MagicEraser: Erasing Any Objects via Semantics-Aware Control

Fan Li[1][*], Zixiao Zhang[1][*], Yi Huang[2], Jianzhuang Liu[2], Renjing Pei[1],
Bin Shao[1], and Songcen Xu[1]

[1] Huawei Noah's Ark Lab
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{lifan61, zhangzixiao3, peirenjing, shaobin3, xusongcen}@huawei.com,
{yi.huang, jz.liu}@siat.ac.cn
* Equal Contribution

Abstract. The traditional image inpainting task aims to restore corrupted regions by referencing surrounding background and foreground. However, the object erasure task, which is in increasing demand, aims to erase objects and generate harmonious background. Previous GAN-based inpainting methods struggle with intricate texture generation. Emerging diffusion model-based algorithms, such as Stable Diffusion Inpainting, exhibit the capability to generate novel content, but they often produce incongruent results at the locations of the erased objects and require high-quality text prompt inputs. To address these challenges, we introduce MagicEraser, a diffusion model-based framework tailored for the object erasure task. It consists of two phases: content initialization and controllable generation. In the latter phase, we develop two plug-and-play modules called prompt tuning and semantics-aware attention refocus. Additionally, we propose a data construction strategy that generates training data specially suitable for this task. MagicEraser achieves fine and effective control of content generation while mitigating undesired artifacts. Experimental results highlight a valuable advancement of our approach in the object erasure task.

Keywords: Object erasure · Diffusion models · Attention refocus

1 Introduction

Image inpainting is a long-standing task that originally completes erased or corrupted regions within an image by incorporating information from their surrounding background and foreground. However, our focus extends beyond traditional inpainting to a more nuanced task—object erasure. While traditional inpainting aims to restore missing or damaged parts, our objective is to generate harmonious background after removing specific objects. The generative models utilized in both tasks share certain similarities, prompting us to delve into the evolution of image inpainting. Subsequently, we identify the challenges posed by existing inpainting algorithms when applied to the object erasure task.
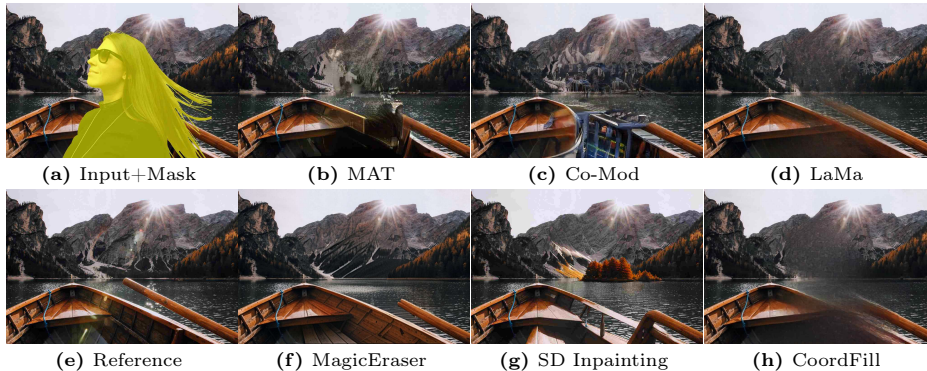
(a) Input+Mask   (b) MAT   (c) Co-Mod   (d) LaMa

(e) Reference   (f) MagicEraser   (g) SD Inpainting   (h) CoordFill

Fig. 1: Comparison with five state-of-the-art inpainting algorithms: MAT [23], Co-Mod [50], LaMa [42], CoordFill [28] and Stable Diffusion (SD) Inpainting [35]. MagicEraser can effectively erase masked objects and achieve the best texture consistency and content fidelity.

Earlier methods in image inpainting, particularly those relying on generative adversarial networks (GANs) [7, 24], encounter challenges in generating high-quality textures for large corrupted regions and struggle with object erasure, which includes approaches like LaMa [42], which introduces large mask inpainting based on fast Fourier convolutions (FFCs), and CoordFill [28], which utilizes parameterized coordinate querying and convolution simplification tricks for efficient high-resolution image inpainting. Both MAT [23] and Co-Mod [50] aim to enhance the performance of inpainting large regions of missing information in images. Despite their advancements, these GAN-based methods still encounter difficulties in generating high-quality textures for complex backgrounds, particularly in large erased regions, as shown in the example of Fig. 1.

Recently, diffusion models [11, 29], including Stable Diffusion [35], DALL-E [33, 34], and Imagen [10], have shown promise in text-to-image generation. Meanwhile, approaches like GLIDE [31] and Stable Diffusion Inpainting [35] (inpainting version of Stable Diffusion) fine-tune diffusion models with random masks, recovering missing regions conditioned on corresponding image captions. Particularly, when applied to inpainting, diffusion models substitute random noise in the background with a noisy version of the original image during the reverse diffusion process. However, this method often yields random and undesirable outcomes due to its heavy reliance on high-quality text prompts. For the example in Fig. 1, SD Inpainting generates a relatively harmonious result based on this long prompt: "*The boat is on a serene lake surrounded by dramatic mountains with rugged textures. The sun is shining directly above the mountain peaks, creating a flare effect in the camera lens. There's a reflection of the sun on the water, suggesting it's a clear day. The trees on the mountainside are tinged with autumn colors, which adds warmth to the scene*". If we use a short prompt

like "*A boat on the lake*", SD Inpainting tends to generate another different boat. Hence, controlling the generation of masked regions becomes a major challenge.

