applied sciences

MDPI

*Article*

# Fragments Inpainting for Tomb Murals using a Dual-Attention Mechanism GAN with Improved Generators

**Meng Wu [1,2,*], Xiao Chang [1,*] and Jia Wang [3]**

[1] School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
[2] Institute for Interdisciplinary and Innovate Research, Xi'an University of Architecture and Technology, Xi'an 710055, China
[3] Shaanxi History Museum, Xi'an 710061, China; wangjia@sxhm.com
* Correspondence: wumeng@xauat.edu.cn (M.W.); 2108811639@xauat.edu.cn (X.C.)

**Abstract:** As the only underground mural in the collection, the tomb murals are subject to damage due to temperature, humidity, and foundation settlement changes. Traditional mural inpainting takes a long time and requires experts to draw it manually. Therefore, the need for digital inpainting is increasing to save time and costs. Due to the scarcity of samples and the variety of damage, the image features are scattered and partially sparse, and the colors are less vivid than in other images. Traditional deep learning inpainting causes information loss and generates irrational structures. The generative adversarial network is, recently, a more effective method. Therefore, this paper presents an inpainting model based on dual-attention multiscale feature aggregation and an improved generator. Firstly, an improved residual prior and attention mechanism is added to the generator module to preserve the image structure. Secondly, the model combines spatial and channel attention with multiscale feature aggregation to change the mapping network structure and improve the inpainting accuracy. Finally, the segmental loss function and its training method are improved.The experimental results show that the results of using signal-to-noise ratio (PSNR), structural similarity (SSIM), and mean square error (MSE) on epitaxial mask, crack mask, random small mask, and random large mask are better than other methods. It demonstrates the performance of this paper in inpainting different diseases of murals. It can be used as a reference for experts in manual inpainting, saving the cost and time of manual inpainting.

**Keywords:** multiscale feature aggregation; double attention; GAN; image inpainting; tomb mural

check for updates

**Citation:** Wu, M.; Chang, X.; Wang, J. Fragments Inpainting for Tomb Murals using a Dual-Attention Mechanism GAN with Improved Generators. *Appl. Sci.* **2023**, *13*, 3972. https://doi.org/10.3390/app13063972

Academic Editor: Ion Sandu

Received: 27 February 2023
Revised: 15 March 2023
Accepted: 17 March 2023
Published: 21 March 2023

**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

The murals in the burial tombs reflect the humanistic and technological development of the time and are of excellent research significance. The tomb murals were generally painted on the mudbrick walls of the burial tombs and suffered from the difficulties of excavation, uncovering them, and the fact that they were easily oxidized after uncovering and could not be easily preserved. Being buried deep underground, they have been damaged by natural factors and human excavation. The surviving murals in the burial tomb are scarce, with little information on the various styles, and are diseased and challenging to inpaint. Many typical diseases are associated with tomb murals, such as cracks in the murals; large areas of the tomb with hollow drums causing the images to peel off; pigment layer warping and peeling; salt efflorescence and mold in the murals; artificial excavation. The cracks and small amounts of detachment are easy to solve. However, the large areas of detachment and salt efflorescence are difficult to solve. Traditional mural inpainting uses manual inpainting, which is irreversible and may result in secondary damage. Digital inpainting makes this problem a solution. Digital image inpainting is also a hot issue in computer vision, aiming to fill in missing areas in the image. The image should be drawn so that a reasonable texture structure is maintained and the interior is drawn more realistically. For

example, Karianakis et al. [1] performed crack detection in missing areas of Tehran murals. An improved TV algorithm was used to detect small cracks. The algorithm combines region- and boundary-based edge detection with the removal of redundant information. It can handle small cracks but is poorer for large-area image inpainting. Jaidiler et al. [2] segmented the scratch locations using self-selected seeded area growth methods and morphological manipulations. They used variation to fill the missing areas with pixels to inpaint the Thai mural. It works well for small areas of damage, such as scratches. Cao et al. [3] proposed a consistency-enhanced generative adversarial network (GAN) to implement ancient mural inpainting. The generative network used a fully convolutional network (FCN) as the basic framework, combining local and global discriminative networks. The algorithm was suitable for small areas of mural exfoliation. Zhou et al. [4] used the mean filter template to inpaint the color of the digital image of the Monastery mural. Then, they used the multimodal feature decomposition method to obtain the depth learning model of the local fuzzy feature repair of the mural, which effectively repaired the image. Cao et al. [5] proposed an adaptive sample block and local search algorithm to inpaint Kaihua Temple murals. Based on the Criminisi algorithm, they introduced a structural tensor, adaptively selected sample block size, and adopted a local search strategy to improve matching efficiency. Priego et al. [6] used 3D laser scanning technology to inpaint a pictorial group atAntonio Palomino. It is based on using a projection that, minimizing metric deformations, reconstructs the original shapes of the images with the greatest possible fidelity. Zeng et al. [7] added the nearest-neighbor-based pixel matching to a convolutional neural network (CNN) to compensate for the missing high-frequency information and used multiscale features to extract more effective information. The experiment repaired the small regular missing featuresof Dunhuang murals. Gupta et al. [8] used automatic mask generation based on mask region-convolution neural networks (Mask R-CNN). U-Net architecture with partial convolution and automatic mask update was used to inpaint the paintings. The problem of creases and minor damages in oil paintings was effectively solved. The existing methods are better for repairing small areas of broken images but weaker for repairing large areas of broken mural images. With the development of deep learning, various image inpainting models have emerged that can effectively perform image painting tasks in certain common situations. However, unlike other image inpainting, mural inpainting is based on the principle of "inpainting the old as the old" and needs to be consistent with the original image. The inpainting of murals also faces many dilemmas.

The acceptance domain is small. Traditional convolutional neural networks (CNNs) process images with a small number of layers and a small receptive domain. After different convolutional layers, it only focuses on small receptive fields and lacks the learning of semantically consistent textures. Although expanded sensory field response is now proposed [9], it still does not handle large broken images well with a small dataset. There is a lack of holistic understanding. When there is no overall understanding of the large image, it is not easy to inpaint the key edges and lines of the scene. For mural data, the case of a small number of crucial sparse features contained in a large area of blank information is challenging for inpainting perfectly. Inconsistent inpainting results. Many current inpainting networks can inpaint murals, but the inpainting results are structurally different from the original reference murals, focusing on the imagery of the model and needing more ability to inpaint the image accurately [10].

This paper proposes a method based on improved generators and dual-attention feature aggregation to handle the missing mural fragments. In addition, this paper also changes the training method of the loss function. The method can reduce the computational cost and eliminate the optimum local problem. The main contributions of this paper are as follows.

1. Proposing a multiscale feature fusion mechanism based on double attention: an improved method with a dual-attention mechanism is used, based on the spatial information extracted by the network at different scales. We reduce the mutual

interference of high intrapixel correlations and improve the impact of attentional feature maps on the overall feature analysis.
2. Improving the generator model: we add improved residual priors and attention mechanisms to improve the accuracy of the repaired structure. Our method addresses the problem of structure loss in the basic model by adding additional image priors to preserve the image structure and it obtains more accurate structural information. This method emphasizes making the model more focused on features relevant to the structure in image inpainting.
3. Constructing a dataset of the tomb murals of the tomb of Prince Zhang Huai of the Tang Dynasty and inpainting them with three kinds of diseases: experimental results show that the model's performance is significantly improved compared with the current mainstream models. The tomb murals have suffered from many typical diseases associated with long-term environmental erosion and human damage. Fissures in the murals, large hollow drums in the floor of the burial tomb leading to the loss of the floor layer and missing images, artificial excavation, and other types of disease are in urgent need of repair, which is the focus of this paper.

This paper discusses related work in Section 2, describes the method in detail in Section 3, gives the results of the experimental configurations and analyzes and discusses them in Section 4, and concludes in Section 5.

## 2. Related Work

Inpainting methods are mainly divided into traditional inpainting methods and deep learning inpainting methods.

