Sketch-guided Image Inpainting with Partial Discrete Diffusion Process

Nakul Sharma[1], Aditay Tripathi[2], Anirban Chakraborty[2], Anand Mishra[1]
[1]IIT Jodhpur [2]IISc, Bengaluru
sharma.86@iitj.ac.in, aditayt@iisc.ac.in, anirban@iisc.ac.in, mishra@iitj.ac.in

Abstract

In this work, we study the task of sketch-guided image inpainting. Unlike the well-explored natural language-guided image inpainting, which excels in capturing semantic details, the relatively less-studied sketch-guided inpainting offers greater user control in specifying the object's shape and pose to be inpainted. As one of the early solutions to this task, we introduce a novel partial discrete diffusion process (PDDP). The forward pass of the PDDP corrupts the masked regions of the image and the backward pass reconstructs these masked regions conditioned on hand-drawn sketches using our proposed sketch-guided bi-directional transformer. The proposed novel transformer module accepts two inputs – the image containing the masked region to be inpainted and the query sketch to model the reverse diffusion process. This strategy effectively addresses the domain gap between sketches and natural images, thereby, enhancing the quality of inpainting results. In the absence of a large-scale dataset specific to this task, we synthesize a dataset from the MS-COCO to train and extensively evaluate our proposed framework against various competent approaches in the literature. The qualitative and quantitative results and user studies establish that the proposed method inpaints realistic objects that fit the context in terms of the visual appearance of the provided sketch. To aid further research, we have made our code publicly available here: https://github.com/vl2g/Sketch-Inpainting.

1. Introduction

Image inpainting is a well-established task in computer vision with diverse applications, including natural photo editing [31, 46, 56] and filling missing data in medical images [1, 2, 47]. Significant progress has been made in image painting in recent years, partly thanks to large neural models [7, 25, 30]. Despite remarkable progress, most current image inpainting methods rely solely on available image regions as context for inpainting. Consequently, these "unconditioned image inpainting methods" lack precise control over semantic object categories or visual attributes such as object shape and pose within the targeted inpainting region. Imagine a scenario where a user intends to inpaint a specific object with precise characteristics like it's size, pose, and shape in the masked image region. While one intuitive approach is to use natural language descriptions for text-guided image inpainting akin to [59], guiding image inpainting using sketch emerges as a promising alternative, especially for users with stronger artistic abilities than linguistic skills. Further, as noted in fine-grained Sketch-based Image Retrieval literature [6, 26], an important characteristic of sketches lies in their ability to capture object appearance and structure intrinsically. This motivates us to propose and study the *sketch-guided image inpainting* as an independent problem. While certain prior studies [22, 28, 56] incorporate partial sketch information for *image manipulation* akin to inpainting, our task distinctively focuses on object-level inpainting utilizing complete object sketches instead of partial sketch strokes. Our goal and a selected result of our approach are illustrated in Figure 1.

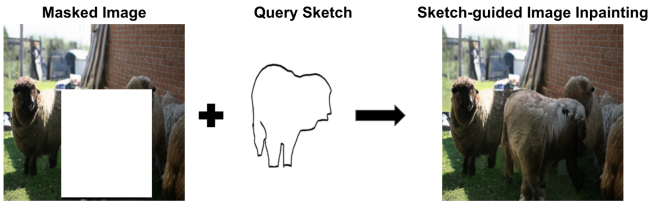

Figure 1. **Sketch-guided Image Inpainting** has been an under-explored task in the literature and is often restricted to partial sketch-based image manipulation [21, 56, 56]. We fill this gap in the literature by proposing a novel partial discrete diffusion process for sketch-guided object-level inpainting. Our proposed approach significantly outperforms other plausible approaches on Sketch-guided Image Inpainting.

Sketch-guided image inpainting requires precise utilization of object-level shape and pose information conveyed by the query sketch while dealing with the large domain gap between hand-drawn sketches and images. We approach this problem by proposing a novel method based on the discrete diffusion process [3] to inpaint the missing regions

conditioned on the hand-drawn query sketch. Our approach involves two main stages: In the first stage, we learn a visual codebook to describe a discrete latent space of images. This codebook enables us to represent images compactly. In the second stage, we model the sketch-guided image inpainting problem in this discrete latent space. Here, we introduce a novel *Partial Discrete Diffusion process* or PDDP tailored for sketch-guided inpainting, which allows for controlled corruption of the image region and subsequent reconstruction guided by the query sketch. Specifically, we propose a *sketch-guided bi-directional transformer* model to reverse the diffusion process, thereby effectively inpainting the missing regions based on the provided sketch. During inference, the inpainting process reduces to reversing the partial discrete diffusion process for the masked area of the image, guided by the user-provided sketch. This enables our method to generate high-quality inpainted images that faithfully capture the intended object shapes and poses specified by the input sketches.

Given the absence of a suitable dataset that can be used to study sketch-guided image inpainting, we curated a dataset specifically for this task by leveraging the rich annotations available in the MS-COCO dataset [27]. Specifically, we segment out objects from the images and sketchify them using an off-the-shelf model [24]. We perform extensive experiments on our curated dataset and compare existing image inpainting approaches adapted for our task with ours.