A high-quality image caption is essential for effective control and tends to emphasize global image features, but it is not very user-friendly. There may be semantic misalignment between the locally erased content and the global text description, potentially leading diffusion models to fill the masked regions with object-level foreground rather than the surrounding background, as illustrated by Stable Diffusion Inpainting in Fig. 1. This emphasizes the necessity for a more nuanced and user-friendly approach to the object erasure task.

To address the aforementioned challenges, we introduce MagicEraser, a new user-friendly diffusion model-based framework designed for object erasure. Broadly, we break down the process into two phases: content initialization and controllable generation. In the former, we employ a pretrained traditional inpainting method to initialize the content within masked regions. The latter has two plug-and-play modules named prompt tuning and semantics-aware attention refocus. The prompt tuning module, employing textual inversion [6] and LoRA [13] fine-tuning techniques, primarily aims to preserve the capability of multi-modal understanding without requiring manual input prompts. This dramatically improves the usage for ordinary people in practical applications. On the other hand, the semantics-aware attention refocus module is effective and training-free. It utilizes semantic cues obtained through panoptic segmentation and then adaptively adjusts the attention values of the background and foreground. This adaptive adjustment contributes to enhanced controllability of the generation process. Additionally, different from traditional training data construction for inpainting, we propose a new data construction strategy for fine-tuning the diffusion model. Experimental results highlight a valuable advancement of our approach in the object erasure task across various scenarios.

Our contributions are summarized in the following: (1) We propose MagicEraser, an effective and user-friendly object-erasing framework based on the diffusion model. (2) We introduce a data construction strategy specifically designed for the object erasure task. (3) We present prompt tuning and training-free semantics-aware attention refocus to enhance the controllability of the generation process. (4) Comprehensive experiments validate that MagicEraser achieves state-of-the-art quantitative and qualitative results.

2   Related Work

2.1   GAN-Based Image Inpainting

Object erasure refers to removing objects from an image and restoring the background behind them, and is often considered a context-driven type of image inpainting. Earlier methods for this task predominantly rely on Generative Adversarial Networks (GANs) [7, 42] that are trained on massive datasets. For instance, Co-Mod [50] harnesses the generative capability of unconditional modulation techniques [17, 18] and employs co-modulation of both conditional and

stochastic style representations to handle large-scale missing regions. LaMa [42] utilizes the Fast Fourier Convolutions (FFCs) to extend the network's receptive field across the entire image at early stages, thereby enhancing perceptual quality and facilitating adaptation to high-resolution images that are not seen during training. MAT [23] presents a transformer-based framework designed for high-resolution inpainting. However, its practical application is limited by the inefficient multi-stage structure. CoordFill [28] proposes a more efficient inpainting decoder utilizing an implicit representation with a multi-layer perceptron (MLP) network. These methods primarily excel in scenarios involving simple backgrounds with repetitive textures, such as grass or sky. However, complex backgrounds characterized by inconsistent textures or lighting conditions pose significant challenges, often leading them to generating content that lacks consistency and exhibits noticeable blurriness. This paper tackles the problem based on diffusion models with semantic awareness of context.

2.2   Diffusion Model-Based Image Inpainting

Recent years have seen a growing interest in diffusion models [11, 39–41] across various vision tasks such as image generation [2, 4, 12, 31, 35, 37], editing [1, 8, 15, 30, 43] and restoration [14, 19, 25, 32] due to their superior capacity to capture complex data distributions and more stable training than GANs. Similar to GAN-based approaches, early studies utilizing diffusion models for inpainting primarily focus on leveraging the surrounding context to fill the missing pixels. For instance, Palette [36] trains a diffusion model by directly concatenating masked images with their original versions as input. Repaint [29] blends masked regions generated from a pretrained unconditional diffusion model and unmasked regions from the original images at each sampling step. However, these methods often fall short in offering precise control over the generated content.

With the advance of text-to-image (T2I) diffusion models, this limitation is being mitigated by incorporating additional conditions, such as text, segmentation maps, and reference images. For instance, Stable Diffusion Inpainting, an adaptation from Stable Diffusion [35], finetunes the pretrained T2I model using randomly generated masks, masked images, and the captions of original images. This approach, however, sometimes fails to maintain relevance to the text prompts, particularly with small masked regions or when only part of an object is covered. To enhance the precision of inpainted content, SmartBrush [46] introduces a precision factor, enabling the generation of masks ranging from fine to coarse by applying Gaussian blur to accurate instance masks. Imagen Editor [44] extends the Imagen [37] model through finetuning with precise object masks, dynamically generated by an object detector, SSD Mobilenet v2 [38], rather than random masking. Although these advancements improve the fidelity of generated content, they tend to introduce new objects into the masked regions rather than restoring the original background, which is crucial for the object erasure task.

