Sketch-Guided Scene-level Image Editing with Diffusion Models

Ran Zuo
Communication University of China
zuoran@cuc.edu.cn

Haoxiang Hu
Institute of Software, Chinese Academy of Sciences
huhaoxiang22@mails.ucas.ac.cn

Xiaoming Deng
Institute of Software, Chinese Academy of Sciences
idengxm@gmail.com

Yaokun Li
Institute of Software, Chinese Academy of Sciences
liyaokun22@mails.ucas.ac.cn

Yu-Kun Lai
Cardiff University
LaiY4@cardiff.ac.uk

Cuixia Ma
Institute of Software, Chinese Academy of Sciences
cuixia@iscas.ac.cn

Yong-Jin Liu
Tsinghua University
liuyongjin@tsinghua.edu.cn

Hongan Wang
Institute of Software, Chinese Academy of Sciences
hongan@iscas.ac.cn

Abstract

Sketch-based image editing allows for intuitive and flexible modification of image details, effectively improving editing efficiency and diversity. When performing the scene-level image editing task where sketches are employed to control multiple objects within the editing region, existing approaches using GAN or diffusion models face limitations in handling complex editing intentions, such as editing scene content with various object attributes including spatial layout, semantics, structure, and number of objects. The challenge lies in effectively utilizing the attributes of multi-objects in the sketch and mapping these sketch attributes to the image editing region. In this work, we propose a Sketch-guided Diffusion Model called SDM, which integrates a global-to-local conditioning strategy to maximize the utilization of each object instance's attributes in the sketch. Specifically, this strategy incorporates a multi-instance guided cross-attention module and modifies attention maps with sketch masks, to help the model capture object semantics, structure, and quantity jointly. Additionally, we optimize the generation of the shared boundary region for overlapped objects to tackle the issue of ambiguous contours and semantics around the boundary. Then we introduce the multi-instance semantic loss to compensate for the diffusion model's limitation of potential semantics comprehension in sketches. Extensive experiments with high-quality editing results show that the proposed method outperforms state-of-the-art methods in the sketch-guided scene-level image editing task.

Keywords: ***Sketch, Scene-level Image Editing, Diffusion Model.***

1. Introduction

Interactive image editing is crucial in reducing manual repetitive operations and streamlining the image creation process, which improves the efficiency and diversity of image content generation. With the popularity of touch screens, free-hand sketching has become an intuitive way to express users' design intent and flexibly modify design details in the process of creation. In recent years, the sketch-based image editing task has been extensively studied, where users input an arbitrary mask to optionally indicate the image editing region, draw the sketch to depict the editing content, and complete the image.

Current research for sketch-based image editing mainly adopts Generative Adversarial Networks (GANs) [5] and uses a few sketch lines to simply edit the partial structure details of the image [20, 11, 36, 35, 15, 37]. Typically, the editing results make subtle refinements to the object's shape based on its unedited structural components and preserve

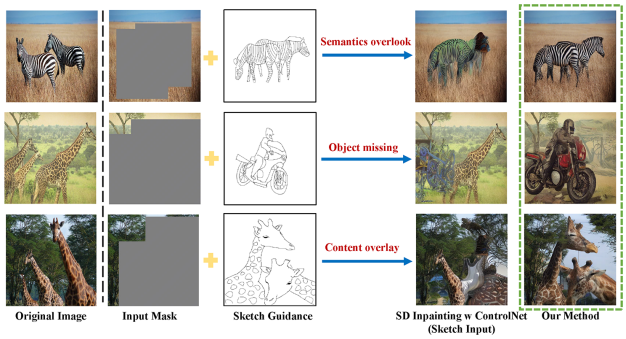

Figure 1. Illustration of the scene-level sketch-guided image editing task and challenges of the diffusion-based methods [23, 38] for this task. The task inputs the enlarged mask to label the foreground regions to be edited and utilizes a multi-instance sketch as key visual clues to guide the structure, semantics, quantity, and spatial layout of objects in the editing region. The major challenge is to fully utilize multi-instance attributes and realize attribute mapping from sketch to the image editing region, which contains three key issues: (1) Semantics overlook: the semantics of objects are unclear. (2) Object missing: one or more objects are missing. (3) Content overlay: the sketch-guided content is overlaid on the original content. We aim to address all three problems and realize high-quality image editing.

the original semantic category. In the real-world application scenario where the editing region often expands to encompass entire objects, existing editing methods lack the key structural and semantic guidance to assist the sketch input to modify the current content, and struggle to achieve satisfactory editing results. We define this challenging task as the sketch-guided scene-level image editing task, where an arbitrary mask to cover objects in the image is used to indicate the editing region, and a sketch containing multiple object instances is drawn to control scene content, including object number, and the attributes of each object (spatial layout, semantics, and structure).

Recently, diffusion-based models have been proposed to produce high-quality generation results [9, 25, 23, 18, 38, 24, 12, 27, 13], and several works [23, 18, 38] are effective for the sketch-guided scene-level editing. For example, by jointly utilizing ControlNet [38] and the Stable Diffusion Inpainting model [23], it can enable the highly challenging image editing task with large-scale occlusion [1]. The editing region is controlled by an input sketch, disregarding the text input in this case. However, these methods only use the structural information of the sketch for editing, which cannot effectively leverage all the object attributes and map them from the sketch to the image editing region. As illustrated in Fig. 1, these methods mainly focus on the shape of the object while overlooking other attributes, especially object categories in the sketch, leading to insufficient sketch comprehension and unrealistic generation. Moreover, they struggle to establish attribute correspondence between multiple objects in the sketch and those in the image and may overlook certain object attribute details or even the entire object omission. Furthermore, the sketch's control capability over the editing region is inadequate. The editing results are redundant where the model attempts to restore the image's original content and then superimpose additional content guided by the sketch.

In this paper, we introduce a Sketch-guided Diffusion Model called SDM to efficiently handle the complex sketch-guided scene-level editing task. To utilize the attributes of each object in the sketch and enable precise control over the editing region, we propose a global-to-local conditioning strategy, which includes a multi-instance guided cross-attention module to integrate the sketch's semantic features with the image and the synchronized semantic and structure control module that utilizes a corresponding sketch mask to adapt the attention map. Moreover, we especially address the generation of well-defined textures and contours in the overlapped boundaries of multiple objects. We focus on optimizing the attention distribution within the boundary region and feature distance between overlapped objects. To further mitigate the diffusion model's limitation in understanding potential sketch semantics, we propose the multi-instance semantic loss during training, which captures and minimizes the semantic features of each instance within the image editing region and the sketch. SDM can effectively utilize the sketch to generate user-desired scene content which integrates well with the remaining content of the original image, enabling efficient editing of scene content in images.

