Pattern Recognition 167 (2025) 111753

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

TRRS-DM: Two-stage Resampling and Residual Shifting for high-fidelity texture inpainting of Terracotta Warriors utilizing Diffusion Models

Xin Cao, Peiyuan Quan, Yuzhu Mao, Rui Cao *, Linzhi Su, Kang Li

School of Information Science and Technology, Northwest University, Xi'an, 710127, Shaanxi, China

ARTICLE INFO

Keywords:
Image inpainting
Diffusion model
Resample and ResShift
Terracotta Warriors
Heritage preservation

ABSTRACT

As a UNESCO World Heritage Site, the Terracotta Warriors face degradation from natural erosion. Traditional restoration is time-consuming, while computer-aided methods provide efficient digital solutions. We propose a **T**wo-stage **R**esampling and **R**esidual **S**hifting framework using **D**iffusion **M**odels (**TRRS-DM**) for texture inpainting. The ResampleDiff module enhances details via perception-weighted learning and lightweight diffusion. The RefineDiff module refines results in latent space by removing noise. Experiments demonstrate that TRRS-DM achieves faster computation, surpasses existing methods in visual quality, and effectively restores damaged artifacts. This approach advances digital heritage restoration and providing scalable supports for archaeological conservation. Our code is available at https://github.com/Emwew/TRRS-DM.

1. Introduction

The Qin Shi Huang Mausoleum and Terracotta Warriors, designated as a UNESCO World Heritage Site in 1987 and hailed as the "Eighth Wonder of the World", showcase the Qin Dynasty's military and cultural traits [1]. Environmental factors have resulted in significant damage and loss, thereby complicating archaeological research and exhibitions. Conventional restoration techniques can be labor-intensive and may pose risks to the artifacts, while computer-assisted restoration presents accurate and non-invasive alternatives. In the restoration of the Terracotta Warriors, two-dimensional restoration technology was employed to reconstruct missing or damaged components. This technology improves the resemblance of the affected areas by analyzing their characteristics, thus facilitating physical repair efforts. Furthermore, digital restoration not only supports the physical preservation of artifacts but also enhances public engagement through virtual experiences, thereby playing a crucial role in raising awareness and safeguarding cultural heritage [2].

Traditional image processing methodologies, including interpolation [3] and block matching [4], exhibit limited efficacy when addressing large or complex lesions, often resulting in unnatural outcomes and a loss of detail. The advent of deep learning [5] has ushered in significant advancements in image repair, particularly through the application of convolutional neural networks (CNN) [6,7] and generative adversarial networks (GAN) [8,9]. CNNs are particularly adept at capturing local features and facilitating precise restoration, while GANs enhance the naturalness of generated images through adversarial training mechanisms. Additionally, autoregressive models that utilize variational autoencoders and transformers can reconstruct various types of missing data on a pixel-by-pixel basis [10,11]. Despite the advantages offered by deep learning techniques in overcoming many of the limitations associated with traditional methods, they also introduce new challenges, including the instability of GAN training, substantial data requirements, and the demand for significant computational resources. Furthermore, in extreme or specialized scenarios, the quality of the repairs may still be constrained.

Recently, diffusion-based methods such as Repaint [12] and LatentPaint [13] have adopted a step-by-step noise removal approach to restore images, demonstrating excellence in repairing complex structures and textures. This brings new opportunities to the field of image inpainting. The rich prior knowledge inherent in diffusion models enables superior performance in detail restoration. However, these models come with significantly higher computational costs compared to GAN-based approaches [14]. Moreover, in the context of the digital restoration of cultural artifacts, challenges such as limited data availability, complex textures, and the necessity to adhere to historical and cultural contexts pose additional hurdles. As a result, the effectiveness of diffusion methods in this particular application remains to be validated.

In response to the challenges associated with elevated model training expenses and prolonged inference durations, we introduce a two-stage image inpainting technique grounded in diffusion models, referred to as TRRS-DM. This methodology is fundamentally composed

* Corresponding author.
E-mail address: cr@nwu.edu.cn (R. Cao).

https://doi.org/10.1016/j.patcog.2025.111753
Received 27 January 2025; Received in revised form 2 April 2025; Accepted 19 April 2025
Available online 5 May 2025

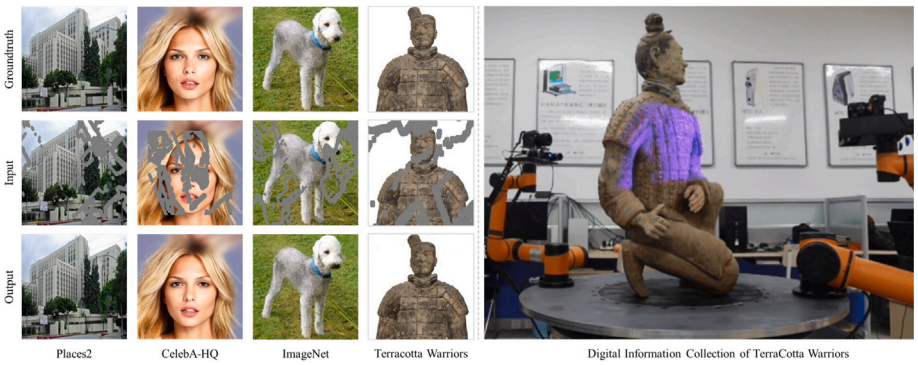

Fig. 1. In the digital restoration of the Terracotta Warriors, (Left) demonstrates the inpainting results of our method on different datasets, achieving efficient and high-fidelity image restoration. (Right) shows the collection of point cloud and image data using a high-definition camera on a robotic arm platform for the digital restoration of the Terracotta Warriors.

of two principal components: ResampleDiff and RefineDiff. The ResampleDiff module initially conducts a preliminary restoration of the image, thereby expediting the generation of a foundational yet precise outcome. The RefineDiff module further mitigates noise and blurriness, enhancing the image's detailed information. Ultimately, this two-stage diffusion process culminates in high-quality image restoration. Empirical findings indicate that our approach significantly diminishes both the parameter count and inference time, while simultaneously achieving superior restoration performance across various datasets (as shown in Fig. 1). This presents a novel and effective strategy for the digital restoration of the Terracotta Warriors.

The main contributions of our paper are as follows:

- We propose a two-stage image restoration method based on diffusion models, referred to as TRRS-DM. This approach integrates ResampleDiff in the pixel space with RefineDiff in the latent space, aiming to achieve high-quality image restoration.
- We employ a lightweight diffusion model, a signal-to-noise ratio (SNR) loss function, DDIM acceleration, and latent space residual shift diffusion techniques to reduce computational costs while ensuring high-quality restoration.
- Our method shows remarkable performance in texture restoration compared to leading techniques. It effectively recovers texture details, yielding realistic and natural results with minimal parameter overhead.

2. Related work

2.1. Image inpainting

Image inpainting is a complex low-level visual task that seeks to reconstruct absent regions of an image by utilizing information from intact pixels. Traditional methods [15], such as gradient interpolation [3], partial differential equations [16], and patch-based filling [4], are capable of generating locally smooth results or filling small areas of missing data. While these methods can effectively complete small missing areas, they face challenges in accurately reconstructing more complex scenes due to limited global understanding of the image.

Compared to traditional methods, deep learning-based image inpainting has achieved tremendous success [17]. Architectures like encoder–decoder [18] and GANs have been proposed, with many novel methods focusing on image inpainting. To ensure that the restored images have semantically reasonable context, researchers have introduced techniques such as dilated convolutions [19] to increase the receptive field, Partial convolutions [7] which guide convolutional kernels according to a mask, improved learnable convolution kernels to generate dynamic soft masks [20], and Fourier-based convolution encoders [21] for image inpainting to avoid generating invalid features within missing regions. Further research focuses on more refined inpainting, using edge guidance with EdgeConnect [22], PEN-Net's [23] pyramid layer-by-layer restoration, DSI [11] based on VQ-VAE diversity inpainting, and CTSDG's [24] dual-stream network architecture combining image texture structure priors. With the success of transformers in natural language processing, ICT [25] and MAT [26] have also applied transformers to image restoration. Spatial self-attention based on this approach can incur high computational costs. To reduce computation, some approaches downsample input images to lower resolutions [27]. Others compute spatial attention after encoding the input image into low-resolution features [28]. RestFormer [29] proposes channel self-attention for image reconstruction in multi-scale representations with linear complexity. These strategies aim to explore how to extract useful information from known areas for hole-filling effectively. These methods, which rely on specific masks during training, exhibit weak generalization to new mask types. Free mask repair within diffusion models represents a promising direction for further research.

2.2. Diffusion inpainting

Diffusion models have achieved significant advancements in image generation, capable of producing high-quality images through an iterative probabilistic process that introduces novel opportunities for image inpainting [30]. The fundamental Denoising Diffusion Probabilistic Model (DDPM) [31,32] executes the transformation of noise into an image through a two-stage process, which includes forward denoising and backward denoising diffusion.

