electronics

MDPI

Article

Ancient Painting Inpainting Based on Multi-Layer Feature Enhancement and Frequency Perception

Xiaotong Liu [1], Jin Wan [2,*], Nan Wang [2] and Yuting Wang [3]

[1] Chinese Painting Department, Xi'an Academy of Fine Arts, Xi'an 710065, China; xiaotongliu01@163.com
[2] Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China; 16112070@bjtu.edu.cn
[3] School of Computer Science, Henan University of Technology, Zhengzhou 450001, China; 221091100214@stu.haut.edu.cn
* Correspondence: 18112033@bjtu.edu.cn

Abstract: Image inpainting aims to restore the damaged information in images, enhancing their readability and usability. Ancient paintings, as a vital component of traditional art, convey profound cultural and artistic value, yet often suffer from various forms of damage over time. Existing ancient painting inpainting methods are insufficient in extracting deep semantic information, resulting in the loss of high-frequency detail features of the reconstructed image and inconsistency between global and local semantic information. To address these issues, this paper proposes a Generative Adversarial Network (GAN)-based ancient painting inpainting method using multi-layer feature enhancement and frequency perception, named MFGAN. Firstly, we design a Residual Pyramid Encoder (RPE), which fully extracts the deep semantic features of ancient painting images and strengthens the processing of image details by effectively combining the deep feature extraction module and channel attention. Secondly, we propose a Frequency-Aware Mechanism (FAM) to obtain the high-frequency perceptual features by using the frequency attention module, which captures the high-frequency details and texture features of the ancient paintings by increasing the skip connections between the low-frequency and the high-frequency features, and provides more frequency perception information. Thirdly, a Dual Discriminator (DD) is designed to ensure the consistency of semantic information between global and local region images, while reducing the discontinuity and blurring differences at the boundary during image inpainting. Finally, extensive experiments on the proposed ancient painting and Huaniao datasets show that our proposed method outperforms competitive image inpainting methods and exhibits robust generalization capabilities.

Keywords: ancient painting inpainting; generative adversarial networks; multi-layer feature enhancement; frequency perception

Citation: Liu, X.; Wan, J.; Wang, N.; Wang, Y. Ancient Painting Inpainting Based on Multi-Layer Feature Enhancement and Frequency Perception. *Electronics* **2024**, *13*, 3309. https://doi.org/10.3390/electronics13163309

Academic Editors: Taehyeon Kim and KyungTaek Lee

Received: 7 July 2024
Revised: 16 August 2024
Accepted: 19 August 2024
Published: 21 August 2024

1. Introduction

As a treasure of traditional art, ancient painting, with its unique aesthetic concepts and expressive methods, not only reflects the people's pursuit of harmonious coexistence with nature, but also contains deep philosophical thinking and spiritual support. The creation of ancient paintings often combines poetry, calligraphy, painting, printing and other art forms, showing the rich connotation and artistic charm of traditional culture [1]. However, many artworks are exposed to various risks of damage due to physical damage, biological erosion, and other environmental factors. Such damage compromises the integrity of the works, diminishes their artistic expression, and erodes their historical significance. To this end, ancient painting inpainting has emerged as a basic task in computer vision which aims to restore the missing parts of an image and repair damaged areas, thereby preserving the visual and historical integrity of the artwork [2,3].

Traditional image inpainting primarily relies on pixel diffusion and pixel block repair methods. Typical algorithms include the variation segmentation-based repair method [4], the

curvature-driven diffusion repair method [5,6], and the fast nearest neighbour algorithm [7]. Segmentation-based methods [4] can usually only repair the image within a relatively small area of the missing area, and are limited by the pixel information around the missing area, which makes it difficult to repair real-world images with complex textures and art. It is difficult to repair real-world images with complex texture detail information. Pixel block-based repair methods [7] are better than diffusion-based methods, but their computational volume increases significantly, and they can only repair missing regions similar to those in the training set, and the repair effect relies heavily on the original dataset. Although the digital inpainting method [8] has improved inpainting efficiency to a certain extent, there are still limitations in maintaining the original style and artistry, and it is difficult to completely restore the original appearance and charm of the work. In short, traditional image inpainting methods [4–7] rely on manual techniques, which are insufficient to meet the needs of contemporary society for efficiency and large-scale processing capabilities.

Recently, with the continuous development of deep learning [9], image inpainting methods [10,11] based on deep learning have achieved remarkable results. Compared with traditional image inpainting methods, Generative Adversarial Network (GAN [12])-based image inpainting methods have good robustness and generalization ability. Among them, Conditional GAN [13] is able to control the generation process more accurately by introducing conditional variables, and provides a new technological path for ancient paintings inpainting. Xue et al. [14] proposed an end-to-end network SAPGAN for the inpainting of ancient paintings, which is the first model to generate ancient paintings from start to finish without conditional inputs. SAPGAN consists of two GANs for generating edge images and edge-to-paint conversion. Yang et al. [15] proposed an ancient painting perception method based on a semantic segmentation approach and designed a novel ancient painting perceptual parsing model to make full use of the intrinsic features of ancient painting images. Based on the multi-scale features of objects in ancient paintings, it enables the model to learn appropriate perceptual fields to extract contextual information. Peng et al. [16] proposed a transfer method from ancient photographs to ancient paintings, incorporating a contour-enhanced translation branch designed to efficiently facilitate the transition from photographs to paintings. The results demonstrate that this method can effectively transform real ancient pictures into paintings. Sun et al. [1] proposed an edge-guided ancient painting inpainting method based on the use of multi-scale channel attention residual blocks to learn different levels of image features, and the design of the self-attention module to improve the attention to the edge and detail features of the image. However, the consistency of semantic information between global and local areas is not fully considered, which leads to inconsistency between the restored image and the original painting.

Although the above methods improve the inpainting capability of ancient paintings to a certain extent, each painter has their unique style and brushwork, and the inpainting needs to accurately mimic the style of the original painter to maintain the consistency of the artwork. Existing methods do not fully utilize the high-level semantic information of ancient painting images, resulting in restored images with poor detail and inconsistencies between local and global styles.

The contributions of this paper can be summarized as:

(1) Innovative Deep Semantic Feature: ExtractionWe introduce a Residual Pyramid Encoder (RPE) network tailored for extracting deep semantic features from ancient paintings. This network innovatively combines a deep feature extraction module with channel attention mechanisms, significantly enhancing the processing of image details and facilitating the identification and reconstruction of missing image segments.
(2) Detail and Texture Feature Reconstruction: We propose a Frequency-Aware Mechanism (FAM) that employs a frequency attention module to capture high-frequency perceptual features. By strategically integrating skip connections between low- and high-frequency components, FAM reconstructs the intricate details and textures of

ancient paintings, effectively addressing the common challenge of high-frequency detail loss in image painting.

(3) Dual Discriminator for Semantic Consistency: To ensure the restored image preserves the semantic integrity of the original in both global structure and local details, we present a dual discriminator approach. This mechanism guarantees semantic consistency across the entire image and local regions during the inpainting process, reducing boundary discontinuities and blur.

(4) Empirical Validation and Generalization: Our experimental and visualization results, conducted on both the proposed ancient painting dataset and the Huaniao dataset, as well as the publicly available CelebAHQ dataset, demonstrate that our method surpasses competitive image inpainting methods, exhibiting robust generalization capabilities.

2. Related Work

2.1. Ancient Painting Inpainting

Ancient painting is an important part of traditional art, and its inpainting requires not only exquisite skills but also a deep cultural heritage [2,3]. Traditional inpainting methods [17] mainly include manual inpainting, copying and color replenishment, which rely on the experience and aesthetics of the restorer. The restorer will carefully analyse the damage to the painting. With the advancement of technology, image inpainting methods [18–21] have introduced digital techniques, such as high-resolution scanning, digital image processing and 3D printing. These technologies can capture and analyze the details of the paintings more precisely, providing a more scientific basis for image inpainting.