Traditional inpainting methods fall into two main areas. On the one hand, the inpainting is carried out using pixel diffusion, sample matching, and sparse representation. The pixel diffusion method will introduce a certain degree of blurring in the inpainting process, making the details of the filled image less clear; the sample matching method requires a large sample library and fine processing and classification of the sample library; the sparse representation method is less effective in inpainting images with low signal-to-noise ratio and requires fine-tuning of the algorithm parameters. For example, Chan et al. [11] proposed a curvature-driven diffusions (CDD) model using curvature diffusion intensity. As it is calculated in the variational space, only structural information is considered, not texture information, and the image is less effective when the texture is complex. On the other hand, the information generation method is based on sample filling, and the current mainstream model is based on Criminisi's sample -filling algorithm as the base model [12], for example, using the fast nearest neighbor algorithm, which can reduce the memory consumption and computational cost during the search [13]. However, the algorithm has difficulty satisfying texture consistency for images with complex textures. It results in possible blurring and distortion of the filled results and requires multiple iterations of the image, which is computationally intensive. Especially when the missing areas in the image are large, or the missing information is important, the filled results may not be accurate enough to achieve the expected inpainting results. In practical applications, the algorithm may require manual adjustment of the filled samples, which requires a high level of expertise and experience.

In artificial intelligence, convolutional neural networks (CNNs) [14] and generative adversarial networks (GANs) [15] are outstanding for their capabilities. Deep learning inpainting methods have demonstrated good results in inpainting images. Their robustness comes from the ability of deep neural networks to learn image texture, structure, and semantic information, enabling automatic or semiautomatic inpainting of defective areas. Several deep learning inpainting methods have continuously improved their effect and quality through different technical means and algorithm improvements, achieving effective image inpainting. Ren et al. [16] proposed a cascade inpainting model with coarse structure and fine texture that can perform effective inpainting. However, as the method requires joint constraints on structural and texture features, it is prone to information loss during mural

inpainting. Liu et al. [17] designed a coherent semantic attention layer, which can learn the semantic correlation between features in the missing regions of an image while preserving the contextual structure by adding it to the encoder of the U-Net structure. However, this model's lack of correspondence between missing and known regions may lead to artifacts in some mural inpainting results. Li et al. [18] proposed a circular feature inference model for image mapping that uses multiple loops and processes to refine the feature mapping and uses the correlation between neighboring pixels to strengthen the constraints on deep pixel prediction, thus improving the inpainting capability of the network at a lower computational cost. However, the method requires consideration of structural and texture differences, which can easily lead to detail ambiguity in the mural inpainting results. Guo et al. [19] considered both texture and structure in the inpainting model, allowing the two modules to complement each other, and adding contextual semantic information to make the generated samples more reasonable. This method addresses the issue of proper interaction with image textures during structural inpainting, but the lack of a priori information leads to missing details in the mural inpainting. Suvorov et al. [20] used fast Fourier convolution to increase the perceptual field, allowing the network to obtain the entire image perceptual field even at shallow layers, thus improving the quality of the model's inpainting. However, in some cases, fast Fourier convolution may introduce a certain amount of error, especially when performing inverse transformations, which should be avoided as much as possible during mural inpainting.

In summary, existing deep learning image inpainting methods are prone to problems such as structural disorder and texture blurring when inpainting murals. The reasons for this are the small number of feature points in the mural and the lack of a priori information on structure and texture to guide the inpainting process with joint constraints. This paper proposes a method for inpainting murals based on improved encoders and integrating dual-attention feature aggregation.

## 3. Methods

This section presents a feature aggregation model based on an improved encoder and fused dual-attention. A joint segmentation loss function is used to optimize the quality of the graphs generated by the network and to eliminate local optimizations. This triple aggregation is the model of this text. The effectiveness of this network depends heavily on the design of the three modules in the model: the generator module, the feature aggregation module, and the joint segmentation loss function.

### *3.1. Improving the Generator Module*

The image inpainting uses the encoder module in the unit network to extract features from the image. The mural image has sparse feature points. Firstly, an improved prior residual module is embedded in the original encoder network to obtain more accurate structural information about the image by adding the image prior. Secondly, a spatial attention mechanism is added to extract more helpful information. The improved a priori residual module general residue channel prior (GRCP) described above is as follows. Three modified residual learning modules, the general residual channel attention block (GRCAB) (Figure 1), are first concatenated. The GRCAB is connected using convolution to extract deeper features. The GRCAB module is formed into a general SE-ResBlock in the residual block (GSRiR) module, as shown in Figure 2. Embedding them into the generator model constitutes the GRCP module, as shown in Figure 3.

The GRCAB module first uses depth-separated convolution to reduce the number of parameters while obtaining the initial feature mapping. The features are fed into channel attention (CA) twice by the Dconv+Mish+Dconv module and then multiplied with the original mapping to obtain the features. This enriches the semantic information and reduces the noise of the initial features. According to Li et al. [21], the improved a priori residual module still works without using color constancy. Based on the observation that tomb murals suffer from achromatic aberrations due to oxidation over time, GRCP can extract

a more complete and accurate object structure. Using this residual structure can train a deeper CNN without gradient disappearance, and the high-frequency information of feature extraction will be present.