To use the diffusion model for image inpainting, RePaint [12] utilizes a pre-trained DDPM as a generative prior, managing the reverse diffusion iterations by sampling exclusively from unoccluded regions, thereby achieving high-quality and diverse image restorations. LatentPaint [13] addresses the challenge of high computational costs associated with conventional techniques by employing latent space conditioning and explicit propagation methods. StrDiffusion [33] reconceptualizes texture denoising through structural guidance and time-dependent sparsity, optimizing the denoising process to reduce semantic discrepancies between occluded and unoccluded areas. M2S [34] introduces a coarse-to-fine sampling strategy to decrease the number of denoising steps and expedite inference. And some guided inpainting methods: SmartBrush [35] employs shape guidance, Paint by Example [36] uses reference images for guidance, and BrushNet [37] and Imagen Editor [38] leverage text-guided image generation to modulate the input and achieve semantically coherent and meaningful results. Additionally, various adaptable diffusion strategies have been developed, including a zero-shot framework for linear image recovery [39], the unsupervised posterior sampling technique DDRM [40], the fusion of forward and reverse diffusion for enhanced efficiency [41], and non-Markovian diffusion acceleration [42]. Despite its promising

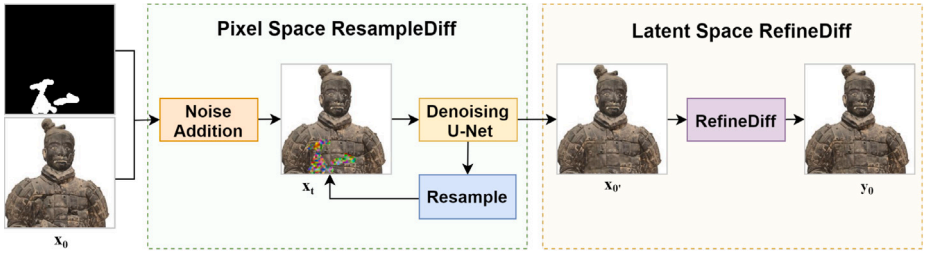

Fig. 2. An overview of the framework. TRRS-DM is structured around a two-tier architecture that comprises a resampling diffusion module and a latent spatial enhancement module.

Table 1
Symbols and Description.

Symbols	Description
x_0	Raw data sample (without added noise)
x_t	The data sample at time step t (with added noise)
β_t	Noise coefficient during diffusion process
$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$	Accumulated denoising factor
$\mu_\theta(x_t,t), \Sigma_\theta(x_t,t)$	Mean & variance predicted in the reverse process
m	Mask, distinguish known and unknown regions
$\odot$	Element-wise multiplication
$\epsilon_\theta(x_t,t)$	Noise predicted by the model at time step t
$\{\eta_t\}_{t=1}^{T}$	Monotonically increasing displacement sequence
α_t	Increment in the displacement sequence
$f_\theta(x_t,y_0,t)$	Deep Neural network predicting x_0 with θ

potential, diffusion-based restoration is computationally intensive and time-consuming. Therefore, accelerating these models is crucial. This study investigates applying diffusion restoration technology to the digital preservation of the Terracotta Warriors, aiming to maintain high-quality cultural heritage conservation while enhancing computational efficiency.

3. Methods

3.1. Preliminary

For the convenience of understanding, the main Symbols and their Description are explained in Table 1.

3.2. Overall architecture

To facilitate the application of diffusion model image restoration techniques to the conservation of the Terra Cotta Warriors' cultural artifacts, we aim to enhance the efficiency of the restoration process while preserving high-quality outcomes. We introduce the **T**wo-stage **R**esampling and **R**esidual **S**hifting using **D**iffusion **M**odels (TRRS-DM). This two-stage method incorporates resampling and residual shifting utilizing diffusion models for two-dimensional texture inpainting, as shown in Fig. 2. TRRS-DM consists of a resampling diffusion module and a latent space enhancement module, which collectively enhance restoration quality through a two-stage processing approach. The ResampleDiff module executes resampling diffusion within the pixel space to initially address the missing areas of the image, producing intermediate results. Subsequently, the RefineDiff module further refines these intermediate outputs, ensuring that the final restoration results are clear and natural.

In the ResampleDiff module, we use a lightweight diffusion network to input the mask and the image x_0 to be repaired, and combine it with the mask to add noise information to the image to generate x_t. Subsequently, a resampling strategy was adopted for denoising, and the preliminary result x_0' was obtained as a reference for the subsequent enhancement stage. We also integrate the denoising diffusion implicit model (DDIM) and perceptual prior weighting strategy to ensure efficient and high-quality output.

In the RefineDiff module, the intermediate result x_0' is input into the RefineDiff enhancement module. Here, a pre-trained encoder encodes this result into latent space, where the residual switching enhancement diffusion method is employed to facilitate the transformation into higher-quality encoded information. The diffusion outcome is then decoded to produce a superior quality repair result, y_0. Our approach leverages latent spatial diffusion to minimize computational expenses while achieving enhanced quality in image conversion.

3.3. ResampleDiff moudle

Our two-stage diffusion inpainting involves the basic diffusion process, forward denoising and backward denoising. The diffusion process in DDPM incrementally adds Gaussian noise to the data at each time step t, described by a Markov chain (see Eqs. (1) and (2)). Using the reparameterization trick, it allows sampling $\mathbf{x}_t$ at any given timestep t in closed form:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

Reverse Process: The aim is to reconstruct the original data distribution from pure noise. In inference, starting from a sample $\mathbf{x}_T \sim \mathcal{N}(0,\mathbf{I})$, we sequentially sample $\mathbf{x}_{t-1}$ until recovering the original image $\mathbf{x}_0$. This is modeled as a Markov chain with conditional distributions:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t,t), \Sigma_\theta(\mathbf{x}_t,t)) \tag{3}$$

For the diffusion-based resampling image inpainting technique, we employed a lightweight pre-trained unconditional DDPM [43] and introduced the masked input image into the reverse diffusion process for conditional generation (see Fig. 3). For known regions, it directly utilizes information from the original image, by Eq. (2), $x_{\text{known}}^{t-1} \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$. For unknown regions, the model generates new content to fill in the gaps, by Eq. (3), $x_{\text{unknown}}^{t-1} \sim \mathcal{N}\left(\mu_\theta(x_t,t), \Sigma_\theta(x_t,t)\right)$. The two components are then combined using a mask operation, as shown in Eq. (4). This result is used as input for the next denoising step, ensuring that the generated content blends harmoniously with existing content, thereby improving the quality and coherence of the inpainted image.

$$\mathbf{x}^{t-1} = \mathbf{m} \odot \mathbf{x}_{\text{known}}^{t-1} + (1-\mathbf{m}) \odot \mathbf{x}_{\text{unknown}}^{t-1} \tag{4}$$

However, a single harmonization resampling step may not adequately incorporate semantic information throughout the denoising process, meaning that the integration between newly generated image parts and the surrounding environment might be insufficient, leading to less natural or coherent results. To better propagate contextual information while avoiding over-processing and enhancing efficiency, we introduce a resampling strategy with parameters such as the number of resampling iterations, jump length, and interval. Fig. 4 presents the specific sampling strategies.

In our method, reducing the diffusion model parameters can accelerate the inference process. However, directly reducing parameters

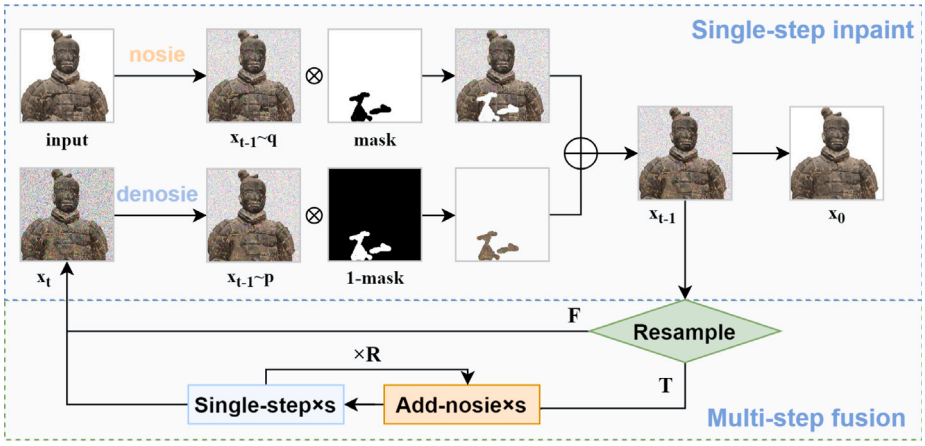

Fig. 3. Schematic of the ResampleDiff Module. This module completes unknown regions by applying diffusion resampling to the input image and mask.