In summary, we make the following contributions:

1. We present a diffusion-based editing framework to realize the sketch-guided scene-level image editing task, which allows for effective scene content editing of images with large occlusions, enabling sketches to flexibly control multi-instance synthesis in the editing region.

2. We incorporate a global-to-local conditioning strategy, empowering sketches to achieve simultaneous object attribute determination including the instance number, structure, spatial layout, semantics, etc. via the cross-attention module.

3. We optimize textures and contours in overlapped boundaries of multiple objects by modifying the attention distribution and feature distance within the boundary regions, and propose the multi-instance semantic loss to further address the diffusion model's incapability of utilizing the sketch's potential semantics.

4. Our method outperforms the state-of-the-art sketch-based image editing methods with diffusion mod-

els and GAN-based methods through quantitative and qualitative evaluations.

2. Related Work

2.1. Sketch-based Image Editing

Recent sketch-based image generation and editing works [10, 32, 3] mainly adopt GAN as the basic framework. Specifically, the editing works [20, 11, 35, 36, 15, 37] focus on modifying the local structure of the editing region while ensuring semantic and style consistency with the remaining original image region. The majority of relevant works [20, 11, 35] primarily concentrate on human face editing. These approaches input face images, arbitrary masks to indicate regions to be edited, and partial strokes to control specific facial components, and then realize targeted edits. Specifically, FaceShop [20] introduces a conditional image completion method for face image manipulation. It enables users to modify local shape and color by utilizing masks, sketches, and a few color strokes. SC-FEGAN [11] enhances image completion capabilities by supporting free-form masks through discriminator optimization to handle large-scale occlusions. Additionally, it incorporates a style loss during training to further improve the quality of the edited results. Deep plastic surgery [35] employs sketches with different levels of abstraction for face image editing, which enables versatile editing of facial attributes. DeepFill-v2 [36], DeflocNet [15], and SketchEdit [37] are methods designed for editing human faces and general scenery. Notably, SketchEdit directly inputs sketches without additional masks to simplify the user input, which leverages sketches to predict the mask for the editing region and generate the final image.

In contrast to these approaches that utilize sketches to edit the original image content's partial structure, we employ a large mask to overlay the image's foreground objects, which completely occlude the semantic and structural information of the current region, leading to an obvious deficiency of image context. Then we incorporate sketches to depict diverse multi-object instances within the editing region and manipulate the scene content, ensuring comprehensive and coherent editing in more complex editing scenarios such as multi-object replacement and scene content composition.

2.2. Diffusion Models

Compared to GAN-based methods using adversarial learning with generators and discriminators for image generation, diffusion models progressively add noise to the data via a Markov Chain and then denoise from the noisy data to generate images, which is more stable for training and significantly improves the generation quality [9, 25, 23, 18, 38, 24, 12, 27, 13].

Sketches can be used as conditional input to determine the structure and spatial layout of the generated image. DiffSketching [31] applies diffusion models to generate object-level images with sketches. DiSS [4] inputs sketches and color strokes to manipulate the structure, color, and realism of generated images via diffusion models. With the development of Stable Diffusion [23], text-based image generation and editing achieve high performance. Current methods leverage text to describe the semantic content, object attributes, and style of generated images. Moreover, they incorporate additional guidance, such as sketches, layouts, keypoints, etc. into the pre-trained text-to-image diffusion model to achieve fine-grained structure control [18, 38, 33, 21, 30, 16, 19, 17]. Specifically, T2I-Adapter [18] and ControlNet [38] modify the text-guided generation results through multiple structural controls mentioned above. Freestyle [33] uses semantic masks as layouts to control objects' shapes and positions through the cross-attention module of the pre-trained text-to-image diffusion model. It utilizes the semantic mask to rectify the attention map calculated with the image and text tokens, which decides the spatial layout of each text token. Several works directly feed sketches into the pre-trained text-to-image diffusion model and utilize sketches to control the generated image structure [30, 16, 19].

The above methods based on Stable Diffusion often rely on two conditional inputs and primarily focus on the task of image generation. However, multiple inputs can impose a burden on novice users and non-native speakers, and the quality of generated results tends to decline when the source image is masked. To address these challenges and achieve a more generalized and user-friendly approach to fine-grained image editing, Paint by Example [34] proposes a novel solution that replaces text with example images as conditional semantic inputs to effectively manipulate the edited image. Motivated by it, we further extend this idea by replacing text with sketches. By capturing attributes such as semantic and structural content, object number, and layout of sketches, we realize precise control over the image editing region and simplify the process of image editing, making it more accessible and intuitive for users.

3. Method

To effectively utilize the sketch to guide the editing of scene content in images, we propose the Sketch-guided Diffusion Model, SDM (shown in Fig. 2). We adopt a global-to-local sketch conditioning strategy to map each object's attributes from the sketch into the image editing region. In this strategy, we first encode semantic features of both the global sketch context and each local object instance, and then adopt the multi-instance guided cross-attention module to utilize the encoded semantic features to generate each instance's attention map, respectively. Then we