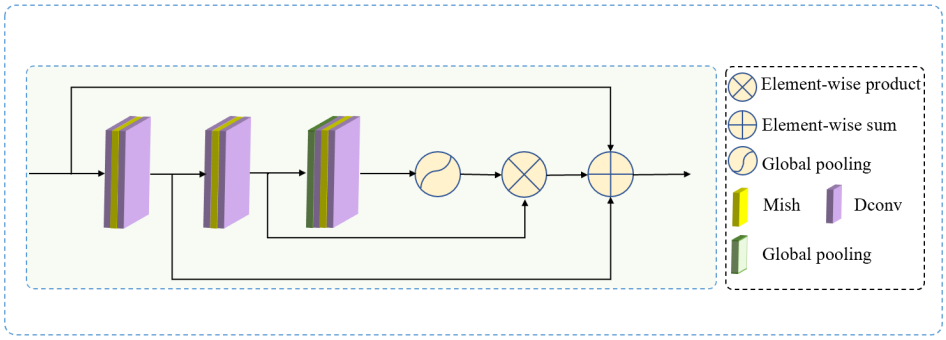


**Figure 1.** Attention module for improving the remaining channels (GRCAB).

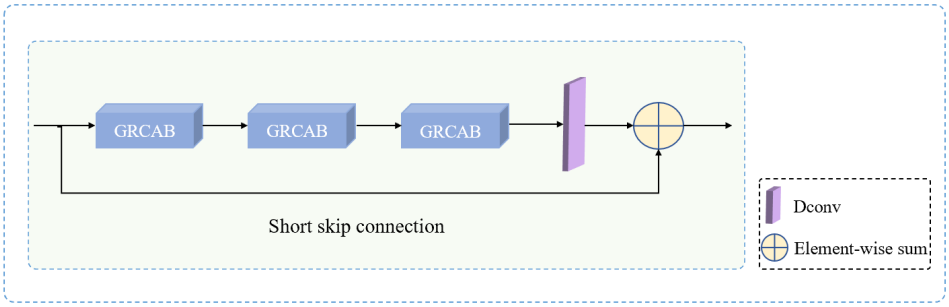


**Figure 2.** The improved residual learning module (GSRiR).

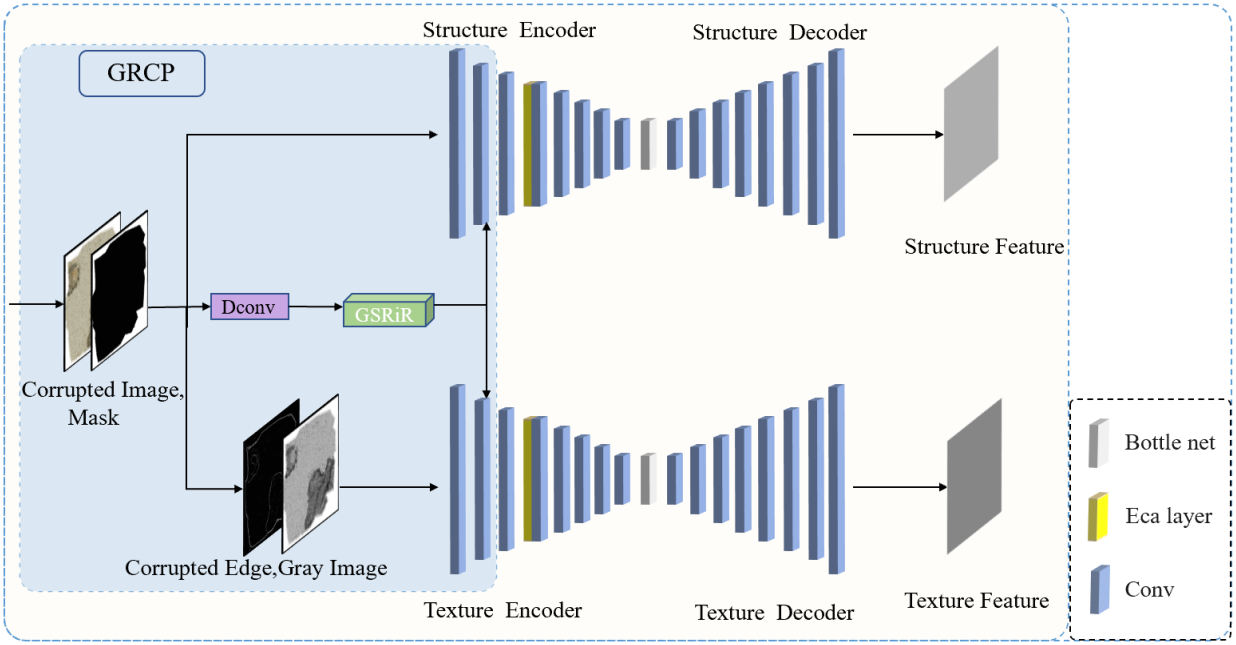


**Figure 3.** Overall generator framework.

In addition, embedding channel attention into the encoder stabilizes the training process and promotes more efficient learning of parameters in the attention layer. As correlations between the generated patches are ignored, this may lead to a lack of continuity

in the final results. This model is inspired by [17], which adds attention to the third layer of the encoder with good results. The efficient channel attention (ECA) module is added to our model, as shown in the yellow module below. Our overall generator framework is shown in Figure 3.

### 3.2. Multiscale Feature Fusion Based on Dual-Attention

In order to obtain features with strong semantic properties, traditional image inpainting models usually use only the feature map output from the last layer of the feature extraction network for object classification and localization. The last feature map corresponds to a significant downsampling rate, typically 16 or 32 times downsampling. This method results in less effective information on the last feature map and reduced detection capability. Multiscale feature fusion is an excellent solution to this problem [22]. It does not only use the last layer of the feature map for detection but selects multiple layers of features to be fused before detection. Image inpainting suffers from information loss in many convolution operations. Feature extraction is even more critical for mural images with scattered and partially sparse features.

In this regard, multiscale feature fusion based on double attention (MFDA) is used to extract features, as shown in Figure 4 below. In Figure 4, RAL refers to region affinity learning, WG refers to weights generator, MSFA refers to multiscale feature aggregation, and GAP refers to global average pooling. The contextual attention therein allows for encoding rich semantic features at multiple scales. It borrows feature information from regions with known distant spatial locations to generate features within the missing regions. It can allow multiscale features to generate more complex predictions, capturing features at different semantic levels to establish long-term spatial dependencies.

A $3 \times 3$ deep separation convolution (Dconv) operation is performed to extract more layers of feature information from the fused multiscale features. A $1 \times 1$ convolution operation is then performed to inpaint the number of channels of the features to the number of channels of the input module features. Dconv combines deep and point-by-point convolution, which can significantly reduce the computational complexity of network learning and the number of parameters and improve network efficiency. Since only spatial attention is incorporated and features in the channels are not well extracted, efficient channel attention (ECA) is introduced. A dimensionless local cross-channel interaction strategy reduces significant complexity while maintaining performance when interacting across channels. The alignment pattern of our modules is inspired by the convolutional block attention module (CBAM) [23]. The relevant alignment order experiments for the modules are discussed in detail in Section 4.

This module first extracts a $3 \times 3$ pixel patch, gives a feature map $\boldsymbol{F}$, and calculates the remaining chord similarity with the following Equation (1):

$$s_{contextual}^{i,j} = \left\langle \frac{f_i}{\|f_i\|_2}, \frac{f_j}{\left\|f_j\right\|_2} \right\rangle \tag{1}$$

where $f_i$ and $f_j$ correspond to the $i$-th and $j$-th patches of the feature map, respectively. Then, the module applies softmax to the similarities to attain the attention scores for each patch, as follows (2).

$$\hat{s}_{contextual}^{i,j} = \frac{\exp\left(s_{contextual}^{i,j}\right)}{\sum_{j=1}^{N} \exp\left(s_{contextual}^{i,j}\right)} \tag{2}$$

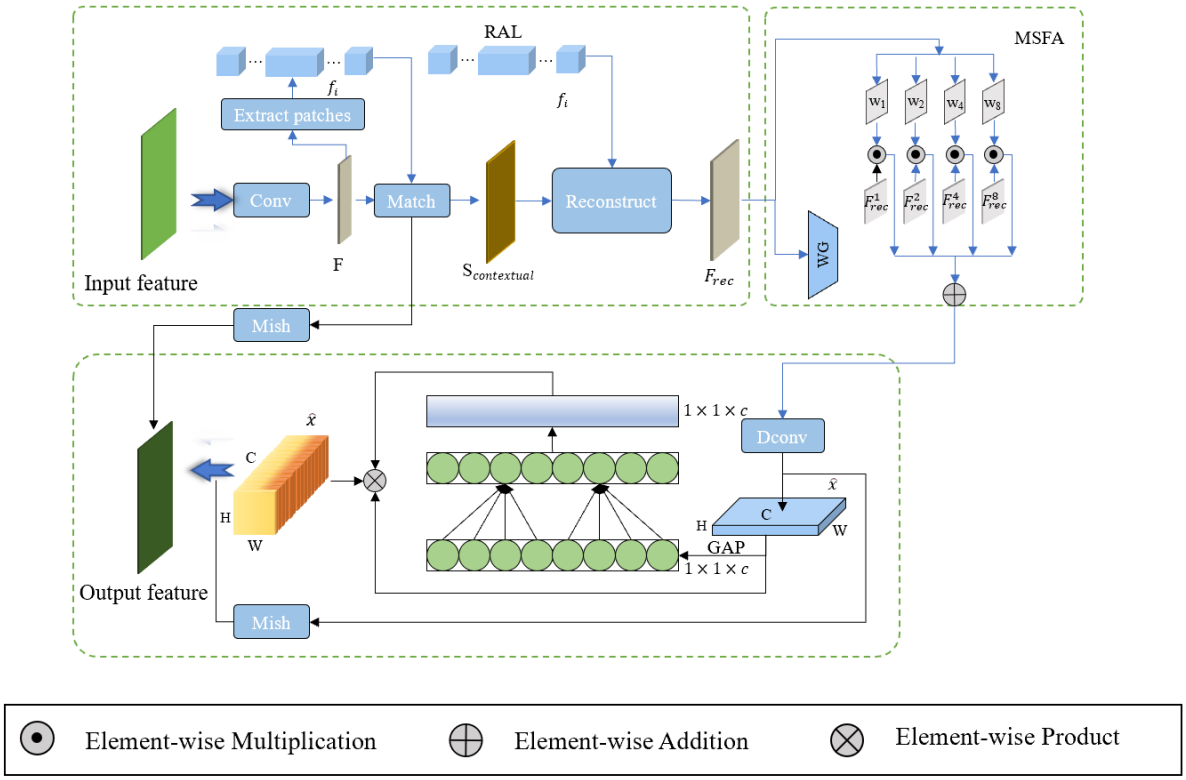


**Figure 4.** Multiscale feature fusion module for dual-attention (MFDA).

Next, using the extracted patches, the feature map is repaired based on the attention map, as in (3).

$$\tilde{f}_i = \sum_{j=1}^{N} f_j \cdot \hat{S}^{i,j}_{contextual} \tag{3}$$

$\tilde{f}_i$ is the $i$-th patch of the repaired feature map $\boldsymbol{F}_{rec}$. The above operations are implemented using convolution, soft channel algorithm, and inverse fold.

In reconstructing the feature map, the model in this paper uses four sets of inflated convolutional layers with different inflation rates to capture multiscale semantic features, as in (4).

$$\boldsymbol{F}^k_{rec} = \mathrm{Conv}_k(\boldsymbol{F}_{rec}) \tag{4}$$

In $\mathrm{Conv}_k(\cdot)$ , $k$ denotes the dilated convolutional layer of $k \in \{1, 2, 4, 8\}$.