Deep learning possesses the capability to autonomously discern and distil key image characteristics, which are essential for tasks such as recognition, classification, and image inpainting. It has made significant progress in image inpainting [18,19]. Currently, the most cutting-edge techniques for image inpainting are deep learning-based methods, which use the proficiency of neural networks to accurately locate and reconstruct damaged or missing image fragments. Common frameworks for these techniques include applying Convolutional Neural Networks (CNNs [20]) to extract underlying features from images, supplemented by Generative Adversarial Networks (GANs [21]) to construct indistinguishable inpainting areas. In addition, some methods use Recurrent Neural Networks (RNNs [22]) and Transformer structures to handle sequence prediction problems, especially when restoring image sequences or time series data. Dutta et al. [23] present an interactive deep network for image restoration. The use of patch interaction terms and quantum-based Hamiltonian operators for non-localized images significantly improves the task of image restoration for each item. The application of deep learning in image inpainting not only improves the accuracy and efficiency of inpainting, but also reduces the reliance on large amounts of labeled data through self-supervised or semi-supervised learning, making the inpainting process more flexible and adaptable.

In the specialized subfield of image inpainting, ancient painting inpainting, deep learning is used to automatically identify and repair damaged parts of paintings, effectively improving the efficiency and accuracy of inpainting work. Zhao et al. [24] developed a multi-channel encoder to capture semantic features at different scales and then applied these fine-grained features to the inpainting of ancient paintings. This method has the problem of high-frequency detail information loss in image restoration, which leads to inconsistency between the restored image and the surrounding area. Zhou et al. [25] enhanced the accuracy of color inpainting in missing areas by extracting color data from deep features. Additionally, they introduced a multi-stage feature refinement framework that effectively transfers feature information from undamaged to damaged areas. However, the lack of depth semantic information leads to poor fusion of deep semantic features and detail features of images. Deng et al. [26] proposed a structure-guided model to use structure to fill in complex and diverse damaged contents for the inpainting of ancient paintings. In this manner, the consistency of semantic information between global and local areas is not fully considered.

Although the above methods improve the image inpainting effect to a certain extent, they are not good at dealing with the problems of insufficient extraction of deep semantic information, loss of high-frequency detail features and inconsistency of semantic information between global and local areas. In order to solve the above problems, this paper makes an improvement. Firstly, a Residual Pyramid Encoder (RPE) network is designed to fully extract the deep semantic features and detail features of landscape painting images and strengthen the ability to deal with image details. In addition, this paper proposes a frequency-aware mechanism to capture and reconstruct the details and texture features of ancient paintings, and provide more perceptual information to solve the problem of high-frequency detail loss in image restoration. Finally, this paper designs a dual discriminator to ensure that the semantic information of the global and local regions is consistent during the restoration process.

2.2. Generative Adversarial Networks

Generative Adversarial Networks (GANs [12]) enable the network to generate high-quality and realistic samples by introducing a unique adversarial training mechanism. GANs have achieved remarkable results in many fields, such as image, speech, and text generation, and have become one of the key technologies in deep learning [27]. In ancient painting inpainting, the application of GANs shows great potential. Ancient paintings are known for their unique artistic style and deep cultural heritage, but with time, many works have suffered damage, fading, and breakage. Traditional inpainting methods [17] often rely on manual copying and empirical judgement, which is not only time-consuming and labor-intensive, but also makes it difficult to ensure that the inpainting effect is consistent with the style of the original painting. Existing CNN-based methods [24–26] cannot generate realistic texture details.

The application of GANs in this field is mainly reflected in their ability to generate high-quality inpainting samples [28]. The generator extracts and reconstructs features by progressively reducing and recovering the spatial dimensions of the image. U-Net [29] utilizes its symmetric U-shape structure and hopping connections to preserve image details and contextual information, while Transformer [30] handles global dependencies through a self-attentive mechanism. The generator learns the deep features and artistic style of ancient paintings and is able to generate restored images that are consistent with the style of the original paintings. The discriminator is responsible for evaluating the quality of the generated images and guiding the generator to optimize them by comparing them with real samples. In the continuous adversarial learning process, the generator gradually improves the authenticity and artistry of its generated images, and ultimately achieves the effect of fake to real. In addition, GANs are able to handle a variety of complex damage situations in the inpainting process, such as irregular broken areas and color degradation.

3. Methods

3.1. MFGAN Architecture

The model architecture of the MFGAN for ancient painting inpainting is depicted in Figure 1. The MFGAN comprises a multi-layer residual pyramid generator and a dual discriminator. Among them, the multi-layer residual pyramid generator is composed of a residual pyramid encoder and a multi-layer decoder. The residual pyramid encoder fully extracts the deep semantic features and detail features of the ancient painting image by effectively combining the deep feature extraction module and the channel attention module, and strengthens the ability to deal with the details of the image. Meanwhile, a frequency-aware mechanism is introduced in the encoder network, and makes use of the frequency attention module to obtain the high-frequency perceptual features of the image, and capture and reconstruct the details and texture features of ancient paintings by increasing the skip connection between low and high frequencies and providing more perceptible information to solve the problem of high-frequency detail loss in image inpainting. The dual discriminator consists of a global discriminator and a local discriminator to safeguard the

semantic consistency of the restored image. During the training process of this network, the multi-layer residual pyramid generator and the dual discriminator constantly interact with each other and, finally, reach a balance to obtain semantically and visually superior images.

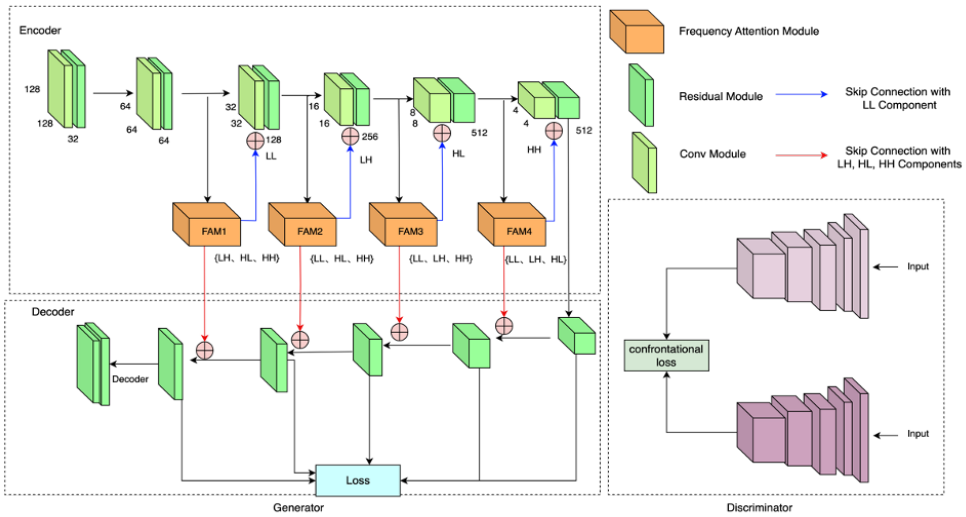

Figure 1. Overall architecture diagram of MFGAN including generator and discriminator.

3.2. Multi-Layer Residual Pyramid Generator

To delve into the profundity of semantic features, existing methods often attempt to deepen the network, addressing the issue of superficiality in information retrieval. However, this method may lead to some problems, such as slow convergence of the network and increased error. Consequently, this paper introduces a multi-layer residual pyramid generator designed to extract deep semantic features from images, comprising a Residual Pyramid Encoder (RPE) and a multi-layer decoder. At the same time, the Frequency-Aware Mechanism (FAM) uses a frequency attention module to obtain the high-frequency perception features of the image and solve the loss of high-frequency details in image inpainting.

