scales from the channel dimension. A standard convolution is used to fuse features at different scales. It can be expressed as:

$$X_m' = Conv_{3\times3}(Concat(ReLU(Dconv_{3\times3}^d(x_{sub}^i)), \cdots,)) \tag{12}$$

where $(d, i) \in \{(1,1),(2,2),(4,3),(8,4)\}$. Spatially-variant features are calculated by a simple gating mechanism[41]. The output of the MSDA module can be obtained by weighted feature operation and collaborative attention[42].

$$X_m = CoorAtt\left(X_2 * \left(1 - Gate\left(X_2\right)\right) + X'_m * Gate\left(X_2\right)\right) \quad (13)$$

Aggregating: the feature maps X_s and X_m from the SwinT module and MSDA module are finally concatenated as the input of a 1×1 convolution which has a residual connection with the input tensor X_{in}. Through the split-transformation-merge strategy, the final output X_{out} of STMA block is expressed as:

$$X_{out} = Conv_{1\times1}\left(Concat\left(X_s, X_m\right)\right) + X_{in} \quad (14)$$

Loss function

It is very meaningful to reconstruct the fine-grained texture of the missing area of mural painting. However, there are multiple possible outcomes for reconstructed missing areas. To generate harmonious reconstruction results, we adopt different loss functions at different stages to guide the training of the model. We first train the coarse-grained network separately in an adversarial manner to generate seemingly complete images, and the discriminator also has preliminary discriminative capabilities during this training process. The objective function in this case is:

$$\min_{G^C} \max_D L_{ad}\left(G^C, D\right) + \alpha_1 L_1(G^C) + \alpha_2 g L_{msssim}(G^C) + \alpha_3 L_{mask}(G^C) \quad (15)$$

where $L_1(G^C) = ||y - G^C(x_L, x_M)||_1$, g is the Gaussian filtering parameter. $L_{msssim}\left(G^C\right) = 1 - MSSSIM(G^C\left(x_L, x_M\right), y)$. α_1 is a weight factor for L_1. α_2 is a weight factor for L_{msssim}. α_3 is a weight factor for L_{mask}. The combination of L_1 and L_{msssim} can ensure the initial quality of the inpainted image[43], but it cannot solve the problem of color change and context consistency. To alleviate this issue, we proposed a simple but effective mask loss L_{mask}, it defined as:

$$L_{mask}(G^C) = \sum_i ||VGG_i\left(y\right) \cdot x_m - VGG_i\left(G^C(x_L, x_M)\right) \cdot x_m||_1 \quad (16)$$

where VGG_i denotes the output of the i-th layer of the pre-trained VGG-19 network. We adopted the corresponding activation output from the ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1 layers.

After the initial training, the fine-grained network is also introduced into the training of the model. Moreover, to reconstruct more refined texture details and colorization, we introduce a histogram loss L_{hist}[44] based on Eq. (15), which is beneficial to guide the pixel distribution of the inpainted image to maintain the original variance and standard deviation. $L_{hist} = \sum_i \tau ||y - h(\hat{y})||$, where $h(.)$ is the histogram matching operation, τ is the control factor. Therefore, the objective function is given by

$$\min_{G^C, G^F} \max_D L_{ad}\left(G^C, G^F, D\right) + \alpha_1 L_1 + \alpha_2 g L_{msssim} + \alpha_3 L_{mask} + \alpha_4 L_{hist} \quad (17)$$

where α_4 is a weight factor for L_{hist}.

Results

Experimental environment and setting

The proposed STMA-GAN is implemented with the Pytorch framework and runs on a platform with a Nvidia ® Tesla V100 SXM2. Part of the datasets for model training and testing were taken from ref. 27 We also obtained 2000 murals through web crawling technology and scanning from the mural album. Finally, the dataset used for model training after manual screening was 3500 murals. 50 murals were used for model testing. For mask images, we adopted the irregular mask dataset publicly available by Nvidia[45]. This dataset contains a large number of irregular masks with randomized shapes, sizes, and positions. The diversity and complexity of mask images can simulate large areas of random-shaped defects on the mural, such as falling-off blocks and mud stains. We multiply the mural with the binary mask image to simulate the damaged mural. To generate the edges of the mural, we first preprocess the mural using bilateral filtering to smooth out the noise, and then use DexiNed[46] to obtain the fine edges of the mural. Figure 4 illustrates some of the murals, line drawings, and binary mask images in the dataset.