Addressing this challenge, specific approaches have been tailored for precise object erasure. Inst-Inpaint [48] allows removal of objects specified by text instructions, bypassing the need for binary masks. It trains a diffusion model

on the self-constructed GQA dataset comprising source images, their ground truths with objects removed, and text instructions. MagicRemover [47] employs an attention guidance strategy within the diffusion model's sampling process to facilitate the erasure of inpainting regions and the restoration of occluded content. PowerPaint [52] finetunes a T2I model with dual task prompts, P_{obj} for text-guided object inpainting and P_{ctxt} for context-aware image inpainting, where P_{obj} serves as a negative prompt with classifier-free guidance sampling for object removal. Despite these advancements, challenges still remain when they confront complex backgrounds, often resulting in unnaturally generated content. This study seeks to overcome such obstacles through semantics-aware control and the construction of high-quality training data.

3   Preliminaries

3.1   Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [11] define a forward noising process following the Markov chain that transforms a data sample x_0 from its real data distribution $q(x)$ into a sequence of noisy samples x_t in T steps with a variance schedule $\beta_1, \ldots, \beta_T$: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I})$. The closed form of the forward process can be expressed as $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To generate images starting with a noisy sample from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, diffusion models learn to reverse the above process through a joint distribution $p_\theta(x_{0:T})$ that follows the Markov chain with parameters θ: $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$. The parameters θ are usually optimized by a neural network $\epsilon_\theta(x_t, t)$ that directly predicts noise vectors ϵ_t instead of μ_θ and Σ_θ with the following simplified objective [11]:

$$L_{simple} = \mathbb{E}_{x_0, t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[||\epsilon_t - \epsilon_\theta(x_t, t)||^2 \right]. \tag{1}$$

As for conditional diffusion models, $e.g.$, T2I generation and inpainting models, the conditions, $e.g.$, text and mask, can be fed into the network ϵ_θ without changing the loss function. Then the model learns to generate images that are consistent with the conditions.

3.2   Stable Diffusion Inpainting

Stable Diffusion Inpainting, a variant of Stable Diffusion [35], is specifically fine-tuned for the image inpainting task using a randomly generated mask, the corresponding masked image, and the caption of the complete image. This adaptation enables the model to utilize information from the unmasked regions effectively during its training phase. Unlike the original Stable Diffusion which processes a 4-channel noisy latent $z_t \in \mathbb{R}^{h \times w \times 4}$ in the Variational Autoencoder (VAE) [21] latent space, Stable Diffusion Inpainting adapts the first convolutional
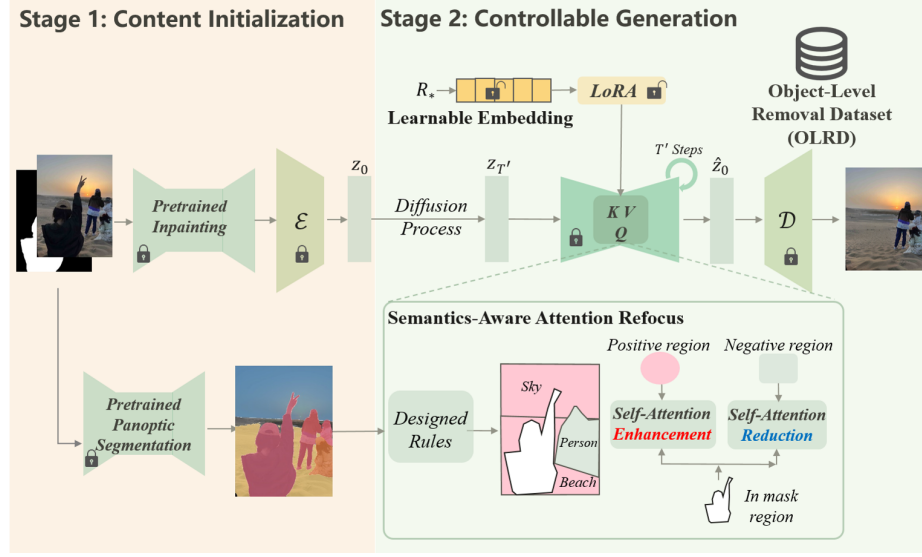
Fig. 2: MagicEraser, built upon Stable Diffusion Inpainting, comprises two main stages: content initialization and controllable generation. Additionally, we construct an object-level removal dataset (OLRD) specifically designed for the object erasure task.

layer of the denoising network to accept a 9-channel input. This expanded input $z'_t \in \mathbb{R}^{h \times w \times 9}$ is the concatenation of the masked image latent $z_{masked} \in \mathbb{R}^{h \times w \times 4}$, $z_t \in \mathbb{R}^{h \times w \times 4}$, and the corresponding randomly generated mask $m \in \mathbb{R}^{h \times w \times 1}$. Therefore, the optimization loss of the 9-channel Stable Diffusion is:

$$L_{9ch} = \mathbb{E}_{z_0, z_{masked}, m, t, y, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[||\epsilon_t - \epsilon_\theta(z'_t, t, \tau(y), m)||^2 \right], \qquad (2)$$

where $\tau(\cdot)$ is a text encoder that maps a text prompt y into a conditional vector.

4   Approach

Given an image x and a binary mask m indicating the target objects for erasure, our objective is to generate an image $\hat{x}$ where the masked regions are seamlessly replaced with harmonious background without introducing new foreground objects. Furthermore, we aim for this erasing process to be achieved without the necessity for extra manual text prompt input.