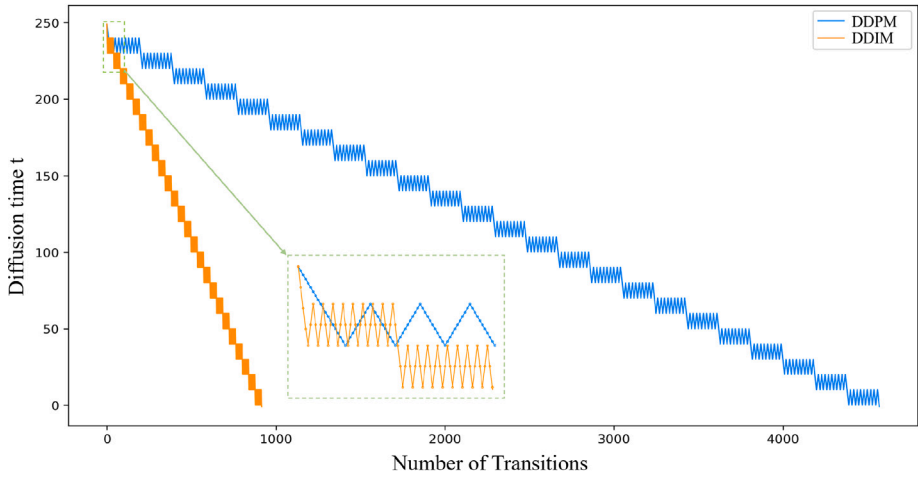

Fig. 4. The resampling process schedule for diffusion inpainting. Blue represents DDPM, orange represents DDIM, and each point in the small image represents a diffusion step. Time T is set to 250, and the number of resampling times, jump length, and jump interval are all set to 10, DDIM=5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

may impair the model's performance. Inspired by [34,43], perceptual priority (P2) weighting aims to prioritize learning from more important noise levels. Unnecessary cleaning stages are assigned the minimum weight, thereby assigning relatively higher weights to the remaining stages. We modify the loss function during the training process to compensate for the parameter reduction.

Diffusion models are trained by optimizing the variational lower bound (VLB), which is the sum of denoising score-matching losses: $L_{\text{vlb}} = \sum_t L_t$. Each step t matching loss L_t measures the distance between two Gaussian distributions and can be rewritten in terms of the noise predictor ϵ_θ as follows:

$$\begin{aligned} L_t &= D_{KL}\left(q\left(x_{t-1} \mid x_t, x_0\right) \parallel p_\theta\left(x_{t-1} \mid x_t\right)\right) \\ &= \mathbb{E}_{x_0,\epsilon}\left[\frac{\beta_t}{(1-\beta_t)(1-\alpha_t)}\left\|\epsilon - \epsilon_\theta(x_t, t)\right\|^2\right] \end{aligned} \tag{5}$$

Within the VLB framework, the objective can be expressed as: $L_{\text{simple}} = \sum_t \lambda_t L_t$, with a weighting scheme $\lambda_t = \frac{(1-\beta_t)(1-\bar{\alpha}_t)}{\beta_t}$. We introduced perception-prioritized weighting into the modified loss function to emphasize context learning:

$$\lambda_t' = \lambda_t k + \text{SNR}(t)^\gamma, \tag{6}$$

Where the original weight λ_t is replaced by λ_t'. The parameter γ controls the strength of down-weighting, and k prevents weight explosion while determining the sharpness of the weighting scheme. The SNR is defined as: $\text{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}$. The overall loss function becomes: $L = \sum_t \lambda_t' L_t$. This modification ensures the model focuses on perceptually rich context during training, leading to richer and more natural results.

Finally, we incorporated the DDIM accelerated sampling strategy into our model [42]. By using a non-Markovian inference process, we obtain denoised result $\mathbf{x}_{\text{known}}^{t-j}$ more efficiently:

$$\mathbf{x}_{\text{unknown}}^{t-j} = \sqrt{\bar{\alpha}_{t-j}}\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_t}\cdot\epsilon_\theta(\mathbf{x}_t, t) + \sqrt{1-\bar{\alpha}_{t-j}-\sigma_t^2}\cdot\epsilon_\theta(\mathbf{x}_t, t) + \sigma_t\epsilon_t, \tag{7}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ represents independent standard Gaussian noise relative to $\mathbf{x}_t$. When $\sigma_t = 0$, DDIM sampling becomes a special case. We then diffuse the input image according to Eq. (2) to align known information with $\mathbf{x}_{\text{unknown}}^{t-j}$. Finally, we concatenate the generated information $\mathbf{x}_{\text{unknown}}^{t-j}$ with the conditional information $\mathbf{x}_{\text{known}}^{t-r}$ as per Eq. (4), ensuring effective integration of conditional information throughout the denoising process. This enhances the consistency and semantic correctness of the inpainted regions with the surrounding image. Finally, the overall algorithm of the RefineDiff module is shown in Algorithm 1.

3.4. RefineDiff moudle

In the RefineDiff section, as shown in Fig. 5, the original image is first compressed four times using VQGAN [44] and encoded into latent space. Then, a Markov chain is constructed to convert high-quality and low-quality images. By controlling the conversion speed and noise

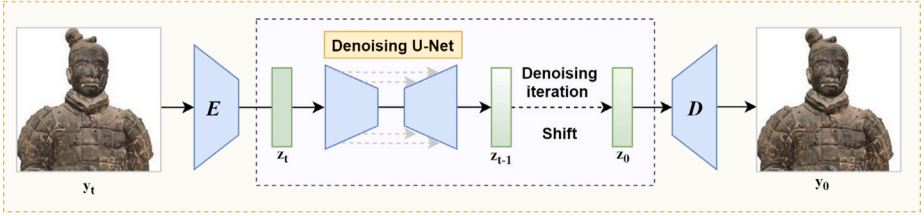

Fig. 5. Schematic of the RefineDiff Module. This module refines low-quality input images by encoding them into a latent space to reduce resource requirements, and then applies residual diffusion methods to achieve higher clarity and accuracy in the results.

Algorithm 1: Inpainting Using ResampleDiff Approach

Input: Image x and Mask m

Initialize the noisy image $x_T \sim \mathcal{N}(0, I)$.
Set r_{sample} (number of iterations), j_{sample} (jump length), and j_{interval} (jump interval), Resample $= [0, 0, 1, \ldots, 0]$, where $R = 1$ for resampling and $R = 0$ for jumping, determined by j.
for $t = T, \ldots, 1$ **do**
  Sample $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, otherwise set $\epsilon = 0$,
  $x^{t-1}_{\text{known}} = \sqrt{\bar{\alpha}_t} x_0 + (1 - \bar{\alpha}_t)\epsilon$;
  Sample $z \sim \mathcal{N}(0, I)$ if $t > 1$, otherwise set $z = 0$,
  $x^{t-1}_{\text{unknown}} = \frac{1}{\sqrt{\alpha_t}} \left[x_t - \beta_t \left(\sqrt{1 - \bar{\alpha}_t} \right) \epsilon_\theta(x_t, t) \right] + \sigma_t z$, use DDIM;
  Update image: $x^{t-1} = m \odot x^{t-1}_{\text{known}} + (1 - m) \odot x^{t-1}_{\text{unknown}}$.
  if *Resample (when* $T \mod j_{\text{interval}} = 0$*)* **then**
    for $i = 1, \ldots, r_{\text{sample}}$ **do**
      if $r < r_{\text{sample}}$ *and* $t > 1$ **then**
        Sample $x_t \sim \mathcal{N}\left(\sqrt{1 - \beta_{t-j-1}} x_{t-r}, \beta_{t-j-1} I \right)$;
        Iterate j_{sample} times Single step inpaint(4-8).
      end
    end
  end
end

Output: Inpainted Image x_0

intensity during the diffusion process, the Markov chain is used for reverse sampling to enhance low-quality images, thereby reducing the computational cost during training and achieving high-quality images.

Specifically, a residual $e_0 = y_0 - x_0$ is defined between the high-quality image y_0 and the low-quality image x_0, from which a Markov chain of length T is constructed to migrate from x_0 to y_0 gradually. For this purpose, we introduce a monotonically increasing displacement sequence $\{\eta_t\}_{t=1}^T$ satisfying $\eta_1 \to 0$ and $\eta_T \to 1$. Based on this sequence, the transition distribution is defined as Eq. (8), where $\alpha_t = \eta_t - \eta_{t-1}$, and k is a hyperparameter controlling the noise variance. It can be further simplified into Eq. (9).