A two-stage training strategy is employed during the training process. However, the multi-stage training strategy can cause the problem of cumulative error in the training process of the model. This is because the results generated by the model during the previous stage of training are used as inputs to the model in the later stage. The output of the model in the previous stage is highly randomized, which can feed bad results into the next stage eventually leading to cumulative errors. To alleviate this issue, we only trained the coarse-grained network in the first stage. After 40,000 epochs, the training process of the coarse-grained network and the discriminator became stable, and the coarse-grained network could generate murals with better structure. Then, the second stage of training is carried out, where the fine-grained network is introduced into the model to train together. Since the coarse-grained network can already generate good results and the discriminator is also well-trained, the fine-grained network also converges faster under their guidance. The proposed model was trained for a total of 1e5 epochs with a mini-batch size of 2. The chosen number of epochs was based on the observation that validation accuracy plateaued after approximately 1e5 epochs, and further training did not significantly improve performance. The Adam optimizer ($\beta_1 = 0.0, \beta_2 = 0.9$) was used for model training with a learning rate of 1e-3. The weight factor α_1 of L_1 is set to 1, the weight factor α_2 of L_{msssim} is set to 10, the weight factor α_3 of L_{mask} is set to 0.1, the weight factor α_4 of L_{hist} is set to 0.0005. The loss term weights were determined by performing a hyperparameter search on 20 validation murals. In the STMA block, MSDA contains four parallel dilated convolutions with the dilation rates d set to 1, 2, 4, and 8, respectively. The SwinT module adopts the default parameter of the Swin Transformer.

Comparison experiment

To validate the mural restoration performance of the model, it was evaluated both quantitatively and qualitatively. We compared the proposed model with Edge Connect[16], CTSDG[47], AOT-GAN[4], ref. 27, and MAT[48].

For quantitative assessment, PSNR[49], SSIM[49], and MSE[50] metrics were used to evaluate the restored murals from the perspective of image quality. Generally speaking, the higher the values of the PSNR and SSIM metrics and the smaller the value of the MSE metric, the better the quality of the recovered image. FID[51] and LIPS[52] metrics to assess the characteristics of restored murals from the perspective of data distribution. The smaller the values of these two metrics, the more similar the restored mural feature distribution is to that of the real mural image. Table 1 lists the results of the quantitative comparison of the relevant methods.

As can be seen from Table 1, our proposed method achieves better results both in terms of distribution-based evaluation metrics and image-based evaluation metrics.

For the qualitative assessment, Figs. 5 and 6 show the results of the restoration of the damaged murals by the relevant methods. As can be seen from Fig. 5, Edge Connect blurs the content of the restored mural, CTSDG causes artifacts in the restored mural, and AOT-GAN do not refine the texture of the restored mural. The ref. 27 method is more effective in the overall restoration of murals. MAT does not restore the color and texture details of the mural well. STMA-GAN achieves better results in the detailing of restored murals. As can be seen from Fig. 6, although Edge Connect, CTSDG, and ref. 27 realize the restoration of the mural, the hanging ornaments next to the figures are not restored. MAT method resulted in the appearance of misplaced image content in the restored murals. AOT-GAN also restored only a small part of the hanging. STMA-GAN not only

Fig. 4 | Murals, line drawings, and binary mask images.

Table 1 | the results of the quantitative comparison of the related methods.

Method	FID↓	LIPS↓	PSNR↑	SSIM↑	MSE↓
Edge Connect	103.96	0.3813	17.21	0.5201	0.0188
CTSDG	93.23	0.3763	18.22	0.5475	0.0168
AOT-GAN	85.01	0.3237	18.29	0.5621	0.0167
Ref. 27	63.14	0.2439	20.38	0.6776	0.0110
MAT	94.21	0.3998	17.16	0.4198	0.0290
STMA-GAN	**59.37**	**0.2136**	**21.31**	**0.7126**	**0.0085**

Bold indicates the best metric.

achieves the restoration of the mural but also the complete recovery of the missing hangings.

Ablation study

To better understand the impact of STMA blocks on the performance of the proposed method, we performed ablation experiments. We trained a model without the STMA block (denoted as STMA^1-GAN), a model with only the MSDA module in the STMA block (denoted as STMA^2-GAN), and a model with only the SwinT module in the STMA block (denoted as STMA^3-GAN). Table 2 shows the comparison results of the performance of different models. It can be seen from Table 2 that the performance of the model decreases when the STMA module is removed, no matter from the perspective of data distribution or image quality. The performance of STMA^2-GAN and STMA^3-GAN is better than STMA^1-GAN, which implies that using only a single component of STMA does help to improve the performance of the model, and the SwinT module improves the model performance more. The performance of the STMA-GAN model is the best, the main reason is that SwinT can capture global context information, and MSDA is good at extracting local features. Combining the two can make full use of their respective advantages and capture global and local information at the same time, to improve the performance of the model.

A visual effect comparison result of ablation experiments is shown in Fig. 7. From the overall visual effect, all models have better completed the repair of the damaged mural. However, it can be seen from Fig. 7d that the murals repaired by STMA^1-GAN have color deviations. The restored area showed more green color. The color deviation of STMA^2-GAN and STMA^3-GAN is not very obvious, but the effect of texture detail restoration is not good. STMA-GAN achieves the best visualization results, which suggests that combining SwinT and MSDA can fully utilize their respective advantages. STMA is beneficial to improve the performance of models in color correction and detail restoration. This demonstrates that the STMA block can not only improve the learning ability and feature representation ability of the proposed model but also correct the color of the inpainted mural to some extent.