In order to better aggregate multiscale semantic features, we design a pixel-level weight mapping generator $G_w$. It predicts pixel-level weight mappings. $G_w$ consists of two convolutional layers with kernel sizes of 3 and 1. Each followed by a rectified linear unit (ReLU) nonlinear activation, and the number of output channels is set to four. The pixel-level weight mapping is computed using (5) and (6).

$$\boldsymbol{W} = \mathrm{Softmax}(G_w(\boldsymbol{F}_{rec})) \tag{5}$$

$$\boldsymbol{W}^1, \boldsymbol{W}^2, \boldsymbol{W}^4, \boldsymbol{W}^8 = \mathrm{Slice}(\boldsymbol{W}) \tag{6}$$

Softmax $(\cdot)$ is channel-wise softmax and Slice $(\cdot)$ is channel-wise slice. Finally, the model in this paper aggregates the multiscale semantic features by weighting the elements and generating a refined feature mapping $\boldsymbol{F}_c$, as in (7).

$$F_c = \left(F_{rec}^1 \odot W^1\right) \oplus \left(F_{rec}^2 \odot W^2\right) \oplus \left(F_{rec}^4 \odot W^4\right) \oplus \left(F_{rec}^8 \odot W^8\right) \tag{7}$$

The dimension of the feature mapping that this essay obtained upfront was H*W*C, based on which this essay used global average pooling GAP to obtain a feature mapping of 1*1*C. The importance between different channels was then learned by 1*1 convolution. At this point, the output dimension was still 1*1*C. Finally, the channel attention was combined, and the channel attention feature map 1*1*C and the original input feature map H*W*C, were multiplied channel by channel to finally obtain the feature map that this paper needs.

We improved the mapping network structure. Based on the original network, inspired by [24], the original convolution was replaced with a depth-separated convolution, and the number of network layers was increased to improve feature extraction. The Mish activation function was used, which is smoother and has no rigid zero bounds compared tothe ReLU activation function. It allows better penetration of more profound information into the neural network, allowing for better accuracy and generalization. It makes the network more optimized, allowing the extracted texture and structure information to be better incorporated into the feature extraction module. The exact structure of our module is shown in Figure 5 below.

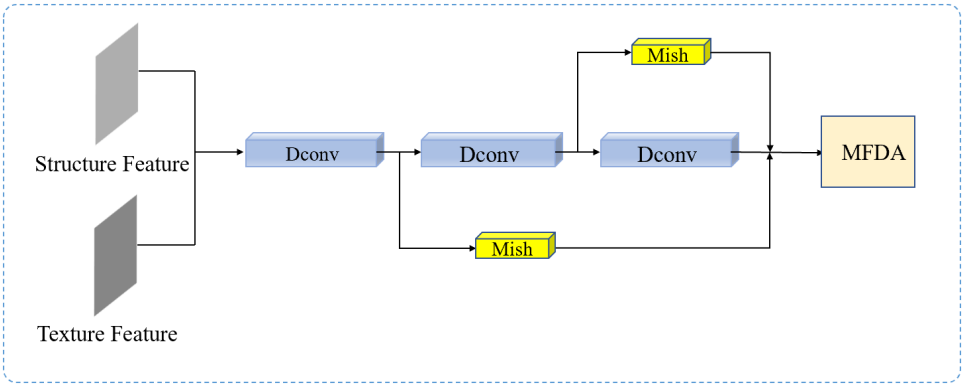


**Figure 5.** Mapping network structure improvement.

### *3.3. Joint Segmented Loss Function*

#### 3.3.1. Joint Loss Function

For the choice of the loss function, this paper uses adversarial loss, perceptual loss, style loss, and reconstruction loss. The loss function is normalized on this basis. Through the union of multiple loss functions, our primary loss function consists of the following components to constrain the model better. Let $G$ be the generator, and $D$ be the discriminator. $I_{gt}$ denotes the actual image, $E_{gt}$ is the complete edge map, and $Y_{gt}$ is the grayscale image.

For the reconstruction loss, this paper uses the L1 distance between $I_{out}$ and $Igt$ as the reconstruction loss, as in (8).

$$\mathcal{L}_{rec} = \mathbb{E}\left[\left\|I_{out} - I_{gt}\right\|_1\right] \tag{8}$$

Perceptual loss is used to evaluate the global structure of the image, where $\phi_i$ denotes the activation map of the ith pooling layer of $VGG-16$ for a given input image $i$. In our implementation, $pool1$, $pool2$, and $pool3$ were used, as in (9).

$$\mathcal{L}_{pre} = \mathbb{E}\left[\sum_i \left\|\phi_i(\boldsymbol{I}_{out}) - \phi_i(\boldsymbol{I}_{gt})\right\|_1\right] \tag{9}$$

Style loss was used to ensure style consistency, and style loss was used to calculate the L1 distance between feature maps, as in (10).

$$\mathcal{L}_{sty} = \mathbb{E}\left[\sum_i \left\| \left(\psi_i(\boldsymbol{I}_{out}) - \psi_i(\boldsymbol{I}_{gt})\right) \right\|_1\right] \tag{10}$$

The Gram matrix constructed from the activation map $\phi_i$ is shown below (11).

$$\psi_i = \phi_i^T \phi_i \tag{11}$$

Adversarial loss, used to ensure the visual fidelity of the repaired image and the consistency of texture and structure, is defined as (12). Each parameter is explained as follows. $\boldsymbol{Y}_{in} = \boldsymbol{Y}_{gt} \odot \boldsymbol{M}_{in}$ indicates a damaged grayscale image. $\boldsymbol{E}_{in} = \boldsymbol{E}_{gt} \odot \boldsymbol{M}_{in}$ is the damaged edge map. $\boldsymbol{I}_{out}, \boldsymbol{E}_{out} = G(\boldsymbol{I}_{in}, \boldsymbol{E}_{in}, \boldsymbol{Y}_{in}, \boldsymbol{M}_{in})$ indicates the output of the generator.

$$\begin{aligned} \mathcal{L}_{adv} = &\min_G \max_D \mathbb{E}_{\boldsymbol{I}_{gt}, \boldsymbol{E}_{gt}} \left[\log D\left(\boldsymbol{I}_{gt}, \boldsymbol{E}_{gt}\right)\right] \\ &+ \mathbb{E}_{\boldsymbol{I}_{out}, \boldsymbol{E}_{out}} \log[1 - D(\boldsymbol{I}_{out}, \boldsymbol{E}_{out})] \end{aligned} \tag{12}$$

Joint loss function: the above losses are normalized into joint losses, as follows (13).

$$\begin{aligned} l_{ss} = &\frac{1}{N} \sum_{m=1}^{4} \sum_{i=1}^{N} (\lambda_{rec} \mathcal{L}_{rec} + \lambda_{pre} \mathcal{L}_{pre} \\ &+ \lambda_{sty} \mathcal{L}_{sty} + \lambda_{adv} \mathcal{L}_{adv}) \end{aligned} \tag{13}$$

The specific parameters are as follows. $\lambda_{\text{rec}} = 10$, $\lambda_{pre} = 0.1$, $\lambda_{sty} = 250$, and $\lambda_{adv} = 0.1$, where $N$ represents the total number of training graphs, and $\lambda$ represents the weight coefficient of each loss function, respectively.

#### 3.3.2. Segmented Loss Function

The reconstruction loss in the loss function is susceptible to local optimum solutions due to the existence of separate parameters. In this paper, the problem is solved using a segmentation function. Based on improving the training effect and optimizing the image inpainting results, the model's complexity is preserved, and the optimal solution is obtained faster. This paper uses two loss functions: the absolute value loss function, also known as the L1 paradigm loss function, and the least square error loss function, also known as the L2 paradigm loss function. The segmented L1 paradigm and L2 paradigm methods were used to implement the training of the loss function. It consists of two main components, reconstructing the loss first using the L1 paradigm and then using the L2 paradigm. The test results for the L1 paradigm and L2 paradigm are shown in Figure 6 below. The red and blue values of the loss graphs use only L1 paradigm and L2 paradigm training, respectively. Figure 7 shows the result graphs of using L1 paradigm and L2 paradigm segmented training. The red one uses the L1 paradigm first and then the L2 paradigm. The blue one uses the L2 paradigm first and then the L1 paradigm, whereas the red one is the training method used in this paper. Through comparison, the training method in this paper has the best effect and reaches convergence faster. The results of the training method in this paper can be found to be the most effective and converge faster. The threshold value is 10,000 times based on convergence, and the function distribution is shown below in (14).