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, y_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1} + \alpha_t e_0, k^2 \alpha_t \mathbf{I}), \quad t = 1, 2, \ldots, T \tag{8}$$

$$q(\mathbf{x}_t|\mathbf{x}_0, y_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0 + \eta_t e_0, k^2 \eta_t \mathbf{I}), \quad t = 1, 2, \ldots, T \tag{9}$$

The transition distribution design adheres to two primary principles: smooth transition and mean parameterization. Smooth transition ensures smooth changes between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$, introducing design flexibility by setting the hyperparameter k when the image data is within the $[0, 1]$ interval. The expected distance between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ can be bounded by $\sqrt{\alpha_t}$. The mean parameter $\mathbf{x}_0 + \alpha_t e_0$ derives the marginal distribution as described in Eq. (9). Moreover, the marginal distributions of $\mathbf{x}_1$ and $\mathbf{x}_T$ converge to $\delta_{\mathbf{x}_0}(\cdot)$ and $\mathcal{N}(\cdot; y_0, k^2 \mathbf{I})$, similar to high-quality and low-quality distributions.

The goal of the reverse process is to estimate the posterior distribution $p(x_0|y_0)$:

$$p(x_0|y_0) = p(x_T|y_0) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y_0) dx_{1:T} \tag{10}$$

where $p(x_T|y_0) \approx \mathcal{N}(x_T|y_0, k^2\mathbf{I})$, and $p_\theta(x_{T-1}|x_T, y_0)$ is the inverse transition kernel from x_T to x_{T-1} with parameters θ. Optimization of θ is achieved by minimizing the negative evidence lower bound:

$$\min_\theta \sum_t D_{\text{KL}}(q(x_{t-1}|x_t, x_0, y_0) \parallel p_\theta(x_{t-1}|x_t, y_0)) \tag{11}$$

where $D_{\text{KL}}[\cdot \parallel \cdot]$ denotes the KL divergence. Combining Eq. (8) and (9), the target distribution $q(x_{t-1}|x_t, x_0, y_0)$ becomes tractable and can be expressed as:

$$q(x_{t-1}|x_t, x_0, y_0) = \mathcal{N}\left(x_{t-1}; \frac{\eta_{t-1}}{\eta_t} x_t + \frac{\alpha_t}{\eta_t} x_0, \kappa^2 \frac{\eta_{t-1}}{\eta_t} \alpha_t \mathbf{I} \right) \tag{12}$$

Considering the variance parameter independence from $\mathbf{x}_t$ and y_0, we set $\Sigma_\theta(\mathbf{x}_t, y_0, t) = k^2 \frac{\eta_{t-1}}{\eta_t} \alpha_t \mathbf{I}$. For the mean parameter, we set $\mu_\theta(\mathbf{x}_t, y_0, t) = \frac{\eta_{t-1}}{\eta_t} \mathbf{x}_t + \frac{\alpha_t}{\eta_t} f_\theta(\mathbf{x}_t, y_0, t)$, where f_θ is a deep neural network with parameters θ predicting $\mathbf{x}_0$. The objective function can be simplified as Eq. (13).

$$\min_\theta \sum_t w_t \| f_\theta(\mathbf{x}_t, y_0, t) - \mathbf{x}_0 \|_2^2 \tag{13}$$

To determine the noise schedule during the diffusion process, we adopt hyperparameters k and the shift sequence $\{\eta_t\}_{t=1}^T$. For the initial state, the noise level under state $\mathbf{x}_t$ is proportional to $\sqrt{\eta_t}$ with a proportionality factor k. Given the constraint $\eta_1 \to 0$, we set $\eta_1 = 0.001$. For the final state, at the last step T, η_T is set to 0.999 to ensure $\eta_T \to 1$. For intermediate states, for $t \in [2, T-1]$, we take a non-uniform geometric scheduling for $\sqrt{\eta_t}$:

$$\eta_t = \eta_1 \times b_0^{\beta_t}, \quad t = 2, \ldots, T - 1 \tag{14}$$

$$\beta_t = \left(\frac{t-1}{T-1} \right)^p \times (T - 1), \quad b_0 = \exp\left(\frac{1}{2(T-1)} \log \frac{\eta_T}{\eta_1} \right) \tag{15}$$

Assumptions are made that $\beta_1 = 0$, $\beta_T = T - 1$, and $\sqrt{\eta_T} = \sqrt{\eta_1} \times b_0^{T-1}$. The parameter p controls the growth rate of $\sqrt{\eta_t}$. By flexibly scheduling noise, setting appropriate k to limit the amplitude of noise, and controlling the speed of noise addition with p, the final state converges to around low-quality images, shortening the length of the Markov chain and improving inference efficiency.

4. Experiments and results

4.1. Datasets

We conducted extensive comparative experiments using three public image datasets: Places2 [45], CelebA-HQ [46], and ImageNet [47], along with a specially designed mask dataset [7]. Additionally, to demonstrate practical application effects, we performed restoration tests on real images of the Terracotta Warriors. The detailed descriptions are provided in Table 2. Details of the Terracotta Warriors data are in Appendix A.1.

Table 2
Dataset Examples and Descriptions.

Dataset	Image	Data Description	Experiment Description
Places2		Complex scene-level inpainting from Paris street scenes	Trained on 180,000 images from 100 categories of Places2-Standard; tested on 10,000 images
CelebA-HQ		High-quality dataset for face inpainting	Trained on 28,000 images; tested on 2000 images
ImageNet		Diverse object inpainting	Trained on 120,000 images from 100 categories; tested on 10,000 images
Mask Dataset		Irregular mask dataset	Tested on 12,000 irregular masked images
Terracotta Warriors		Terracotta Warrior images, including 4,170 intact and 4000 damaged	Trained on 4000 intact images; tested on 170 intact and 4000 damaged images

4.2. Baseline

In this work, we compared our method with four relevant and currently relatively advanced restoration methods:

DF-v2 [20]: Utilizes gated convolution to effectively handle free-form missing regions, allowing for precise and natural inpainting.

DSI [11]: Generates diverse and plausible structural information through a hierarchical VQ-VAE, enhancing the realism of the inpainted areas.

CTSDG [24]: Ensures global and local consistency in inpainted regions using conditional texture and structure dual generators, achieving coherent results.

RePaint [12]: Implements high-quality image inpainting based on denoising diffusion probabilistic models, offering iterative refinement and strong controllability.

The results of these comparison methods in the experimental table are derived from the results we reproduced using the optimal checkpoint they provided.

Table 3
Quantitative comparison of five methods on datasets Places2, CelebA HQ, and ImageNet, where "↑" indicates that the larger, the better, and "↓" indicates that the smaller, the better.

Metrics		PSNR ↑			SSIM ↑			LPIPS ↓		
	Methods	0-20	20-40	40-60	0-20	20-40	40-60	0-20	20-40	40-60
Places2	DF-v2	30.297	24.725	18.756	0.956	0.857	0.674	0.045	0.136	0.289
	DSI	31.418	24.199	19.295	0.952	0.858	0.684	0.042	0.135	0.260
	CTSDG	32.145	**25.707**	20.642	0.963	0.883	0.725	0.033	0.118	0.257
	RePaint	32.445	25.091	**20.664**	0.965	**0.894**	**0.775**	0.031	**0.108**	**0.212**
	Ours	**32.466**	25.103	20.662	**0.968**	0.885	0.764	**0.030**	0.110	0.243
CelebA-HQ	DF-v2	32.074	27.330	22.694	0.953	0.866	0.796	0.034	0.106	0.244
	DSI	31.752	26.643	22.540	0.949	0.893	0.785	0.037	0.178	0.262
	CTSDG	32.170	28.480	23.208	0.958	0.902	0.847	0.035	**0.082**	0.218
	RePaint	32.653	**29.850**	23.534	0.963	0.925	**0.861**	0.026	0.089	0.192
	Ours	**32.734**	29.641	**23.940**	**0.968**	**0.926**	0.858	**0.025**	0.085	**0.181**
ImageNet	DF-v2	31.458	24.428	19.536	0.955	0.861	0.714	0.047	0.146	0.278
	DSI	31.902	23.647	20.019	0.961	0.853	0.734	0.043	0.145	0.236
	CTSDG	32.667	25.127	20.640	0.953	0.893	0.725	**0.030**	0.128	0.273
	RePaint	32.713	28.192	20.494	0.960	0.898	**0.818**	0.034	**0.093**	**0.171**
	Ours	**32.821**	**28.973**	**20.679**	**0.962**	**0.902**	0.813	0.031	0.095	0.176

4.3. Evaluation metric

In the experiment, we used the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual similarity (LPIPS) to measure the objective quality and similarity of images from different perspectives. PSNR is a commonly used method for measuring image quality, SSIM is an image quality assessment method based on the human visual system that considers the similarity of brightness, contrast, and structure, and LPIPS is a deep learning based image quality assessment method aimed at simulating the human visual system's perception of image differences. The formulas for these indicators are as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{16}$$

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{17}$$

A higher PSNR value indicates less image quality loss, while SSIM ranges from −1 to 1, with higher values showing greater similarity in structure, brightness, and contrast. Lower LPIPS scores mean the images are more perceptually similar. For real-world applications like the Terracotta Warriors restoration, where original images are unavailable, we use no-reference metrics — BRISQUE, NIQE, and PI — to assess naturalness and overall perceptual quality. BRISQUE predicts quality via natural statistics, NIQE estimates naturalness based on scene statistics, and PI evaluates comprehensive perceptual quality. For all three metrics, lower scores indicate better image quality and closer resemblance to natural images.