The loss function also plays an important role in guiding model training, and we performed ablation experiments on the loss function as well. Since L_1 and L_{msssim} are common combinations, we only explore L_{hist} and L_{mask}. A model without L_{hist} is denoted as STMA-GAN^1 and a model without L_{mask} is denoted as STMA-GAN^2. As can be seen in Fig. 8, STMA-GAN^1 and STMA-GAN^2 are weaker than the performance of STMA-GAN, which indicates that the missing L_{mask} and L_{hist} reduces the performance of the model. L_{mask} calculates the differences between mask feature maps, which helps the model to more accurately localize the feature areas that need to be repaired during training. It avoids causing unnecessary modifications to undamaged areas. L_{hist} calculates the difference in the distribution of gray levels of the murals, which is beneficial to guide the model to optimize the contrast of the murals during training. It ensures that the restored mural is consistent with the ground truth in the grayscale distribution, making the restored mural more natural and realistic. Therefore, L_{mask} and L_{hist} are beneficial to guide the training process of the model.

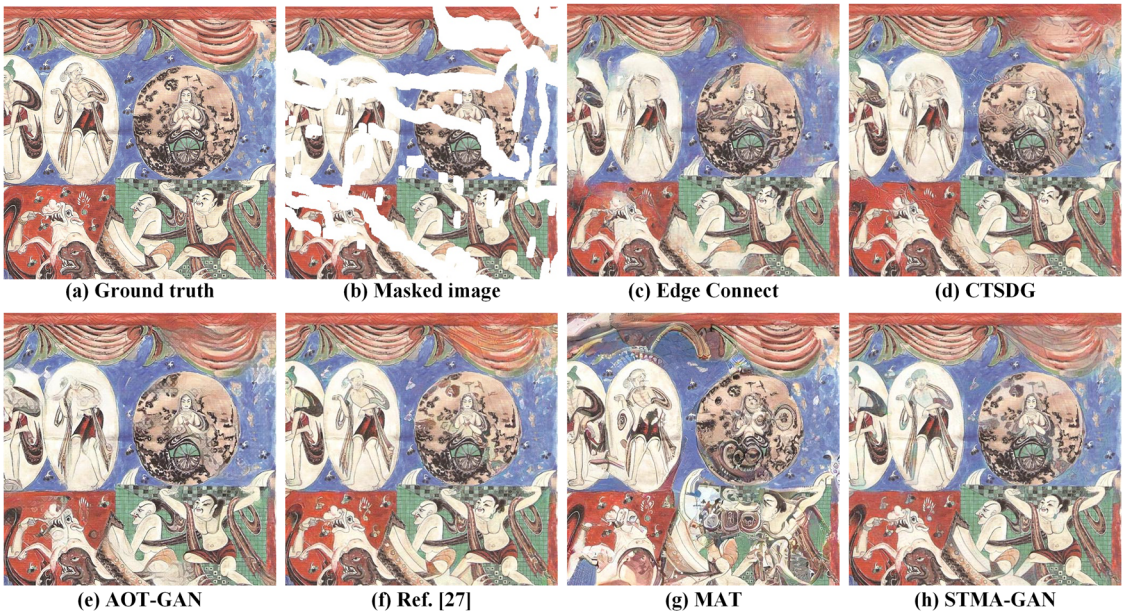

Fig. 5 | Qualitative comparisons of the related methods. **a** Ground truth. **b** Masked image. **c** Repair result of Edge Connect. **d** Repair result of CTSDG. **e** Repair result of AOT-GAN. **f** Repair result of ref. [27]. **g** Repair result of MAT. **h** Repair result of the proposed method.