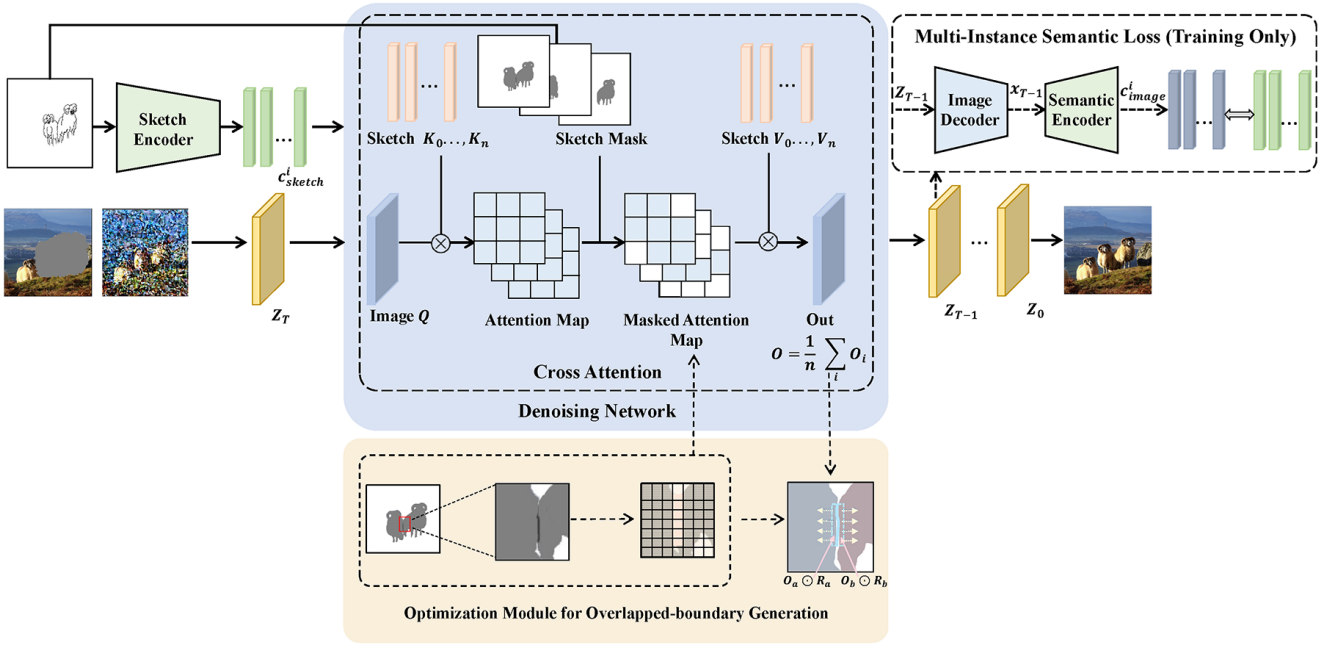

Figure 2. The framework of Sketch-guided Diffusion Model (SDM) for sketch-guided scene-level image editing. The model proposes a global-to-local sketch conditioning strategy that leverages the multi-instance guided cross-attention module to generate attention maps with the sketch's global and local instance semantic features and then uses the instance sketch mask to modify attention maps to realize synchronized semantics and structure guidance. Additionally, the model incorporates an optimization module for overlapped-boundary generation through maximizing attention distribution and feature distance in the boundary region, and then utilizes the multi-instance semantic loss to enhance its understanding of the sketch's potential semantics.

synchronously incorporate the corresponding sketch mask into the cross-attention to manipulate the attention map and modify the spatial layout and structure of each instance. To achieve distinct textures and precise contours in the overlapped boundaries of multiple objects, we optimize the model's attention mechanism in these regions and maximize the feature distance between overlapped objects. To remedy the diffusion model's deficiencies in potential semantic understanding of sketches and promote better coherence in the final output, we propose the multi-instance semantic loss to minimize each object's semantic features of the sketch and image editing region.

The primary processing of the model is formulated as follows. Given the image $x \in \mathbb{R}^{H\times W\times 3}$ with height H and width W, and the masked region $x \odot m$ with $m \in \{0,1\}^{H\times W}$, the objective of this task is to learn a mapping function between x and $\{x \odot \overline{m}, x_s\}$ to manipulate the edited region with the sketch and synthesize desired realistic image automatically, where $\overline{m}$ represents complementary matrix of mask m, and x_s represents the input sketch. We consider Stable Diffusion [23] as the basic framework and progressively add the Gaussian noise to obtain noisy latent code z_t of image x in the forward process. Then we train the network $\epsilon_\theta(z_t, x \odot \overline{m}, t, m_s, \phi_s(x_s))$, apply ϵ_θ to predict the noise, and gradually denoise z_t to z_0, where m_s is the sketch mask and ϕ_s is the sketch encoder. After decoding z_0, the final image x_0 is generated.

3.1. Global-to-Local Sketch Conditioning

Due to the sparsity and abstraction of sketches, directly utilizing the global sketch may overlook the fine-grained features of each object instance. Therefore, we train the instance segmentation model [7] to segment the multi-instance sketch S, obtaining a set of sketch instances S_i and corresponding sketch mask $m^i_s (i = 0, 1, ..., n)$, where S_0 represents the global sketch input and $\{S_i\}(i = 1, ..., n)$ represents each sketch object instance. Then we encode the sketch and manipulate the cross-attention module with a global-to-local strategy, which effectively determines the instance number, semantics, layout, and structure of the image editing region through the multi-instance guided cross-attention module and synchronized control of semantics and structure module.

Multi-Instance Guided Cross-Attention Motivated by recent advancements in image-to-sketch and sketch-to-image generation [29, 2] that leverage CLIP [22] to optimize feature mapping between sketch and image, we adopt the pre-trained CLIP image encoder to extract features of each sketch instance $S_i \in \mathbb{R}^{224\times 224\times 3}$ and retain only the

class token (denoted as $CLIP_{cls}(\cdot)$) to represent the sketch semantic features. In the resulting sketch token sequence, we have a class token representing global information and 256 patch tokens representing local features. Unlike text where each token carries specific semantics, the patch tokens in a sketch primarily represent abstract local lines which may increase the model's difficulty in understanding the sketch and potentially decrease its performance. Instead, the class token summarizes the content of the sketch from a global perspective and provides a better representation of the semantic information conveyed by the sketch. Then we adopt an MLP layer to match the conditional input dimensions and feed the sketch semantic features c^i_{sketch} into the diffusion process through cross-attention:

$$c^i_{sketch} = MLP(CLIP_{cls}(S_i)), i = 0, 1.., n \quad (1)$$

where the dimension of c^i_{sketch} is $B \times 1 \times 768$, B represents the batch size.

The sketch features are compressed with the CLIP image encoder and control the semantics of editing results through the cross-attention module. To fully utilize each sketch instance on the image, we propose a multi-instance guided cross-attention mechanism that generates multiple Keys $K_0, K_1, ..., K_n$ and Values $V_0, V_1, ..., V_n$ corresponding to c^i_{sketch} and obtain the attention map M_i for each sketch instance:

$$K = [f_K(c^0_{sketch}), f_K(c^1_{sketch}), ..., f_K(c^n_{sketch})], \quad (2)$$

$$V = [f_V(c^0_{sketch}), f_V(c^1_{sketch}), ..., f_V(c^n_{sketch})], \quad (3)$$

$$M_i = \frac{QK_i^T}{\sqrt{d}} \in \mathbb{R}^{C \times H \times W}, \quad (4)$$

where the dimension of K_i is $B \times 1 \times C$ ($C = 40, 80, 160...$), C varies with different channel size. The image feature z_t is converted to Q by $Q = f_Q(z_t)$, $f_Q(\cdot)$, $f_K(\cdot)$ and $f_V(\cdot)$ are the learnable linear transformations.