In summary, our contributions are as follows: (i) We study the task of sketch-guided image inpainting, which involves completing the missing region in an image while considering the shape and pose details of the object provided in the accompanying hand-drawn sketch. This work can be seen as a first attempt at studying sketch guidance in image painting at the object level. (ii) To tackle this challenging problem, we first learn the latent representation of the natural images and model the forward diffusion process only on the masked image regions. We propose the Partial Discrete Diffusion Model to learn the sketch conditional reverse diffusion process to complete the image by incorporating the visual information from the provided hand-drawn sketch. (iii) We compare the performance of our model with suitable baselines and establish a new state-of-the-art for our proposed sketch-guided image inpainting task. We have made our code publicly available here: https://github.com/vl2g/Sketch-Inpainting.

2. Related Work

In recent years, we have seen rapid progress in the deep learning applications for the sketch domain; these include sketch-based image retrieval [6, 26], object localization [48, 49], sketch generation [51], scene-level sketch-based image retrieval [15], etc. In this work, we shall focus on inpainting and sketch-to-image generation literature.

Image Inpainting: Traditional approaches of image inpainting rely on propagating low-level features from surrounding image content to reconstruct the missing regions [5, 10]. More recently, deep learning-based methods have significantly progressed by leveraging semantic image representations learned by convolutional neural networks. Context encoders [37] introduced an encoder-decoder architecture to generate the contents of an irregularly shaped hole based on the surrounding image context. Subsequent works have expanded on this approach with attention mechanisms [55], adversarial training [56], and improved network architectures [32]. Another paradigm leverages diffusion models, which can synthesize high-quality images by learned reverse diffusion processes [18, 44]. RePaint [31] exploits pre-trained unconditional DDPMs [18] to improve the inpainting process by diffusion models.

Guided image inpainting began with Zhang et al.'s work [59] via TDANet, employing a dual attention mechanism to utilize textual cues for inpainting by comparing text with the corrupted and original images. Diffusion-based text-to-image generation models like [12, 16, 34, 40–42] can be used directly for text-guided image inpainting; however, these methods can produce sub-optimal results as they are trained on image-level captions instead of object instance-specific descriptions. Recently, [33] proposes a text-guided image inpainting framework leveraging a defect-free VQ-GAN version for improved inpainting results. Zeng et al. [57] proposed shape-guided object inpainting in images and subsequently, [35, 54] proposed frameworks to guide the inpainting process using the shape of the mask region along with the text. Our problem setup involves providing a user sketch of an object which should be inpainted. This gives the user better control over specifying the object's shape, pose, and size.

Sketch-to-Image Generation: Sketch2Photo [8] and PhotoSketching [24] synthesized whole images by compositing the retrieved foreground and background images using a given sketch. Gao et al. [14] introduced a two-stage method using EdgeGAN to generate realistic images from scene-level sketches. Initial works on object-level image generation from sketches include [9, 29]. More recent works AODA [53] and [23] propose methods for open-set object-level image generation from sketches. With the rise of text-to-image diffusion models, interest has been in controlling the generation using sketches. Voynov et al. [50] trains a latent-guided predictor module that maps latent features of noisy images to spatial maps for providing sketch-guidance to text-to-image diffusion models. ControlNet [60] utilizes pre-trained StableDiffusion [42] by learning parallel architectures for different modalities (edge, scribble, depth maps, pose, etc.) and uses it along with modality-specific guidance to control the image generation process.ControlNet can guide the StableDiffusion model using scribbles for

scene-level image synthesis. Since diffusion models are capable of doing inpainting inherently [18, 34, 44], we can utilize ControlNet for our task as a plausible approach by providing sketch input of only the region that needs to be inpainted. We experimentally compare against such a baseline in Section 4.

Sketch-based Image Manipulation: Previous research on sketch-based image manipulation is primarily based on a conditional image inpainting framework. For instance, DeepFill-v2 [56] enables the manipulation of both general and facial images via partial sketches. Meanwhile, FaceShop [38] permits localized shape and color adjustments in facial images through sketch-based manipulation accomplished via conditional image completion. SC-FEGAN [21] delves into facial manipulation via sketches and color strokes by integrating free-form masks and style loss into the image completion model. Yang et al. [20] adapt a face manipulation model trained on sketches generated by edge detection to human-drawn sketches, and propose a refinement strategy that dilates and refines user-drawn sketches to resemble edge detection results. These methods combine an input image and these low-level controls for CNN inputs. However, the corresponding feature representations are not sufficient to convey user intentions. DeFLOCNet [28] deals with this problem by proposing a new architecture capable of preserving these control features in the deep feature representations. SketchEdit [58] introduces a mask-free image manipulation framework using partial strokes. While our work draws inspiration from these methods, we focus on object-level inpainting.

3. Methodology

Given an image I with missing regions defined by a mask M and a hand-drawn sketch S containing high-level details of an object's appearance, our objective is to generate a completed image I' that contains the generated object that precisely follows the visual, e.g., shape, pose information provided in the hand-drawn sketch. To achieve this, a two-stage methodology is proposed in this work. In the first stage, images are represented as a sequence of codebook indices following [13]. In the second stage, the codebook representation of images is utilized to inpaint using a partial discrete diffusion approach conditioned on the hand-drawn sketches. In the next section, we formally introduce the problem statement of sketch-guided image inpainting (Section 3.1), briefly explain the discrete diffusion model (Section 3.2), and finally describe our approach (Refer to Section 3.3) to address the problem.