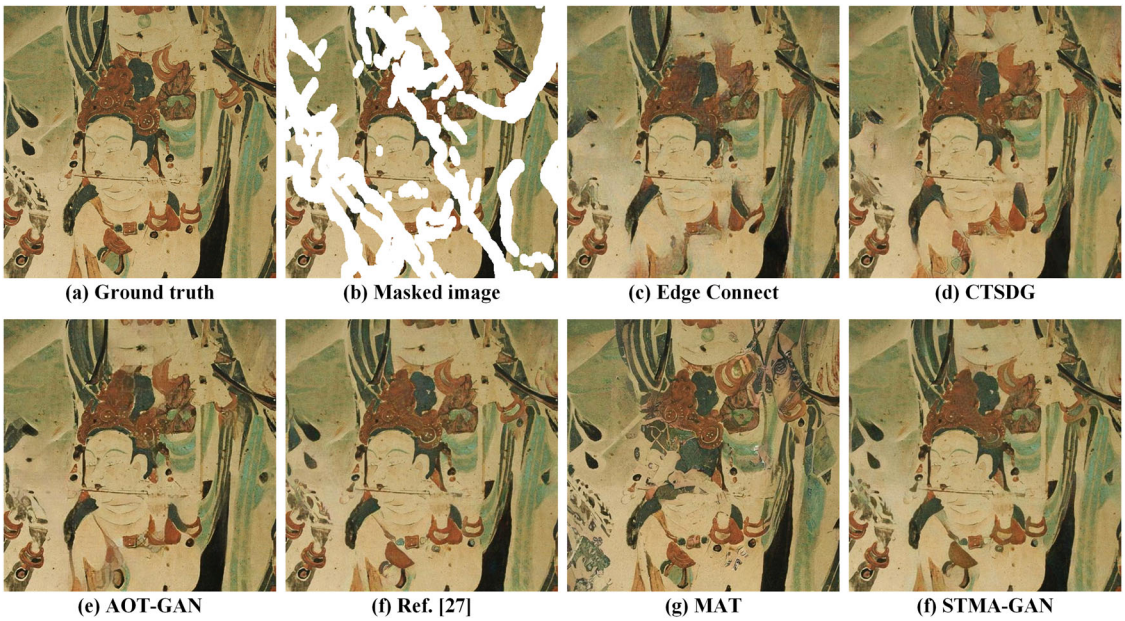

Fig. 6 | Qualitative comparisons of the related methods. **a** Ground truth. **b** Masked image. **c** Repair result of Edge Connect. **d** Repair result of CTSDG. **e** Repair result of AOT-GAN. **f** Repair result of ref. [27]. **g** Repair result of MAT. **h** Repair result of the proposed method.

Table 2 | Comparative results of ablation experiments.

Method	FID↓	LIPS↓	PSNR↑	SSIM↑	MSE↓
STMA[1]-GAN	62.72	0.2182	21.06	0.7034	0.0090
STMA[2]-GAN	64.64	0.2165	**21.44**	0.7132	0.0097
STMA[3]-GAN	59.60	0.2143	21.17	**0.7155**	0.0088
STMA-GAN	**59.37**	**0.2136**	21.31	0.7126	**0.0085**

Bold indicates the best metric.

User perceptual study

We also conducted a user study to evaluate the restoration performance of different methods for murals. We invited 20 users to evaluate 50 restored murals restored by different methods. The user scores the restored mural on a scale from 1 to 5 (from worst to best) concerning the original mural provided by us. The final average rating values are presented with two decimal places. Figure 9 shows the user ratings of the restored murals from the perspective of the overall visual effect and detailed textures. The visual effect evaluation of mural restoration is mainly on the naturalness and fidelity of the restored area. From an overall visual perspective, our proposed method received the highest ratings compared to the comparison methods, indicating that the mural restoration results of our proposed method are more acceptable to users. The evaluation of detail texture in mural restoration is mainly to restore the details of the repaired area and realize the smooth transition of the edge. As can be seen from Fig. 9, users also gave high ratings to our proposed method in terms of mural detail restoration,

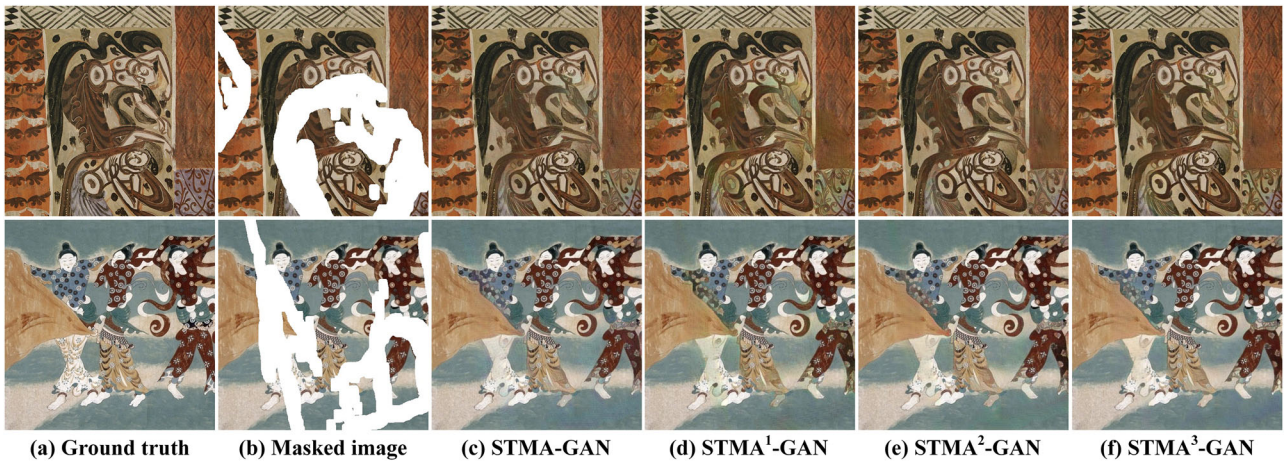

Fig. 7 | Comparison results of ablation experiments on STMA blocks. a Ground truth. **b** Masked image. **c** Repair result of STMA-GAN. **d** Repair result of STMA^1-GAN. **e** Repair result of STMA^2-GAN. **f** Repair result of STMA^3-GAN.

