Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053232

ARTICLE

Research on Restoration of Murals Based on Diffusion Model and Transformer

Yaoyao Wang[1], Mansheng Xiao[1,*], Yuqing Hu[2], Jin Yan[1] and Zeyu Zhu[1]

[1]School of Computing, Hunan University of Technology, Zhuzhou, 412000, China

[2]School of Computing, Hunan Software Vocational and Technical University, Xiangtan, 411100, China

*Corresponding Author: Mansheng Xiao. Email: xiaomansheng@hut.edu.cn

Received: 28 April 2024 Accepted: 03 August 2024 Published: 12 September 2024

ABSTRACT

Due to the limitations of a priori knowledge and convolution operation, the existing image restoration techniques cannot be directly applied to the cultural relics mural restoration, in order to more accurately restore the original appearance of the cultural relics mural images, an image restoration based on the denoising diffusion probability model (Denoising Diffusion Probability Model (DDPM)) and the Transformer method. The process involves two steps: in the first step, the damaged mural image is firstly utilized as the condition to generate the noise image, using the time, condition and noise image patch as the inputs to the noise prediction network, capturing the global dependencies in the input sequence through the multi-attention mechanism of the input sequence and feed-forward neural network processing, and designing a long skip connection between the shallow and deep layers in the Transformer blocks between the shallow and deep layers using long skip connections to fuse the feature information of global and local outputs to maintain the overall consistency of the restoration results; In the second step, taking the noisy image as a condition to direct the diffusion model to back sample to generate the restored image. Experiment results show that the PSNR and SSIM of the proposed method are improved by 2% to 9% and 1% to 3.3%, respectively, which are compared to the comparison methods. This study proposed synthesizes the advantages of the diffusion model and deep learning model to make the mural restoration results more accurate.

KEYWORDS

Transformer; deep learning; noise estimation network; diffusion model; mural restoration

1 Introduction

As a cultural heritage of a country or region, ancient frescoes carry a wealth of social, religious, and historical information. However, due to the long-term influence of environmental factors, mural images are usually faded or even mutilated. With the rapid development of artificial intelligence, intelligent restoration technology has attracted extensive attention from researchers [1,2].

In the field of image restoration, traditional deep learning techniques like Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) [3] have achieved significant results. CNNs, with their hierarchical feature extraction and reconstruction capabilities, can effectively restore local structures and textures of images. GANs, through adversarial training between the generator and

the discriminator, can produce repair results that are highly like the original images. However, these traditional methods still have limitations when dealing with complex image structures and textures. For example, CNNs primarily focus on local information when processing images, making them relatively weak in capturing global dependencies. This can lead to difficulties in maintaining consistency between the restored areas and the surrounding regions in images with complex structures and textures.

To solve these problems, the Transformer [4] structure emerges due to its powerful global information-capturing ability and self-attention mechanism. The Transformer overcomes the limitations of CNNs in terms of restricted receptive fields and uses attention mechanisms to achieve dynamic interaction and computation of features from different regions of the image. This improves the quality and efficiency of image restoration. Notably, the ViT (Vision Transformer) model proposed by Dosovitskiy et al. [5] applies the Transformer to images. Through global attention mechanisms, it captures global information in images, providing the model with a larger receptive field and greater flexibility in the restoration process.

In addition, the diffusion model, as an emerging generative model, also shows great potential in the field of image restoration. Diffusion models transfer pixel information from neighboring regions by designing diffusion functions to fill in missing pixels and restore image integrity. Compared with GAN, diffusion models have better generalization ability and stability, and they can generate high-quality restoration results [6]. In particular, the Denoising Diffusion Probabilistic Model (DDPM) [7] generates samples through an iterative denoising process [8], making the generated images more coherent and realistic in both detail and global structure [9].

Currently, research combining diffusion models [10] and Transformers [11] is relatively limited and still faces issues such as insufficient detail, limitations in capturing global structure, and lengthy training processes. In this study, we propose a Transformer and DDPM-based picture restoration model. This model combines the global information-capturing ability of the Transformer with the high-quality generation capability of DDPM. It aims to address key issues in the restoration of ancient mural images, such as the recovery of missing textures and the improvement of incomplete data coverage. By introducing ViT as the core structure and incorporating DDPM's denoising diffusion process, our model can adaptively adjust the scale of the attention mechanism during the restoration process. This enables better handling of image restoration tasks of varying sizes and complexities. Next, we describe in detail the structure, working principle, and experimental validation results of the model.

In conclusion, the main contribution of this work is four-fold:

1. We propose a method of cultural relics image restoration based on the diffusion model and ViT for the problems of missing structure and texture and incomplete data coverage of cultural relics murals due to improper preservation.
2. We design forward diffusion and backward sampling to guide the restoration, where in forward diffusion the information is better extracted through VIT to capture the image to generate a clear textured restoration image to restore the original appearance of the heritage mural image in a more reasonable way. Additionally, we utilize long skip connections between the shallow and deep layers of ViT, enabling the model to use low-level features more effectively for pixel prediction training.

3. Before the output, we add a 3×3 convolution block to prevent artifacts that might appear in images generated by the Transformer. By adjusting different forms of model parameters, we address issues of image quality degradation and excessive processing time that deep learning models might encounter in image restoration.
4. We optimized the loss function and dynamically adjusted the focus on different image regions to better measure the difference between the restoration results and the original murals. Extensive experiments validate that our proposed framework significantly outperforms the existing state-of-the-art methods.