4.1   Overall Framework

Existing diffusion model-based inpainting models, similar to their text-to-image counterparts, often heavily rely on high-quality text prompt input [30,35], which is not intuitive to obtain, especially for ordinary users. These models demonstrate

limitations in understanding multi-modal inputs, struggling to interpret semantic information and to achieve seamless background completion. Therefore, we propose a diffusion model-based framework, MagicEraser, as highly suitable and user-friendly for the object erasure task. Illustrated in Fig. 2, MagicEraser comprises two main phases: content initialization and controllable generation. The former initializes the content of the erasure regions, while the latter governs denoising generation based on learnable text prompts and training-free semantics-aware attention refocus. Additionally, we propose a specialized data construction strategy for fine-tuning the diffusion model.

4.2   Content Initialization

Latent initialization plays a crucial role in high-resolution image generation, particularly within the latent diffusion model (LDM) [35]. The noising and denoising processes typically take places in the latent space $\mathcal{Z}$, with a parameter called denoising strength ($s \in (0, 1]$) controlling the entire procedure. Specifically, in the sampling process of the diffusion model with a maximum of T sampling steps, the actual number of sampling steps is given by $T' = \lfloor T \cdot s \rfloor$. This means that the denoising process starts from $z_{T'}$, which can be calculated as:

$$z_{T'} = \sqrt{\bar{\alpha}_{T'}}z_0 + \sqrt{1 - \bar{\alpha}_{T'}}\epsilon, \tag{3}$$

where $z_0 = E(x_0)$, x_0 is the given image and E is the encoder of Variational Autoencoder (VAE) [21] within LDM.

When $s = 1$, the generation starts from standard Gaussian noise $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, often resulting in significant deviations from the original image x_0. In the object erasure task, where the objective is to generate a harmonious background, a smaller s (e.g., $s = 0.75$) is capable of enhancing texture harmony. However, this can lead to the generation of undesired new objects similar to those being erased in the masked regions.

To address this problem, we employ a pretrained traditional inpainting model such as LaMa or CoordFill to roughly initialize the content of the erasing regions in the pixel space. The pre-processed image $\tilde{x}_0$ is then passed through the VAE encoder to obtain $\tilde{z}_0 = E(\tilde{x}_0)$. Subsequently, the initial noisy latent vector $z_{T'}$ can be calculated by Eq. 3. Under this initialization, we set $s = 0.9$ in MagicEraser, maintaining fine and harmonious texture while mitigating the generation of undesired artifacts.

4.3   Controllable Generation

Prompt Tuning. The growing need for object removal in photography is primarily driven by ordinary people. They often lack the expertise to acquire professional-grade prompts (see the footnote on page 2), which are essential for accurately directing diffusion models to remove unwanted objects. To address this, we design a prompt tuning method for object erasure based on our object-level removal dataset (OLRD[3]) detailed in Section 4.4, which only tunes a small

---

[3] https://github.com/lifan724/magic_eraser

amount of the parameters added to the U-Net in Stable Diffusion Inpainting, avoiding destroying the capability of the pre-trained model.

Specifically, our objective is to obtain a tuned prompt, which can teach the model a new concept, "background completion", using Textual Inversion [6]. We designate a placeholder string "R_*" to represent this new concept, whose corresponding token embedding added to the vocabulary is denoted as v_*. We initialize v_* using Textutal Inversion on a small random subset of OLRD. To condition the generation, we utilize the background tag (e.g., "sky" or "beach") to obtain a short text prompt y in the form of "A photo of R_* sky" or "A photo of R_* beach", where the tag is from the results of a pretrained panoptic segmentation algorithm. Note that R_* can be considered as a *universal* "background completion" concept that is expected to force the model to focus more on background generation (e.g., "sky" or "beach", etc.). If we only use the text prompt without R_* (e.g., "A photo of sky", "A photo of beach", etc.), the diffusion model tends to generate new objects similar to those to be erased.

We find that relying solely on Textual Inversion is not enough to capture this intricate concept. So we further tune it together with the model fine-tuning on OLRD using the low-rank method LoRA [13]. The additional parameters ϕ of LoRA are added to the U-Net of the diffusion model and simultaneously trained with v_*, enhancing the model's understanding of the concept of object erasure. The optimization of the diffusion model fine-tuning is then defined as:

$$v_*, \phi_* = \arg\min_{v,\phi} \mathbb{E}_{z_0, z_{masked}, m, t, y, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[||\epsilon_t - \epsilon_{\theta,\phi}(z'_t, t, \tau(y), m)||^2 \right]. \quad (4)$$

Additionally, to avoid degrading the generation quality by only using the above simple text prompt during fine-tuning (e.g., "A photo of R_* sky"). We only use them with a 50% chance and use image captions detailed in Section 4.4 with another 50% chance. During inference, the model only uses the above simple text prompts that are automatically constructed by the panopatic segmentation algorithm and the learned R_* (v_*), without the need of user input.