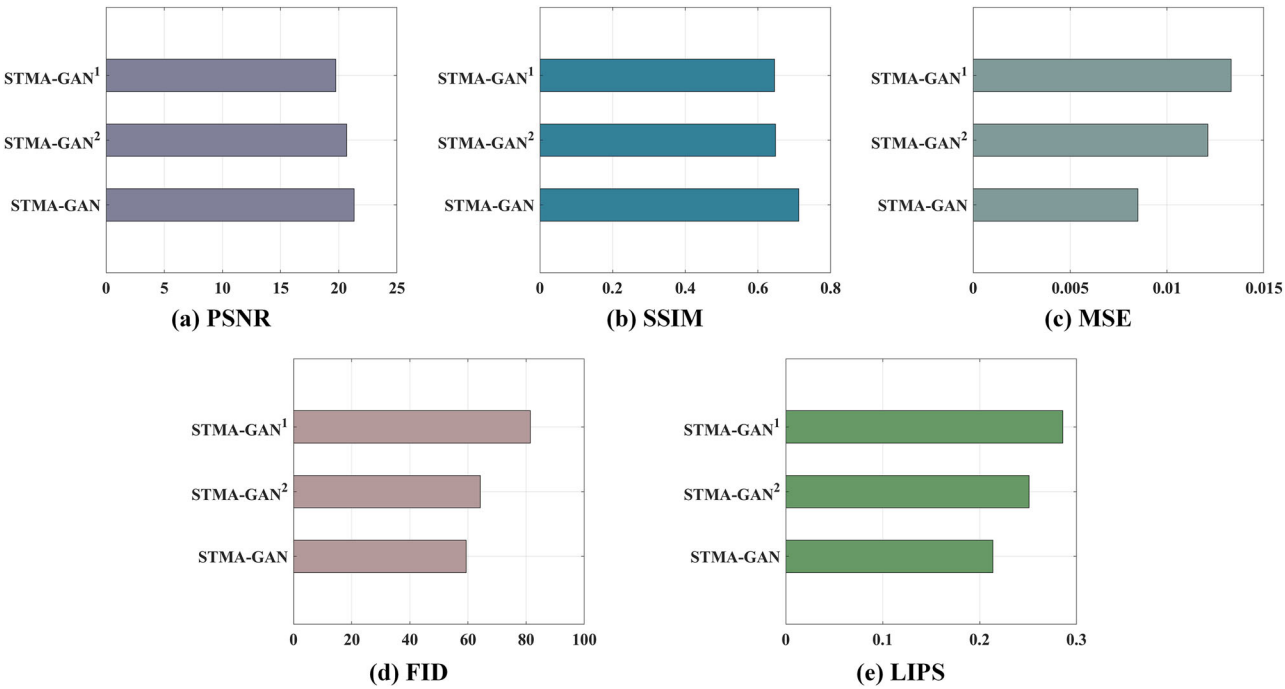

Fig. 8 | Comparative results of ablation experiments on loss function. a Comparison results of different models in PSNR. **b** Comparison results of different models in SSIM. **c** Comparison results of different models in MSE. **d** Comparison results of different models in FID. **e** Comparison results of different models in LIPS.

Fig. 9 | User evaluation. a Evaluation performed by users on the restoration results of different methods in terms of visual quality. **b** Evaluation of the inpainting results of different methods by users in terms of detail textures.

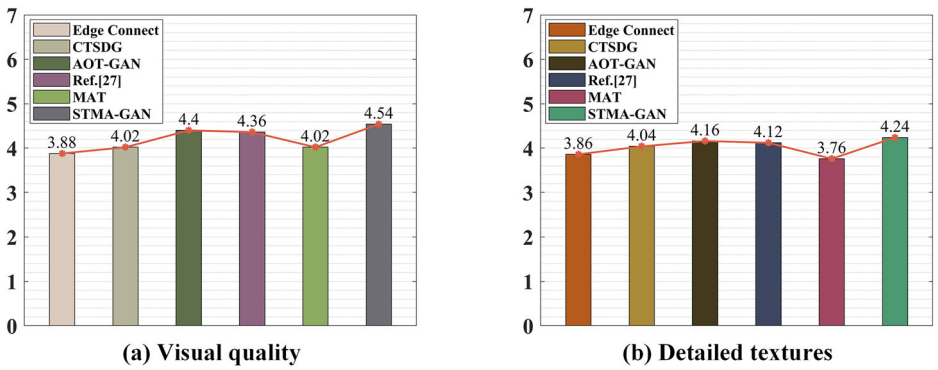

Fig. 10 | An example. a Masked image. **b** Ground truth. **c** Repair result of the proposed method.

which indicates that our proposed method ensures the clarity of the details of the restored mural.