4.4. Implementation details

Both diffusion models utilized in our methodology underwent 500,000 training iterations, commencing with an initial learning rate of 2×10^{-5}. For the ReampleDiff model, the parameters were set as follows: $T = 1000$, $\alpha = 0.5$, $k = 1$, and an exponential moving average (EMA) rate of 0.999. In the case of the RefineDiff model, a pre-trained VQGAN model was employed as the encoder–decoder module, with the parameters configured to $T = 4$, $p = 0.3$, and $k = 2$. During our experimental procedures, both images and masks were resized to dimensions of 256×256 pixels. The entire model was developed using the PyTorch 2.5.0 framework and trained on an NVIDIA® A6000 GPU.

4.5. Main results

4.5.1. Quantitative comparisons

To ensure the reliability and consistency of our findings, we carried out three independent tests on different repair masks and diverse image data, thereby effectively eliminating the potential influence of

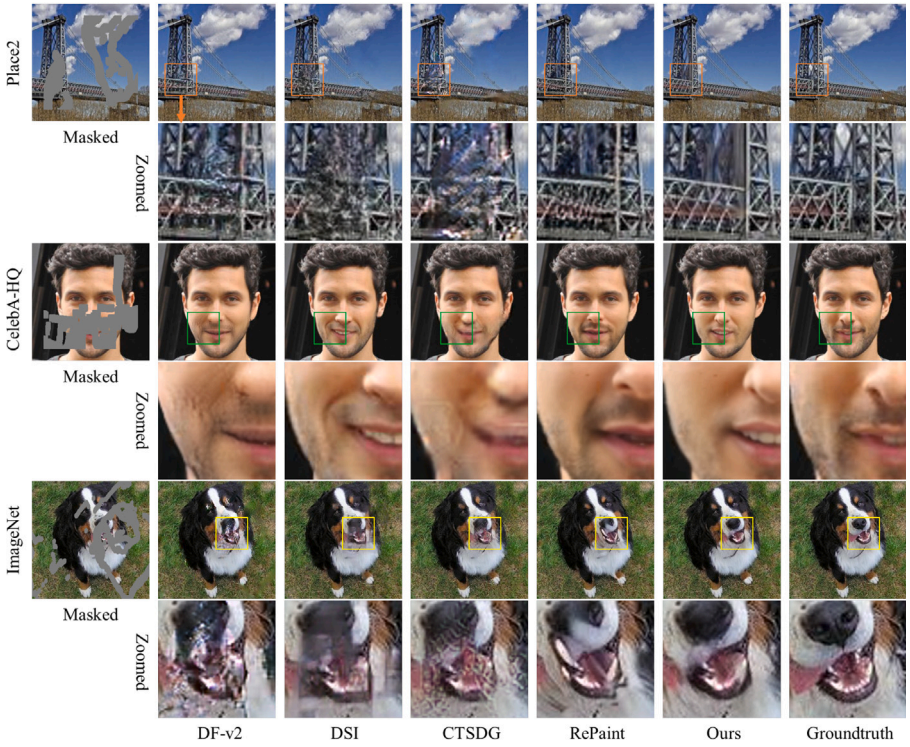

Fig. 6. Qualitative comparisons of five methods on three datasets show the results on Places2, CelebA-HQ, and ImageNet, as well as the details of the repair results. The last column represents GruondTruth.

randomness. We then assessed the repair outcomes using PSNR, SSIM, and LPIPS metrics to obtain quantitative indicators across three distinct datasets and various mask intervals. An analysis of the overall results presented in Table 3 reveals that our proposed method achieves the highest PSNR and SSIM values while demonstrating the lowest LPIPS score among the evaluated methods. Our approach showcases superior repair efficacy and adaptability across multiple masks and datasets. Although the performance of our method in restoring large masks on the Places2 dataset is relatively lower than that of RePaint, this can likely be ascribed to the inherent complexity of the scene restoration content. Nevertheless, our method significantly improves the efficiency of diffusion computations while maintaining a competitive edge. This highlights the effectiveness of our two-stage progressive diffusion repair technique in achieving high-quality structural and textural restoration results.

4.5.2. Qualitative comparisons

Fig. 6 presents examples of both overall and localized inpainting outcomes achieved through five different methodologies applied to the Places2 (scene images), CelebA HQ (facial images), and ImageNet (natural images) datasets. The observations depicted in the figure indicate that the visual quality of the DF-v2, DSI, and CTSDG methods is relatively subpar, which aligns with the findings from the quantitative assessments. In contrast, our proposed approach demonstrates superior capabilities in restoring structural integrity and intricate details. For instance, the mesh structure of the bridge illustrated in the first row exhibits more defined textures, the facial restoration in the second row appears more aesthetically pleasing and harmonious, and the depiction of the dog's teeth in the third row retains an overall aesthetic appeal while maintaining a reasonable structural representation.

Furthermore, Fig. 7 shows the inpainting results associated with two asymptotic inpainting processes. By implementing additional denoising and deblurring techniques on the initial inpainting outputs, the image quality is significantly improved, thereby underscoring the efficacy of the asymptotic inpainting method and the RefineDiff module. In conclusion, these qualitative comparisons substantiate the effectiveness of our methodology across various mask configurations and datasets.

Table 4
Ablation Study of SNR Loss on Places2.

Loss (γ)	PSNR ↑	SSIM ↑	LPIPS ↓
SNR-0	19.654	0.814	0.145
SNR-1	**20.313**	0.795	0.161
SNR-0.5	20.239	**0.816**	**0.116**

4.6. Ablation studies

This section analyzes the effectiveness of quality improvement methods and accelerated inference strategies in two-stage diffusion inpainting. By comparing different SNR loss functions and RefineDiff control parameters, we verify the impact of parameters on repair quality. By comparing different sampling methods of ResampleDiff, combining jump parameters with RefineDiff, and analyzing the impact of acceleration strategies on the results, we aim to obtain a more efficient sampling strategy.

4.6.1. SNR loss function

To assess the efficacy of the signal-to-noise ratio (SNR) loss function, we performed comparative experiments utilizing various gamma (γ) parameters within a lightweight diffusion model implemented in the ResampleDiff module. The parameter γ modulates the extent of weight reduction, thereby enhancing the emphasis on the inpainting of regions with high noise levels. We compared the SNR loss against the effects of parameters $\gamma = 0$, $\gamma = 1$, and $\gamma = 0.5$ on the inpainting outcomes. Specifically, $\gamma = 0$ corresponds to the mean squared error (MSE) loss. Table 4 and Fig. 8 illustrate that the optimal results were achieved with $\gamma = 0.5$ across all three evaluation metrics. In contrast, $\gamma = 1$ diminished the weight allocated to the latter portion of the data, resulting in pronounced noise artifacts in the outcomes. The SNR loss function prioritizes the learning of initial details by recalibrating the training weight distribution, which subsequently enhances the inpainting quality of our lightweight model.

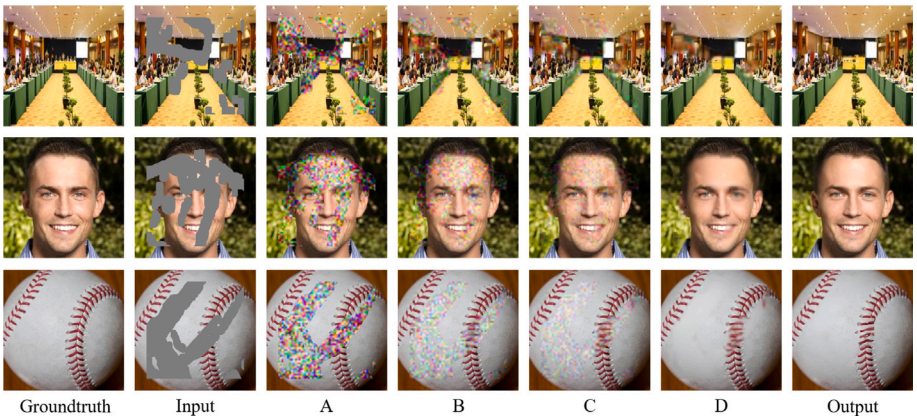

Fig. 7. Visualization of diffusion in the painting process. Figures A to D show the gradual denoising process of the first stage, and the output shows the result after enhancement in the second stage.