Semantics-Aware Attention Refocus. The self-attention layers in Stable Diffusion are crucial components that reorganize intermediate features to ensure globally coherent generated content. Previous research [5, 20, 47] has demonstrated that appropriately modulating the self-attention layers can enhance the controllability of T2I models. In the context of object erasure, pixels outside the mask can be considered as a type of "visual prompts", influencing the content generation within the mask. Therefore, modulating self-attention layers to focus more on desired regions outside the mask and to ignore undesired ones can improve the generation of coherent content and suppress the generation of incongruent content. Consequently, we propose a training-free semantics-aware attention refocus module, which utilizes semantic cues obtained through panoptic segmentation as guidance for modulating the self-attention layers. Our experiments show that this module significantly enhances the controllability of our diffusion model, thereby boosting the quality of generated images.
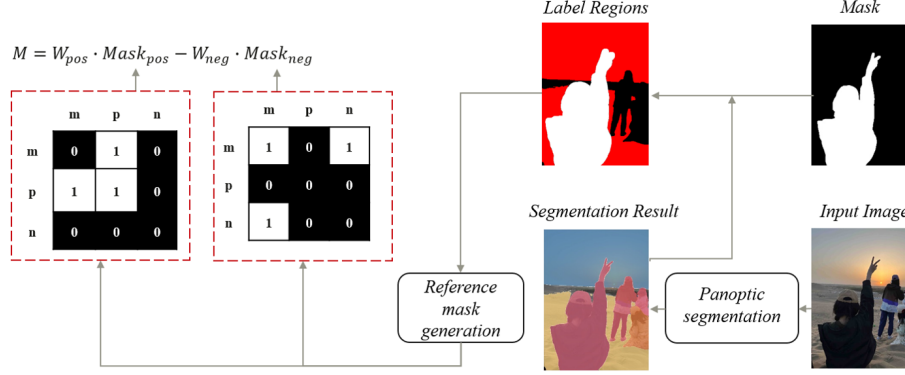
Fig. 3: Semantics-aware attention refocus. We combine the panoptic segmentation result of the input image with the input mask to generate $Mask_{pos}$ and $Mask_{neg}$. With the input mask and the panoptic segmentation results, we obtain the labels (l) of different regions (white for *mask* (m) regions, red for *positive* (p) regions and black for *negative* (n) regions).

Unlike [5, 47] which use well-designed losses to optimize attention maps, we opt for a direct way to modify them. Inspired by [20], which modulates the attention values by:

$$A' = \text{softmax}(\frac{QK^T + M}{\sqrt{d}}), \tag{5}$$

we design M as follows:

$$M = W_{pos} \cdot Mask_{pos} - W_{neg} \cdot Mask_{neg}, \tag{6}$$

where the binary masks $Mask_{pos}, Mask_{neg} \in \mathbb{R}^{|\text{queries}| \times |\text{keys}|}$, indicating which self-attention values should be modulated, and $W_{pos}, W_{neg} \in \mathbb{R}^{|\text{queries}| \times |\text{keys}|}$ are their corresponding modulation weights.

We design $Mask_{pos}$ and $Mask_{neg}$ to be semantics-aware by utilizing the panoptic segmentation results, as shown in Fig. 3. Specifically, based on their semantic categories and the objects to be erased, we assign each latent pixel with a label $l \in \{mask, positive, negative\}$ (m, p, n for short) standing for *mask* regions, *positive* regions and *negative* regions, respectively. Here, a positive region is one whose semantics belong to background, while a negative region is one whose semantics are similar to the objects to be erased. During denoising process, using Eq. 5, we increase the self-attention values of the *mask* regions with the *positive* regions while decreasing them with both the *negative* regions and the *mask* regions. Let $l[i]$ be the label of pixel i. Then for each query pixel i and key pixel j in the self-attention maps, we define

$$Mask_{pos}[i, j] = \begin{cases} 1, & \text{if } (l[i] = m \text{ and } l[j] = p) \text{ or } (l[i] = p \text{ and } l[j] \in \{m, p\}), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

Original Image I Mask m Masked $\tilde{I}$

(a)

Original Image I Mask m Object o Shifted Mask $\tilde{m}$ Shifted Object $\tilde{o}$ Blended Image $\tilde{I}$
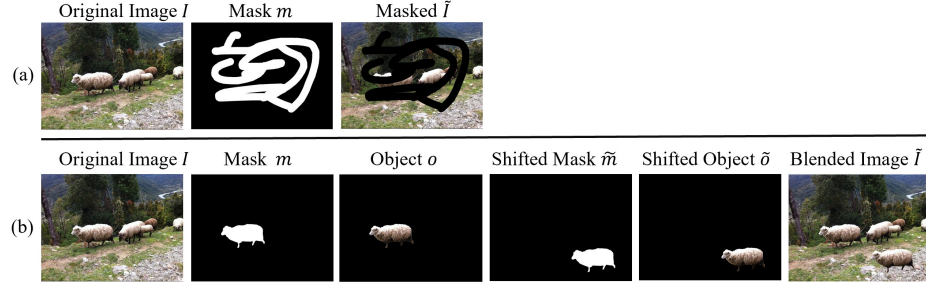
(b)

Fig. 4: Training data comparison between the traditional inpainting and object erasure. (a) Traditional inpainting methods use random mask m and the masked image $\tilde{I}$ to recover the original image I. (b) Our model uses the shifted mask $\tilde{m}$ and the blended image $\tilde{I}$ to recover I.

$$Mask_{neg}[i,j] = \begin{cases} 1, & \text{if } (l[i] = m \text{ and } l[j] \in \{m, n\}) \text{ or } (l[i] = n \text{ and } l[j] = m), \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

As for W_{pos} and W_{neg}, we first apply max and min operations to the similarity matrix QK^T, obtaining the maximum and minimum values for each query, then replicate these values along the key-axis to obtain final $S_{max}, S_{min} \in \mathbb{R}^{|\text{queries}| \times |\text{keys}|}$, and finally define

$$W_{pos} = (1 - \lambda_{pos}) \cdot S_{min} + \lambda_{pos} \cdot S_{max}, \quad W_{neg} = \lambda_{neg} \cdot S_{max}, \quad (9)$$

where λ_{pos} and λ_{neg} are empirically set to 0.8 and 1.0, respectively. In addition, as discussed in [20], we only modulate the self-attention layers at the initial denoising steps ($t = 1 \sim 0.7$).