Discussion

Image inpainting is essentially an ill-posed inverse problem. Inpainting images with large damaged areas is still a very difficult challenge. According to the above experimental results, our proposed method achieves a relatively ideal visual effect, but there are still some problems when inpainting the more seriously damaged images. From Fig. 10, it can be seen that our proposed method can only repair the general contour of the face and fails to refine the texture details of the face (yellow-bordered area). Moreover, the proposed method also fails to estimate the original image content well (green-bordered area).

This means that our proposed model still has some limitations in the image details of refined restoration and model reasoning ability. The marked areas in Fig. 10 are areas where the mural is more severely damaged or where the structural information is more complex. The large amount of missing information leads to the fact that our proposed model cannot rely on the contextual information of the image and the surrounding pixel information to infer the content of the missing area. Although our proposed model can generate restoration results that are consistent with the surrounding area, when the missing area is large and contains complex textures or objects, the model may not be able to generate textures and refined structures that exactly match the surrounding area, resulting in unnatural restoration results. In future work, image prior technology or advanced network backbone can be introduced to further improve the degree of refinement of the repaired image.

In addition, we use supervised learning to train the proposed model. However, the ground truth of the mural is lacking in the actual scene. Creating ground truth for murals is a labor-intensive and time-consuming process that requires expertise in art history, archeology, and possibly digital restoration. This makes it difficult to scale the supervised learning approach to a larger number of murals. Even if ground truth murals are available, they may not always be accurate or comprehensive due to human error, incomplete information, or changes over time. Inaccuracy in labels can lead to poor model performance and misleading results. Here are some potential solutions: (1) Some prior information of the mural is integrated into the training process of the model, to better guide the repair process of the model. (2) Exploring unsupervised or semi-supervised learning methods can mitigate the dependence on labeled data. (3) Utilizing transfer learning from related domains (e.g., image semantic information in general art) can help improve the model's ability to generalize to complex mural restoration problems. Pre-trained models can be fine-tuned on limited labeled mural data to adapt to specific mural features. (4) Multi-modal or cross-modal learning techniques can leverage information from multiple sources to provide a more comprehensive understanding of murals.

The proposed method is mainly aimed at solving the large area of irregular shape damage on the mural, such as falling-off blocks and mud stains. This makes our method have certain limitations in the task of repairing different types of damaged murals. Murals, as invaluable cultural artifacts, face a wide range of degradation issues, including cracking, peeling, fading, and biological growth. These damages are often compounded by historical factors such as the materials used in their creation, environmental conditions, and human activities. In future work, we will design masks for different types of damaged murals and a unified restoration model for different types of degraded murals.

Inspired by the process of manual mural inpainting, we propose a novel two-stage GAN for mural inpainting. The main goal of the first stage is to use a coarse-grained network to achieve a coarse-grained semantic reconstruction of the damaged mural, to provide a basis for the subsequent finer restoration. The second stage focuses on achieving fine-grained feature reconstruction to restore local features and detailed information in the mural. The proposed STMA block combines the advantages of the Swin transformer module and the multi-scale dilated convolution attention module to enhance the overall learning and restore ability of the model. The training process of the model is guided by introducing different loss functions in the two-stage training strategy to ensure the restored performance of the model. Abundant experiments show that our proposed method can achieve a good restoration of murals.

In our future work, we will further improve the quality of mural restoration in terms of a model learning approach, model training strategy, and multimodal information fusion to promote the digital preservation and communication of cultural heritage. At the same time, to repair murals subjected to different types of damage, we will also design new masks and mural restoration models.

Data availability

The mural dataset in this study is available at https://1drv.ms/u/s!AittnGm6vRKLzXorf1nkiDPRQB4D?e=Avv27i .