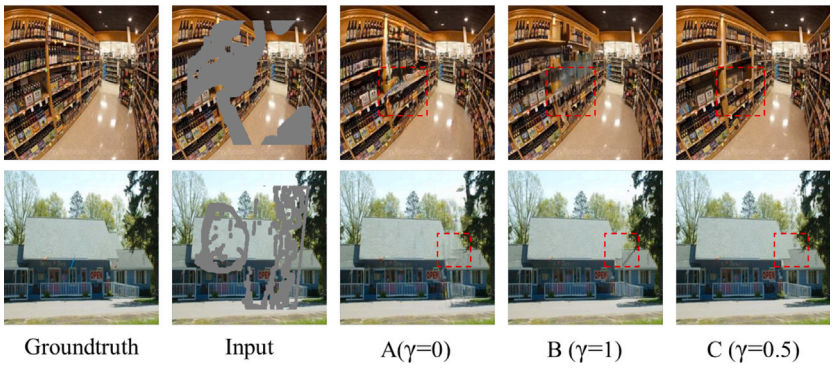

Fig. 8. Qualitative Comparison of SNR Loss on Places2 (Panels A, B, and C show the inpainting results for γ=0, 1, and 0.5).

Table 5
Comparison of Different RefineDiff Configurations on ImageNet.

P	k	PSNR ↑	SSIM ↑	LPIPS ↓
0.3	1	19.987	0.656	0.348
0.3	2	**22.031**	**0.838**	**0.110**
0.5	2	19.978	0.764	0.284
1.0	2	19.767	0.742	0.216

Table 6
Comparison of Different Resampling Configurations on CelebA-HQ.

R_{sample}	J_{interval}	PSNR ↑	SSIM ↑	LPIPS ↓	Time (s)
5	10	29.147	0.932	0.049	21.49
10	10	29.060	0.934	0.048	26.29
15	10	**30.395**	0.939	**0.047**	32.64
10	5	29.651	**0.941**	0.048	40.50
10	15	30.243	0.931	**0.047**	**18.51**

4.6.2. *RefineDiff Residual Control Parameters*

To improve the conversion of initially repaired images into higher-quality images, comparative experiments focused on various parameters related to noise intensity and conversion speed within the RefineDiff module. Specifically, within the diffusion step range denoted as T, the parameter k regulates the final state noise intensity, facilitating convergence towards the low-quality image spectrum. In contrast, the parameter p allows for precise modulation of the noise conversion speed. As illustrated in Table 5 and Fig. 9, which pertain to the case when $T = 4$, enhancing the initially repaired image A yielded superior results for image C when p was set to 0.3 and k to 2. Conversely, using $k = 1$ with a larger p-value resulted in a blurred output. When k was set to 2, the noise intensity approximated that of the initial repair outcome. Furthermore, with p at 0.3, the conversion speed was deemed optimal, thereby significantly improving the quality of the resultant image.

4.6.3. *ResampleDiff schedule parameters*

To enhance the integration of the two components of diffusion repair and to examine the effects of various accelerated repair processes within a single stage on the final outcomes, a comparative experiment was conducted focusing on the resampling parameters while utilizing the same RefineDiff module enhancement. Typically, more resampling steps and reduced intervals yield superior preliminary repair results. By setting the jump length to 10, we analyzed the influence of varying resampling frequencies and jump intervals on both quality and efficiency. As illustrated in Table 6 and Fig. 10, the findings indicate that a resampling frequency of $r = 10$ and a jump interval of $j = 15$ resulted in the most rapid repair speed alongside high-quality outcomes. This demonstrates that an effective combination of the two components can improve performance and enhance visual quality.

4.7. *Efficiency comparison*

In this section, we assess the computational complexity of the TRRS-DM model by analyzing the architectural configurations, parameter counts, and average inference times associated with various repair methodologies. As illustrated in Table 7, the diffusion-based model generally demonstrates a greater number of parameters and extended inference times when juxtaposed with the GAN-based model. This discrepancy arises from the necessity of numerous sampling steps in the denoising process of the diffusion model to attain high-quality image generation. Nevertheless, in contrast to the diffusion-based RePaint method, our approach markedly decreases the parameter count by

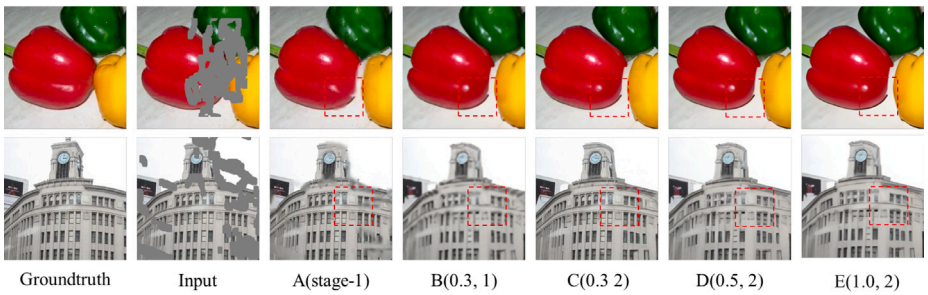

Fig. 9. Qualitative Comparison of Different RefineDiff Configurations on ImageNet (Panel A shows coarse inpainting results; Panels B, C comparep = 0.3, k=1, 2; Panels C, D, E compare k=2,p = 0.3, 0.5, 1).

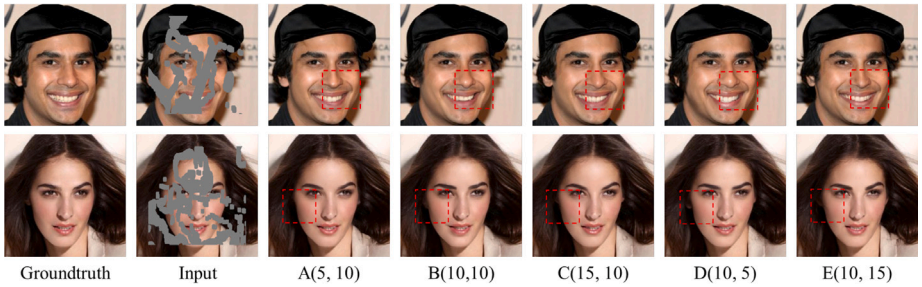

Fig. 10. Qualitative Comparison of Different Resampling Configurations on CelebA-HQ (Panels A, B, C compare j=10, r=5, 10, 15; Panels B, D, E compare (r=10, j=10, 5, 15).

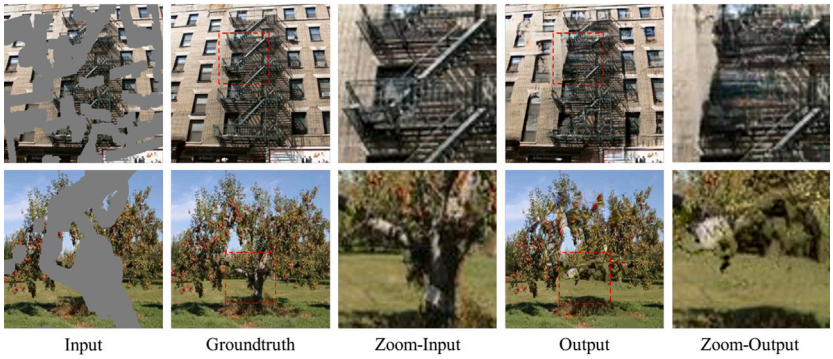

Fig. 11. Restricted inpainting example on Places2 Dataset.

Table 7
Comparison of Complexity and Efficiency Among Different Methods.

Methods	Architecture	Params	Inference time (s)
DF-v2	GAN	4M	0.65
DSI	GAN	76M	1.14
CTSDG	GAN	52M	2.23
RePaint	Diffusion	552M	492.04
Ours	Diffusion	191M	19.52

two-thirds and reduces the inference time to one-twentieth of what is required by RePaint. Furthermore, the modular design of our two components allows for independent training, which further alleviates computational demands during the training phase and facilitates a more efficient process for diffusion-based image inpainting.

4.8. Limitation and discussion

Limitations: As shown in Fig. 11, the complexity of large-area masks, combined with the relatively small proportion of sample categories, makes it difficult to adequately generate images that perform well in terms of both structural texture and semantic coherence. This phenomenon is prevalent among various image inpainting techniques and is also a problem faced in the restoration of the Terracotta Warriors. The possible reasons for this include the limited available information in the damaged images and the insufficient representativeness of sample categories in the training dataset.

Discussion: In order to improve the quality of image inpainting, it is advisable to investigate the integration of some supplementary reference information or guiding methods. Adopting high-quality data for pre-training and implementing data augmentation on small datasets can enhance the data quality and optimize the performance of diffusion models in complex scenarios. Furthermore, future research should delve deeply into the application of diffusion models in high-resolution image inpainting and consider increasing user engagement, so that the inpainting results can more effectively meet user requirements.