2 Related Work

Among the image restoration based on diffusion modeling, the U-Net network architecture with CNN as the core is often used for the noise estimation of inverse generated images. In the forward process, the diffusion model aims to convert the original image to a full Gaussian noise image, however, this approach suffers from the problem of multiple sampling steps and long sampling time during the inference process, which leads to the high time cost of the inference process and restricts the scope and effectiveness of its application [12]. Resolving how to converge to a specific prior distribution in the expected time [13] as well as incorporating adaptive mechanisms are key issues that need to be addressed nowadays. In noise prediction by diffusion modeling, it is usually necessary to model complex data including spatial and temporal variations as well as possible sources of noise, and traditional methods may be limited by the fact that feature extraction does not allow for good restoration of mural images, whereas the use of ViT is able to better capture global information and local structure in the image. We propose an improved denoising diffusion probabilistic model that combines the forward diffusion and backward generation processes, denoising by iteration, and generating samples using a standard Gaussian distribution so that they gradually evolve into samples that conform to the empirical distribution. In the improved denoising diffusion probabilistic model, the input data in the forward diffusion phase corrupts the original data by gradually adding Gaussian noise, and the added noise level is dynamically estimated by means of a Transformer-based noise estimation network. While in the reverse diffusion stage, the task of the generative model is to learn the reverse diffusion process to recover the original input data from the noisy data, through introducing the idea of reverse denoising and combining the noise estimation network, it can capture the potential distribution of the original data more efficiently, so as to improve the performance of the generative model and the quality of the generated samples. This integration not only improves the quality of restoration but also accelerates the restoration process, making our method more efficient and practical for handling large-scale mural restoration tasks.

2.1 Forward Diffusion

The forward diffusion process is defined as a Markov chain, where Gaussian noise is continuously added to successive nodes to obtain noisy samples, which in turn gradually transforms the Gaussian noise distribution into a distribution of data on which the generative model is trained. Specifically, as shown in Fig. 1, given a data sample $x_0 \sim x_N$, where x_0 is the original image and x_N is the image corresponding to the Nth moment of added noise.

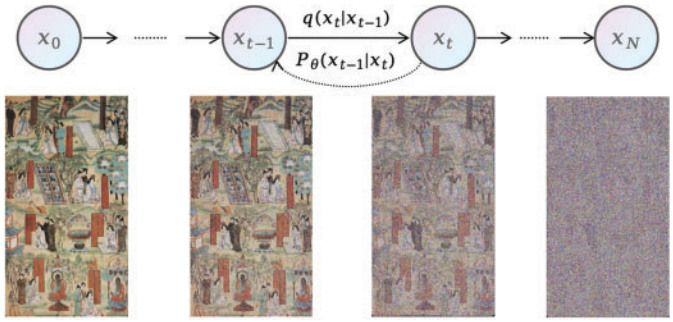

Figure 1: Image noise schematic map

The forward additive noise t-moment process is defined as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon, \alpha_t = 1 - \beta_t \tag{1}$$

where x_t is the added noise data at moment t, $t \in \{0, 1, 2, \cdots, T\}$, T is the number of times the noise is added, is the noise added at moment t, obeys Gaussian distribution, α_t is the initialized value, which is the empirical value, β_t increases linearly from 0.0001 to 0.002 during forward diffusion, ε is the noise, obeys Gaussian distribution. In the forward diffusion process, the later the moment, the more closely the noise data is related to the noise increased in the previous moment. According to the Markov chain, the state x_{t-1} is denoted as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon \tag{2}$$

$$x_t = \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon\right) + \sqrt{1-\alpha_t}\varepsilon \tag{3}$$

where $\varepsilon \sim N(0, I)$ can also be expressed as:

$$x_{t-1} = \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\varepsilon \tag{4}$$

further the relationship between x_t and x_0 is obtained as:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon. \tag{5}$$

where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

2.2 Reverse Generation

The forward diffusion process results in a data that nearly obeys a Gaussian distribution, and the inverse diffusion process recovers the original data from Gaussian noise, generating original images for learning a parameterized posterior distribution $p_\theta(x_0|x_t)$ through x_t. Assuming that the inverse process $q(x_{t-1}|x_t)$ is obtained, it is possible to gradually reduce an image through random noise x_t. The DDPM uses the neural network $p_\theta(x_0|x_t)$ to fit the inverse process $q(x_{t-1}|x_t)$, with the formula:

$$q(x_{t-1}|x_t, x_0) = N\left(x_{t-1}|\overline{\mu}(x_t, x_0), \tilde{\beta}_t I\right) \tag{6}$$

can be deduced:

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}|\mu_\theta(x_t, t), \sum_\theta (x_t, t)\right) \tag{7}$$

where $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \cdot \beta_t$, solve the equation by Bayesian formula:

$$\overline{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_{t-1}} x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} x_0 \tag{8}$$

α_t and $\overline{\alpha}_t$ depend only on β_t and it follows from the forward diffusion to express x_0 as:

$$x_0 = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon\right), \varepsilon \sim N(0, I) \tag{9}$$

by combining Eqs. (7) and (8), a mean value, that depends only on x_t, results:

$$\overline{\mu}_t(x_t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}\varepsilon}\right) \tag{10}$$

thus, a neural network $\varepsilon_\theta(x_t, t)$ can be used to approximate ε and obtain the average of the following:

$$\overline{\mu}_t(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t, t)\right) \tag{11}$$

where ε is the noise value predicted by the trained model. Actually, to predict the noise more accurately, a noise prediction network $\varepsilon_\theta(x_t, t)$ is used to learn $E[\varepsilon | x_t]$ by minimizing the objective function $minE_{t,x_0,\varepsilon} \|\varepsilon - \varepsilon_\theta(x_t, t)\|_2$, where t is uniformly distributed between 1 and T. To learn the conditional diffusion model, this paper further inputs the conditional information c into the noise prediction objective function:

$$minE_{t,x_0,c,\varepsilon} \|\varepsilon - \varepsilon_\theta(x_t, t, c)\|_2 \tag{12}$$

The image recovered after the noise value is obtained using the noise estimation training model in this paper, as shown in Fig. 2.