3.1. Problem Setup

Let $I \in \mathbb{R}^{H\times W\times 3}$ be a three-color channel input image, where H and W are the height and width of the image, respectively. Further, let $M \in \mathbb{R}^{H\times W\times 1}$ be a binary mask that indicates the missing regions of the input image. Each element $M_{ij} \in \{0,1\}$ on the binary mask M represents whether the pixel at the location (i,j) on the image I is missing. Let $S \in \mathbb{R}^{H_s\times W_s\times 1}$ be a free-hand sketch that provides high-level visual guidance for the appearance and pose of an *object* in the missing regions. In the proposed sketch-guided image inpainting task, we aim to learn a generative model G that takes an image I with the missing region represented by the mask M and a hand-drawn sketch S as input and outputs a completed image $I' \in \mathbb{R}^{H\times W\times 3}$ such that the object inpainted in the missing region is visually consistent with other regions of the image and is visually coherent with the provided hand-drawn sketch.

3.2. Preliminary: Discrete Diffusion

D3PM [3] introduced a general diffusion framework in discrete space for categorical variables. We will first describe the forward diffusion process for a discrete diffusion model with total time-steps $T \in \mathbb{N}$. For a discrete random variable at time $t \in [1,T]$, $z_t \in \{1,2,3,\ldots,C-1,C\}$, the transition matrix $\mathbf{Q_t} \in [0,1]^{C\times C}$ defines transition probabilities associated with each state that z_t can take. More formally, it defines the probabilities that z_{t-1} transits to z_t, $[Q_t]_{mn} = q(z_t = m|z_{t-1} = n)$ in a single time step. It is mathematically described as $q(z_t|z_{t-1}) = v(z_t)^\top \mathbf{Q_t} v(z_{t-1})$, where $v(.)$ denotes a function that encodes a nominal value to one-hot vector over C categories, i.e., $v(z) \in \{0,1\}^C$. Assuming the Markov property, the t step transition probabilities can be obtained as $q(z_t|z_0) = v(z_t)^\top \bar{\mathbf{Q}}_\mathbf{t} v(z_0)$, where $\bar{\mathbf{Q}}_\mathbf{t} = \prod_{i=t}^{1} \mathbf{Q_i}$. This analysis can be extended to the N-dimensional random variables $z_t \in \{1,2,3,\ldots,C\}^N$, and the transition matrix is used for each variable in the random vector z_t independently. From here onwards, we consider z_t as an N-dimensional random variable representing N discrete tokens. D3PM, inspired by the masked language modeling task in NLP, proposes *absorbing state* formulation of this transition matrix and introduces a `[MASK]` token and argues that this special token helps identify corrupted and non-corrupted regions. VQ-Diffusion [16] advances this formulation of matrix $\mathbf{Q_t}$ with their mask-and-diffuse strategy for image generation by introducing three probabilities γ_t of replacing the current token with the `[MASK]` token, β_t of replacing the current token with another token, and α_t describing the probability of token to retain its state. The transition matrix $\mathbf{Q_t} \in [0,1]^{(C+1)\times(C+1)}$ is then given by:

$$\mathbf{Q}_t = \begin{pmatrix} \alpha_t+\beta_t & \beta_t & \ldots & \beta_t & 0 \\ \beta_t & \alpha_t+\beta_t & \ldots & \beta_t & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta_t & \beta_t & \ldots & \alpha_t+\beta_t & 0 \\ \gamma_t & \gamma_t & \ldots & \gamma_t & 1 \end{pmatrix}. \quad (1)$$

The reverse process is parameterized by a neural network that models the z_{t-1} distribution given z_t. Specifically, z_{t-1}