As the length of K_i is 1, the length of attention map M_i also becomes 1 with Eq. 4. Since directly using the softmax function makes all the values of the attention map become 1, we adopt the simplified softmax function for valid attention embedding to make sure all values of the attention map are non-negative. Finally, we average the different outputs O_i to obtain O and feed it into downstream layers to generate the final image x:

$$O_i = softmax_{simplified}(M_i)V_i = e^{M_i}V_i \quad (5)$$

$$O = \frac{\sum_{i=0}^{n} O_i}{n} \quad (6)$$

Synchronized Control of Semantics and Structure Given that the attention map plays a crucial role in determining the spatial layout of the current token [28, 33], we modify the weight distribution of the attention map by incorporating the mask m^i_s of each sketch object instance to achieve structural control over the current instance. As the denoising network is designed based on the U-Net architecture, the size of the attention map varies across different attention layers. To address this, we resize the mask to align with the size of the attention map. Since the channel of attention map M_i is 1, we modify its distribution with the 1-channel rearranged mask tensor m^i_s. Specifically, the value of M_i at position (k, j) remains unchanged when $m^i_s(k, j) = 1$, and it is set to negative infinity when $m^i_s(k, j) = 0$. We insert this operation to update M_i before its calculation with V_i in Eq. 5. This ensures the attention map is guided by the sketch mask by assigning weights to the regions of interest, effectively incorporating synchronous structure and semantic control through cross-attention during the denoising process.

3.2. Optimization Module for Overlapped-boundary Generation

To tackle the issue of overlapped objects in the image editing region which often results in blurred textures and ambiguous contours along the shared boundaries, we present an optimization module for overlapped boundary generation. The module first selects k pairs of neighboring object instances with overlapped boundaries according to the distribution of the sketch mask $m_s = \{m^i_s\}(i = 1, \ldots, n)$. The overlapped boundary between adjacent object instances a and b is denoted as $B_{a,b}$. We first enhance the attention distribution of the boundary to ensure the model places greater emphasis on learning its generation. After obtaining the attention maps M_a and M_b of the object instance a and b according to the Eq. 4, the attention value of the overlapped boundary is set to the maximum value identified within the attention map of the current object instance:

$$M_a(i, j) = max(M_a),\ M_b(i, j) = max(M_b),\ if\ (i, j) \in B_{a,b} \quad (7)$$

Furthermore, we propose the boundary loss to maximize the feature distance between overlapped objects. We utilize the coordinate of the boundary $B_{a,b}$ to obtain the rectangular region R occupied by the boundary. The region is divided into region R_a and region R_b, which belong to the object instance a and the object instance b respectively. After obtaining the output feature O_a and O_b in Eq. 5, the boundary loss is adopted to maximize the feature distance within region R_a and R_b:

$$\mathcal{L}_{boundary} = -\frac{1}{k}\sum_{i=1}^{k} d(O^i_a \odot R^i_a, O^i_b \odot R^i_b) \quad (8)$$

where d denotes the Euclidean distance of the computed

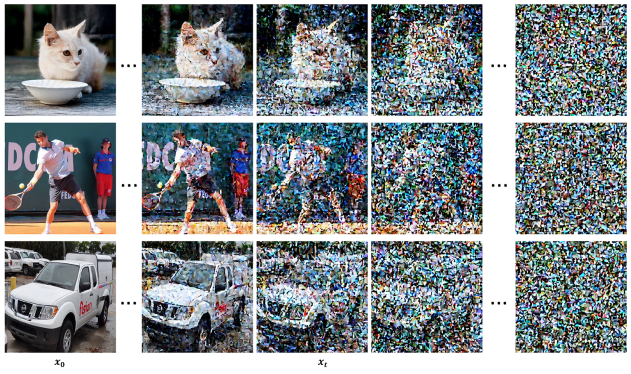

Figure 3. The visualization of image x_t in the denosing process of SDM.

features. This module can effectively enhance the visual difference between object instances in the overlapped region, and improve their generation quality around the boundary.

3.3. Enhanced Sketch Semantics Comprehension

In the process of image editing with sketches, the diffusion model's capability for learning the semantics of the sketch affects whether the editing results will be consistent with the sketch. Otherwise, the model will rely on vague and inaccurate semantic guidance to complete the editing region and overlay the structural control of the sketch on it, resulting in unrealistic editing results. Therefore, in order to address existing diffusion models' insufficient exploration of sketch's potential semantics, we propose the multi-instance semantic loss function. We first decode the latent code z_t to generate the image x_t and then utilize the sketch mask m_s^i to segment object instances in the image editing region. As depicted in Fig. 3, the intermediate image x_t during the denoising process retains certain semantic features. This allows for the measure of semantic similarity between sketch and image at each step of the denoising process. Using a semantic encoder similar to the sketch encoder (see Sec. 3.1), we encode the semantic representation c_{image}^i of each instance within the image editing region where $i = 1, ...n$. Finally, we optimize the model by minimizing the distance between the semantic features c_{image}^i and c_{sketch}^i of multiple instances. The semantic loss function is formulated as follows:

$$\mathcal{L}_{semantic} = \frac{1}{n}\sum_{i}^{n} \|c_{image}^i - c_{sketch}^i\|_2^2 \tag{9}$$

$$c_{image}^i = MLP(CLIP_{cls}(x_t \odot m_s^i)), i = 0, 1, .., n \tag{10}$$

In the model training stage, the objective of scene-level sketch-based image editing is to minimize the combination of denoising loss, semantic loss, and boundary loss as follows:

$$\mathcal{L} = \mathbb{E}_{z,m_s,x_s,\epsilon\sim N(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, x \odot \overline{m}, t, m_s, \phi_s(x_s)\|_2^2\right] \tag{11}$$

$$\mathcal{L}_{total} = \mathcal{L} + \lambda_1\mathcal{L}_{semantic} + \lambda_2\mathcal{L}_{boundary} \tag{12}$$

where $\mathcal{L}$ is the denoising loss, λ_1 and λ_2 control the relative weights of semantic loss and boundary loss and are fixed throughout our experiment.