4.4   Training Data Construction and Model Finetuning

As shown in Fig. 4(a), traditional inpainting methods often generate random mask m and recover the original image I from the masked image $\tilde{I}$. However, the objective of the erasure task is to erase objects and generate a harmonious background. Because there is currently no large-scale object-level removal dataset suitable for object erasure training, we propose a new data construction strategy and build an object-level removal dataset (OLRD) based on Place2 [51].

Specifically, given an original image I, we utilize a pretrained panoptic segmentation network, such as Mask2Former [3], to label the entire image. We then randomly select an object o (e.g., "sheep" in Fig. 4(b)). Subsequently, the object o and its mask m are shifted to a region marked as background (e.g., "grass" or "gravel") based on the segmentation results, obtaining in $\tilde{o}$ and $\tilde{m}$. Finally, $\tilde{o}$ is blended into the original image I to generate $\tilde{I}$:

$$\tilde{I} = \tilde{o} + \tilde{m} * I. \quad (10)$$

Our MagicEraser is obtained by fine-tuning Stable Diffusion Inpainting using $\tilde{m}$ and $\tilde{I}$ with the ground truth I. In practice, common data augmentation tricks such as scaling and rotation and color change can be applied to $\tilde{o}$. In this work, we do not perform them.

Additionally, to obtain the textual description of I, we add a prompt to guide a Vision-Language model (VLM) (e.g. LLAVA [27]) to focus more on the background regions rather than the objects. For instance, adding a prompt like "Describe the grass and gravel in the image" to the VLM yields a response such as "The grass is green and lush and the gravel is scattered throughout the scene", placing more emphasis on the "grass" and "gravel" regions identified by panoptic segmentation. This textual description of I, together with $\tilde{m}$ and $\tilde{I}$, is utilized to fine-tune the T2I Stable Diffusion Inpainting model for object erasure. The optimization of the model fine-tuning using LoRA is represented in Eq. 4.

5   Experiments

5.1   Experimental Setup

Implementation Details. Our framework, MagicEraser, is built on the Stable Diffusion Inpainting model v1.5[4]. As for the content initialization module, a traditional pretrained inpainting model Big-LaMa[5] is leveraged. And we apply Mask2Former[6] to obtain the panoptic segmentation results for the semantics-aware refocus module. We use Adam optimizer with the learning rate being 1e-4 in the prompt tuning process, which takes around 50K steps. The training data is constructed based on Place2 [51], where the erasing masks are also generated by Mask2Former and the image captions are produced by LLaVA[7]. All images and their corresponding masks are resized to 512×512 during training.

Evaluation Datasets and Metrics. We assess the performance of our MagicEraser on three different datasets: OpenImages [22], COCO [26] and RealHM [16]. From OpenImages and COCO, we respectively sample 200 representative examples with side resolution higher than 512 and then construct the erasing pairs with the strategy detailed in Section 4.4. As for RealHM, it is collected for self-supervised image harmonization, containing 215 high-quality examples with side resolution higher than 4000. Specifically, every example of RealHM already contains an original image I, an object mask m and a blended result $\tilde{I}$ which can directly be applied to assess the performance of the erasure task by using $\tilde{I}$ and m to recover I. During testing, every image is transformed to the size of 512×512 by linearly mapping its long side to 512 and then padding the short side with values 0. By comparing the erasure results with their corresponding reference images (i.e., the original images), we report PSNR [45], SSIM [45], LPIPS [49], and FID [9] as the quantitative evaluation metrics.

---

[4] https://github.com/runwayml/stable-diffusion
[5] https://github.com/advimman/lama
[6] https://github.com/facebookresearch/Mask2Former
[7] https://github.com/haotian-liu/LLaVA

Table 1: Quantitative comparison with five SOTA methods on three datasets.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
Open-Images	MAT [23]	<u>26.994dB</u>	**0.949**	**0.030**	31.30
	Co-Mod [50]	26.446dB	0.941	0.033	<u>30.40</u>
	LaMa [42]	21.618dB	0.936	0.055	37.10
	CoordFill [28]	22.072dB	0.934	0.081	35.94
	SD Inpainting [35]	26.096dB	0.942	0.036	31.10
	MagicEraser	**28.123dB**	<u>0.947</u>	<u>0.032</u>	**30.02**
COCO	MAT [23]	<u>24.758dB</u>	<u>0.903</u>	0.056	<u>41.33</u>
	Co-Mod [50]	19.444dB	0.757	0.101	43.33
	LaMa [42]	20.675dB	0.897	0.087	44.24
	CoordFill [28]	20.966dB	0.897	<u>0.094</u>	46.68
	SD Inpainting [35]	22.248dB	0.892	0.079	42.85
	MagicEraser	**24.766dB**	**0.908**	**0.062**	**39.55**
RealHM	MAT [23]	21.484dB	0.843	<u>0.107</u>	51.73
	Co-Mod [50]	20.777dB	0.801	0.117	54.43
	LaMa [42]	19.053dB	0.825	0.150	55.70
	CoordFill [28]	19.239dB	0.827	0.177	56.92
	SD Inpainting [35]	<u>21.758dB</u>	<u>0.846</u>	0.116	**45.05**
	MagicEraser	**23.620dB**	**0.861**	**0.101**	<u>46.56</u>