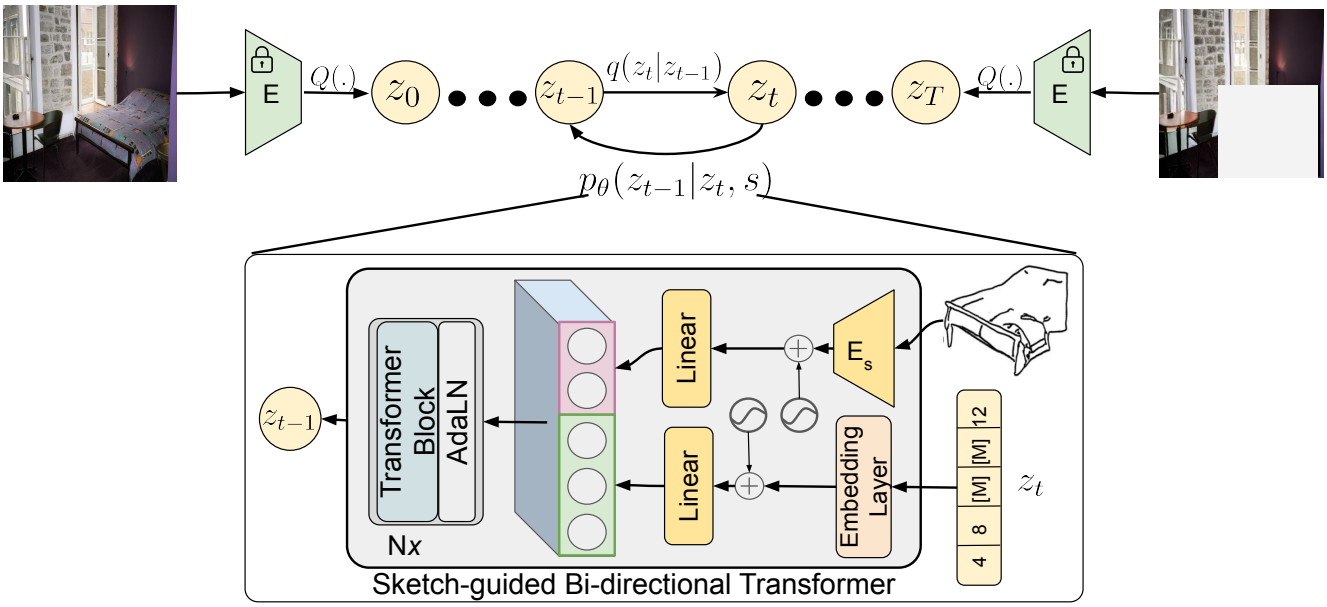

Figure 2. Our method involves obtaining a discrete latent space representation of the original image and its masked counterpart using a pretrained VQ-VAE. The image is first converted to noise by iteratively adding noise to the masked region in the forward process of the proposed Partial Discrete Diffusion Process (Section 3.2). In the reverse process, sketch-guided inpainting is performed iteratively using a sketch-guided bi-directional transformer model that takes the masked image tokens and the query sketch. It predicts the tokens of the missing regions (Section 3.3.2). By iteratively refining the inpainted image using the sketch and the available information from the original image, the proposed method can generate high-quality inpainted images with correct visual and pose details. (Best viewed in color).

is sampled from $p_\theta(z_{t-1}|z_t) \in [0,1]^{(C+1)\times(C+1)}$. During inference, image synthesis tasks initialize all tokens of z_T as `[MASK]`, and then iteratively sample denoised latents from $p_\theta(z_{t-1}|z_t)$ to obtain z_0 [3, 16].

3.3. Sketch-guided Image Inpainting using Partial Discrete Diffusion Process

This section describes the two-stage model we have designed for the sketch-guided image inpainting task. In the first stage, we train an encoder E, a codebook Z, and a decoder D to learn a perceptually compressed discrete latent space of images using the method described by Esser et al. [13]. Any image can then be represented as a sequence of indices of latent vectors from the codebook as $z_0 \in \{1, 2, \ldots, C\}^K$, where K denotes the number of visual tokens representing the image in the discrete space. Our novelty lies in the second stage where we first project the ground truth image and the masked image in the discrete latent space using the learned encoder and the codebook and then use these discrete representations in the discrete diffusion process. For each ground truth image $I_{GT} \in \mathbb{R}^{H\times W\times 3}$, we randomly mask an object in the image using mask $M \in \{0,1\}^{H\times W\times 1}$ where $M(i,j) = 1$ means that the image regions $I_{GT}(i,j)$ is masked. Let $S \in \mathbb{R}^{H\times W\times 1}$ be the hand-drawn sketch corresponding to the masked object containing visual details of the object to be inpainted. Since the image encoder $E(.)$ is a CNN model, the latent code related to any patch in an image is affected by the pixel values of its neighbors. Hence, directly masking the input image in the pixel space $I_M = M \odot I_{GT}$ is erroneous for obtaining latent representation (where $\odot$ represents element-wise multiplication). A better way to obtain the masked image I_M is to firstly encode I_{GT} to a sequence of latent codebook entries, z_0, and then to replace tokens corresponding to masked regions with a special `[MASK]` token. The original image mask M is transformed to obtain the mask for the discrete latent image representation $M_L \in \{0,1\}^K$ that represents masked tokens in the latent space. It is important to note here that an embedding corresponding to the `[MASK]` token does not exist in the codebook Z, but since we aim to represent the image as a sequence of indices, we assign the index $(C+1)$ to this special token. To sum up, we encode the image I_M in discrete latent space as $z_m = (C+1)M_L + (1 - M_L) \odot z_0$.

3.3.1 PDDP: Partial Discrete Diffusion Process

In order to obtain the denoised latent image representation z_0 from the noisy representation z_m, iterative denoising of z_m is performed until we have a visually plausible z_0 [16]. Yet, it does not allow explicit training for the inpainting problems as it does not align with the inference where we have to diffuse from an intermediate state with the desired masked region to obtain z_0. In this work, we introduce a novel inpainting model called Partial Discrete

Diffusion (PDD), which can be used to train general inpainting models in the discrete latent space. With PDD, we aim to align the forward and backward processes of discrete diffusion for image inpainting. Specifically, we propose a forward process that gradually corrupts z_0 to z_m in $T \in \mathbb{N}$ timesteps, i.e., in our formulation, $z_T = z_m$. During inference for inpainting, this means denoising our corrupted latent representation z_m to z_0 in T timesteps. The fact that the single-step transition probability for each token at each timestep is independent of each other allows us to incorporate the position of masked regions into the transition distribution $q(z_t|z_{t-1})$ by augmenting it as follows: $M_L \odot v(z_t)^\top \mathbf{Q_t} v(z_{t-1}) + (1 - M_L) \odot v(z_{t-1})$. Furthermore, the t step transition probability $q(z_t|z_0)$ can be obtained as follows: $M_L \odot v(z_t)^T \bar{\mathbf{Q}}_\mathbf{t} v(z_0) + (1 - M_L) \odot v(z_0)$.