3.2.1. Residual Pyramid Encoder

To solve the deficiency of ancient painting images in the acquisition of deep semantic information, this paper designs a Residual Pyramid Encoder (RPE) network, which fully extracts the deep semantic features and detail features of ancient painting images by effectively combining the deep feature extraction module and the channel attention module. As shown in Figure 2, RPE is designed to enhance the capacity for handling intricate image details and to facilitate the detection and reconstruction of missing image segments. The depth of RPE is six layers, which captures feature maps from the highest layer to the lowest layer, labeled f1 to f6. Each layer of RPE integrates convolutional blocks and residual modules to support feature extraction and refinement of image details, enhancing the feature extraction capability of the image. Each layer of residual block uses a depth-separable convolution operation and residual connection, which reduces the parameters and calculation of the model and improves the image inpainting efficiency. When processing high-resolution images in time, the image inpainting efficiency will not change much.

The residual modules consist of a Bottleneck block and an Attention block. The Bottleneck block splits the input feature map channel into left and right parts equally. The right half of the split is convolved by 1×1 convolution to transform the channel dimension to half of the output channel, and the feature scale is balanced by a BN layer and a Silu activation function. The left half extracts fine-grained features through a deep convolutional structure, and then the outputs of the left and right paths are spliced together, and the channel dimension is reduced by deep separable convolution.

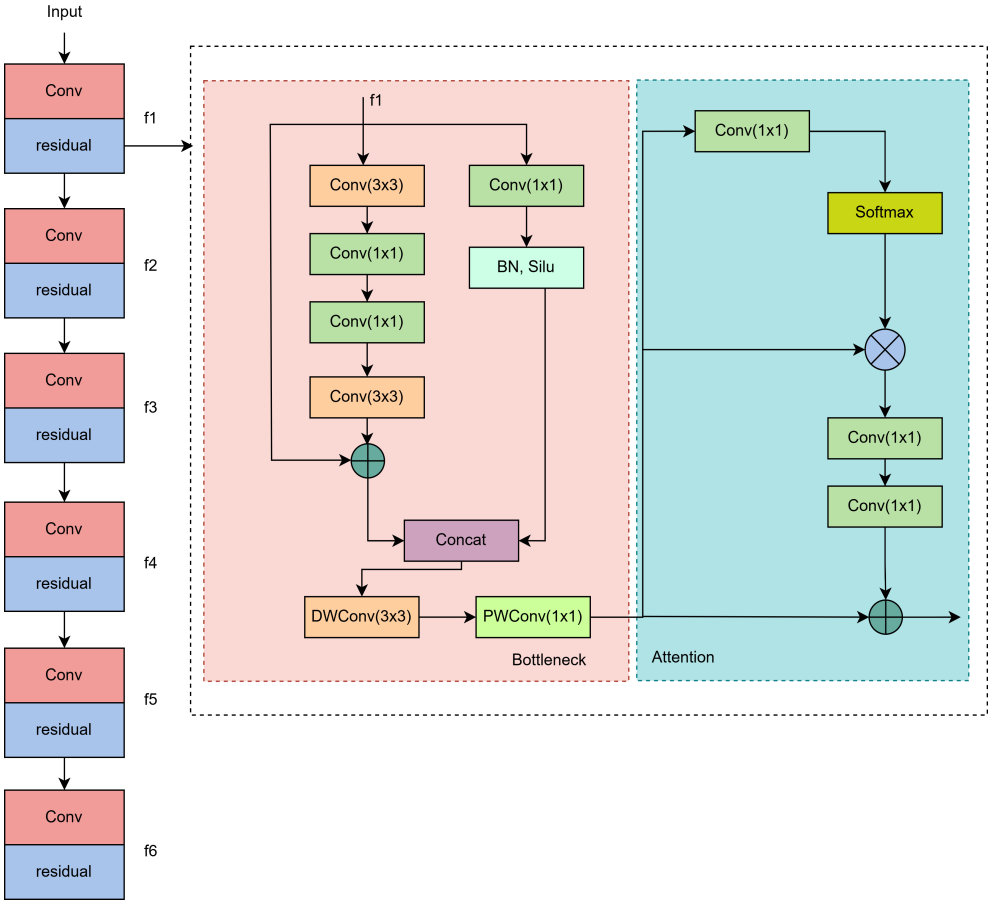

Figure 2. Architecture of the residual pyramid encoder.

The Bottleneck block enhances the overall learning ability of the network to learn richer and more abstract feature representations, thus capturing the details and texture information in ancient paintings. Meanwhile, the Attention block is used to strengthen the importance between channels and dynamically adjust the channel weights, which enables the network to focus more on the feature channels that have a significant impact on image inpainting, and enhances the RPE encoder's ability to process local features. The Bottleneck block is represented by

$$F_1^1 = \mathrm{Conv}_3(\mathrm{Conv}_1(\mathrm{Conv}_1(\mathrm{Conv}_3(\mathrm{F})))) \oplus \mathrm{F} \tag{1}$$

$$F_1^2 = \mathrm{Silu}(BN(\mathrm{Conv}_1(\mathrm{F}))) \tag{2}$$

$$F_1 = PW\,\mathrm{Conv}_1\left(DW\,\mathrm{Conv}_3\left(\mathrm{Concat}\left(F_1^1, F_1^2\right)\right)\right) \tag{3}$$

where F is the input feature of the previous residual module, $DW\,\mathrm{Conv}_3$ is a 3×3 deep convolution, $PW\,\mathrm{Conv}\,1$ is a 1×1 point-by-point convolution, Concat is a splicing operation, Silu is an activation function, and BN is a batch normalization operation.

After extracting the deep semantic features of the image through the Bottleneck block, the Attention module is then used to enable the network to focus more on the feature channels that have a significant impact on image inpainting and enhance the image inpainting capability. The Attention block can be represented as:

$$F_2 = \mathrm{Softmax}(\mathrm{Conv}_1(F_1)) \otimes F_1 \tag{4}$$

$$F_{\mathrm{out}} = \mathrm{Conv}_1(\mathrm{Conv}_1(F_2)) \oplus F_1 \tag{5}$$

where Softmax is the activation function, $Conv_1$ is the 1×1 convolution module, $\otimes$ is element-by-element multiplication, and $\oplus$ is element-by-element addition.

3.2.2. Frequency-Aware Mechanism

To reconstruct the details and texture features of ancient painting images and maintain the consistency and coherence between the repaired area and the surrounding area, this paper proposes a Frequency-Aware Mechanism (FAM), which uses the frequency attention module to enhance the frequency perception ability of the model, and uses frequency-based skip connections to improve the extraction ability of high-frequency information. The frequency-aware mechanism module is shown in Figure 3. Features are dissected into an array of frequency bands through the application of a Discrete Wavelet Transform (DWT [31]). By utilizing the Haar wavelet analysis technique, the input features are separated into four distinct frequency components LL, LH, HL, and HH. In the frequency domain, the texture and details of the painting are presented in the form of different frequency components, with the high-frequency components usually associated with details and edges, while the low-frequency components are associated with the overall structure and smooth regions. By designing novel encoders, different frequency features can be selectively enhanced or suppressed to highlight or smooth specific texture features. To this end, we design skip-connected frequency-aware mechanisms to more accurately predict and repair missing parts, maintaining the naturalness and realism of the image.

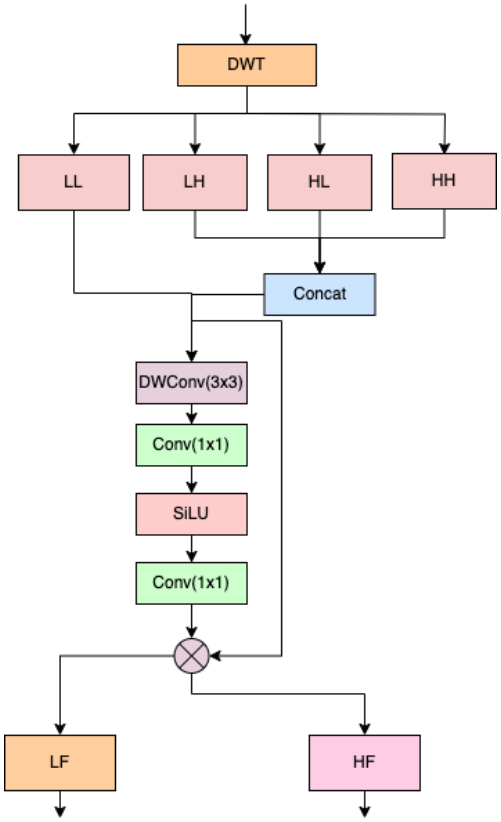

Figure 3. Frequency attention module.