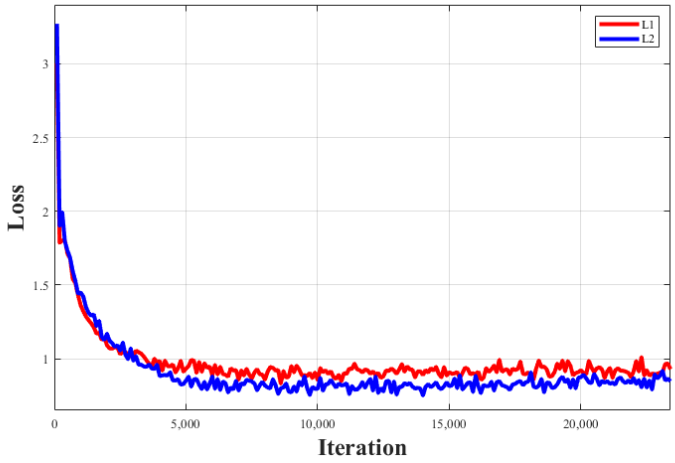


**Figure 6.** Using L1 and L2 paradigms alone.

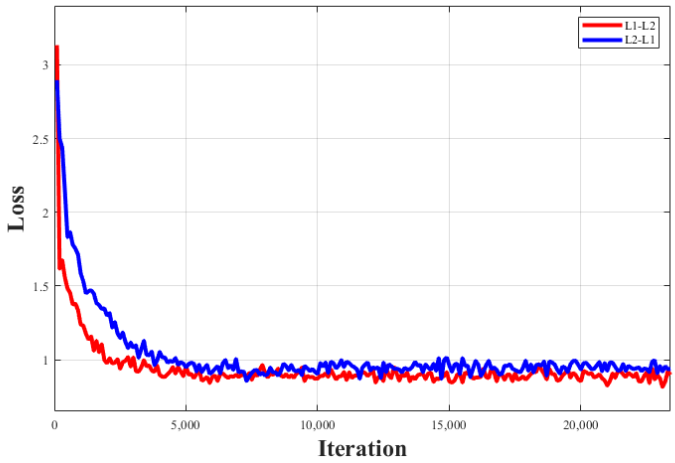


**Figure 7.** Using the L1 and L2 paradigms in stages.

The loss curves for the different training methods are shown in (14). $\mathcal{L}_{\text{rec1}}$ involves using the L1 paradigm before the L2 paradigm, and $\mathcal{L}_{\text{rec2}}$ involves using the L2 paradigm before the L1 paradigm.

$$\begin{cases} \mathrm{l}_{ss} = \frac{1}{N}\sum_{m=1}^{4}(\lambda_{rec}\mathcal{L}_{rec1}+\lambda_{pre}\mathcal{L}_{pre} \\ \qquad +\lambda_{sty}\mathcal{L}_{sty}+\lambda_{adv}\mathcal{L}_{adv}) & iter \le 10,000 \\ \mathrm{l}_{ss'} = \frac{1}{N}\sum_{m=1}^{4}(\lambda_{rec}\mathcal{L}_{rec2}+\lambda_{pre}\mathcal{L}_{pre} \\ \qquad +\lambda_{sty}\mathcal{L}_{sty}+\lambda_{adv}\mathcal{L}_{adv}) & iter > 10,000 \end{cases} \tag{14}$$

Different training methods are used here for experimental comparison. The segmentation function in this paper can converge faster, and the running time is shorter than the original running time, which is better than the other two methods. Figure 8 compares the inpainting images with the epitaxial mask used on both the training and test sets. The loss function used in the original network is the L1 paradigm loss, and the loss function used in this paper is the L1–L2 split-paradigm loss form. The comparison reveals that our method also slightly outperforms the other forms of the function in terms of subjective effect.




**Figure 8.** Subjective comparison of different training methods.

The Table 1 comparison shows that using this training method is slightly better than the original network, proving that the method is feasible. The best indicators are shown in bold.

**Table 1.** Comparison of objective indicators using different training methods.

| Method | L1 | L2 | L2–L1 | L1–L2 |
|---|---|---|---|---|
| PSNR | 22.6238 | 22.2908 | 22.3710 | **22.8600** |
| SSIM | 0.8523 | 0.8558 | **0.8654** | 0.8629 |
| MSE | 0.0085 | 0.0074 | **0.0050** | 0.0068 |

## 4. Experiments

### *4.1. Dataset and Implementation Details*

This paper uses the polo of the tomb of Prince Zhang Huai of the Tang Dynasty as a dataset. It uses a Sinar *P*2 large-format technical camera and Sinar 75LV digital back from Switzerland for the dataset, with a color gamut model of Adobe RGB and a Schneider APO lens, illuminated by a German Bach spotlight, in the same row, from left to right, with an overlap of 40–50% between two adjacent images. The whole process was noncontact, with the parallel acquisition in separate shots. In order to enhance the generalization of the model, the dataset was processed as follows: the methods used to broaden the data [25] were rotation, mirror flip, and chroma transformation. The 710 murals with rich texture details were sieved and expanded to 7100, including 5680 training data, 710 test data, and 710 validation data. Some of the datasets are shown in Figure 9 below.



**Figure 9.** Dataset used in this paper (partial).

Mask selection: masks are determined according to the different types of disease to be addressed by the mural. Loss caused by the falling off of the ground floor in the tomb: random large mask and random small mask are used to represent the different inpainting areas inside. Missing due to human chiseling: irregular epitaxial masks drawn by hand are used to represent these. Missing due to fissures in the mural: there are a large number of cracks in the mural, and due to the lack of accurate crack datasets corresponding to the mural, this paper used a CNN method to generate the crack maps, ensuring that the two could be mapped back to the original and that there were no pixel shifts between them. Among many methods, this paper found that TMCrack-Net [26] can generate crack maps suitable for tomb murals. The generated partial crack mask map is shown below in Figure 10.

This paper implemented our model using Cuda 11.1 on an Nvidia RTX 2080Ti GPU with Ubuntu OS. The batch size is 8, the initial learning rate is set to $2 \times 10^{-4}$, and our iteration number is $3 \times 10^5$.

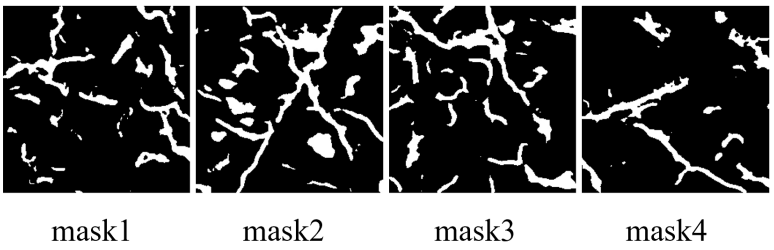


**Figure 10.** Crack masks diagram.

### *4.2. Assessment Indicators*

Objectively, peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and mean square error (MSE) are used, while subjectively, color consistency, texture similarity, and structural similarity are used to discriminate. The higher the value of PSNR, the better the mural information generation. SSIM calculates the brightness, contrast, and structural similarity between the generated and actual mural. The closer the value is to 1, the higher the similarity between the two murals. MSE measures the interpixel difference, and the smaller the value, the better the generation effect.

### *4.3. Multiscale Feature Fusion Module Ranking for Dual-Attention*

Experiments were conducted here to compare the various modules of MFDA in different orders and to replace the corresponding channel attention with pixel attention to compare their effects. The flowchart corresponding to each set of experiments is shown in Figure 11 below: J1 is the mapping network before improving MFDA only, and J2, J3, J4, J5, and J6 are the adjusted order. Due to the small dataset, this paper used the random large mask dataset as the training mask and the epitaxial mask as the test mask to obtain experimental results suitable for the large mask.

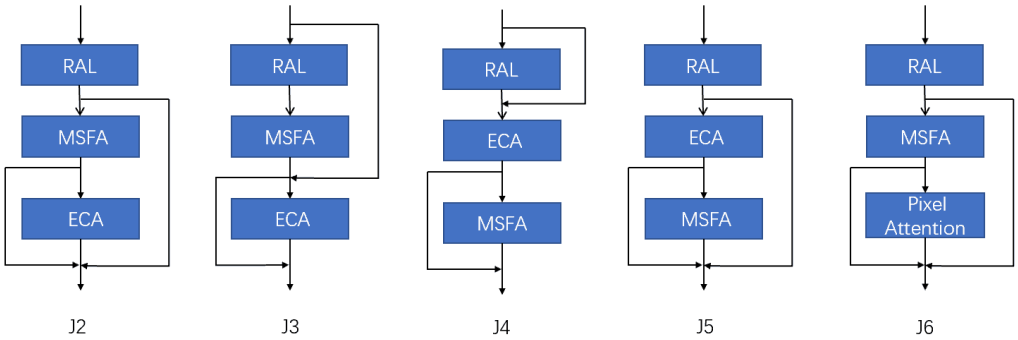


**Figure 11.** Flow chart of the different sequential module arrangements.

Figure 12 below shows the subjective effects obtained in different orders. Where J1 is an individual effect of changing only the structure of the mapping network and differs little from the original network in terms of subjective effect, the structure J2 used in this paper better depicts the detail on the lower right side of the garment compared to the original network.