Algorithm 1 The training process of SDM

Input: Image x, mask m, instance-segmented sketch $x_s = \{x_s^i\}(i = 1, ..., n)$, sketch masks $m_s = \{m_s^i\}(i = 1, ..., n)$, learning rate r etc.;
Output: Model parameters $\hat{\theta}$
Initialize model parameters θ using a pre-trained SD model
repeat
  $z = \mathcal{E}(x)$, $z_{inpaint} = \mathcal{E}(x \odot \overline{m})$, $z_0 =$ concatenate$(z_0, z_{inpaint}, m)$
  $t \sim U(1, ..., T)$
  $\epsilon \sim N(0, \mathbf{I})$
  $z_t = \alpha_t z_0 + \sqrt{1-\alpha_t}\epsilon_t$
  Calculate the semantic features $\{c_{sketch}^i(i = 1, ..., n)\}$ of multiple object instances in the sketch
  for each cross-attention layer in ϵ_θ **do**
    for each instance in $\{c_{sketch}^i\}$ **do**
      Obtain the latent feature output z_t from the preceding layer
      $Q \leftarrow f_Q(z_t), K_i \leftarrow f_K(c_{sketch}^i), V_i \leftarrow f_V(c_{sketch}^i)$
      Obtain the attention map M_i using Q and K_i
      Modify the weight distribution of attention map M_i with sketch mask m_s^i
      if instance is overlapped with others **then**
        Update the weights of overlapped boundaries in the attention map
      end if
      Obtain the result O_i through M_i and V_i
    end for
    Obtain the final output O and feed it into the subsequent attention layer
  end for
  Update model parameters using gradient descent $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{total}(\theta)$
until meets the convergence condition of the algorithm

return the parameters θ as the trained parameters $\hat{\theta}$

The specific training process of the SDM model is shown in Algorithm 1. This model first processes the original image x and the image region without mask $x \odot \overline{m}$ through the Autoencoder $\mathcal{E}$ to obtain the latent features z and $z_{inpaint}$.

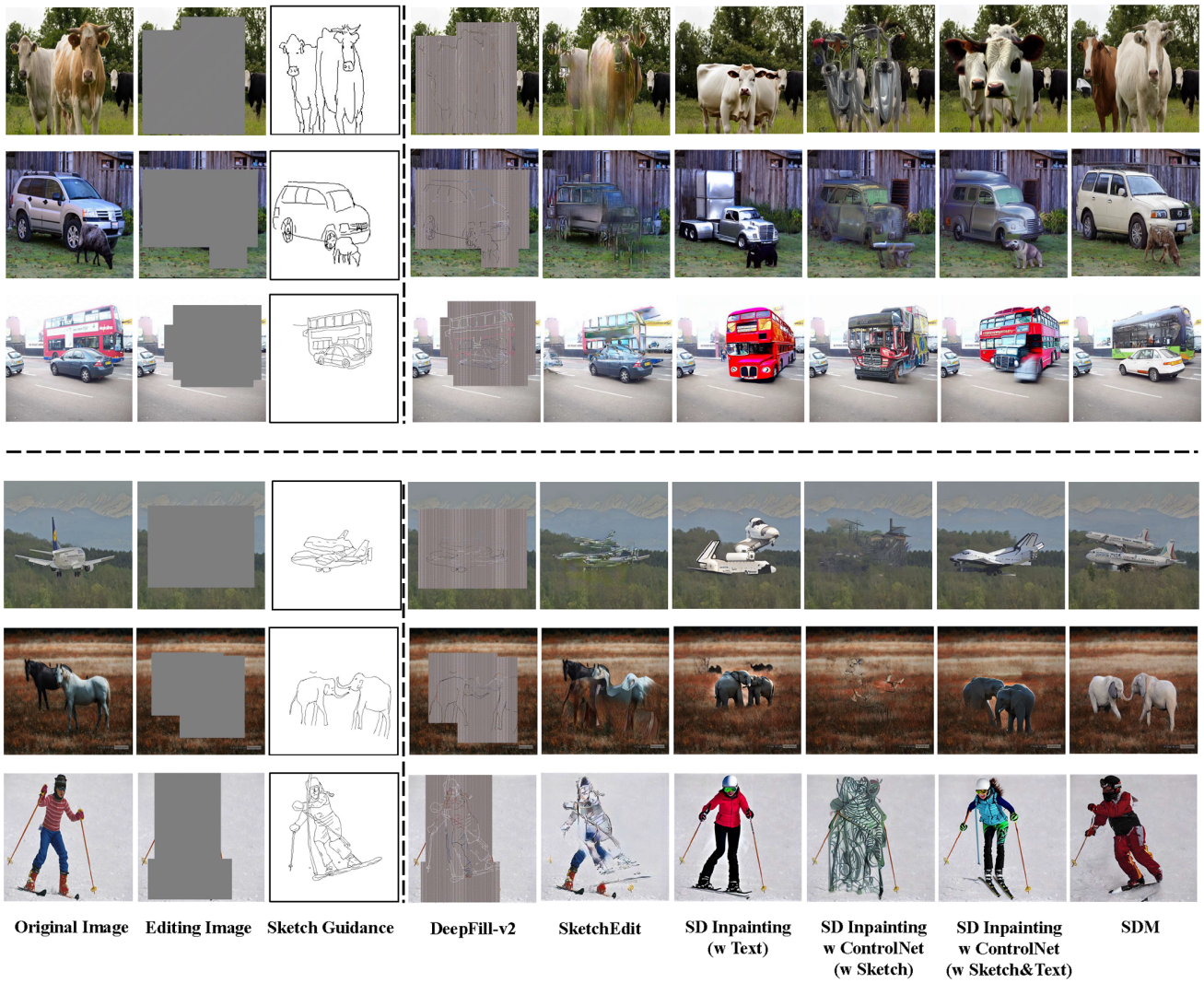

Figure 4. Visualization of comparison with SOTA sketch-based image editing methods with GAN and diffusion models on SFSD 512×512. For both inpainting the image with the corresponding sketch (see Rows 1-3) or editing the image region with a new sketch (see Rows 4-6), SDM demonstrates superior controllability over the editing region and produces higher-quality results compared to the other methods.