DWT facilitates the breakdown of an image's signal across a spectrum of distinct frequency bands, where the LL component represents the approximate information of the image and contains the main structural and appearance features, and LH, HL, and HH represent the high-frequency components in the image and contain details and edge features. The four frequency components are represented by the following equation:

$$(\mathrm{LL_i}, \mathrm{LH_i}, \mathrm{HL_i}, \mathrm{HH_i}) = \mathrm{DWT}(F_i) \tag{6}$$

where F_i is the feature obtained from the i-th residual module of the encoder, DWT denotes the Discrete Wavelet Transform, and LL_i, LH_i, HL_i, and HH_i are four components of different frequencies.

Then, the Frequency Attention Module (FAM) uses the deep convolution of 3×3 and two 1×1 to obtain fine-grained frequency feature information, which is multiplied with the original input features to obtain the individual channel weights. The original style of painting is preserved through residual connection, which helps to alleviate the problems of gradient disappearance and explosion, thus improving the stability and convergence of the model. Thus, the network can identify and enhance the channels containing important texture and detail features. The acquisition of high-frequency features and low-frequency features can be expressed by

$$\mathrm{HF}_1 = \mathrm{Cat}(\mathrm{LH}, \mathrm{HL}, \mathrm{HH}) \tag{7}$$

$$\mathrm{HF} = \mathrm{Conv}_1(\mathrm{SiLU}(\mathrm{Conv}_1(\mathrm{DWConv}_3))) \otimes \mathrm{HF}_1 \tag{8}$$

$$\mathrm{LF} = \mathrm{Conv}_1(\mathrm{SiLU}(\mathrm{Conv}_1(\mathrm{DWConv}_3))) \otimes \mathrm{LL} \tag{9}$$

where HF denotes the obtained high-frequency features, LF denotes the obtained low-frequency features, Cat denotes the splicing operation, DWConv_3 denotes the 3×3 depth convolution, Conv_1 denotes the 1×1 convolution operation, and SiLU is the activation function.

Finally, as shown in Figure 1, we use the skip connection with the LL component in the encoding stage to ensure that accurate and realistic image features are captured throughout the feature extraction process. This approach effectively utilizes the global information of the LL component to maintain the details and integrity of feature extraction. Secondly, we introduce the skip connection with LH, HL, and HH components in the decoding stage to ensure that the utilization of high-frequency features is effectively strengthened throughout the feature reconstruction process, restoring more realistic texture details.

3.3. Multi-Layer Decoder

The model structure of the multi-layer decoder (MD) is shown in Figure 4. The multi-layer decoder first takes the semantic feature map f6 of the highest layer in the residual pyramid encoder as the input to the decoder. Next, the feature map f5 in the residual pyramid encoder is combined with the decoded feature map f6 to obtain a new feature map F6. After that, these feature maps are sequentially used as inputs to the decoder and they are decoded to obtain the restored images F5, F4, F3, F2, and F1 at different scales. Finally, the multi-layered residual pyramid generator model is optimized using pyramidal loss, i.e., the output images at different scales are optimized by computing the normalized chemical distance between the output image and the original image at each scale to progressively refine the final output.

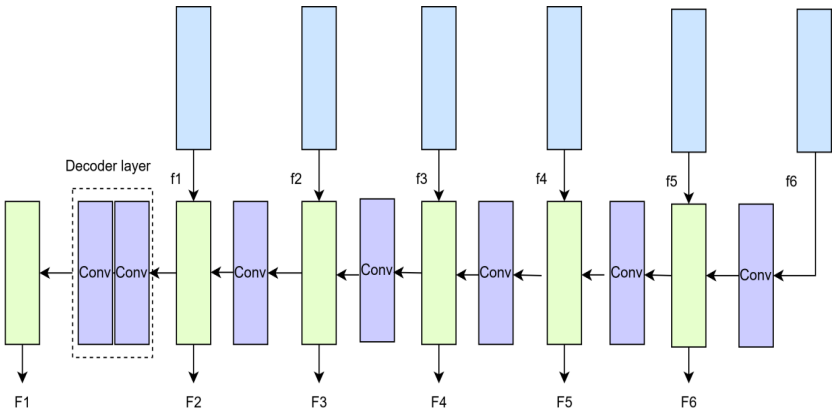

Figure 4. Multi-layer decoder architecture.

3.4. Dual Discriminator

To ensure that the restored image maintains the same semantic information as the original painting in terms of global structure and local details, we designed a dual discriminator to avoid the semantic fragmentation that may occur in traditional methods, while reducing the blurriness of the restored area and improving the visual quality of the image.

As shown in Figure 1, a Dual Discriminator (DD) contains a global discriminator and a local discriminator. The global discriminator operates by receiving both the entire reconstructed image and the original image as inputs, aiming to produce a similarity probability value that reflects the overall consistency between the restored and the authentic images in terms of semantics, color, and other visual aspects. On the other hand, the local discriminator is designed to process the specific missing segments of both the restored and the real images, generating a similarity probability value for these areas. This output measures the likeness of the missing regions between the reconstructed and the actual images, ensuring that the details in these localized areas are also accurately reflected. It serves to ensure that the boundary has the same color. Unlike traditional discriminators, the input in the dual discriminator can be not only square missing regions but also irregular missing regions. The design ensures the overall consistency and regional consistency of the generated image. Table 1 illustrates the architecture of the global discriminator, comprising five convolutional layers and a single fully connected layer, culminating in the generation of a 1×1024 dimensional feature vector. Table 2 presents the configuration of the local discriminator, including four convolutional layers and one fully connected layer, similarly yielding a 1×1024 dimensional feature vector.

Table 1. Global discriminator network architecture.

Layer	Kernel	Stride	Output Channels	Activation Function
Conv	5×5	2	64	ReLU
Conv	5×5	2	128	ReLU
Conv	5×5	2	256	ReLU
Conv	5×5	2	512	ReLU
Conv	5×5	2	512	ReLU
FC	-	-	1024	Sigmoid

Table 2. Local discriminator network architecture.

Layer	Kernel	Stride	Output Channels	Activation Function
Conv	5×5	2	64	ReLU
Conv	5×5	2	128	ReLU
Conv	5×5	2	256	ReLU
Conv	5×5	2	512	ReLU
FC	-	-	1024	Sigmoid

3.5. Loss Function

In the adversarial training process of the generator and discriminator, the loss function [32] plays a crucial role. The generator within the network is responsible for creating smaller images that correspond to the missing parts of the original. These smaller images, along with the application of an embedding mask over the voids, are then integrated to form the fully restored image, effectively concealing the areas that were previously absent. Therefore, the global $L1$ loss is considered to be added to constrain the overall consistency of the image, while TV loss is introduced to reduce the image noise and maintain the smoothness of the boundary for the boundary prominence problem of the repaired region. The generator losses include the adversarial loss, local $L1$ loss, global $L1$ loss, and TV loss as follows:

$$\text{Loss(G)} = u_{adv}\,\text{Loss}_{adv} + u_{L-L1}\,\text{Loss}_{L-L1} + u_{G-L1}\,\text{Loss}_{G-L1} + u_{TV}\,\text{Loss}_{TV} \quad (10)$$

where Loss_{adv}, Loss_{L-L1}, Loss_{G-L1}, and Loss_{TV} are the adversarial loss, local $L1$ loss, global $L1$ loss, and TV loss. u_{adv}, u_{L-L1}, u_{G-L1} and u_{TV} correspond to the weights of the four loss functions, respectively. The TV loss helps to maintain the local smoothness and coherence, while the global $L1$ loss ensures that the overall image consistency is preserved, and important details, including the artist's unique style, are not lost. This hybrid loss function strategy aims to strike a balance between the competing goals of spatial smoothness and detail preservation.