5.2   Comparison with State-of-the-Arts

To evaluate the effectiveness of MagicEraser, we conduct a comprehensive comparison with state-of-the-art (SOTA) methods in the field of image inpainting, including four traditional GAN-based approaches (MAT [23], Co-Mod [50], LaMa [42] and CoordFill [28]) and a diffusion model-based method SD Inpainting [35]. We utilize LLaVA [27] to craft detailed textual prompts for SD Inpainting. Table 1 lists the quantitative results of the compared methods across four metrics. It shows that MagicEraser outperforms others by a large margin in terms of PSNR and obtains competitive performance in SSIM, indicating its superior effectiveness in erasing objects and recovering backgrounds. Furthermore, MagicEraser excels in LPIPS and FID, demonstrating its capability to maintain visual reality and aesthetic quality while effectively removing objects from images. This conclusion can also be validated in the visual comparison of Fig. 5. Through this comparison, we observe distinct limitations in the performance of other methods. MAT and Co-Mod fail to completely erase the masked objects, resulting in ghost remnants of the objects or the introduction of artifacts (see all the rows). LaMa and CoordFill tend to induce severe blurriness within the masked regions, particularly in scenes with complex textures (see all the rows). While SD Inpainting shows an improvement over the aforementioned GAN-based approaches by avoiding such artifacts and blurriness, it struggles with instability and sometimes generates unwanted elements unrelated to the original back-
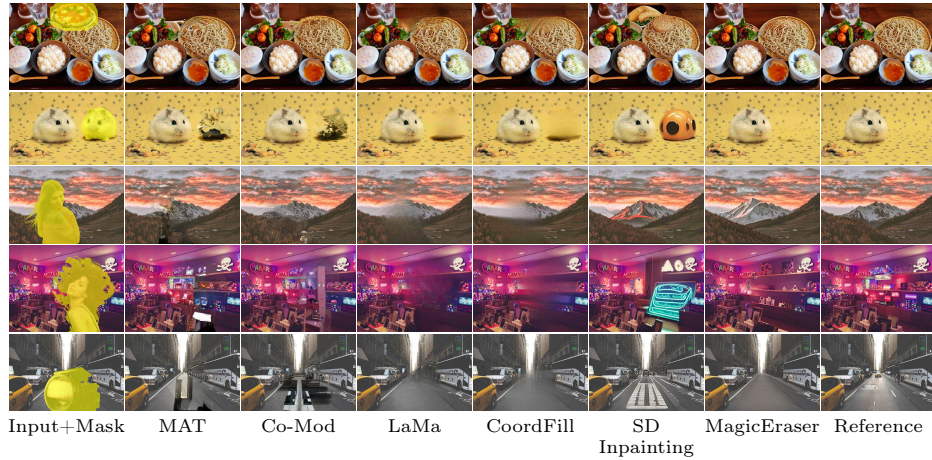
Input+Mask   MAT   Co-Mod   LaMa   CoordFill   SD Inpainting   MagicEraser   Reference

Fig. 5: Visual comparison with five SOTA algorithms.

Table 2: Quantitative comparison with two commercial products on RealHM.

Model	PSNR↑	SSIM↑	LPIPS↓	FID↓
Adobe PhotoShop's Generative Fill	22.913dB	0.851	0.113	49.07
Google Photos Eraser	20.310dB	0.822	0.173	53.55
MagicEraser	**23.620dB**	**0.861**	**0.101**	**46.56**

grounds (see the second row). Compared with them, our MagicEraser can stably generate highly realistic content harmonious with the surrounding context and obtain the most visually pleasing results.

Moreover, we also compare MagicEraser with two commercial products, Adobe Photoshop's Generative Fill[8] and Google Photos Eraser[9]. The quantitative results on the RealHM dataset are shown in Table 2, which demonstrate that our method achieves better performance.

5.3 Ablation Study

We perform a comprehensive ablation study to assess the impact of each component in MagicEraser on RealHM with 512×512 images. The quantitative results are listed in Table 3, where the baseline is Stable Diffusion Inpainting.

Comparing (i) and (ii), we see that our dataset OLRD achieves better performance than traditional random mask training. This is because during model training, the random masking scheme in the traditional inpainting task often leads to recovering the missing regions no matter they are background or objects, which is not suitable for object erasure. Comparing (ii) and (iii), we see

---

[8] https://www.adobe.com/products/firefly.html, May 11, 2024
[9] Google Pixel8 Build Number AP1A.240305.019.A1

Table 3: Ablation study on the RealHM dataset with 512×512 images.