By concatenating z, $z_{inpaint}$, and the mask m, the dimension of the image is expanded from 4-d to 9-d. Then we augment the U-Net architecture of the noise prediction network ϵ_θ by incorporating 5 additional channels into the first convolutional layer. Specifically, 4 channels are allocated for $z_{inpaint}$, and 1 channel is dedicated to m. Subsequently, Gaussian noise ϵ is gradually added to the latent features z, with noise ϵ_t being added at step t to obtain the latent feature z_t. Then, z_t is input into the network ϵ_θ during the denoising process. Afterward, a global-to-local strategy is utilized to control the structure and semantics of the image editing region through a cross-attention mechanism, and the generation of overlapped boundaries is improved by optimizing the attention map distribution of multiple objects' overlapping boundaries. Finally, the model parameters are updated by minimizing the loss function through gradient descent.

4. Experiment

4.1. Datasets and Evaluation Metrics

Since the available sketch-image datasets for training diffusion models are limited in size, we use MSCOCO [14] containing $164K$ images as the training dataset. We employ the edge detector [26] to generate sketches. Specifically, we utilize each object's mask to extract them from the original image and generate the corresponding sketch for each extracted object. Then we apply Gaussian Blurring to reduce noise in the generated sketches, making them more suitable for subsequent use in our model. As for the

testing benchmark, we utilize the free-hand sketch-image dataset SFSD [39], which is created by leveraging the images of the MSCOCO dataset and consists of over $12K$ scene-level hand-drawn sketch-image pairs. We exclude images appearing in the SFSD dataset during the training phase, which ensures that our model is tested on unseen data, and we can evaluate the effectiveness and robustness of our model in handling image editing with free-hand sketches.

To evaluate the performance of our model, we employ four widely-used evaluation metrics, including FID (Fréchet Inception Distance) [8], QS (Quality Score) [6], SSIM (Structural Similarity Metric) and PSNR (Peak Signal-to-Noise Ratio). FID evaluates the similarity of deep feature distributions between generated images and real images. The lower FID score means the similarity between the generated images and real images is higher, indicating higher quality of image generation. Meanwhile, QS assesses the realism and quality of individual images while a higher QS score signifies better generation performance. SSIM evaluates the structural similarity between generated images and real images, with higher scores indicating greater structural similarity. PSNR measures the degree of image distortion, with higher scores suggesting higher quality of the generated image.

4.2. Implementation Details

Our method for sketch-guided scene-level image editing is built upon the text-driven image generation model Stable Diffusion [23]. To initialize our Sketch-guided Diffusion Model (SDM), we leverage the publicly released v1-4 model of Stable Diffusion, which sets a solid foundation for our model's training and contributes to high-quality generation. We resize the input images to 512×512 and train 40 epochs with a batch size of 8, taking 4 days on 4 NVIDIA A100 GPUs. The loss weight λ_1 and λ_2 in Eq. 12 are set to 0.1. To generate the mask m that indicates the region for image editing, we randomly select n bounding boxes of objects ($n = 2$ in the experiment) in the image and combine them as the mask input. The number of objects in the sketch is also set to 2. We train the sketch instance segmentation model proposed in [7] using large-scale generated sketches derived from the MS COCO dataset [14]. The model achieves an accuracy of 81.4% in segmenting object instances within free-hand scene sketches of the SFSD dataset [39].

4.3. Comparisons

For GAN-based methods, we choose SketchEdit [37] and DeepFill-v2 [36] to compare sketch-based image editing with our method. Notably, SketchEdit eliminates the need for an additional mask input, allowing us to directly use the sketch to edit the target region. For diffusion-based methods, we have selected three baseline methods: (1) We utilize the pre-trained Stable Diffusion (SD) inpainting model [23] and employ the text prompt to represent the sketch, enabling us to inpaint the masked region. (2) We augment the SD inpainting model [23] with the pre-trained ControlNet [38] to realize sketch-guided image inpainting [1], which only inputs the sketch to maintain consistency with our approach. (3) We integrate both sketch and text prompts as conditioning inputs into the combined ControlNet and SD inpainting model.

Fig. 4 and Tab. 1 show the qualitative and quantitative comparison of editing images on the SFSD dataset of these methods. We can observe that the quality of the edited images and their fidelity decrease significantly when using GAN-based editing methods. The SD inpainting method exhibits higher plausibility according to the metrics. However, it solely relies on semantic guidance from text input, which may result in discrepancies with actual expectations regarding other attributes (e.g. structure, number, layout) of the generated objects. Given the SD inpainting model with ControlNet, when only providing sketch guidance, the model is restricted to utilizing the structural information of the sketch which often leads to confusing editing results. In the case of inputting both sketch and text guidance, the model encounters challenges due to the instance-level feature misalignment between sketch and text, and the model is incapable of precisely manipulating attribute details for each object instance. In contrast, SDM presents a robust capability to effectively determine attributes of objects within the editing region and exhibits a significant increase in evaluation scores, which enables SDM to generate more accurate scene content that aligns closely with their corresponding sketch depictions.

Table 1. Quantitative comparison of editing images on SFSD [39] with state-of-the-art (SOTA) GAN-based and diffusion-based sketch-guided image editing methods.

Method	FID (↓)	QS (↑)	SSIM (↑)	PSNR (↑)
DeepFill-v2 [36]	14.05	47.39	0.9065	31
Deep Plastic Surgery [35]	13.73	57.27	0.91	33.72
SketchEdit [37]	12.84	63.68	0.9124	35.71
SD Inpainting [23]	4.85	80.28	0.919	40.66
SD Inpainting w ControlNet (Sketch Input) [23, 38]	12.13	61.28	0.92	41.22
SD Inpainting w ControlNet (Sketch & Text Input) [23, 38]	5.5	75.66	0.9187	40.79
Ours (SDM)	**4.64**	**81.57**	**0.93**	**43.99**

4.4. Ablation Study

To validate the effectiveness of the major components in SDM, we progressively incorporate key components from SDM onto the baseline to investigate their effectiveness: (1) we construct the baseline based on the Stable Diffusion [23] by replacing text with sketch as the condition and using the sketch's class token via the CLIP image encoder to guide the image editing through the cross-attention module. Based on the global-to-local strategy, we (2) incorporate the multi-instance guided cross-attentions (GSC_w_MCA), and