4. Experimental Setup

4.1. Datasets

We constructed two themed ancient painting datasets for ancient painting inpainting. The ancient painting dataset includes 5798 ancient paintings with ancient themes, covering ancient paintings from multiple historical periods, including different styles, such as ink painting. All image works are scanned and quality-screened to ensure image resolution and clarity. The size of each image is uniformly 256 × 256 pixels. The training set accounts for 80%, which is 4638 images, and the test set accounts for 20%, which is 1160 images. The huaniao painting dataset contains 95 Huaniao paintings with flower and bird themes. These paintings also come from different historical periods and styles. The size of each image is also unified to 1024 × 1024 pixels. This dataset is also divided into an 80% training set and a 20% test set, 76 and 19 images, respectively. We used a specific script to generate the mask image. This script can adjust the image to the specified size, and automatically generate multiple masks, pairing different mask images for each image, and finally form an image, mask pair.

To verify the generalization of our method, we also employ the publicly available CelebAHQ dataset [33] for further validation. CelebAHQ is a high-quality dataset comprising 30,000 face images, all uniformly resized to 256 × 256 pixels. Following standard practice, we allocated 80% of the dataset to the training set, which amounts to 24,000 images, and reserved 20% for the test set, totaling 6000 images. To ensure fairness, all numerical experimental results in this experiment are obtained by training and testing with the same dataset.

4.2. Experimental Setup

The system platforms used in this experiment are torch2.0.0, CUDA11.8, and the RTX 4090 (24 GB) graphics card. The parameters are set to a batch size of 16. The experiment is divided into three phases: 100 epochs for T1, 100 epochs for T2, and 200 epochs for T3. The hyperparameter is 0.00001, the learning rate is 0.0001, and Adam is used as the optimizer.

4.3. Evaluation Metrics

In this experiment, we used the Mean Absolute Error (MAE), the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index Measure (SSIM [34]), and the Learned Perceptual Image Patch Similarity (LPIPS [35]) as model performance evaluation indicators.

MAE is a simple image quality evaluation metric that is calculated. MAE is expressed by

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{11}$$

where x_i and y_i are the pixel values in the original image and reconstructed image, respectively, and N represents the total number of pixels.

PSNR calculation is based on the mean square error (MSE). MSE is expressed by

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[I(i,j) - K(i,j)]^2 \tag{12}$$

where $I(i,j)$ and $K(i,j)$ are the pixel values of images I and K at position (i,j), respectively. M represents the height of the image or dataset, and N represents the width of the image or dataset. Once the value of MSE is available, $PSNR$ is expressed by

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{13}$$

where, MAX_I is the maximum possible value of the image pixel.

SSIM can expressed by

$$\text{SSIM} = [l(x,y)c(x,y)s(x,y)] = \frac{(2\mu_x\mu_y + C_1)(2\tau_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\tau_x^2 + \tau_y^2 + C_2\right)} \tag{14}$$

where $l(x,y)$, $c(x,y)$, and $s(x,y)$, respectively, represent the similarity of luminance, contrast, and structure. u_x and u_y are the average values of x and y, respectively, τ_x and τ_y are their standard deviations, τ_{xy} is their covariance, and C_1 and C_2 are constants to maintain stability.

LPIPS is an advanced measurement method for measuring the visual similarity between images. When evaluating the quality of generated images, LPIPS provides an evaluation method that is closer to human visual perception. It solves the problem that traditional crispy chicken measurements cannot reflect differences in human visual perception. The lower the value, the smaller the perceived difference between the two images. It is expressed by

$$\text{LPIPS}(x,y) = \sum_l w_l \|f_l(x) - f_l(y)\|_2^2 \tag{15}$$

where $f_l(x)$ and $f_l(y)$ are features extracted from images x and y through the pre-trained deep neural network at layer l, $\|\cdot\|$ is between $f_l(x)$ and $f_l(y)$ Euclidean distance, w_l is the weight of the l-th layer, and the summation is performed on the selected l-layer feature layer.

5. Experiments

5.1. Experiments on Ancient Painting Dataset

5.1.1. Quantitative Comparison

To validate the effectiveness of our method, we conducted a comparison experiment on the ancient painting dataset, using four models for validation comparison PI [36], RFR [37], EC [38], and FcF [39]. The results of the experiment are shown in Table 3. The proposed models are quantitatively evaluated using PSNR, SSIM, MSE, and LPIPS as evaluation metrics. As can be seen from Table 3, the PSNR and SSIM values of the PI model are the lowest, which is because the PI model is insufficient in extracting deep semantic information, resulting in the loss of the high-frequency detail features of the reconstructed images. The insufficient extraction of detail features leads to a poor image restoration effect. The performance of the EC and LG models in China landscape painting image restoration has been improved. The LG model produces better PSNR and SSIM because it uses a large receptive field network to repair the overall structure and local texture details and a small receptive field local refinement network to eliminate artifacts. However, the acceptance area of the local refinement network is small, which means that when the damaged area is large, it cannot capture the global information, resulting in inconsistency between the global and local semantic information. It leads to a more serious plaque effect and a higher LPIPS value. Although the EC and FcF models have improved the performance of ancient painting restoration, it can be seen from the visual comparison of the five methods in Figure 5 that compared with other methods, the proposed method has achieved the best results, and the training has stabilized at the 300–400 epoch stage. In addition, our MFGAN, with a size of 15 M, demonstrates a promising inference time of approximately 65 ms for each image of size 256 × 256. In comparison, the RFR [37] model, which has a

size of 33 M, requires an inference time of 95 ms. As a result, our method not only has a lower computational cost but also a faster run time, indicating its potential for practical applications. The PSNR and SSIM indicators reached the highest values of 34.59 and 0.929. This is because MFGAN uses the RPE residual encoder to effectively combine the deep feature extraction module and the channel attention mechanism to fully extract the deep semantic features and detail features of the ancient painting images. Furthermore, MFGAN incorporates the frequency perception mechanism to capture and reconstruct the details and texture features of the ancient paintings by increasing the skip connection between low frequency and high frequency, and effectively restoring the high-frequency details. Finally, this paper uses the double discriminator network to ensure that the semantic information of the restored image is consistent with that of the original painting in terms of global structure and local details, thus ensuring that the semantic information of the global and local areas of the image is consistent during the restoration process.

Table 3. Comparison of different models on the ancient painting dataset.

Model	PSNR↑	SSIM↑	MSE↓	LPIPS↓
PI [36]	32.41	0.892	0.0009	0.0762
RFR [37]	33.42	0.921	0.0008	0.0720
EC [38]	33.53	0.913	0.0007	0.0523
FcF [39]	33.72	0.914	0.0008	0.0532
MFGAN (Ours)	34.59	0.929	0.0006	0.0512

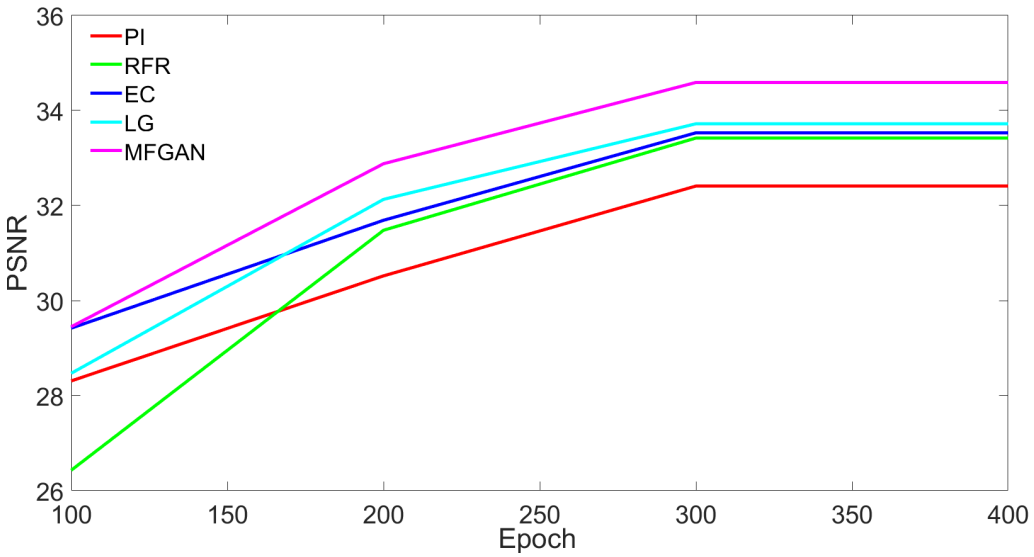

Figure 5. Comparative visualization of five methods on PSNR metrics.