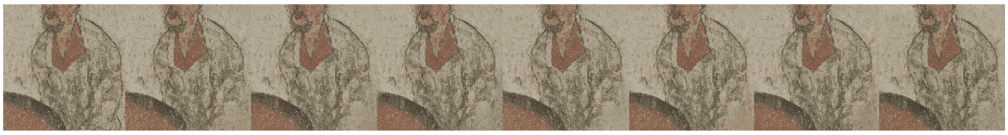


**Figure 12.** Subjective results for different sequences of modules. By comparison, the best results are obtained with J2, which is more similar to the image ground truth.

The specific test metrics are shown in Table 2 below by testing the objective data with different module sequences. Objective data show better results using an improved mapping network and structure. The best indicators are shown in bold.

**Table 2.** Comparison of test metrics with different module orders.

| Method | Baseline | J1 | J2 | J3 | J4 | J5 | J6 |
|---|---|---|---|---|---|---|---|
| PSNR | 22.6238 | 23.2908 | **24.3710** | 23.5600 | 23.0815 | 23.0912 | 23.3141 |
| SSIM | 0.8523 | 0.8558 | **0.8654** | 0.8629 | 0.8463 | 0.8470 | 0.8590 |
| MSE | 0.0085 | 0.0074 | **0.0055** | 0.0068 | 0.0075 | 0.0074 | 0.0071 |

A comparison of objective and subjective experimental results shows that our use of the MDFA module proposed in this paper results in a better network structure after changing the mapping network and using a specific order of modules. The experiment results are more in line with our needs subjectively and objectively.

### *4.4. Ablation Experiments*

In this section, the conditional texture and structure dual-generation (CTSDG) is used as the base model, and ablation experiments are conducted on the three modules added to the overall network model. By comparing the ablation experiments, the model in this paper outperforms the base model with a significant performance improvement. Model A is the original network model. Model B is a model with segmented loss functions. Model C is a model based on an improved encoder. Model D is a model with multiscale feature fusion with double attention, and model E is a combination of model A, model B, and model C. From the objective values in the figure below, this paper found that model A, model B, and model C all have an improved effect when added individually, while the objective and subjective viewpoint results obtained after adding them all are better than those added individually. The inpainting is better and more in line with our needs. The objective results are shown in Table 3 below, and the subjective results are shown in Figure 13 below. The best indicators are shown in bold.

Figure 13 shows the ablation comparison experiments. Input is the image to be repaired, and ground truth is the original image. The original network model A loses the detail of the trunk in the lower right corner on the first row of the trunk inpainting, which has complex textures. There is significant blurring in the lower right corner on the third row of the repaired horse head with sparse feature information. Model B, model C, and model D are all better repaired in detail than model A. All models have similar results in inpainting the second row of stones. Model E is a combination of B, C, and D and is as close as possible to the detail of the original drawing in the presence of the complexly textured trunk in the first row and the sparse information features in the third row.

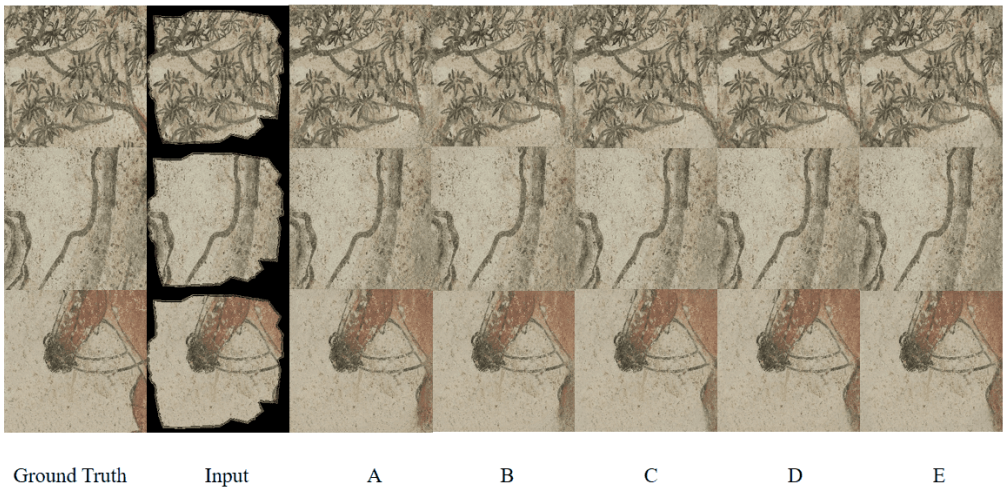


**Figure 13.** Comparative image of the ablation experiment. Module E compares the model used in this paper, in three rows of images, with superior results.

**Table 3.** Comparison of test indicators for ablation experiments.

| Method | A | B | C | D | E |
|---|---|---|---|---|---|
| PSNR | 22.6238 | 22.8600 | 23.7136 | 23.6139 | **23.7365** |
| SSIM | 0.8523 | 0.8629 | 0.8548 | 0.8657 | **0.8659** |
| MSE | 0.0085 | 0.0068 | 0.0047 | 0.0048 | **0.0046** |

### *4.5. Comparison Experiments*

This section compares the model proposed in this paper with some mainstream inpainting models. The main models compared are coherent semantic attention for image inpainting (CSA) [17], recurrent feature reasoning net (RFR) [18], resolution-robust large mask inpainting with Fourier convolutions (LaMa) [20], and image inpainting via conditional texture and structure dual-generation (CTSDG) [19]. The experimental mask is the same as the test mask, and different types of masks are used for inpainting images. The specific inpainting results are shown in Figures 14–17 below.

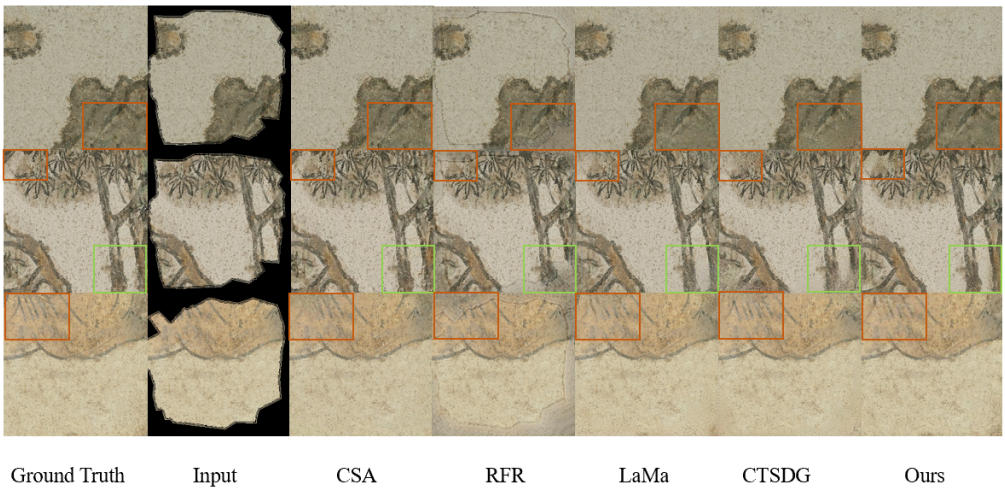


**Figure 14.** Comparison of inpainting results for each model on the epitaxial masks. Our model works best and looks more similar to the original after the comparison.

On the epitaxial mask, this paper's model has fewer image artifacts, and on the crack mask, the inpainting effect of each model is similar. On the random small mask, this paper's

model will not produce textures that do not conform to the image information. Due to the small number of datasets and sparse features, the inpainting is imperfect for the random large mask. This network can better represent the image information for the mural dataset and shows precise details. The details are shown below.