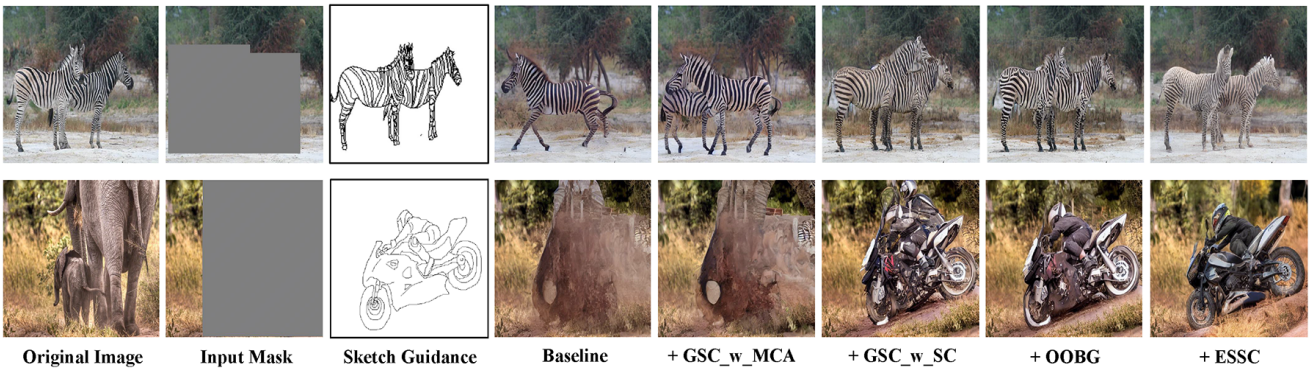

Figure 5. Visualization of ablation study about components of SDM. We incrementally add each component to the baseline for validation.

then (3) achieve multi-instance synchronized control over semantics and structure (GSC_w_SC) via modifying the attention map of each instance. (4) We introduce the optimization module to improve the generation performance of multiple objects' overlapped boundaries (OOBG) and then (5) add multi-instance semantic loss to enhance the model's comprehension capability over potential semantics (ESSC).

As shown in Fig. 5 and Tab. 2, the baseline model initially determines the object semantics but fails to generate the correct number and structure of objects. Then we introduce the global-to-local strategy into the model, ensuring the generated output aligns with the input sketch in terms of quantity, semantics and structure. We additionally adopt the OOBG module due to the synthesized texture and contour are blurry at the overlapped boundary between objects. The ESSC module further improves the model's awareness of the texture, semantics, and even structure of each object and eliminates artifacts, resulting in a more realistic representation. In the more challenging editing scenario, we replace the original image scene content with the sketch from a new category (Row 2 in Fig. 5). The baseline model yields confusing editing results that make it difficult to discern the influence of the sketch conditioning and the original context. After employing a complete global-to-local strategy (GSC_w_MCA and GSC_w_SC), the sketch exerts complete control over the editing region but the result is only coarsely aligned with the sketch. The OOBG and ESSC module gradually improve the texture distribution of specific instances and explore more control around object boundaries, leading to realistic editing results.

Table 2. Quantitative comparison on the configuration of SDM. By employing all of these techniques, we can achieve the optimal performance.

GSC_w_MCA	GSC_w_SC	OOBG	ESSC	FID (↓)	QS (↑)	SSIM (↑)	PSNR (↑)
				6.44	78.56	0.9163	40.79
✓				5.79	81.10	0.9168	40.99
✓	✓			4.91	80.82	0.921	41.45
✓	✓	✓		4.83	81.22	0.9273	41.82
✓	✓	✓	✓	**4.64**	**81.57**	**0.93**	**43.99**

4.5. Application and Limitation

Our method facilitates precise and flexible control of multi-instance sketches over scene content within the editing region, extending the sketch-based image editing task into new application scenarios. As illustrated in Fig. 6 (a), SDM provides the capability to modify the number of objects n in the sketch (e.g. $n = 1, 3$), or gradually add the object number to flexibly edit the image. Furthermore, SDM enables us to specify the desired scene by replacing multiple instances within the editing region with sketches from various categories (see Fig. 6 (b)), or compositing the sketch input with diverse scene backgrounds to generate new realistic composited images (see Fig. 6 (c)). SDM also allows users to input simple sketches with sparse lines and vague contours. It effectively captures the essential information within these sketches and generates the desired editing results (see Fig. 6 (d)). However, our method still has several limitations shown in Fig. 6 (e). Specifically, the model excels at controlling human poses but faces challenges in manipulating human faces, and struggles to generate the corresponding instance with a partial sketch input. Therefore, future work should enhance the model's capacity for more detailed feature mapping between sketch and image, and comprehend partial sketches for efficient editing.

5. Conclusion

We investigate the complex sketch-guided scene-level image editing task, which inputs multi-instance sketches to modify the image scene content, encompassing the number, structure, semantics, and spatial layout of objects. To achieve efficient and high-quality editing, we propose the Sketch-guided Diffusion Model (SDM) as our framework. SDM builds upon the Stable Diffusion model and incorporates a global-to-local sketch conditioning scheme to effectively leverage the overall attributes conveyed by the sketch and enhance the influence of the sketch over the editing region. We additionally incorporate the optimization module for overlapped-boundary generation to retain each