5.1.2. Qualitative Comparison

Figure 6 shows the visualization results of different methods on the ancient painting dataset. It can be seen from the figure that Figure 6c,d have loss of detail features in the image inpainting process, Figure 6e,f have noise and distortion in the inpainting process, whereas the proposed method generates a higher-quality restored image in comparison to the other methods, and the image inpainting effect is more prominent in terms of the edge details, which achieves the best inpainting effect.

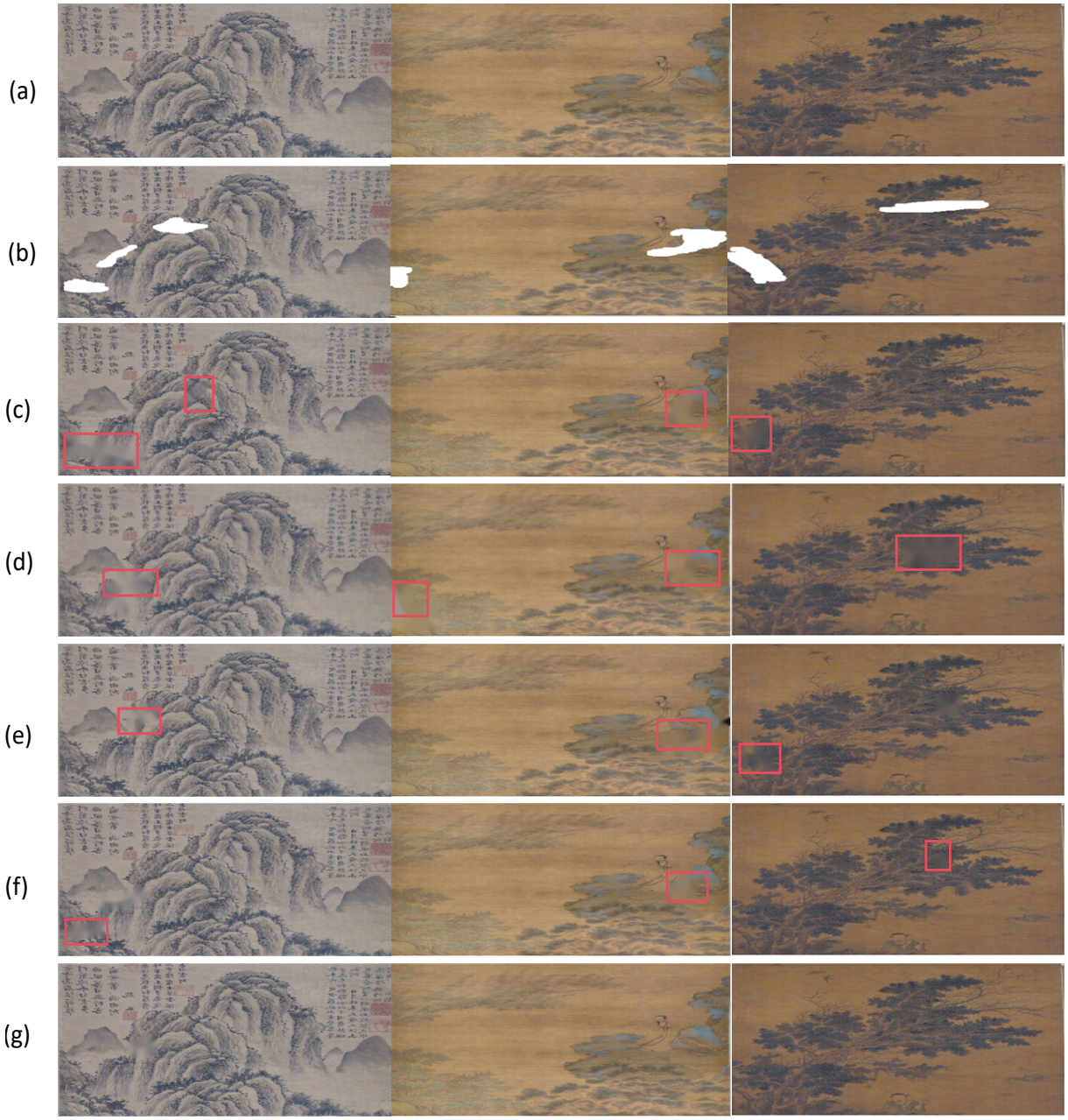

Figure 6. Image inpainting results of different methods on the ancient painting dataset. (**a**) is the groundtruth image, (**b**) is the input image, and (**c**–**g**) are the image inpainting results for PI, RFR, EC, FcF and the proposed method, respectively.

5.2. Experiments on the Huaniao Painting Dataset

5.2.1. Quantitative Comparison

To verify the generalization of the models in this paper, we conducted a comparison experiment on the Huaniao dataset, using four mainstream models for validation and comparison. The results of the experiment are shown in Table 4. From Table 4, it can be seen that our method reaches the maximum value in the PSNR and SSIM metrics, which are 22.85 and 0.835, respectively, and the minimum value in MSE and LPIPS, which are 0.0042 and 0.0976, respectively. From the comparison results of the five methods on the Huaniao dataset in Figure 7, it can be seen that, before 100 epochs, the MFGAN model's PSNR value is lower than EC and FcF, and with increase in training times, the MFGAN method obtains the highest PSNR accuracy with a large difference, which fully demonstrates the excellent performance of our method in the image inpainting of ancient paintings, and substantially

improves the task of image inpainting of ancient paintings. The PSNR value of our method is still on the rise at 400 epochs and reaches saturation at about 500 epochs of training.

Table 4. Comparison of different models on the Huaniao painting dataset.

Model	PSNR↑	SSIM↑	MSE↓	LPIPS↓
PI [36]	19.15	0.723	0.0182	0.1896
RFR [37]	20.12	0.745	0.0151	0.1637
EC [38]	19.52	0.724	0.0154	0.1756
FcF [39]	21.34	0.793	0.0095	0.1874
MFGAN (Ours)	22.85	0.835	0.0042	0.0976

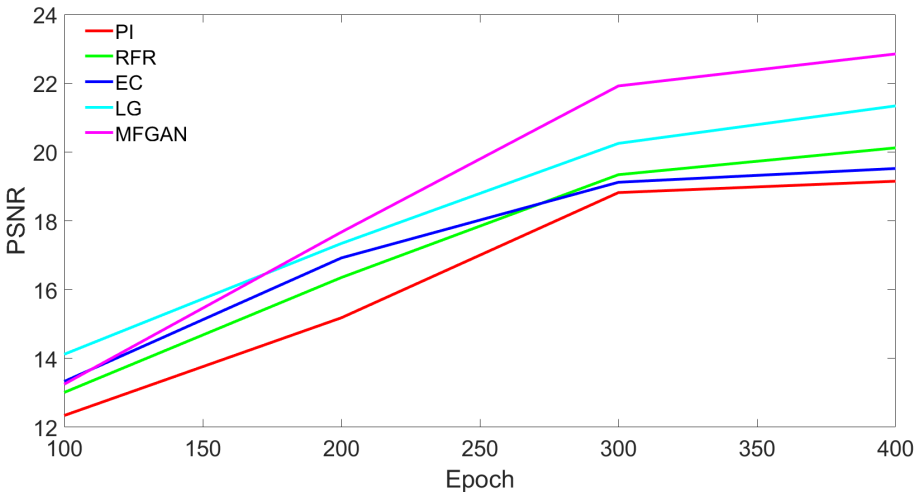

Figure 7. Comparative visualization of five methods on PSNR metrics for the Huaniao dataset.