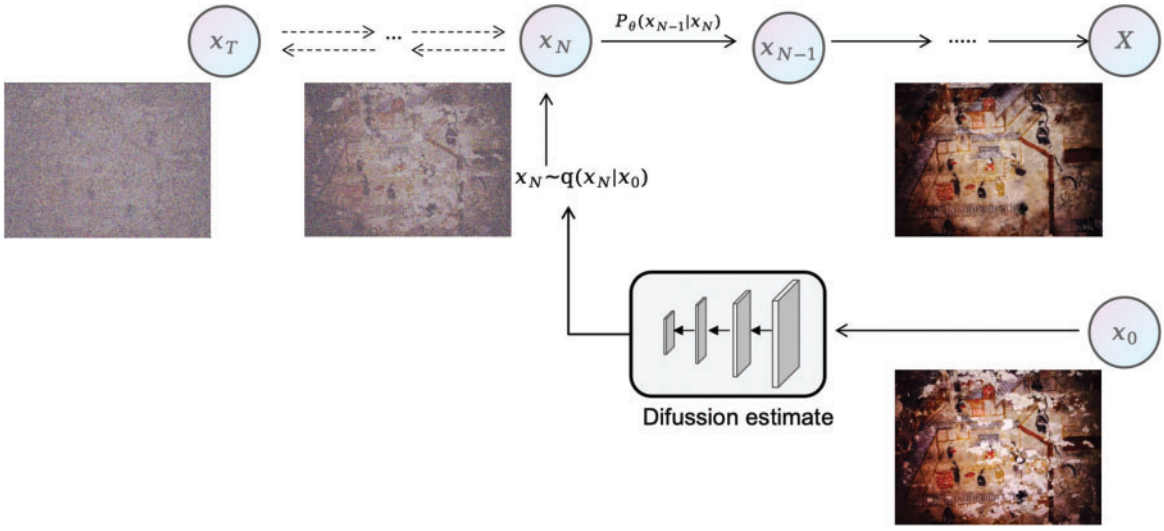

Figure 2: Image conditional diffusion model map

3 Noise Estimation Based on Transformer

ViT can better capture the global information and local structures in images. Additionally, ViT offers better generalization capabilities because instead of relying on fixed-size filters like traditional convolutional neural networks, it processes input sequences through a self-attentive mechanism, which makes it more flexible with respect to the length and size of the inputs. For this purpose, a simple and generalized ViT-based architecture noise estimation network is improved in this paper, as shown in Fig. 3. The model follows the transformer's design approach, taking all inputs, including temporal, conditional, as well as noisy picture patches as markers, using long jump branch between shallow and deep layers. The architecture allows for more efficient training of pixel prediction targets using low level features. In addition, to prevent possible artifacts in the image generated by the Transformer [14], we added a 3×3 convolution block before the output. From experiments, the visual quality of the model-repaired images was improved.

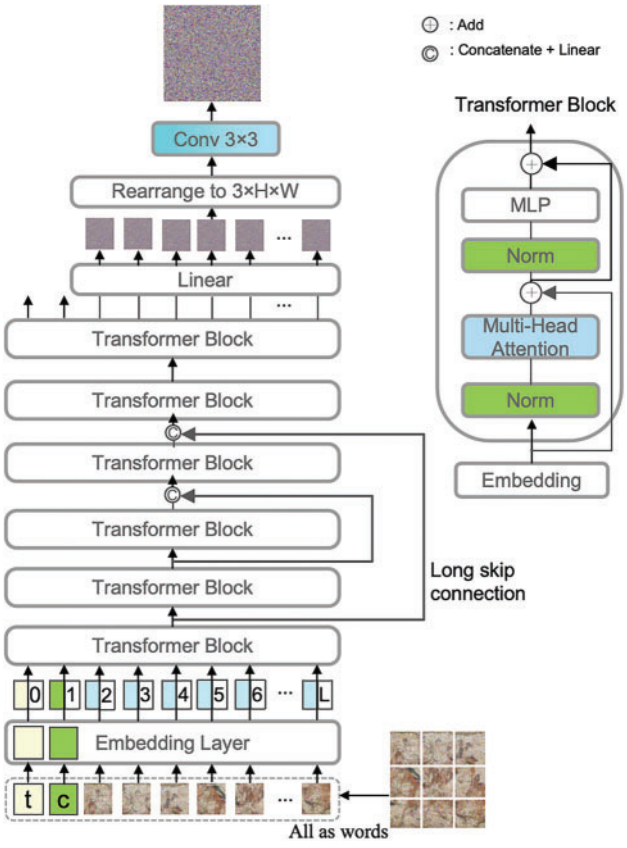

Figure 3: Noise estimation network structure map

The model is a simple generalized backbone network for image generation diffusion models is shown in Fig. 3. Given time t, condition c (discrete text converted to embedded sequences via CLIP encoder, aligned with stabilized diffusion, and input as a tagged sequence), alongside the noisy images x_t as input, and estimates the noise infused within x_t. Following ViT's design, images are segmented into homogeneous blocks, transformed into a sequence, and combined with time, condition, and blocks as inputs. The input sequence is then processed with a multi-head attention mechanism and a feed-forward neural network to enable the model to capture global dependencies in the input sequence and

use stacking of multi-layer Transformer blocks to incrementally improve the feature representation at higher levels. To enhance the information delivery and reduce the information loss, we apply long jump branch between Transformer blocks by learning the ideas in U-Net of CNN. This branch allows shallow information to be passed directly to deeper layers, effectively preserving the low-level feature information of the image, and providing efficient paths for low-level features, thus simplifying the training process of the noise prediction network. The output of Transformer blocks is mapped to a spatial representation of the noisy image and features are further processed through a 3×3 convolutional layer to improve the model's ability to capture image details.

In Section 3.1, we present a specific instantiation of our model, detailing its core components and architecture. Subsequently, in Section 3.2, we delve into evaluating the model's scalability potential, focusing on the influence of architectural dimensions such as depth, width, and patch size on its performance.

3.1 Concrete Realization

To make the model more effective, this paper conducts a systematic empirical study on its key elements and conducts ablation experiments on this paper's dataset, evaluating the FID scores of 1 K generated samples every 5 K training iterations, and selecting the optimal effect through the experiments.