3.3.2 Modelling the Reverse Process

We learn to reverse the partial discrete diffusion process to obtain z_0 from z_m iteratively. Typical parameterization of the reverse process comprises of predicting un-normalized log probabilities $\log p_\theta(z_{t-1}|z_t)$. But the recent works [3, 19] have found that directly predicting the noiseless target variable $q(z_0)$ results in a better quality of generated images. This formulation is achieved by using the following reparameterization trick, which results from the Markovian nature of the forward discrete diffusion process:

$$q(z_{t-1}|z_t, z_0) = \frac{q(z_t|z_{t-1}, z_0) q(z_{t-1}, z_0)}{q(z_t|z_0)}. \tag{2}$$

Building on this, we design our neural network $p_\theta(.)$ to predict the distribution of noiseless target variable $\tilde{z_0}$, estimated at each reverse step conditioned on sketch embeddings s. Using this $p_\theta(\tilde{z_0}|z_t, s)$, we can compute the one-step reverse transition [16] using the following equation which combines it with the posterior $q(z_{t-1}|z_t, z_0)$:

$$p_\theta(z_{t-1}|z_t, z_0, s) = \sum_{\tilde{z_0}} q(z_{t-1}|z_t, z_0) p_\theta(\tilde{z_0}|z_t, s). \tag{3}$$

We train the network to minimize the Variational Lower Bound (VLB) objective [45] for one-step reverse prediction of $q(z_{t-1}|z_t, z_0)$ and a denoising objective following [3, 16], which encourages the model to predict a better noiseless $\tilde{z_0}$. The total loss is given by: $\mathcal{L}_0 = \mathcal{L}_{VLB} + \lambda \mathcal{L}_{z_0}$, where λ is a hyper-parameter used to balance the contributions from the two losses, VLB loss is defined as:

$$\begin{aligned} \mathcal{L}_{VLB} = &- \log p_\theta(z_0|z_1, s) \\ &+ \sum_T D_{KL}(q_\theta(z_{t-1}|z_t, s)||p_\theta(z_{t-1}|z_t, s)) \end{aligned} \tag{4}$$

where $D_K L(x||y)$ denotes the KL-Divergence between the random variables x and y. $\mathcal{L}_{z_0}$ is the denoising objective defined as: $\mathcal{L}_{z_0} = -\log p_\theta(z_0|z_t, s)$.

3.3.3 Conditioning the Reverse Process on Sketches

We use a sketch of the object to be inpainted as a conditioning signal to guide the image inpainting process. The sketch image $S \in \mathbb{R}^{224 \times 224 \times 1}$ represents a rough outline of the input image's missing content, indicating the object's overall shape and pose needs to be inpainted. Recent studies have utilized AdaLN [4] and AdaGN [11] parameters for conditioning the generation process. Yet, considering the intricate nature of the sketches with varying spatial information (i.e., shape and pose), conditioning using AdaLN is inadequate. Thus, we propose incorporating the sketch's visual information into our inpainting model through a simple yet effective method. We first pass the hand-drawn sketch through a sketch encoder E_s, which we realize as a ResnNet50 [17]. The ResNet50 extracts features from the sketch and produces a feature map $f_s \in \mathbb{R}^{7 \times 7 \times 2048}$. We add 2D-learnable positional embeddings to the feature map f_s to further incorporate the positional information into the sketch features and obtain a final flattened feature map $s \in \mathbb{R}^{49 \times 2048}$ for representing the sketch information for further inpainting by reversing the diffusion process. We linearly project these representations before feeding them to our bi-directional transformer.

3.3.4 Model Architecture

We propose to realize $p_\theta(.)$ as a bidirectional encoder-only transformer to estimate the distribution $p_\theta(\tilde{z_0}|z_t, s)$. As shown in Figure 2, our model consists of a sketch encoder E_s, and the diffusion decoder $p_\theta(.)$. We use a ResNet50 [17] as a sketch encoder E_s and it takes in a hand-drawn sketch $S \in \mathbb{R}^{W_s \times H_s \times 1}$ and maps it to a set of latent features s after adding positional embeddings. At any timestep $t \in \{1, 2, \ldots, T\}$, the goal of our network p_θ is to take in z_t and s and predict the distribution $q(z_{t-1}|z_0)$. To condition the denoising step on the sketch embeddings s, we concatenate a linear projection of latent representation s with the vector representation of z_t, i.e., v_{z_t} and then we feed the concatenated representation $[s; v_{z_t}]$ through a series of bi-directional transformer blocks to finally predict $p_\theta(\tilde{z_0}|z_t, s)$.