4.9. Application of terracotta warriors

To better assess the efficacy of our proposed methodology, we conducted a comparative analysis using the authentic Terracotta Warriors

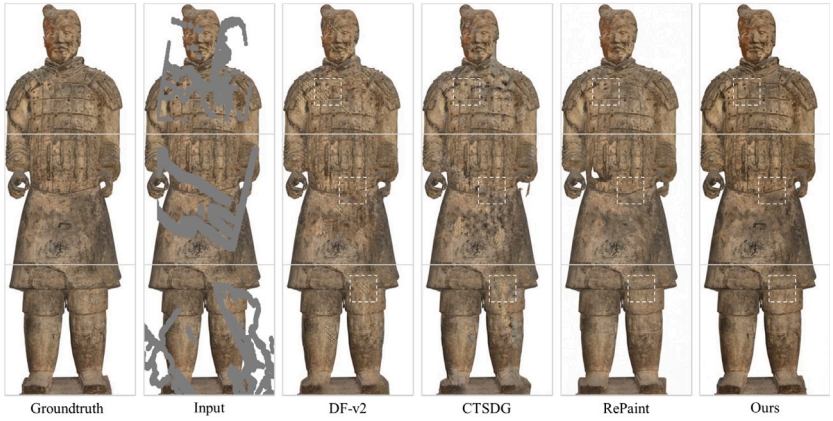

Fig. 12. The complete Terracotta Warriors inpainting with reference, the white box focuses on local differences, zoom in for optimal results.

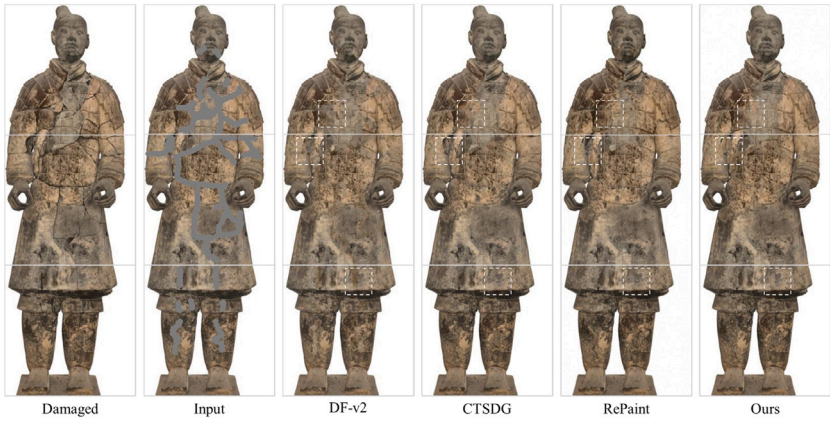

Fig. 13. The damaged Terracotta Warriors inpainting without reference. The white box focuses on local differences, zooming in for optimal results.

Table 8
Comparison of Full-Reference and No-Reference Inpainting Applied to Terracotta Warriors.

Method	Full-Reference			No-Reference		
	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	NIQE ↓	PI ↓
DF-v2	26.096	0.857	0.114	15.505	5.568	3.603
CTSDG	24.824	0.848	0.111	18.543	6.202	4.015
RePaint	26.250	0.900	0.069	14.532	5.957	3.636
Ours	**29.211**	**0.929**	**0.048**	**11.141**	**3.511**	**2.895**

dataset. The inpainting of complete Terracotta Warriors with reference is illustrated in Fig. 12, while the inpainting of damaged Terracotta Warriors without reference is depicted in Fig. 13. The corresponding performance metrics are presented in Table 8. Our approach demonstrated superior performance, achieving the highest scores in reference metrics such as PSNR, SSIM, and LPIPS, as well as in non-reference metrics including BRISQUE, NIQE, and PI. Additionally, it is evident from the figures that both DF-v2 and CTSDG exhibit texture-blurring artifacts. In contrast, our method effectively restores intricate features and textures, yielding more precise and more coherent results than RePaint's. This experiment underscores the potential of our approach in effectively restoring significant cultural heritage, thereby contributing to the preservation and appreciation of historical artifacts.

5. Conclusion

The two-stage diffusion inpainting method (TRRS-DM) introduced in this study integrates pixel-spatial resampling (ResampleDiff) and latent-spatial refinement (RefineDiff). This approach adopts lightweight acceleration techniques and latent-spatial diffusion methods. While minimizing computational requirements, it incorporates signal-to-noise ratio weighting and secondary enhancement to maintain restoration quality. As a result, TRRS-DM can achieve high-fidelity image restoration under various occlusion scenarios. Evaluations on multiple public datasets indicate that TRRS-DM outperforms existing advanced methods in several performance metrics. Specifically, it demonstrates outstanding performance and enhanced visual quality, especially in restoring damaged Terracotta Warrior images.

In conclusion, TRRS-DM presents a novel strategy for diffusion-based image inpainting, demonstrating its ability to achieve high-quality restoration with fewer computational resources. This method holds great potential for applications in cultural heritage conservation. For future Terracotta Warriors restoration, we are overcoming more challenges by integrating 3D point cloud and 2D image restoration/registration, using new ATR methods like learning from good and bad samples, and combining damage recognition and classification to advance intelligent 3D restoration. It is expected that TRRS-DM will promote the development of digital restoration techniques applicable to

Fig. 14. Multi-view 2D images obtained from 3D scanned Terracotta Warriors for dataset collection and preprocessing, in preparation for the restoration work.

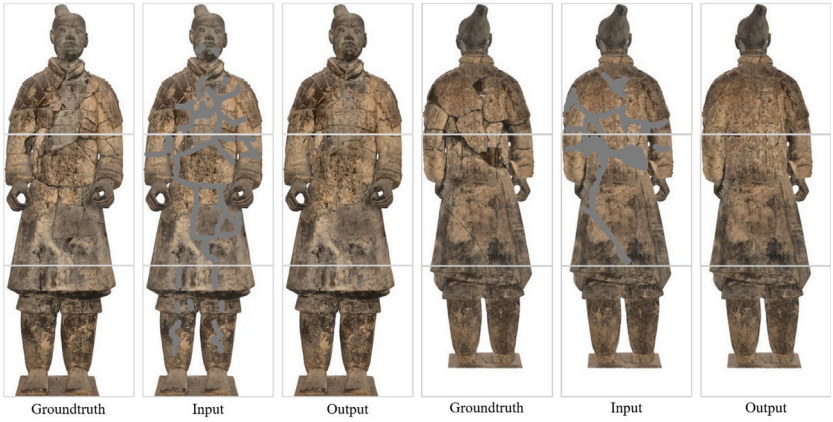

Fig. 15. Restoration results of damaged Terracotta Warriors, showcasing both front and backside restoration. Zoom in for a detailed view.

the Terracotta Warriors, thus providing essential resources for archaeologists and restoration specialists and making significant contributions to the broader field of cultural heritage conservation.

CRediT authorship contribution statement

Xin Cao: Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Peiyuan Quan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation. **Yuzhu Mao:** Validation, Supervision, Project administration, Methodology, Data curation. **Rui Cao:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Data curation, Conceptualization. **Linzhi Su:** Funding acquisition, Formal analysis, Data curation, Conceptualization. **Kang Li:** Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Youth Fund of the National Natural Science Foundation of China (No. 62406247), in part by the National Natural Science Foundation of China (No. 61701403, No. 61806164), and in part by the Key R&D Program of Shaanxi Province (No. 2024SF-YBXM-681, No. 2019GY215, No. 2021ZDLSF06-04).

Appendix

A.1. Dataset of TerraCotta Warriors

The Terracotta Warriors dataset is provided by the National and Local Joint Engineering Research Center for the Digitalization of Cultural Heritage at Northwest University. During the digital protection project, the center deploys an advanced 3D robotic arm to conduct comprehensive scans of the Terracotta Warriors, precisely acquiring their 3D information. Subsequently, multi-view 2D images with a resolution of 4096 × 4096 are generated from the 3D models using professional software (multiple perspectives are shown in Fig. 14).

Strict adherence to cultural heritage protection norms is maintained throughout the scanning process. Rigorous control is exercised over lighting conditions, angles, and scanning duration. A meticulous multi-angled scan of 70 distinct Terracotta Warrior individuals is carried out, amassing a substantial volume of original images. Once the raw data is obtained, screening is promptly initiated based on clarity and integrity criteria. Images that do not meet the requirements are discarded. The remaining images are then cropped in line with experimental needs, and their resolution is adjusted to 256 × 256. Eventually, a usable dataset can be acquired, comprising 4,170 intact and 4000 damaged images.

A.2. More comparative examples

See Figs. 15–17.