We investigate various strategies for integrating the long skip branch within our Transformer architecture. Specifically, we consider the main branch embedding h_m and the long skip branch embedding and the long skip branch embedding h_s, both of which reside in $R^{H\times W}$. Prior to passing these embeddings to the subsequent Transformer block, we explore five fusion approaches: (1) concatenating h_m and h_s followed by a linear projection (Linear (Concat(h_m, h_s))), (2) direct element-wise summation ($h_m + h_s$), (3) applying a linear projection to h_s prior to summation (h_m + Linear(h_s)), (4) summing the embeddings and then applying a linear projection (Linear($h_m + h_s$)), and (5) a baseline scenario without the long skip branch for comparative analysis. Notably, the direct summation of h_m and h_s (i.e., $h_m + h_s$) solely modulates the contribution of h_s in a linear fashion, leaving the fundamental network architecture unaltered. In contrast, all alternative fusion strategies involving the long skip branch demonstrate enhanced performance compared to the absence of such a connection. As shown in Fig. 4a, the approach that concatenates h_m and h_s followed by a linear projection emerges as the most effective, suggesting that this method is particularly adept at leveraging the complementary information from both branches.

In investigating the enhancement of our model, we evaluated two methodologies for integrating an additional convolutional layer following the Transformer block: (1) The first method involved appending a 3×3 convolutional block following the linear projection step, which converts token embeddings into image patches, as illustrated in Fig. 4b. (2) Before linear projection, adding a 3×3 convolutional layer, and the one-dimensional sequence of label embeddings needs to be rearranged into two-dimensional feature map of dimensions H/P $\times$ W/P $\times$ D, with P represents the size of the patches. (3) Additionally, comparisons were made for the situation in which no additional convolutional layers were added. According to the results in Fig. 4b, the method of adding 3×3 convolutional layers following the linear transformation exhibits marginally superior performance to the remaining two alternatives.

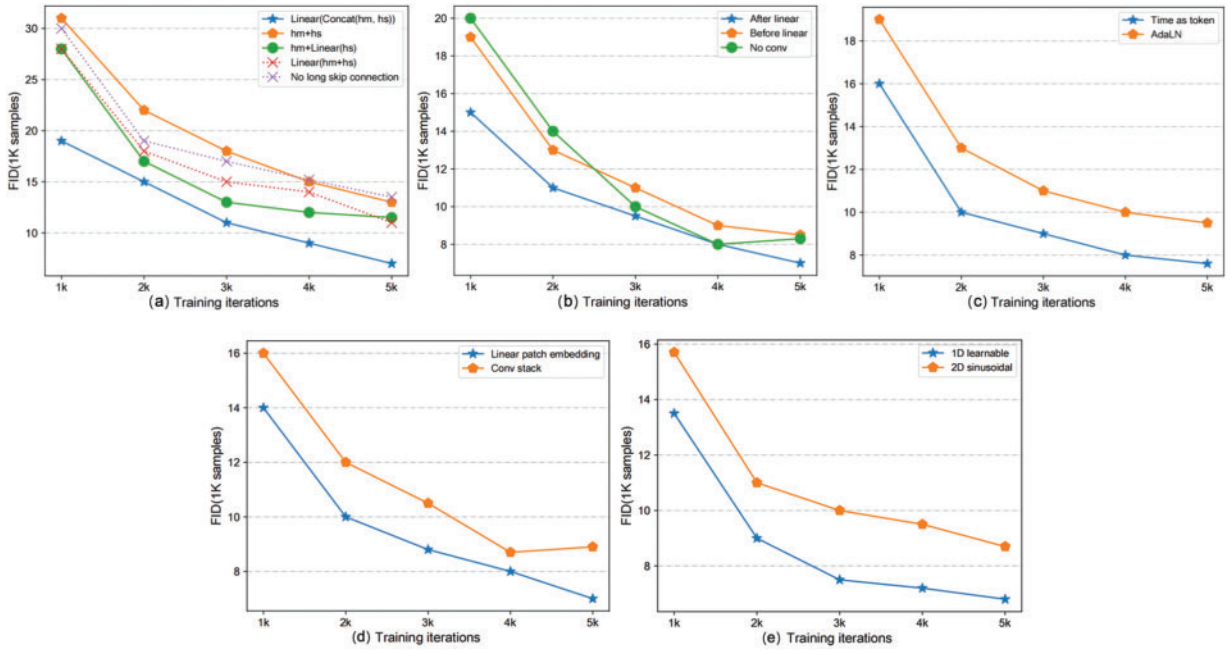

Figure 4: Alate design choices map, (a) Combine long skip branch (b) Add an extra convolutional block (c) Feed time into the network (d) Different forms of patch embedding (e) Different forms of position embedding

To incorporate temporal information into the network, we evaluate two different methods for inputting the time variable t: (1) Consider it as a marker, as shown in Fig. 4c. (2) Merge the layer-normalized times into the Transformer block [15], analogous to the adaptive group normalization employed within U-Net [16]. Another method uses adaptive layer normalization (AdaLN). Use the normalization operation: $AdaLN\,(h, y) = y_s LayerNorm\,(h) + y_b$, within the Transformer block, we embed the input as h, and subsequently derive y_s and y_b through a linear projection of the temporal embedding. Despite the relative simplicity of the AdaLN approach, the first approach, which treats time as a marker, outperforms AdaLN as shown in Fig. 4c.

We delve into the nuances of patch embedding by examining two distinct variants. (1) The original approach employs a linear projection to transform each patch into a labeled embedding, as illustrated in Fig. 4d. (2) We employed a sequence of 3×3 convolutional blocks, followed by a 1×1 convolutional block, to map images into token embeddings. In contrast, the results indicate that the conventional patch embedding method outperforms this approach.

We delve into the realm of positional embedding variants, exploring two distinct methodologies tailored for our image restoration framework: (1) Firstly, we adopt the ubiquitous one-dimensional learnable positional embedding, which is the default setting used in this paper and proposed in the original ViT, (2) The second variant utilizes a 2-dimensional sinusoidal positional embedding, constructed by concatenating the sinusoidal embeddings of coordinates i and j for each patch located at (i, j). According to the results in Fig. 4e, the former performs better than the latter, and it is found that the model is unable to generate meaningful images after trying it without any positional embedding, which proves that positional information is crucial in image generation.