5.2.2. Qualitative Comparison

Figure 8 shows the visualization results of the different methods on the ancient painting dataset. As can be seen from Figure 8, the methods of Figure 8c,d have a certain degree of blurring, the method of Figure 8e has a discontinuity in the inpainting at the image boundary, the method of Figure 8f has improved the details of image inpainting but still has problems such as noise, while the experimental results of our method are better compared to the other image inpainting methods, demonstrating semantically and color-consistent inpainting results, and achieving excellent inpainting results in terms of image details and edge texture.

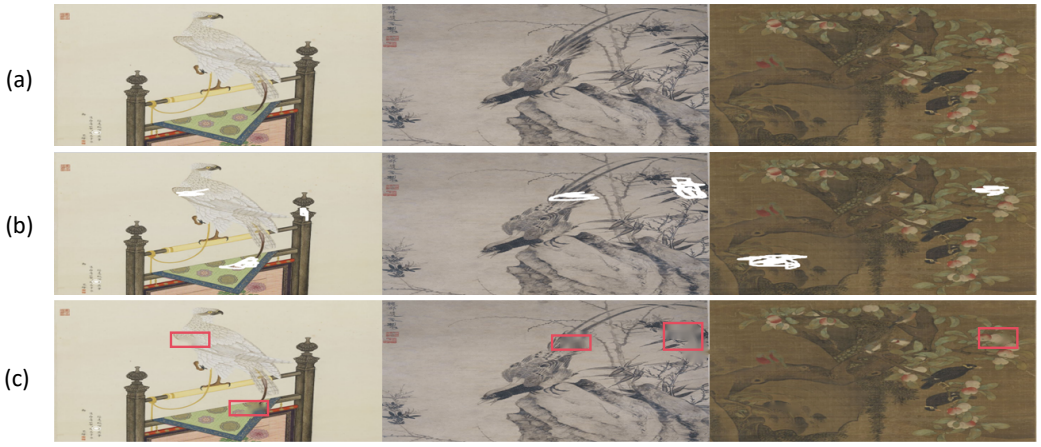

Figure 8. *Cont.*

Figure 8. Image inpainting results of different methods on the Huaniao painting dataset. (**a**) is the groundtruth image, (**b**) is the input image, (**c**–**g**) are the ancient inpainting results for PI, RFR, EC, FcF and the proposed method, respectively.

5.3. Experiments on CelebAHQ Dataset

In order to verify the generalization of our method, we conducted a comparative experiment on the CelebAHQ dataset, and the experimental results are shown in Table 5. As can be seen from Table 5, compared with the other competitive methods, our method obtains the highest PSNR and SSIM values, which are 23.52 and 0.855, respectively. Compared with PI, RFR, EC, and FcF, the PSNR score of our MFGAN is improved by 3.23, 2.26, 2.04, and 1.24, respectively. At the same time, the lowest LPIPS value of 0.0534 was obtained. From the experimental results, it can be seen that the proposed MFGAN has also achieved excellent performance in high-quality face image inpainting, which proves the strong generalization of our method.

Table 5. Comparison of different models on the CelebAHQ dataset [33].

Model	PSNR↑	SSIM↑	MSE↓	LPIPS↓
PI [36]	20.29	0.773	0.0171	0.1357
RFR [37]	21.26	0.783	0.0142	0.1105
EC [38]	21.48	0.817	0.0129	0.0842
FcF [39]	22.28	0.834	0.0075	0.0714
MFGAN (Ours)	23.52	0.855	0.0054	0.0534

5.4. Ablation Study

Effect of Residual Pyramid Encoder (RFE). Table 6 presents an evaluation of the impact of the various components discussed in this study. From the following table, it can be seen that the use of the residual pyramid encoder structure RFE ensures the semantic consistency of the restored images of ancient paintings, effectively combines the deep feature extraction module and the channel attention module, fully extracts the deep semantic features of the images of ancient paintings, strengthens the ability to process the details of the images, and helps to identify and reconstruct the missing parts of the images. The PSNR metrics and the SSIM metrics are, respectively, improved by 1.8 and 0.049. It can be seen from Figure 9b that there is local ambiguity in the image inpainting without using the RFE module.

Table 6. Ablation study of different modules of this paper on the Huaniao painting dataset.

Model	PSNR↑	SSIM↑	MSE↓	LPIPS↓
Baseline	18.78	0.746	0.0068	0.2071
w/o RFE	21.05	0.786	0.0052	0.1301
w/o MD	21.32	0.753	0.0054	0.1221
w/o FAM	20.76	0.812	0.0051	0.1201
w/o DD	20.73	0.801	0.0049	0.1015
MFGAN (Ours)	22.85	0.835	0.0042	0.0976

Effect of Multi-layer Decoder (MD). Using the multi-layer decoder structure MD, the filling content of the missing regions of different images at each scale is refined by calculating the normalized distance between the output image and the original image at each scale to progressively refine the final output. As can be seen in Figure 9c, the lack of use of the MD module suffers from the disadvantage of incomplete inpainting results, proving the effectiveness of the MD module in image inpainting.

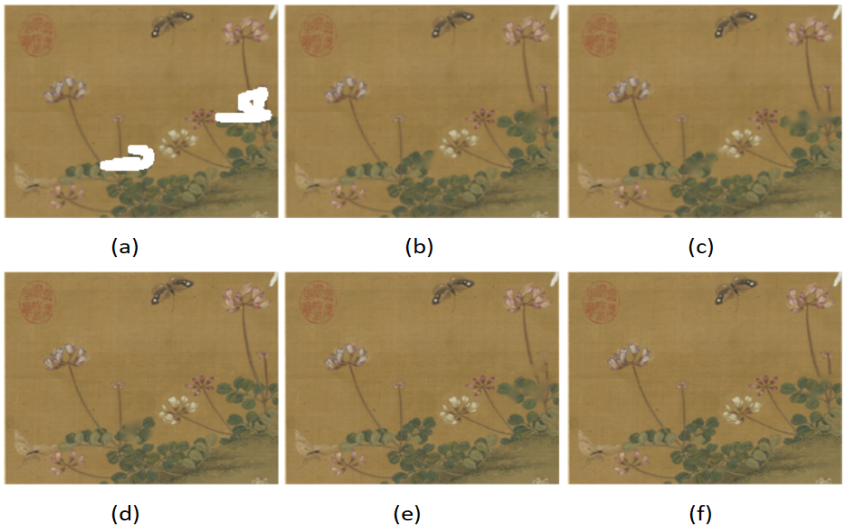

Figure 9. Comparison of results before and after using the RFE module, where (**a**) is the input image, (**b**) is the restored image without using the RFE module, (**c**) is the restored image without using the MD module, (**d**) is the restored image without using the FAM module, (**e**) is the restored image without using the DD module, and (**f**) is the restored image using the proposed method.

Effect of Frequency-Aware Mechanism (FAM). Using the FAM mechanism, the frequency attention module is used to obtain high-frequency perceptual features of the image, capture and reconstruct the details and texture features of ancient paintings by increasing the skip connections between low and high frequencies, provide more perceptible information to solve the problem of high-frequency detail loss in image inpainting, and improve the quality of image inpainting, with an improvement in the PSNR metrics and SSIM metrics by 2.09 and 0.023. It proves the effectiveness of the FAM mechanism in capturing and reconstructing the detail and texture features of ancient paintings, and obtains high-quality ancient painting images.