The comparison in Figure 14 shows that when the CSA model is used, the resulting trunk is thicker in the lower right corner of the image in the second row at the green box compared to the original image, which repairs the image but does not accurately inpaint the image. The third row of the image has a blurred texture in the fold in the top left corner of the red box. When the RFR model was used, more obvious artifacts were produced. When using the LaMa model, the internal texture of the stone is missing in the red box in the bottom right corner of the first row of images. The repaired tree trunk in the green box in the bottom right corner of the second row of images needs to be structurally sound, with the trunk shape not matching the original image. When using the CTSDG model, artifacts are present, and in both the top left red box and the bottom correct green box in the second row, there are blurred details. With the model in this paper, there are fewer artifacts, and the detail recovery is more similar to the original image in both the red box in the bottom right corner of the first row and the green box in the bottom right corner of the second row. In summary, our network generates fewer image artifacts on the epitaxial mask and produces a structure and texture similar to the original image.

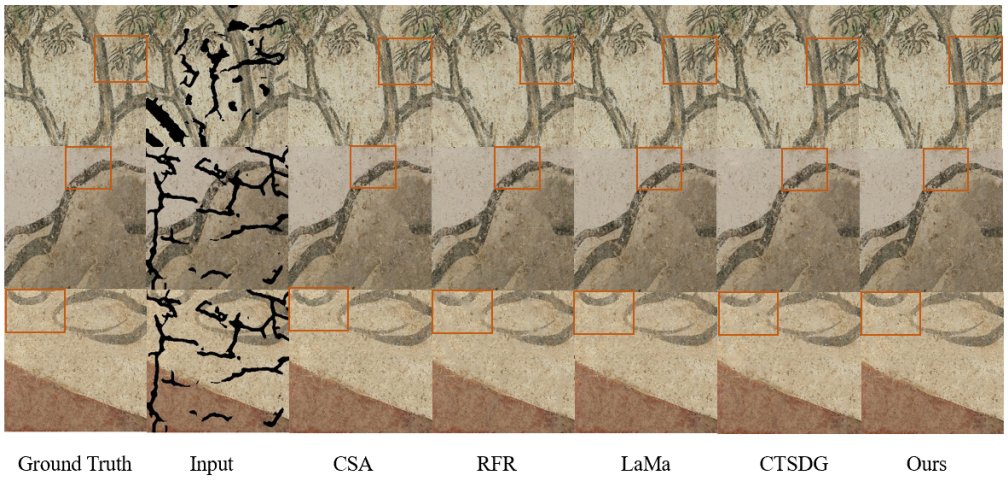


**Figure 15.** Comparison of inpainting results for each model on the crack masks. Our model works best and looks more similar to the original after the comparison.

The comparison in Figure 15 shows that the inpainting results are similar across the networks on the crack masks, with our network generating relatively good images when the difference is not significant. When the CSA model is used, the inpainting results are similar to the original image. With the RFR model, the inpainting of the leaves in the first row of the image in the red box is blurred, and the details are not precise. With the RFR model, the artifacts in the red box in the third row are more prominent. With the LaMa model, the inpainting is similar to the original image. When using the CTSDG model, the inpainting of the leaves and trunk in the first row of the image is blurred and lacks detail. Our model's overall inpainting is better, and the detail repaired at the red box in the first row is more similar to the original image. In summary, our network can produce a structure and texture more similar to the original image on the crack mask.

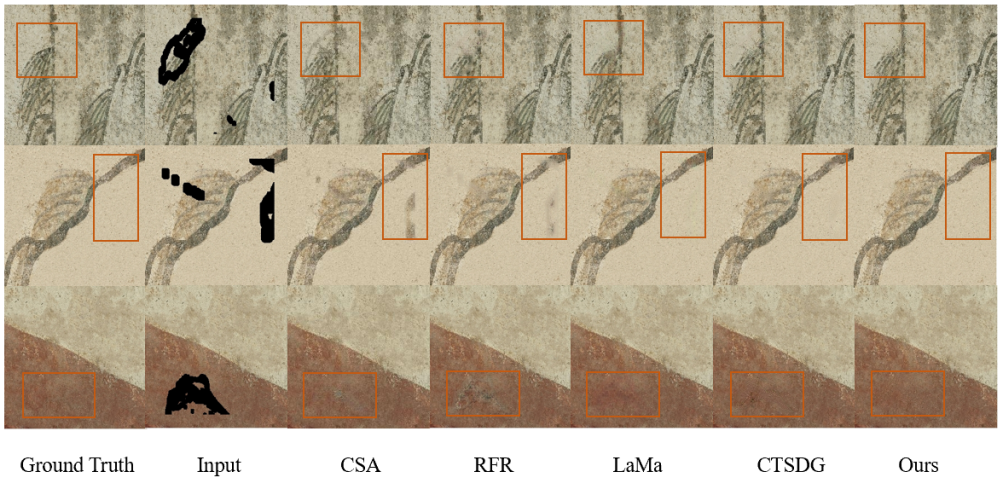


**Figure 16.** Comparison of inpainting results for each model on the small masks. Our model works best and looks more similar to the original after the comparison.

The comparison in Figure 16 shows that our image has fewer artifacts on the small mask, and the image is more similar to the original. With the CSA model, the image is repaired at the red box in the first row and the red box in the second row, but the inpainting structure does not resemble the original image structure. With the RFR model, artifacts were more evident in the second and third rows of the image at the red box, and the defect at the trunk was not repaired. With the LaMa model, the inpainting of the trunk in the red box in the first row of the image is different from the original structure, and the inpainting of the red box in the third row is reddish and has a color different from the original image. When the CTSDG model is used, the repaired leaf structure in the red box in the first row of the image is different from the original image. The trunk defect in the red box of the image in the second row is not fully repaired. Using the model in this paper, the inpainting of the details is more similar to the original image in detail, except for the trunk defect at the red box in the second row, which is not entirely repaired, and the overall inpainting is better. In summary, on the small mask, our network generated a structure and texture more similar to the original image.

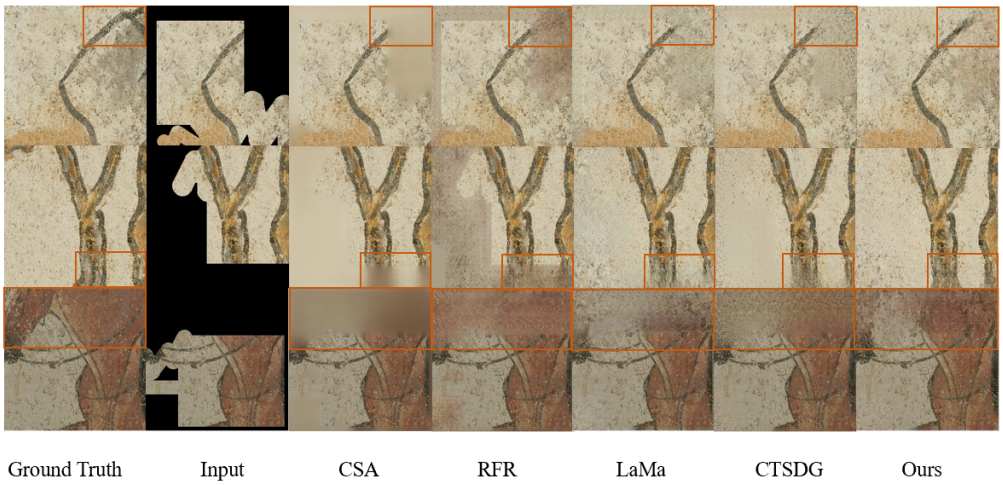


**Figure 17.** Comparison of inpainting results for each model of the large masks. Our model works best and looks more similar to the original after the comparison.

The comparison in Figure 17 shows that none of the paintings on the large masks are good due to the sparse dataset. When the CSA model was used, the artifacts produced were

more pronounced. When the RFR model was used, it produced more apparent artifacts, and the inpainting results were blurred. With the LaMa model, the inpainting of the horse's head is blurred in the third row of the image in the red box. With the CTSDG model, the repaired structure in the second row of the trunk is different from the original image and is blurred. The inpainting results are more precise and detailed with the model in this paper than the other models. However, the repaired trunk structure in the second row does not match the original image, and its overall inpainting effect is better. In summary, our network can produce structures and textures more similar to the original image on the large mask.

To sum up, each model compared in this paper has disease types that are good at inpainting. However, many kinds of mural diseases need to be repaired comprehensively. Our network can inpaint many kinds of mural diseases and is superior to some mainstream methods in detail and overall understanding of images when drawing different types of mural diseases, proving this network's feasibility.