3.2 Influence of Depth, Width, and Patch Size

We demonstrate the scalability of our proposed model through a meticulous investigation encompassing its depth, quantified by the number of layers, its width, defined by the hidden layer size D, and the granularity of its input, characterized by the patch size. As shown in Fig. 5, the best performance was achieved at a depth of 14 in the 5 K iterations of the experiment, which shows that the depth is not positively correlated with the performance of the model, i.e., the model does not benefit from a greater depth. Analogously, augmenting the width dimension, specifically the hidden size, from 256 to 512 yields a noticeable performance enhancement. However, further escalation to 768 does not yield any discernible gains, indicating a saturation point in the model's capacity to leverage additional width for improved performance. On the other hand, a smaller patch size consistently improves the performance. In contrast, high-level tasks (e.g., classification) may require larger patches. In practice, due to the high dimensionality of the image data, there may be an increase in cost when using smaller patch values for training, and therefore it is recommended to downscale the image data before using the model.

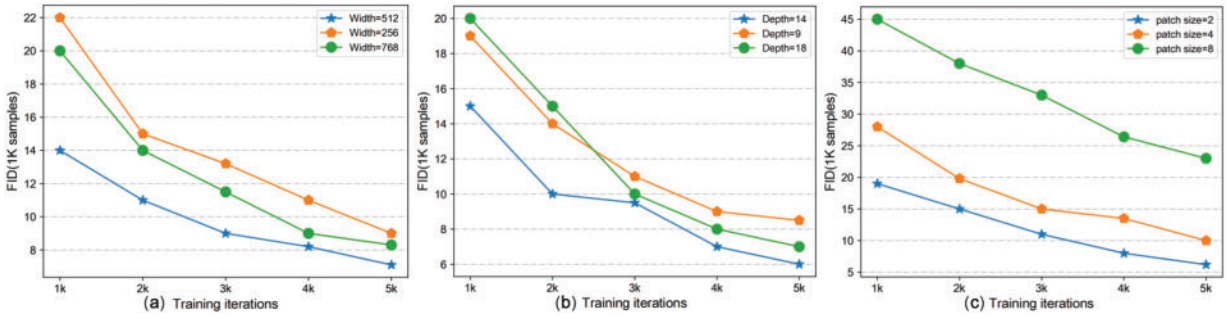

Figure 5: Influence diagram of different factors, (a) Width (b) Depth (c) Patch size

3.3 Optimizing Denoising Losses and Estimating Distribution Functions

To better accomplish the repair task based on the diffusion model, in this paper, we refine our diffusion model's learning process by incorporating two distinct objective functions. The primary objective function implements a straightforward denoising loss, which is computed given a reference output image x and a randomly selected time step t, the reference image with a noisy version is generated as follows:

$$x' = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\varepsilon \tag{13}$$

take T to be 500. We train our conditional diffusion model to faithfully reconstruct the reference image x under the influence of the conditional feature c and the time step t as follows:

$$\mathcal{L}_{simple} = E_{t,x_0,c,\varepsilon}[\|_{\varepsilon-\varepsilon_\theta(x_t,t,c)}\|_2] \tag{14}$$

based on the enhanced denoising diffusion model, we further train the network to predict the variance $\sum_\theta(x', c, t)$, which not only improves the fidelity of the reconstructed image, but also helps to improve its log-likelihood. The conditional diffusion model additionally outputs the interpolated coefficients s for each dimension and converts the output to variance as follows:

$$\sum_\theta(x', c, t) = exp\left(v\log\beta_t + (1-s)\log\tilde{\beta}_t\right) \tag{15}$$

where β_t and $\tilde{\beta}$ denote the upper and lower limits of the variance, respectively. The second objective function directly optimizes the KL scatter between the estimated distribution $p_\theta\left(x_{t-1}|x_t, x_0\right)$ and the diffusion posterior $q\left(x_{t-1}|x_t, x_0\right)$ with the following formula:

$$\mathcal{L}_{vlb} = KL\left(p_\theta\left(x_{t-1}|x_t, x_0\right) \| q\left(x_{t-1}|x_t, x_0\right)\right) \tag{16}$$

The total loss function is the weighted sum of the two objective functions, formulated as follows:

$$\mathcal{L} = \eta\mathcal{L}_{simple} + \lambda\mathcal{L}_{vlb} \tag{17}$$

where η and λ are the weight parameters of the balanced loss function, the improved loss function improves the performance of the network model, accelerates the convergence speed of the algorithm, which improves the efficiency of training when $\eta = 0.4$, $\lambda = 0.6$ is adjusted through experiments.

4 Experiments and Analysis

To verify the restoration effect of the restoration method proposed in this paper on ancient mural images, and to compare and analyze it with existing restoration methods we conducted experiments on the dataset of this paper, the specific experimental process is as follows.

4.1 Data Sets and Experimental Environments

Due to the scarcity of cultural relics mural data set, this paper selects the training data set from the official website of Dunhuang Research Institute and the official website of Shanxi Museum to provide 4000 images of cultural relics mural with different resolution synthesized into a training dataset, through data augmentation to ultimately obtain 10,000 cultural relics images with varying resolution. We firstly manually screened 4000 images with different resolutions, eliminated images with too much single color and too much irrelevant content, and then augmented the images to generate a large dataset of 10,000 images of cultural relics.

4.2 Evaluation Indicators

We used 2 types of subjective and objective evaluations to validate the method. Subjective evaluation is done by observing the texture and color information of the generated image, objective methods are evaluated by peak signal-to-noise ratio and structural similarity (SSIM), peak signal-to-noise ratio (PSNR) and Fréchet Inception Distance (FID) evaluating the strengths and weaknesses of each algorithm. PSNR mainly estimates the noise fidelity of the reconstructed image, the higher the value, the better the quality of the reconstructed image. SSIM combines three factors: brightness, contrast, and structure. The mean is used as an estimate of brightness, the standard deviation as an estimate of contrast, and the covariance as an estimate of structural similarity. The value range is between [0, 1]. The closer the result is to 1, the better the reconstructed image quality is. FID calculates the similarity between advanced features of the image, and the smaller the value, the higher the degree of similarity.