4. Experiments and Results

4.1. Dataset and Performance Metrics

We curate a dataset from the widely used MS-COCO dataset [27]. We begin by isolating objects of interest and then removing any irrelevant background information by segmenting all objects from the MS-COCO images using

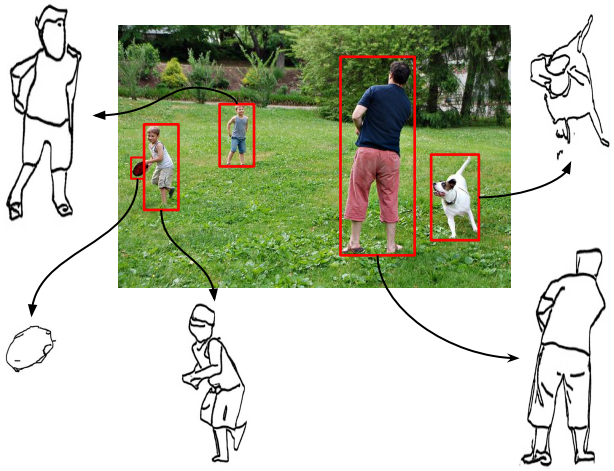

Figure 3. An example of our dataset. We randomly mask a bounding box shown using red color and provide the masked image along with the corresponding sketch as input to the image inpainting method. Please refer to Section 4.1 for more details.

available annotations. To improve the resolution of the segmented objects, we employed a pretrained super-resolution method, namely ESR-GAN model [52]. Finally, we apply PhotoSketching [24] to produce sketchy versions of the high-quality segmented objects and generate a dataset of images and corresponding object sketches. The resulting dataset contains 860K object sketches in the training set and 36K object-level sketches in the validation set. Please refer to Figure 3 for a sample image of this dataset.

Traditional image inpainting metrics like mean squared error, peak signal-to-noise ratio (PSNR), or structural similarity index (SSIM) are not well-suited for sketch-guided image inpainting as the input sketches only provide a rough outline of the missing regions and do not specify the exact content or color palette to be inpainted, resulting in different valid inpainting outputs. Thus, we use the FID score [36] to measure the quality of the inpainted images and the inpainted region, respectively. We also adopt the LPIPS metric [61], commonly used in recent inpainting works [30, 31, 44], to measure the similarity between the inpainted image and ground truth. However, given that these metrics take the entire image into account rather than solely the inpainted region, we additionally present Local-LPIPS and Local-FID score, which measures the LPIPS similarity and FID score between the inpainted region and its corresponding region in the ground truth image. Please note that local metrics are more effective at capturing performance than global metrics when the masked region is small. We evaluate inpainting methods by randomly masking an object using bounding box annotations in 5K images from the MS-COCO validation set, and further conduct user studies to measure the photorealism of the output images and their consistency with the sketch query.

4.2. Competing Approaches

To assess our model's performance, we adapted closely related methods by training them on the trainset of our dataset. A brief overview of these approaches is provided below (i) Sketch-Colorization GAN: We implement a simple DCGAN [39], which takes a corrupted image I with a sketch S pasted into the missing region through channel-wise concatenation and generates a completed image. The model is then trained to synthesize an image to match the ground-truth image. **(ii) DeFLOCNet [28]** demonstrated state-of-the-art performance on facial attribute editing tasks using partial sketches while generating realistic results. We train the SC-FEGAN baseline with object sketches instead of partial sketches to generate inpainted images. **(iii) DeepFillv2 [56]** is an image inpainting method aimed at filling in missing regions of images with free-form masks. We train this model from scratch by adapting it to our problem setup, where the inpainting is produced by concatenating the sketch with the binary mask and the corrupted image. **(iv) Palette [44]** is based on Denoising Diffusion Probabilistic Models. We adapt this model to our proposed setup by training it for the inpainting task by conditioning the inpainting process on the latent representation of the sketch obtained through a ResNet-50 encoder, using the AdaLN layer [4]. **(v) ControlNet [60]** introduces conditional control to the Stable Diffusion [42] model. In our early experiments, the pre-trained models demonstrated poor performance for our task because these models are designed for text-to-image generation and not image inpainting. Therefore, we train ControlNet on our dataset by providing a scene image containing masked region, sketch query, and a default caption, "A photo-realistic image".

4.3. Results and Discussion

We conducted experiments to evaluate the models' performance for sketch-guided image inpainting and present the results in Table 1. It is evident from the Table that the Sketch-colorization GAN method performs inferior to all other methods. This is because Sketch-colorization GAN is a naïve approach that generates an inpainted image from a corrupted image with a sketch pasted into the missing region. DeFLOCNet and DeepFillv2 also fail to encapsulate the information from crude, object-level sketches. The Palette uses AdaLN [4] to condition the inpainted region on the sketch, which produces poor results since the conditioning mechanism cannot encapsulate accurate information about the shape and pose of the object from the hand-drawn sketch. The pre-trained ControlNet performs poorly at our proposed task, achieving an FID score of 29.52, which is attributed to the fact that it utilizes a pre-trained text-to-image StableDiffusion model, which makes its output heavily dependent on the text prompt. Therefore, we train a ControlNet from scratch on our curated