Effect of Dual Discriminator (DD). The use of a dual discriminator to solve the problem of local regions of ancient painting images, and to eliminate the problems of discontinuity and blurring at the boundary of the missing regions of the image, significantly improves the image inpainting task of ancient paintings. The PSNR metrics and SSIM metrics are improved by 2.12 and 0.034, respectively. The following Figure 9e shows the inpainting effect of the image without the use of the SSIM module, with problems of discontinuity, blurring, etc., present at the boundary of the regions. On the contrary, as shown in Figure 9f, the inpainted image generated by our method is not only highly

similar to the original visually, but also consistent with the original in artistic style and expressive force.

6. Conclusions

In this paper, a Generative Adversarial Network (GAN)-based ancient painting inpainting method, named MFGAN, which utilizes multi-layer feature enhancement and frequency perception, is proposed to address the issue of insufficient extraction of deep semantic information in ancient paintings inpainting. MFGAN effectively enhances the extraction of deep semantic features of ancient paintings through the introduction of a residual pyramid encoder and significantly strengthens the processing of image details by the effective combination of the deep feature extraction module and channel attention, which are effectively combined to significantly enhance the processing of image details. Furthermore, in order to reconstruct the high-frequency details and textures of the image, this paper proposes a frequency-aware mechanism to obtain the high-frequency-aware features of the image by using the frequency attention module, and capture and reconstruct the high-frequency details and texture features of the ancient paintings by establishing a skip connection between the low-frequency and high-frequency features. In addition, a dual discriminator is proposed, which effectively ensures the semantic consistency of the restored image in global and local regions, significantly reduces the discontinuities and blurring differences at the boundary, and improves the overall quality of the restored image. In the future, we will explore the effect of the frequency transformation method DWT, as well as other alternatives (e.g., the quantum mechanics-based adaptive basis) on the detailed restoration of ancient paintings, to further improve the performance of ancient painting inpainting.

Author Contributions: Project administration, X.L.; data curation, X.L.; writing—draft preparation, X.L.; methodology, X.L. and J.W.; investigation, Y.W.; writing—review and editing, J.W., Y.W., and N.W.; validation, X.L. and N.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fundamental Research Funds for the Central Universities (2020YJS031).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: The authors would like to thank all the reviewers for their insightful comments and constructive suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, Z.; Lei, Y.; Wu, X. Chinese Ancient Paintings Inpainting Based on Edge Guidance and Multi-Scale Residual Blocks. *Electronics* **2024**, *13*, 1212. [CrossRef]
2. Moreira, F.; Queiroz, A.; Aronson, J. inpainting principles applied to cultural ancients. *J. Nat. Conserv.* **2006**, *14*, 217–224.
3. Skokanov, H.; Slach, T.; Havlıek, M.; Halas, P.; Divek, J.; Pinlerov, Z.; Kouteck, T.; Ebesta, J.; Kallabov, E. ancient painting in the Research of ancient Changes. *J. Anc. Ecol.* **2021**, *14*, 110–127.
4. Chen, X.; You, S.; Tezcan, K.; Konukoglu, E. Unsupervised lesion detection via image inpainting with a normative prior. *Med. Image Anal.* **2020**, *64*, 101713
5. Yin, X.; Zhou, S. Image Structure-Preserving Denoising Based on Difference Curvature Driven Fractional Nonlinear Diffusion. *Math. Probl. Eng.* **2015**, *2015*, 930984.
6. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.* **2023**, *56*, 1–39.
7. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212.
8. Amudha, J.; Pradeepa, N.; Sudhakar, R. A survey on digital image inpainting. *Procedia Eng.* **2012**, *38*, 2378–2382.
9. Mathew, A.; Amudha, P.; Sivakumari, S. Deep learning techniques: An overview. In *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 599–608.

10. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image inpainting using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1833–1844.
11. Banham, M.; Katsaggelos, A. Digital image inpainting. *IEEE Signal Process. Mag.* **1997**, *14*, 24–41.
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.
13. Ding, X.; Wang, Y.; Xu, Z.; Welch, W.; Wang, Z. Ccgan: Continuous conditional generative adversarial networks for image generation. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
14. Xue, A. End-to-end ancient painting creation using generative adversarial networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Event, 5–9 January 2021; pp. 3863–3871.
15. Yang, R.; Yang, H.; Zhao, M.; Jia, R.; Wu, X.; Zhang, Y. Special perceptual parsing for Chinese ancient painting scene understanding: A semantic segmentation approach. *Neural Comput. Appl.* **2024**, *36*, 5231–5249.
16. Peng, X.; Peng, S.; Hu, Q.; Peng, J.; Wang, J.; Liu, X.; Fan, J. Contour-enhanced CycleGAN framework for style transfer from scenery photos to Chinese ancient paintings. *Neural Comput. Appl.* **2022**, *34*, 18075–18096.
17. Zha, Z.; Yuan, X.; Zhou, J.; Zhu, C.; Wen, B. Image inpainting via simultaneous nonlocal self-similarity priors. *IEEE Trans. Image Process.* **2020**, *29*, 8561–8576.
18. Deng, X.; Dragotti, P. Deep convolutional neural network for multi-modal image inpainting and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3333–3348.
19. Zamir, S.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.; Yang, M.; Shao, L. Multi-stage progressive image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14821–14831.
20. Alzubaidi, L.; Zhang, J.; Humaidi, A.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamarıa, J.; Fadhel, M.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 1–74.
21. Jabbar, A.; Li, X.; Omar, B. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–49.
22. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306.
23. Dutta, S.; Basarab, A.; Georgeot, B.; Kouamé, D. DIVA: Deep unfolded network from quantum interactive patches for image restoration. *Pattern Recognit.* **2024**, *155*, 110676.
24. Zhao, L.; Ji, B.Y.; Xing, W.; Lin, H.Z.; Lin, Z.J. Ancient painting inpainting algorithm based on multi-channel encoder and dual attention. *J. Comput. Res. Dev.* **2023**, *50*, 334–341.
25. Zhou, Z.H.; Liu, X.R.; Shang, J.Y.; Huang, J.C.; Li, Z.H.; Jia, H.P. Inpainting digital Dunhuang murals with structure-guided deep network. *ACM J. Comput. Cult. Herit.* **2022**, *15*, 77.
26. Deng, X.C.; Yu, Y. Ancient mural inpainting via structure information guided two-branch model. *Herit. Sci.* **2023**, *11*, 131.
27. Baasch, G.; Rousseau, G.; Evins, R. A Conditional Generative adversarial Network for energy use in multiple buildings using scarce data. *Energy AI* **2021**, *5*, 100087.
28. Zamir, S.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.; Yang, M.; Shao, L. Learning enriched features for fast image inpainting and enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1934–1948.
29. Yan, L.; Zhao, M.; Liu, S.; Shi, S.; Chen, J. Cascaded transformer U-net for image inpainting. *Signal Process.* **2023**, *206*, 108902.
30. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
31. Salah, E.; Amine, K.; Redouane, K.; Fares, K. A Fourier transform based audio watermarking algorithm. *Appl. Acoust.* **2021**, *172*, 107652.
32. Wang, Q.; Ma, Y.; Zhao, K.; Tian, Y. A comprehensive survey of loss functions in machine learning. *Ann. Data Sci.* **2020**, *9*, 187–212.
33. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
34. Brunet, D.; Vrscay, E.; Wang, Z. On the mathematical properties of the structural similarity index. *IEEE Trans. Image Process.* **2011**, *21*, 1488–1499.
35. Talebi, H.; Milanfar, P. Learned perceptual image enhancement. In Proceedings of the 2018 IEEE International Conference on Computational Photography (ICCP), Pittsburgh, PA, USA, 4–6 May 2018; pp. 1–13.
36. Zheng, C.X.; Cham, T.J.; Cai, J.F. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 1438–1447.
37. Li, J.Y.; Wang, N.; Zhang, L.F.; Du, B.; Tao, D.C. 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.
38. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
39. Jain, J.; Zhou, Y.; Yu, N. Keys to better image inpainting: Structure and texture go hand in hand. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 208–217.