4.3 Experimental Results Analysis

For objective comparison of restoration results of image restoration methods, the comparison methods in this paper use the same input data. The following experiment is the method of this paper with the hierarchical Transformer-based image restoration method [17]; Based on generative adversarial networks to generate high quality restored images by matching and correlating background

patches method Contextual Attention [18]; Shift-Net, a deep learning method that combines a priori information [12]; Global Uniform and Local Continuity (GU&LC) combining global uniformity and local continuity based on the relationship implied between linear systems and image restoration [19] and a comparison of the probabilistic diverse GAN method PD-GAN [20] for image repair on irregularly corrupted images of this dataset.

4.3.1 Repair of Scratches and Damages of Different Sizes

This section compares the restoration results of each method on murals with different scratches that conform to realistic scenarios. Combining the characteristics of irregular area and discontinuous damaged area of cultural relics, Fig. 6 shows the restoration results of each method.

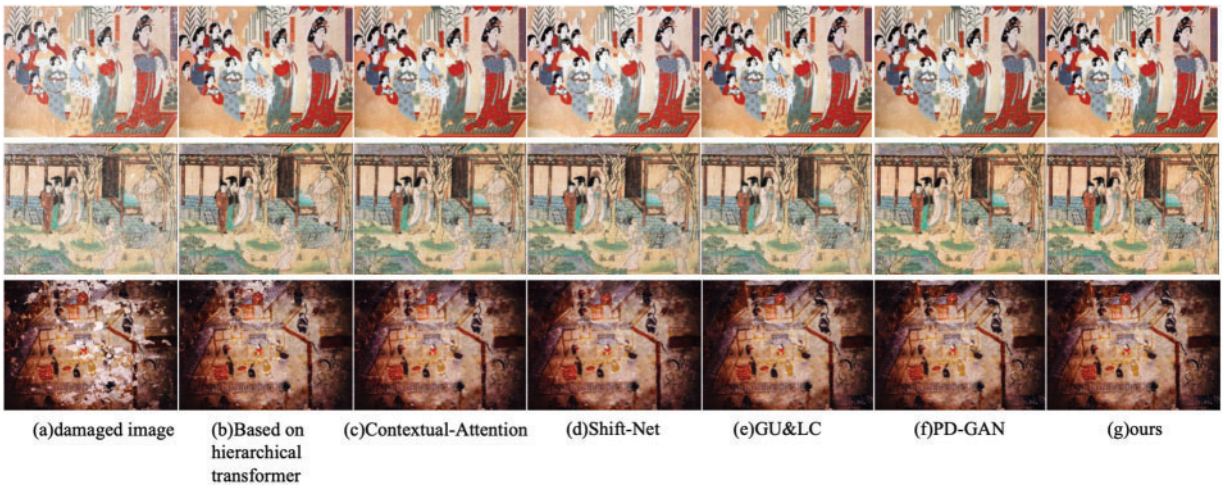

Figure 6: Comparison of repair results of different algorithms for real damaged scenes

We compare the image reconstruction effect when the methods are of the same magnification by subjective evaluation method, which mainly observes the texture information and color brightness of the generated image. As seen in Fig. 6b, the hierarchical Transformer-based method repairs scratches with larger area better, however, it still fails to repair smaller scratches completely. As shown in Fig. 6c, compared to the previous method, the Contextual Attention method scratch repair is more complete, but there are still details that are partially repaired that are not reasonable, probably because the model produces inaccurate results in predicting the structure of larger damaged areas. As can be seen in the third image, this is particularly evident when the known region does not provide enough a priori information. As seen in Fig. 6d, the Shift-Net method performs well overall, successfully repairing the basic scratches and broken parts, and the color remains largely unchanged significantly. however, the lack of contextual semantic information in the repair region rubs off the texture of the image, resulting in detail not being visible. Especially in the first image, the Shift-Net model has changed less based on the original color and still shows a dark and old feeling, while in the third image the basic scratches of the image are all restored, but the detailed part of the restoration does not look natural. As seen in Fig. 6e,f, the restoration quality of the GU&LC method and the PD-GAN method is relatively high, and the detail restoration of the GU&LC is still not as natural as that of this paper despite the elimination of artifacts. The PD-GAN method excels in the completeness and rationality of the restoration results in terms of context, while more detailed comparisons and improvements in detail and color are still needed. Further observation of Fig. 6g, this paper's method in the realization of

the texture becomes clear at the same time, it also highlights the vivid colors of the image, so that the original dim image of cultural relics regained its former glory. The method in this paper is better than other methods for image restoration of this dataset, and the restoration results are semantically coherent, with fewer artifacts and duplicate textures, and better metrics and visual effects are achieved.

The quantitative evaluation of repair indexes of each method is shown in Table 1, the method of this paper is optimal in PSNR, SSIM performance, compared with other methods, the PSNR indexes were improved respectively by 9.01%, 5.42%, 2.11%, 3.32% and 1.51%. SSIM indicator has improved by 1.87%, 3.31%, 2.67%, 3.48% and 1.89%. The above results show that the method of this paper has outstanding structural recovery ability for scratched and damaged cultural relics, and the recovery of texture and color is also more reasonable, which has a very good restoration effect.

Table 1: This method and other methods for scratch damage repair results

Method	PSRN/dB	SSIM
Based on hierarchical Transformer [17]	28.6570	0.8756
Contextual-Attention [18]	29.6303	0.8634
Shift-Net [12]	30.5930	0.8688
GU&LC [19]	30.2321	0.8620
PD-GAN [20]	30.7720	0.8754
Proposed	**31.2376**	**0.8920**