In terms of objective data, the test indicators of the network in this paper are superior to the current mainstream network effect in general. Its objective data are shown in Table 4 below. In the comparison of multiple groups of data, it can be found that the network in this paper is more effective, and the small mask test data can best show our model's superiority. The best indicators are shown in bold.

**Table 4.** Comparative experimental test indicators for each model.

| Mask Type | Method | Ours | CSA | RFR | LaMa | CTSDG |
|---|---|---|---|---|---|---|
| epitaxial mask | PSNR | **25.2366** | 25.0880 | 23.9347 | 25.1932 | 24.7238 |
| | SSIM | 0.8509 | 0.7990 | 0.8108 | 0.8468 | **0.8524** |
| | MSE | 0.0062 | 0.0053 | 0.0067 | **0.0050** | 0.0086 |
| crack mask | PSNR | **31.8855** | 28.6886 | 31.1276 | 31.2935 | 30.8397 |
| | SSIM | 0.9132 | 0.8262 | 0.9102 | 0.9114 | **0.9150** |
| | MSE | 0.0025 | 0.0035 | 0.0030 | **0.0019** | 0.0022 |
| small mask | PSNR | **35.6748** | 28.9423 | 34.4726 | 35.1649 | 32.2429 |
| | SSIM | **0.9597** | 0.8491 | 0.9562 | 0.9550 | 0.9538 |
| | MSE | **0.0008** | 0.0032 | 0.0010 | 0.0009 | 0.0020 |
| large mask | PSNR | **25.2643** | 23.2033 | 23.5682 | 25.0319 | 24.9831 |
| | SSIM | 0.7311 | **0.7323** | 0.7201 | 0.7124 | 0.7306 |
| | MSE | **0.0077** | 0.0094 | 0.0113 | 0.0091 | 0.0080 |

## 5. Conclusions

In order to reduce costs and assist in manual inpainting, we conducted this study. In this paper, we proposed an improved generative adversarial network of fragmentation of tomb murals. The protection of the image structure was enhanced by improving the generator to include prior knowledge and channel attention. The inpainting accuracy of the model was improved by using a new module of feature aggregation, which combines multiscale feature aggregation with a dual-attention mechanism and switches to depth-separable convolution. A joint loss function was used to improve the inpainting results further. The improved model improved the ability to accurately inpaint the original murals, satisfying the requirement of "inpainting the old to the old" as far as possible. Our model showed an overall improvement in the effectiveness of inpainting different disease types. Objectively, on the epitaxial mask, PSNR increased by 0.5128, and MSE decreased by 0.0024. On the crack mask, PSNR increased by 1.0458. On the random small mask, PSNR increased by 3.4319, SSIM increased by 0.0059, and MSE decreased by 0.0012. On the random large mask, PSNR increased by 0.2812, SSIM increased by 0.0005, and MSE decreased by 0.0003. In terms of objective metrics, the largest improvement was seen in PSNR. In terms of objective metrics, the most remarkable improvement was seen in PSNR. In aggregate, the greatest numerical improvement was seen in random small masks, followed by random

large masks. From a subjective point of view, the graphs we generated were also found by experts to be more compatible with the inpainting. The results showed that our method can effectively address three types of diseases to tomb murals: long-term environmental erosion and human damage; fissures in the murals and large hollow drums in the floor of the tomb murals lead to the loss of the floor layer and missing images as well as artificial excavation.

In the future, we will further address the limitations of our model and will continue to improve the quality of the inpainting model of diversified mural deterioration. As the number of tomb murals is small and precious, we will continue to collect more tomb mural images to build a larger dataset to enable the model to learn and inpaint better.

**Author Contributions:** Conceptualization, M.W.; methodology, M.W.; writing—review and editing, M.W.; software, X.C.; writing—original draft preparation, X.C.; validation, J.W.; resources, J.W.; data curation, J.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (No. 61701388), the Interdisciplinary Foundation of Xi'an University of Architecture and Technology (No. X20220082), and the Natural Science Foundation of Xi'an University of Architecture and Technology (No. ZR21032).

**Data Availability Statement:** The Mural dataset presented in this study is available from the Shaanxi History Museum.

**Acknowledgments:** We are thankful for the help of Huaidong Zhao from the School of Art of Xi'an University of Architecture and Technology. Zhao was responsible for the verification part, verifying that the inpainting results followed the artistic standards. He assisted in the inpainting of the murals under the ancient artistic style.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Karianakis, N.; Maragos, P. An integrated system for digital restoration of prehistoric theran wall paintings. In Proceedings of the 2013 18th International Conference on Digital Signal Processing (DSP), Santorini, Greece, 1–3 July 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1–6
2. Jaidilert, S.; Farooque, G. Crack detection and images Inpainting method for Thai mural painting images. In Proceedings of the 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, China, 27–29 June 2018; IEEE: Piscataway, NJ, USA, 2013; pp. 143–148
3. Cao, J.; Zhang, Z.; Zhao, A.; Cui, H.; Zhang, Q. Ancient mural restoration based on a modified generative adversarial network. *Herit. Sci.* **2020**, *8*, 1–14. [CrossRef]
4. Zhou, S.; Xie, Y. Intelligent Restoration Technology of Mural Digital Image Based on Machine Learning Algorithm. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 4446999. [CrossRef]
5. Cao, J.; Li, Y.; Zhang, Q.; Cui, H. Restoration of an ancient temple mural by a local search algorithm of an adaptive sample block. *Herit. Sci.* **2019**, *7*, 1–14. [CrossRef]
6. Priego, E.; Herráez, J.; Denia, J. L.; Navarro, P. Technical study for restoration of mural paintings through the transfer of a photographic image to the vault of a church. *J. Cult. Herit.* **2022**, *58*, 112–121. [CrossRef]
7. Zeng, Y.; Gong, Y.; Zeng, X. Controllable digital restoration of ancient paintings using convolutional neural network and nearest neighbor. *Pattern Recognit. Lett.* **2020**, *133*, 158–164. [CrossRef]
8. Gupta, V.; Sambyal, N.; Sharma, A.; Kumar, P. Restoration of artwork using deep neural networks. *Evol. Syst.* **2021**, *12*, 439–446. [CrossRef]
9. Wang, Y.; Tao, X.; Qi, X.; Shen, X.; Jia, J. Image inpainting via generative multi-column convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
10. Zhou, X.; Xu, Z.; Cheng, X.; Xing, Z. Restoration of Laser Interference Image Based on Large Scale Deep Learning. *IEEE Access* **2022**, *10*, 123057–123067. [CrossRef]
11. Chan, T. F.; Shen, J. Nontexture inpainting by curvature-driven diffusions. *J. Vis. Commun. Image Represent.* **2001**, *12*, 436–449. [CrossRef]
12. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef] [PubMed]
13. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D. B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]

15. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.
16. Ren, Y.; Yu, X.; Zhang, R.; Li, T.H.; Liu, S.; Li, G. Structureflow: Image inpainting via structure-aware appearance flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 181–190.
17. Liu, H.; Jiang, B.; Xiao, Y.; Yang, C. Coherent semantic attention for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 4170–4179.
18. Li, J.; Wang, N.; Zhang, L.; Du, B.; Tao, D. Recurrent feature reasoning for image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7760–7768.
19. Guo, X., Yang, H., Huang, D. Image inpainting via conditional texture and structure dual generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14134–14143.
20. Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; Lempitsky, V. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2149–2159.
21. Li, R.; Tan, R.T.; Cheong, L.F.. Robust optical flow in rainy scenes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 288–304.
22. Zhang, G.; Gao, X.; Yang, Y.; Wang, M.; Ran, S. Controllably deep supervision and multi-scale feature fusion network for cloud and snow detection based on medium-and high-resolution imagery dataset. *Remote Sens.* **2021**, *13*, 4805. [CrossRef]
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
24. Qian, G.; Abualshour, A.; Li, G.; Thabet, A.; Ghanem, B. Pu-gcn: Point cloud upsampling using graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11683–11692.
25. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In 2018 international interdisciplinary PhD workshop (IIPhDW), Swinoujscie, Poland, 9–12 May 2018; pp. 117–122.
26. Wu, M.; Jia, M.; Wang, J. TMCrack-Net: A U-Shaped Network with a Feature Pyramid and Transformer for Mural Crack Segmentation. *Appl. Sci.* **2022**, *12*, 10940. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.