Received: 17 September 2024; Accepted: 12 April 2025;
Published online: 19 May 2025

References

1. Xu, W. & Fu, Y. Deep learning algorithm in ancient relics image colour restoration technology. *Multimedia Tools Appl* **82**, 23119–23150 (2023).
2. Liu, W. et al. Multi-stage progressive reasoning for dunhuang murals inpainting. In *Proc. 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*. 211–217 (IEEE, 2023).
3. Zheng, C., Cham, T. J., Cai, J. & Phung, D. Bridging Global Context Interactions for High-Fidelity Image Completion. In *Proc. 2022 IEEE/*

CVF Conference on Computer Vision and Pattern Recognition (CVPR). 11502–11512 (IEEE, 2022).
4. Zeng, Y., Fu, J., Chao, H. & Guo, B. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transac. Visualiz. Comput. Gr.* **29**, 3266–3280 (2023).
5. Quan, W. et al. Image Inpainting with Local and Global Refinement. *IEEE Transac. Image Process.* **31**, 2405–2420 (2022).
6. Li, Y. et al. Efficient and explicit modelling of image hierarchies for image restoration. In *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada. 18278–18289 (IEEE, 2023).
7. Yang, J., Ruhaiyem, N. I. R. & Zhou, C. A 3M-Hybrid model for the restoration of unique giant murals: a case study on the murals of Yongle Palace. *IEEE Access* **13**, 38809–38824 (2025).
8. Liang, J., Cao, J., Sun, G., Zhang, K. & Timofte, R. SwinIR: image restoration using swin transformer. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Montreal, BC, Canada, 1833–1844, (IEEE, 2021).
9. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada, 9992–10002 (IEEE, 2021).
10. Chen, G., Gao, Z., Zhu, P. & Chen, Z. Learning a multi-scale deep residual network of dilated-convolution for image denoising. In *Proc. 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, Chengdu, China, 348–353 (IEEE, 2020).
11. Ople, J. J. M., Yeh, P. Y., Sun, S. W., Tsai, I. T. & Hua, K. L. Multi-scale neural network with dilated convolutions for image deblurring. *IEEE Access* **8**, 53942–53952 (2020).
12. Burger, H. C., Schuler, C. & Harmeling, S. Learning how to combine internal and external denoising methods. In *Proc. Pattern Recognition: 35th German Conference*, GCPR 2013, Saarbrücken, Germany, **8142**, 121–130 (2013).
13. Zhang, K. et al. Practical blind image denoising via Swin-Conv-UNet and data synthesis. *Mach. Intell. Res.* **20**, 822–836 (2023).
14. Goodfellow, I. et al. Generative adversarial networks. *Commun ACM* **63**, 139–144 (2020).
15. Liu, H. et al. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA, 9367–9376 (IEEE, 2021).
16. Nazeri, K., Ng, E., Joseph, T. Qureshi, F. & Ebrahimi, M. EdgeConnect: Structure-Guided image inpainting using edge prediction. In *Proc. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South), 3265–3274 (IEEE, 2019).
17. Wan, Z., Zhang, J., Chen, D. & Liao, J. High-fidelity pluralistic image completion with transformers. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada, 4672–4681 (IEEE, 2021).
18. Jiang, J. et al. Research on digital image restoration technology of Xizang Murals Based on CDD Model. *Electron. Design Eng.* **22**, 177–179 (2014).
19. Chen, Y., Ai, Y. & Guo, H. Inpainting Algorithm for Dunhuang mural based on improved curvature-driven diffusion model. *J. Comput. Aided Design Comput. Gr.* **32**, 787–796 (2020).
20. Wu, M. & Wang, H. Research on multi-scale detection and image inpainting of Tang dynasty tomb murals. *Comput. Eng. Appl.* **52**, 169–174 (2016).
21. Cao, J., Li, Y., Zhang, Q. & Cui, H. Restoration of an ancient temple mural by a local search algorithm of an adaptive sample block. *Heritage Sci.* **7**, 39 (2019).
22. Yang, X. P. & Wang, S. W. Dunhuang mural inpainting in intricate disrepaired region based on improvement of priority algorithm. *J. Comput. Aided Design Comput. Gr.* **23**, 284–289 (2011).
23. Chen, Y., Chen, J. & Tao, M. F. Mural inpainting progressive generative adversarial networks based on structure guided. *J. Beijing Univ. Aeronaut. Astronaut.* **49**, 1247–1259 (2023).
24. Chen, Y., Zhao, M. X. & Tao, M. F. Mural inpainting algorithm for group sparse based on multi-scale contourlet transform decomposition. *J. Xidian Univ.* **49**, 120–128 (2022).
25. Chen, Y., Ai, Y. P. & Chen, J. Algorithm for Dunhuang Mural inpainting based on Gabor transform and group sparse representation. *Laser Optoelectron. Progress* **57**, 175–184 (2020).
26. Cao, J., Zhang, Z., Zhao, A., Cui, H. & Zhang, Q. Ancient mural restoration based on a modified generative adversarial network. *Heritage Sci.* **8**, 7 (2020).
27. Li, L. et al. *Line Drawing Guided Progressive Inpainting of Mural Damages*. arXiv preprint arXiv: 2211.06649 (2022).
28. Chen, Y., Chen, J. & Tao, M. F. Mural inpainting with generative adversarial networks based on multi-scale feature and attention fusion. *J. Beijing Univ. Aeronaut. Astronaut.* **49**, 254–264 (2023).
29. Li, J., Shi, Y., Liu, W., Wang, J. & Du, S. Progressive dunhuang murals inpainting based on attention mechanism. In *Proc. 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*. Urumqi, China, 350–356 (IEEE, 2023).
30. Wang, N., Wang, W., Hu, W., Fenster, A. & Li, S. Thanka mural inpainting based on multi-scale adaptive partial convolution and stroke-like mask. *IEEE Transac. Image Process.* **30**, 3720–3733 (2021).
31. Deng, X. & Yu, Y. Ancient mural inpainting via structure information guided two-branch model. *Heritage Sci.* **11**, 131 (2023).
32. Khaled, A. BCN: Batch Channel Normalization for Image Classification. In: Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, CL., Bhattacharya, S., Pal, U. (eds) *Pattern Recognition*. ICPR 2024. Springer, Cham,**15311**, 295–308 (2025).
33. Oyedotun, O. K., Aouada, D. & Ottersten, B. Going deeper with neural networks without skip connections. In *Proc. 2020 IEEE International Conference on Image Processing (ICIP)*. Abu Dhabi, United Arab Emirates, 1756–1760 (IEEE, 2020).
34. Mao, X. J., Shen, C. & Yang, Y. B. Image restoration using convolutional auto-encoders with symmetric skip connections. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Barcelona, Spain, 2810–2818 (2016).
35. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral Normalization for Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations (ICLR,* 2018).
36. Qi, G. J. Loss-sensitive generative adversarial networks on lipschitz densities. *Int. J. Comput. Vision* **128**, 1118–1140 (2020).
37. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 5987–5995 (IEEE, 2017).
38. Tolstikhin, I. O. et al. MLP-Mixer: An all-MLP Architecture for Vision. *Adv. Neural Inform. Process. Syst.* **34**, 24264–24272 (2021).
39. Xiang, X., Liu, M., Zhang, S., Wei, P. & Chen, B. Multi-scale attention and dilation network for small defect detection. *Pattern Recognit. Lett.* **172**, 82–88 (2023).
40. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*. 1–9 (ICLR, 2016).
41. Zhang, Y. et al. Image super-resolution using very deep residual channel attention networks. *In Proc. Eur Conf Comput Vis (ECCV).* **11211**, 294–310 (2018).
42. Hou, Q., Zhou, D. & Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA, 13708–13717 (IEEE, 2021).