Figure 4. Qualitative comparison of the proposed sketch-guided inpainting method with the competitive baselines. The results show that the proposed model effectively utilizes visual information in the sketch query, producing inpainting results with high visual fidelity and query faithfulness. Refer to Section 4.3 for more details.

data, achieving an FID score of 10.77. This is our closest-performing model that produces better visual results than the other approaches. However, in many cases, as seen in Figure 4, the faithfulness of the inpainted region with the provided hand-drawn sketch remains an unresolved issue with this model, too. In contrast, our proposed technique achieves state-of-the-art performance in sketch-guided image inpainting, as demonstrated by a lower (i.e., better) FID score in Table 1. It is worth noting that the Palette, ControlNet, and our framework are based on diffusion models. Palette performs diffusion in latent space, and inpainting is conditioned only through a single sketch embedding. On the other hand, ControlNet uses diffusion in the continuous space and StableDiffusion's U-Net encoder [43] to guide the pre-trained StableDiffusion model in generating the inpainted images. Our proposed solution exploits discrete diffusion and benefits from a more robust sketch-conditioning mechanism where the features are first extracted from a ResNet50 model and then combined with image features in the self-attention blocks of the transformer.

The qualitative results of our model in comparison to competitive approaches for the proposed task are shown in Figure 4. Our proposed model outperforms all the competent approaches and successfully utilizes the visual shape and pose information from the query sketch to generate high visual fidelity, semantic consistency, and faithfulness to the provided sketch when inpainting the missing region. We omit the results of sketch-colorization GAN because of poor quantitative performance.

Method	**FID** (↓)	**LPIPS** (↓)	**LLPIPS** (↓)	**LFID** (↓)
Sketch-coloring GAN	37.23	0.79	0.98	152.64
DeFLOCNet [28]	30.68	0.17	0.57	76.16
DeepFillv2 [56]	27.19	0.16	0.55	105.49
Palette [44]	25.87	0.14	0.53	98.56
ControlNet [60]	<u>10.77</u>	<u>0.11</u>	<u>0.49</u>	<u>21.98</u>
Ours	**7.72**	**0.11**	**0.42**	**21.91**

Table 1. Performance comparison of the proposed model on the curated MS-COCO dataset. Lower is better (see Section 4.3). LLPIPS and LFID denote "Local LPIPS" and "Local FID", respectively. Bold and underlined numbers refer to the best and second-best performances for their respective metrics.

Ablation study: Recent studies [3, 16, 42] have shown

Timesteps (T)	1	2	10	25	50
FID	12.55	10.40	8.17	7.89	7.72
LPIPS	0.124	0.110	0.109	0.109	0.107
LLPIPS	0.498	0.480	0.438	0.429	0.414

Table 2. The analysis of the effect of the number of inference steps on the quality of the inpainted images (refer to Section 4.3).

Method	User Preference (%)
DeFLOCNet [28]	2.72
DeepFillv2 [56]	2.72
Palette [44]	0.01
ControlNet [60]	25.45
Ours	**68.54**

Table 3. The study involved 50 masked images randomly selected from the validation split of our dataset. A group of 22 human participants were presented with the inpainted images generated by our method and competing approaches. They were then asked to express their preferences, focusing on photorealism.

that the quality of images generated by diffusion models is affected by the number of diffusion timesteps ($T \in \mathbb{N}$). To study this, we conducted a study to measure the effect of the number of inference timesteps on the quality of inpainted images. The results in Table 2 demonstrate that the quality of the generated inpainted images increases as the number of inference steps increases with the highest FID of **7.72** for 50 steps. Even the images generated in two timesteps have a better quality than the closest baseline model (FID score of **10.77** for ControlNet v.s. **10.40** FID for our method).

User Study: In addition to quantitative evaluations, we also conducted a subjective assessment of the proposed inpainting method using a human preference metric. To carry out the assessment, we randomly selected 50 masked images and inpainted them using four top-performing competitive approaches and our proposed framework. We recruited 22 human users to participate in the evaluation. Each user was presented with all sets of masked images and corresponding inpainted versions. The subjects were then instructed to carefully examine each set and choose the inpainted image they perceived as the most natural-looking. The human preference evaluation results, indicating the participants' choices, are reported in Table 3. As shown, 68.54% of the time, the users preferred the inpainted results generated by our method. This analysis indicates the naturalness and visual fidelity of the generation. Additionally, we performed another user study to quantitatively evaluate how well the inpainted region of our and the baseline models align with the provided user sketch. We randomly select 20 images from the validation split and show 22 human subjects the inpainting results produced by competing baselines and our method along with the query sketch, and ask them to rate the consistency and alignment of the inpainted region with the given sketch from 1 (poor alignment and consistency) and 5 (best alignment and consistency). The results in Table 4 indicate the superior performance of our proposed method.

Method	Consistency Score (mean±std)
DeFLOCNet [28]	1.40 ± 0.65
DeepFillv2 [56]	2.24 ± 1.07
Palette [44]	1.09 ± 0.33
ControlNet [60]	3.75 ± 1.20
Ours	$\mathbf{4.34 \pm 0.77}$

Table 4. The study involved 20 masked images and corresponding sketch queries randomly selected from the validation split of our dataset. A group of 22 human participants were presented with the inpainted images generated by our method and competing approaches. They were to score the consistency of the inpainted region with the sketch on a scale of **1 (poor) to 5 (best)**.