4.3.2 Restoration of Natural Weathering Dislodgement

The experiments in this section focus on two types of restoration effects: large weathering and shedding repair and small weathering and shedding repair. Observe whether the restored image outlines the object of the painting clearly, whether the color contrast is sharp, and whether its texture is relatively reasonable. Therefore, the experiments in this section are trained using the images of figure murals in the dataset with large weathering and shedding areas of 30% to 40% and small weathering and shedding areas of 5% to 10%. This was done to ensure consistency and reproducibility of the experiments and to focus on specific types of artifact images. The experimental results are shown in Figs. 7 and 8. From Fig. 7, the method of this paper grasps the global information of the large weathered and detached parts very well, and the texture is clear, but a small part of it will be slightly distorted. As shown in Fig. 8, in the small area of detachment area, this paper's method also well restored the missing part of the image, perfected the integrity of the mural, improved the color contrast compared to a larger area of broken, no distortion, the restoration results of the structure of the coherent and in line with the context of the semantic information. From the data in Table 2, the PSRN index of large weathering shedding loss compared to other methods. It respectively increased to 2.5523, 2.3094, 0.3762, 0.9981, and 0.5502 dB. The SSIM metrics improved by 0.0290, 0.0490, 0.0199, 0.0287, and 0.0164, respectively. The PSRN metrics for small-area weathering and shedding loss increased by 1.9977, 2.6687, 0.6686, 1.3537, and 0.9058 dB, and the SSIM metrics improved by 0.0307, 0.0537, 0.0246, 0.0334, and 0.0211. We can see that small areas are better restored than large areas.

Figure 7: Large area off repair effect diagram

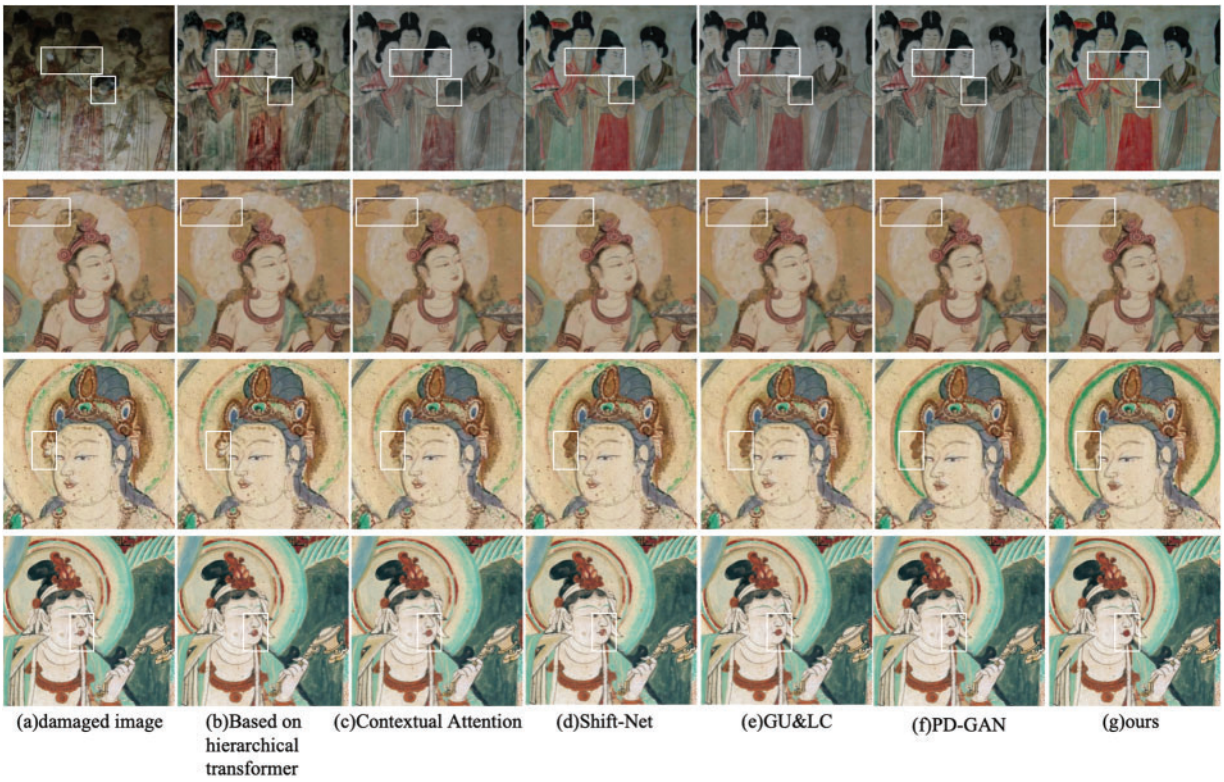

Figure 8: Small area off repair effect diagram

Table 2: The results of this method and other methods are different in size of needle damage repair

Method	Large area off		Small area off	
	PSRN/dB	SSIM	PSRN/dB	SSIM
Based on hierarchical Transformer [17]	28.3241	0.8630	29.2343	0.8660
Contextual-Attention [18]	28.5670	0.8430	28.5633	0.8430
Shift-Net [12]	30.5002	0.8721	30.5634	0.8721
GU&LC [19]	29.8783	0.8633	29.8783	0.8633
PD-GAN [20]	30.3262	0.8756	30.3262	0.8756
Proposed	**30.8764**	**0.8920**	**31.2320**	**0.8967**

4.3.3 Ablation Experiment

To further validate the effectiveness of our proposed image restoration method for artifacts and the contribution of each component, we conducted a series of ablation experiments. This experiment aims to analyze the effect of different modules and steps on the restoration results, first, we evaluate the effect of using only the denoising diffusion model for the image restoration of artifact murals. In this experiment, long skip connections were removed and additional added 3×3 convolutional machine layers were removed to determine the repair ability of the denoising diffusion model itself and to analyze the effect of having or not having long skip branch and adding additional convolutional layers on the repair effect. To verify its ability to improve the accuracy of noise estimation. The experimental results are shown in Table 3 for the performance of different modules on this dataset.

Table 3: The performance of different modules

	PSRN/dB	SSIM	FID
Base model	28.302	0.834	7.32
Removal of LSC	28.653	0.854	6.78
Remove of extra coiler layer	29.042	0.876	5.95
Full-scale model	**30.030**	**0.908**	**5.48**

The experimental results are shown in Table 3, the addition of long skip connections in the noise estimation network is 0.74 dB better than no long skip connections in PSRN and 0.042 better in SSIM. Adding the extra coiler layer is going to improve over the baseline model by 0.351 dB on PSRN and 0.02 on SSIM. Both add improvements on PSRN of 1.728 dB and 0.074 on SSIM. FID was reduced by 1.84. As shown in Fig. 9, After removing the LSC module, it is evident from the comparative images that while the repaired result image exhibits clear color contrast and higher saturation, there are still structural deficiencies present. Compared to LSC, the impact of the additional convolutional layer is relatively minor. From the comparative images, it is apparent from the restoration of the cultural relic mural that although the removal of this module results in a coherent image with well-done detail restoration, the color contrast is not as distinct. The repair result of adding 2 modules at the same time is optimal in all 3 metrics, and the repair result is optimal.