Model	PSNR↑	SSIM↑	LPIPS↓	FID↓
i. Baseline + Traditional Random Mask Traning	21.331dB	0.815	0.134	52.10
ii. Baseline + OLRD	22.130dB	0.834	0.119	50.73
iii. Baseline + OLRD + Content Initialization	22.891dB	0.840	0.109	48.91
iv. Baseline + OLRD + Content Initialization + Semantics-Aware Attention Refocus	23.277dB	0.844	0.110	48.93
v. Baseline + OLRD + Content Initialization + Prompt Tuning	23.311dB	0.858	0.104	47.94
vi. Baseline + OLRD + Content Initialization + Prompt Tuning + Semantics-Aware Attention Refocus (MagicEraser)	**23.620dB**	**0.861**	**0.101**	**46.56**

an increase of PSNR around 0.7dB when content initialization is employed. Because the content initialization utilizes the traditional pretrained GAN-based method to initialize the latent of Stable Diffusion, the model without it generates the images from random noise, which easily leads to unwanted artifacts. Comparing (iii), (iv), and (v), both Prompt Tuning and Semantics-Aware Attention Refocus further improve the model's performance. Moreover, based on the ablation results (iv) and (v) in Table 3, Prompt Tuning is more important than Semantics-Aware Attention Refocus. While both modules help the diffusion model utilize background information to fill masked regions, they work differently. Prompt Tuning globally encodes the semantic clues through the learnable text embedding and LoRA to guide content generation aligned with the overall background concept. Semantics-Aware Attention Refocus locally modulates self-attentions to generate content spatially consistent with the background.

These results demonstrate that the three proposed components are vital to our MagicEraser framework and all have obvious contributions.

6   Limitation and Conclusion

Although notable advantages are demonstrated by our proposed framework, there are still some limitations. Following Stable Diffusion v1.5, MagicEraser works with 512×512 images, where the original high-frequency details of high-resolution images (e.g., 2k, 4k and 8k) may not be preserved. On the other hand, the semantics-aware attention refocus module is sensitive to the results of the pretrained segmentation model. For example, if the background region is not properly segmented, the generated content may appear discordant.

We have proposed a diffusion model-based framework MagicEraser especially suitable for the object erasure task which is recently in increasing demand. It utilizes a traditional inpainting algorithm to roughly initialize the content, and leverages the significant generation capacity of Stable Diffusion by fine-tuning the model with a new dataset OLRD. To further control the generation, we develop a universal prompt tuning module and a semantics-aware attention refocus module. The experiments show that MagicEraser performs best on several datasets compared with several state-of-the-art methods.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
2. Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z.: A survey on generative diffusion models. IEEE TKDE (2024)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
4. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE TPAMI (2023)
5. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation. arXiv preprint arXiv:2306.00986 (2023)
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
8. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2023)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS. vol. 30 (2017)
10. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
12. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR **23**(1), 2249–2281 (2022)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
14. Huang, Y., Huang, J., Liu, J., Dong, Y., Lv, J., Chen, S.: Wavedm: Wavelet-based diffusion models for image restoration. IEEE TMM (2024)
15. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Chen, S., Cao, L.: Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525 (2024)
16. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: ICCV. pp. 4832–4841 (2021)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948 (2019)
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv preprint arXiv:1912.04958 (2020)
19. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. In: NeurIPS. vol. 35, pp. 23593–23606 (2022)
20. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. arXiv preprint arXiv:2308.12964 (2023)

21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
22. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)
23. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. arXiv preprint arXiv:2203.15270 (2022)
24. Li, W., Yu, X., Zhou, K., Song, Y., Lin, Z.: Image inpainting via iteratively decoupled probabilistic modeling. In: ICLR (2024)
25. Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X., Chen, Z.: Diffusion models for image restoration and enhancement–a comprehensive survey. arXiv preprint arXiv:2308.09388 (2023)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
28. Liu, W., Cun, X., Pun, C.M., Xia, M., Zhang, Y., Wang, J.: Coordfill: Efficient high-resolution image inpainting via parameterized coordinate querying. arXiv preprint arXiv:2303.08524 (2023)
29. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR. pp. 11461–11471 (2022)
30. Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Image synthesis and editing with stochastic differential equations. In: ICLR (2022)
31. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML. pp. 16784–16804 (2022)
32. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE TPAMI (2023)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
34. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831 (2021)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
36. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH. pp. 1–10 (2022)
37. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
38. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)
39. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)

42. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: WACV. pp. 2149–2159 (2022)
43. Wang, K., Yang, F., Yang, S., Butt, M.A., van de Weijer, J.: Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. In: NeurIPS (2023)
44. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., et al.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In: CVPR. pp. 18359–18369 (2023)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
46. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: CVPR. pp. 22428–22437 (2023)
47. Yang, S., Zhang, L., Ma, L., Liu, Y., Fu, J., He, Y.: Magicremover: Tuning-free text-guided image inpainting with diffusion models. arXiv preprint arXiv:2310.02848 (2023)
48. Yildirim, A.B., Baday, V., Erdem, E., Erdem, A., Dundar, A.: Inst-inpaint: Instructing to remove objects with diffusion models. arXiv preprint arXiv:2304.03246 (2023)
49. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. arXiv preprint arXiv:1801.03924 (2018)
50. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)
51. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI (2017)
52. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. arXiv preprint arXiv:2312.03594 (2023)