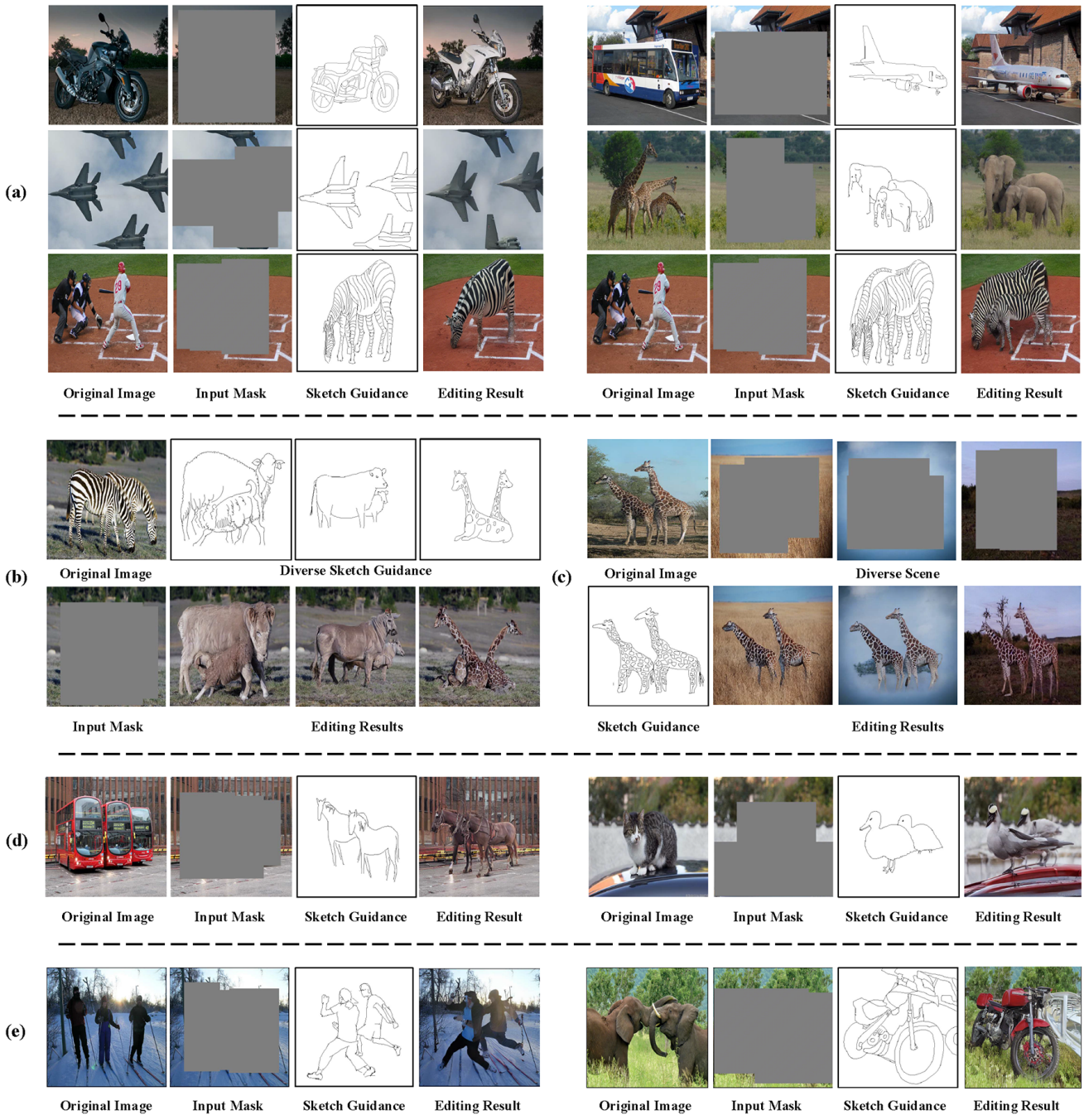

Figure 6. Typical applications and limitations of sketch-guided scene-level image editing. (a) Image editing with various numbers of objects in the sketch; (b) Multi-object replacement within a fixed scene; (c) Scene content composition through combining the sketch with diverse scene backgrounds; (d) Image editing with simple hand-drawn sketches; (e) Limitation of the method.

object's distinct characteristics in the overlapped boundary, and multi-instance semantic loss to further capture the sketch's potential semantics. Our method achieves state-of-the-art performance in sketch-based image editing and demonstrates promising potential for real-world applications.

6. Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.62272447. Cuixia Ma is the corresponding author of this paper.

References

[1] https://github.com/mikonvergence/ControlNetInpaint. 2, 8

[2] D. Bashkirova, J. Lezama, K. Sohn, K. Saenko, and I. Essa. Masksketch: Unpaired structure-guided masked image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1879–1889, 2023. 4

[3] S.-Y. Chen, F.-L. Liu, Y.-K. Lai, P. L. Rosin, C. Li, H. Fu, and L. Gao. Deepfaceediting: Deep face generation and editing with disentangled geometry and appearance control. *arXiv preprint arXiv:2105.08935*, 2021. 3

[4] S.-I. Cheng, Y.-J. Chen, W.-C. Chiu, H.-Y. Tseng, and H.-Y. Lee. Adaptively-realistic image generation from stroke and sketch with diffusion model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4054–4062, 2023. 3

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 1

[6] S. Gu, J. Bao, D. Chen, and F. Wen. Giqa: Generated image quality assessment. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 369–385. Springer, 2020. 8

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 4, 8

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 8

[9] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 3

[11] Y. Jo and J. 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

[12] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2, 3

[13] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov, and J. Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 740–755. Springer, 2014. 7, 8

[15] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, J. Liao, B. Jiang, and W. Liu. Deflocnet: Deep image editing via flexible low-level controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10765–10774, 2021. 1, 3

[16] A. MaungMaung, M. Shing, K. Mitsui, K. Sawada, and F. Okura. Text-guided scene sketch-to-photo synthesis. *arXiv preprint arXiv:2302.06883*, 2023. 3

[17] S. Mo, F. Mu, K. H. Lin, Y. Liu, B. Guan, Y. Li, and B. Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 3

[18] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3

[19] Y. Peng, C. Zhao, H. Xie, T. Fukusato, and K. Miyata. Difffacesketch: High-fidelity face image synthesis with sketch-guided latent diffusion model. *arXiv preprint arXiv:2302.06908*, 2023. 3

[20] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018. 1, 3

[21] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8748–8763. PMLR, 2021. 4

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4, 8

[24] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3

[25] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[26] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5117–5127, 2021. 7

[27] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[28] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image

translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 5

[29] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics*, 41(4):1–11, 2022. 4

[30] A. Voynov, K. Aberman, and D. Cohen-Or. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022. 3

[31] Q. Wang, D. Kong, F. Lin, and Y. Qi. Diffsketching: Sketch control image synthesis with diffusion models. *arXiv preprint arXiv:2305.18812*, 2023. 3

[32] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3

[33] H. Xue, Z. Huang, Q. Sun, L. Song, and W. Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023. 3, 5

[34] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[35] S. Yang, Z. Wang, J. Liu, and Z. Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 601–617. Springer, 2020. 1, 3, 8

[36] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 3, 8

[37] Y. Zeng, Z. Lin, and V. M. Patel. Sketchedit: Mask-free local image manipulation with partial sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5951–5961, 2022. 1, 3, 8

[38] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 8

[39] Z. Zhang, X. Deng, J. Li, Y. Lai, C. Ma, Y. Liu, and H. Wang. Stroke-based semantic segmentation for scene-level free-hand sketches. *The Visual Computer*, pages 1–13, 2022. 8