Figure 9: Visual effects of different modules in the restoration of cultural relics murals

5 Conclusions

This paper proposes and achieves a new method for restoring images of cultural relics, which aims at restoring the original appearance of mural images of cultural relics in a more rational way through fine steps and techniques. The noise is mainly estimated by the noise prediction network in the forward diffusion model. The improved Transformer module proposed can process the image information efficiently, due to the long skip connection that can reduce the problem of information loss brought about by the process of multiple up-sampling and down-sampling, which enables the efficient Transformer module to improve the effect of intelligent restoration of the broken image while maintaining the global attention. The experimental results show that our method achieves excellent results in both breakage repair experiments and large area breakage repair experiments, which is not only validated in subjective assessment but also performs well in objective assessment.

Acknowledgement: The authors would like to express their heartfelt gratitude to the editors and reviewers for their detailed review and insightful advice.

Funding Statement: This financial support from Hunan Provincial Natural Science and Technology Fund Project (Grant No. 2022JJ50077), Natural Science Foundation of Hunan Province (Grant No. 2024JJ8055).

Author Contributions: The authors contribute to this paper in the following capacities: Conception and design of the study: Mansheng Xiao, Yuqing Hu; data collection: Yaoyao Wang; analysis and

interpretation of results: Yaoyao Wang, Jin Yan, Zeyu Zhu; preparation of draft manuscript: Yaoyao Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data and materials are available upon request from authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Tang, G. H. Geng, and M. Q. Zhou, "Application of digital processing in relic image restoration design," *Sens. Imaging*, vol. 21, no. 1, 2020, Art. no. 6. doi: 10.1007/s11220-019-0265-8.

[2] W. N. Xu and Y. L. Fu, "Deep learning algorithm in ancient relics image colour restoration technology," *Multimed. Tools. Appl.*, vol. 82, no. 15, pp. 23119–23150, 2023. doi: 10.1007/s11042-022-14108-z.

[3] I. Goodfellow *et al.*, "Generative adversarial nets," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020. doi: 10.1145/3422622.

[4] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 4–9, 2017, pp. 5998–6008. doi: 10.48550/arXiv.1706.03762.

[5] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, Millennium Hall, Addis Ababa, Ethiopia, Apr. 26–30, 2020, pp. 1–21. doi: 10.48550/arXiv.2010.11929.

[6] J. Y. Peng, Z. Yu, S. Y. Qu, Q. Y. Hu, and J. Wang, "A research review on deep learning based image restoration methods," (in Chinese), *J. Northwest. Univ. (Nat. Sci. Ed.)*, vol. 53, no. 6, pp. 943–963, 2023. doi: 10.16152/j.cnki.xdxbzr.2023-06-006.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 6–12, 2020, pp. 6840–6851. doi: 10.48550/arXiv.2006.11239.

[8] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte and L. V. Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 11461–11471. doi: 10.1109/CVPR52688.2022.01117.

[9] Z. H. Yan, C. B. Zhou, and X. C. Li, "Review of generating diffusion models," (in Chinese), *Comput. Sci.*, vol. 51, no. 1, pp. 273–283, 2024. doi: 10.11896/jsjkx.230300057.

[10] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "UNIT-DDPM: Unpaired image translation with denoising diffusion probabilistic models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Nashville, TN, USA, Jun. 20–25, 2021, pp. 14125–14134. doi: 10.48550/arXiv.2104.05358.

[11] F. Bao *et al.*, "All are worth words: A ViT backbone for diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Vancouver, BC, Canada, Jun. 18–22, 2023, pp. 22669–22679. doi: 10.1109/CVPR52729.2023.02171.

[12] Z. Y. Yan, X. M. Li, M. Li, W. M. Zuo, and S. G. Shan, "Shift-Net: Image inpainting via deep feature rearrangement," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 8–14, 2018, pp. 3–19. doi: 10.1007/978-3-030-01264-9_1.

[13] V. D. Bortoli, J. Thornton, J. Heng, and A. Doucet, "Diffusion schrödinger bridge with applications to score-based generative modeling," in *Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 6–14, 2021, pp. 17695–17709. doi: 10.48550/arXiv.2106.01357.

[14] B. W. Zhang *et al.*, "StyleSwin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 11304–11314. doi: 10.1109/cvpr52688.2022.01102.

[15] S. Y. Gu *et al.*, "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 10696–10706. doi: 10.1109/CVPR52688.2022.01043.

[16] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 6–12, 2021, pp. 8780–8794. doi: 10.48550/arXiv.2105.05233.

[17] Y. T. Kang, "Image inpainting model based on gated deformable convolution and hierarchical transformer and its application," Donghua Univ., Shanghai, China, 2023. doi: 10.27012/d.cnki.gdhuu.2022.001329.

[18] J. H. Yu, Z. Lin, J. M. Yang, X. H. Shen, X. Lu and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 5505–5514. doi: 10.48550/arXiv.1801.07892.

[19] H. Wang, L. Li, Q. Li, J. Y. Deng, and H. M. Shang, "A mural restoration method combining global consistency and local continuity," (in Chinese), *J. Hunan Univ. (Natural Sci. Ed.)*, vol. 49, no. 6, pp. 135–145, 2022. doi: 10.16339/j.cnki.hdxbzkb.2022292.

[20] H. Y. Liu, Z. U. Wan, W. Huang, Y. B. Song, X. T. Han and J. Liao, "PD-GAN: Probabilistic diverse GAN for image inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, TN, USA, Jun. 19–24, 2021, pp. 9367–9376. doi: 10.1109/CVPR46437.2021.00925.