Limitations: Our method achieves state-of-the-art performance in sketch-guided image inpainting. However, there is significant room for improving the visual quality of the inpainted images. Our work represents a small step towards object-level sketch-guided image inpainting. One area to enhance is our sketch information embedding, which currently uses a straightforward ResNet50 encoder for extracting embeddings from rasterized sketches. Future research could explore more sophisticated sketch embeddings capturing stroke-level details. Another area of exploration involves refining the conditioning mechanisms that merge sketch embeddings with image representations to synthesize the inpainted image. Furthermore, we aim to develop diffusion models for their generative capabilities and leverage transformer models for their robust modeling capabilities, building upon the discrete diffusion process.

5. Conclusion

In this work, we investigated sketch-guided image inpainting, where a query sketch and non-missing regions of the image provide cues for filling in the missing regions. The proposed approach alleviates the problem of limited control over inpainted objects in traditional image inpainting and text-to-image inpainting, thereby making it more practical for image manipulation applications. Despite challenges such as the significant domain gap between hand-drawn sketches and images, our proposed approach achieved state-of-the-art results and generated photo-realistic objects that fit the context in terms of the shape and pose of the object in the provided sketch. Both quantitative and qualitative analyses demonstrate that our approach significantly outperforms other relevant approaches.

References

[1] Karim Armanious, Youssef Mecky, Sergios Gatidis, and Bin Yang. Adversarial inpainting of medical image modalities. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 1

[2] Karim Armanious, Vijeth Kumar, Sherif Abdulatif, Tobias Hepp, Sergios Gatidis, and Bin Yang. ipa-medgan: Inpainting of arbitrary regions in medical imaging. In *2020 IEEE International Conference on Image Processing (ICIP)*, 2020. 1

[3] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *ArXiv*, abs/2107.03006, 2021. 1, 3, 4, 5, 7

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5, 6

[5] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000. 2

[6] Ayan Kumar Bhunia, Aneeshan Sain, Parth Hiren Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. In *ECCV*, 2022. 1, 2

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[8] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Trans. Graph.*, 2009. 2

[9] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018. 2

[10] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004. 2

[11] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 5

[12] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[13] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4

[14] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5174–5183, 2020. 2

[15] Ce Ge, Jingyu Wang, Qi Qi, Haifeng Sun, Tong Xu, and Jianxin Liao. Scene-level sketch-based image retrieval with minimal pairwise supervision. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 650–657. AAAI Press, 2023. 2

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 5, 7

[17] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[19] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *CoRR*, abs/2102.05379, 2021. 5

[20] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[21] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1745–1753, 2019. 1, 3

[22] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[23] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6850–6861, 2023. 2

[24] Mengtian Li, Zhe L. Lin, Radomír Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 2, 6

[25] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 1

[26] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. 2014. 1, 2

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in

context. In *European Conference on Computer Vision*, 2014. 2, 5

[28] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, Jing Liao, Bing Jiang, and Wei Liu. Deflocnet: Deep image editing via flexible low level controls. In *CVPR*, 2021. 1, 3, 6, 7, 8

[29] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *Proceedings of the European conference on computer vision (ECCV)*, pages 205–220, 2018. 2

[30] Zeyu Lu, Junjun Jiang, Jun Huang, Gang Wu, and Xianming Liu. Glama: Joint spatial and frequency loss for general image inpainting. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 1, 6

[31] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022. 1, 2, 6

[32] K Nazeri, E Ng, T Joseph, FZ Qureshi, and M Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. arxiv 2019. *arXiv preprint arXiv:1901.00212*, 2020. 2

[33] Minheng Ni, Xiaoming Li, and Wangmeng Zuo. Nuwa-lip: language-guided image inpainting with defect-free vqgan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14192, 2023. 2

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[35] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2

[36] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 6

[37] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[38] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018. 3

[39] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 6

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[42] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6, 7

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 7

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 2021. 2, 3, 6, 7, 8

[45] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, 2015. 5

[46] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022. 1

[47] Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Multi-task learning for medical image inpainting based on organ boundary awareness. *Applied Sciences*, 11, 2021. 1

[48] Aditay Tripathi, Rajath R Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 532–547. Springer, 2020. 2

[49] Aditay Tripathi, Anand Mishra, and Anirban Chakraborty. Query-guided attention in vision transformers for localizing objects using a single sketch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1083–1092, 2024. 2

[50] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[51] Qiang Wang, Haoge Deng, Yonggang Qi, Da Li, and Yi-Zhe Song. Sketchknitter: Vectorized sketch generation with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[52] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang.

Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 6

[53] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P. Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1434–1444, 2022. 2

[54] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 2

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[56] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2018. 1, 2, 3, 6, 7, 8

[57] Yu Zeng, Zhe Lin, and Vishal M Patel. Shape-guided object inpainting. *arXiv preprint arXiv:2204.07845*, 2022. 2

[58] Yu Zeng, Zhe Lin, and Vishal M. Patel. Sketchedit: Mask-free local image manipulation with partial sketches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[59] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 1, 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 6, 7, 8

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6