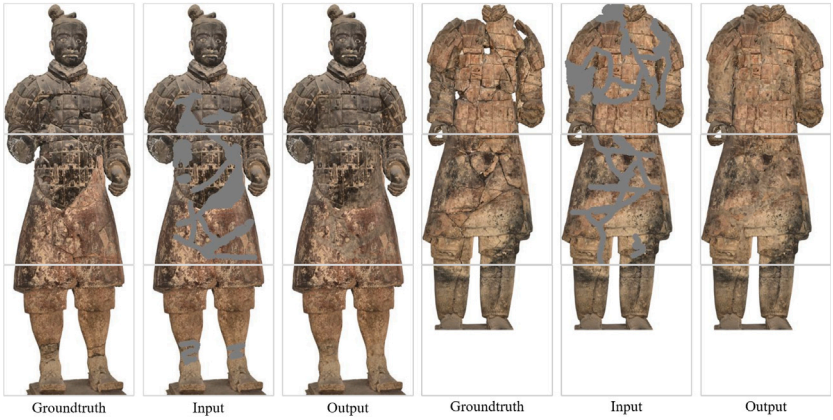

Fig. 16. Restoration of Terracotta Warriors with extensive damage and large missing parts. Zoom in to observe the meticulous restoration work.

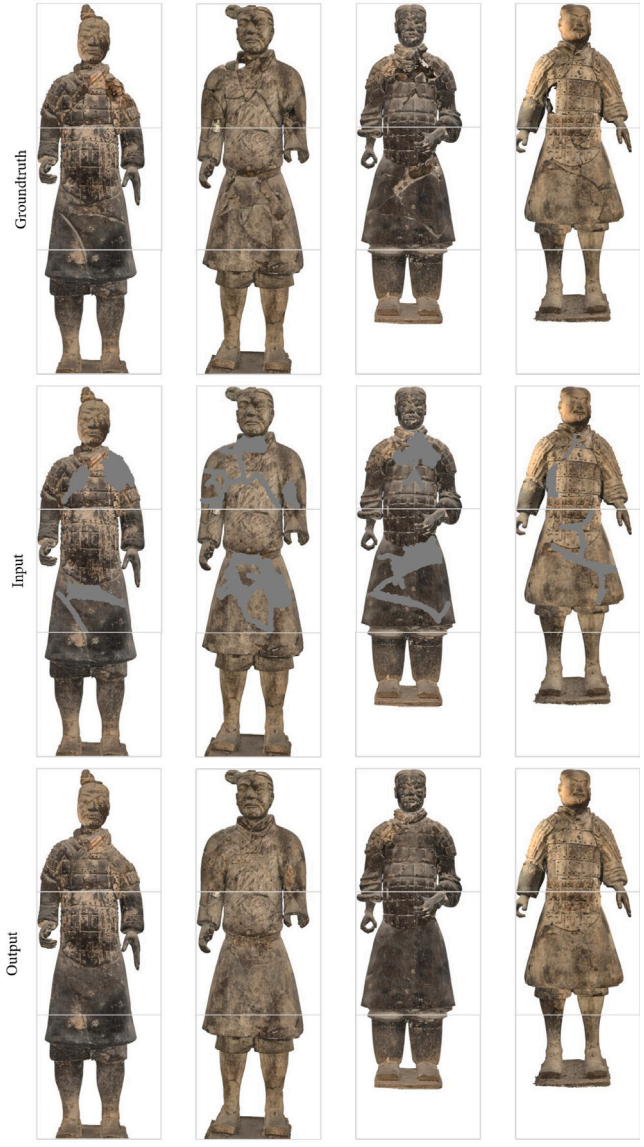

Fig. 17. Additional restoration results of various types of Terracotta Warriors in full body. Explore more detailed restoration effects by zooming in.

Data availability

Data will be made available on request.

References

[1] W. Zilin, The museum of qin shi huang terracotta warriors and horses, Mus. Int. 37 (3) (1985) 140–147.
[2] Z. Qin, Q. Zeng, Y. Zong, F. Xu, Image inpainting based on deep learning: A review, Displays 69 (2021) 102028.
[3] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, J. Verdera, Filling-in by joint interpolation of vector fields and gray levels, IEEE Trans. Image Process. 10 (8) (2001) 1200–1211.
[4] C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, PatchMatch: A randomized correspondence algorithm for structural image editing, ACM Trans. Graph. 28 (3) (2009) 24.
[5] H. Xiang, Q. Zou, M.A. Nawaz, X. Huang, F. Zhang, H. Yu, Deep learning for image inpainting: A survey, Pattern Recognit. 134 (2023) 109046.
[6] D. Ulyanov, A. Vedaldi, V. Lempitsky, Deep image prior, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9446–9454.
[7] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 85–100.
[8] L. Ding, J. Zhang, C. Wu, C. Cai, G. Chen, Real-time image inpainting using PatchMatch based two-generator adversarial networks with optimized edge loss function, in: 2022 IEEE International Symposium on Circuits and Systems, ISCAS, IEEE, 2022, pp. 3145–3149.
[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.
[10] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, C. Miao, Diverse image inpainting with bidirectional and autoregressive transformers, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 69–78.
[11] J. Peng, D. Liu, S. Xu, H. Li, Generating diverse structure for image inpainting with hierarchical VQ-VAE, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10775–10784.
[12] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, L. Van Gool, Repaint: Inpainting using denoising diffusion probabilistic models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11461–11471.
[13] C. Corneanu, R. Gadde, A.M. Martinez, Latentpaint: Image inpainting in latent space with diffusion models, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 4334–4343.
[14] Y. Peng, A comparative analysis between gan and diffusion models in image generation, Trans. Comput. Sci. Intell. Syst. Res. 5 (2024) 189–195.
[15] P. Buyssens, M. Daisy, D. Tschumperlé, O. Lézoray, Exemplar-based inpainting: Technical review and new heuristics for better geometric reconstructions, IEEE Trans. Image Process. 24 (6) (2015) 1809–1824.
[16] D. Tschumperlé, R. Deriche, Vector-valued image regularization with PDEs: A common framework for different applications, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 506–517.
[17] J. Yang, N.I.R. Ruhaiyem, Review of deep learning-based image inpainting techniques, IEEE Access (2024).
[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.

[19] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Trans. Graph. (ToG) 36 (4) (2017) 1–14.
[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Free-form image inpainting with gated convolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4471–4480.
[21] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with fourier convolutions, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2149–2159.
[22] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, M. Ebrahimi, Edgeconnect: Structure guided image inpainting using edge prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[23] Y. Zeng, J. Fu, H. Chao, B. Guo, Learning pyramid-context encoder network for high-quality image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1486–1494.
[24] X. Guo, H. Yang, D. Huang, Image inpainting via conditional texture and structure dual generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14134–14143.
[25] Z. Wan, J. Zhang, D. Chen, J. Liao, High-fidelity pluralistic image completion with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4692–4701.
[26] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, J. Jia, Mat: Mask-aware transformer for large hole image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10758–10768.
[27] S. Xu, W. Xiang, C. Lv, S. Wang, G. Liu, Diversified image inpainting with transformers and denoising iterative refinement, IEEE Access (2024).
[28] Y. Zhang, Y. Liu, R. Hu, Q. Wu, J. Zhang, Mutual dual-task generator with adaptive attention fusion for image inpainting, IEEE Trans. Multimed. 26 (2023) 1539–1550.
[29] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
[30] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.
[31] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[32] A.Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.
[33] H. Liu, Y. Wang, B. Qian, M. Wang, Y. Rui, Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8038–8047.
[34] L. Zhang, X. Du, L. TomyEnrique, Y. Wang, Y. Zheng, C. Jin, Minutes to seconds: Speeded-up ddpm-based image inpainting with coarse-to-fine sampling, in: 2024 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2024, pp. 1–6.
[35] S. Xie, Z. Zhang, Z. Lin, T. Hinz, K. Zhang, Smartbrush: Text and shape guided object inpainting with diffusion model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22428–22437.
[36] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, F. Wen, Paint by example: Exemplar-based image editing with diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18381–18391.
[37] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, Q. Xu, Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion, 2024, arXiv preprint arXiv:2403.06976.
[38] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D.J. Fleet, R. Soricut, et al., Imagen editor and editbench: Advancing and evaluating text-guided image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18359–18369.
[39] Y. Wang, J. Yu, J. Zhang, Zero-shot image restoration using denoising diffusion null-space model, 2022, arXiv preprint arXiv:2212.00490.
[40] B. Kawar, M. Elad, S. Ermon, J. Song, Denoising diffusion restoration models, Adv. Neural Inf. Process. Syst. 35 (2022) 23593–23606.
[41] Z. Yue, J. Wang, C.C. Loy, Resshift: Efficient diffusion model for image super-resolution by residual shifting, Adv. Neural Inf. Process. Syst. 36 (2024).
[42] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, 2020, arXiv preprint arXiv:2010.02502.
[43] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, S. Yoon, Perception prioritized training of diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11472–11481.
[44] P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12873–12883.
[45] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1452–1464.
[46] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.