43. Zhao, H., Gallo, O., Frosio, I. & Kautz, J. Loss functions for image restoration with neural networks. *IEEE Transact. Comput. Imaging* **3**, 47–57 (2017).
44. Risser, E., Wilmot, P. & Barnes, C. Stable and controllable neural texture synthesis and style transfer using histogram losses. arXiv preprint arXiv:1701.08893 (2017).
45. Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A. & Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. **11215**, 89–105 (2018).
46. Poma, X. S., Riba, E. & Sappa, A. Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection. In *Proc. IEEE/CVF winter conference on applications of computer vision (WACV)*. Snowmass, CO, USA, 1912–1921 (IEEE, 2020).
47. Guo, X., Yang, H. & Huang, D. Image Inpainting via Conditional Texture and Structure Dual Generation. In *Proc. IEEE/CVF international conference on computer vision (ICCV)*, Montreal, QC, Canada, 14134–14143 (IEEE, 2021).
48. Li, W. et al. MAT: Mask-aware transformer for large hole image inpainting. In *Proc. IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. New Orleans, LA, USA, 10758–10768 (IEEE, 2022).
49. Horé, A. & Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In *Proc. 20th International Conference on Pattern Recognition*, 2366–2369 (IEEE, 2010).
50. Allen, D. M. Mean square error of prediction as a criterion for selecting variables. *Technometrics* **13**, 469–475 (1971).
51. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* **30**, 6629–6640 (2017).
52. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE conference on computer vision and pattern recognition(CVPR)*. Salt Lake City, UT, USA, 586–595 (IEEE, 2018).

Acknowledgements

This work was supported by the Youth Project of Humanities and Social Sciences Research of the Ministry of Education(No. 24YJCZH199), the key scientific research projects of universities in Henan Province (No. 24A520034, No. 24A520033), Henan Xing culture Project Cultural research special project (No. 2024XWH198), Henan Province colleges and universities young backbone teacher training program (No. 2023GGJS148), the Doctoral Research Start-up Fund of Pingdingshan University (No. PXY-BSQD-2023019, No. PXY-BSQD-2024008), the Henan Province Key R&D Promotion Special Project (No. 232102210011), Project of International Cultivation of Henan Advanced Talents.

Author contributions

Qiongshuai Lyu is responsible for implementing the model code, algorithm support, and original manuscript, Na Zhao contributed to the review and revision of the manuscript, Junke Song contributed to data management and visualization, Yu Yang performed the analysis with constructive discussions, Yuehong Gong polished the manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Qiongshuai Lyu or Na Zhao.