DiffSketching: Sketch Control Image Synthesis with Diffusion Models

Qiang Wang
wanqqiang@bupt.edu.cn
Di Kong
dikong@bupt.edu.cn
Fengyin Lin
fylin@bupt.edu.cn
Yonggang Qi✉
qiyg@bupt.edu.cn

Beijing University of Posts and Telecommunications, Beijing, China

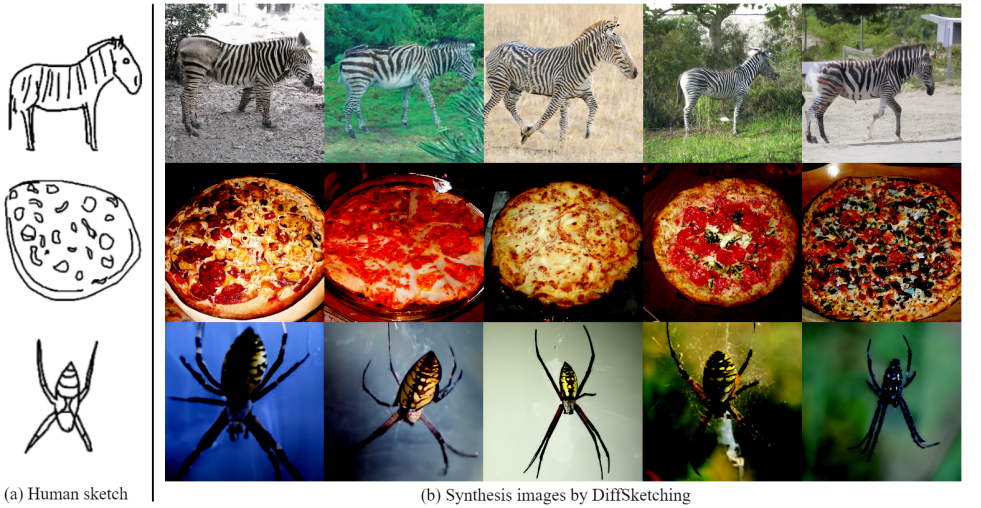
Abstract

Creative sketch is a universal way of visual expression, but translating images from an abstract sketch is very challenging. Traditionally, creating a deep learning model for sketch-to-image synthesis needs to overcome the distorted input sketch without visual details, and requires to collect large-scale sketch-image datasets. We first study this task by using diffusion models. Our model matches sketches through the cross domain constraints, and uses a classifier to guide the image synthesis more accurately. Extensive experiments confirmed that our method can not only be faithful to user's input sketches, but also maintain the diversity and imagination of synthetic image results. Our model can beat GAN-based method in terms of generation quality and human evaluation, and does not rely on massive sketch-image datasets. Additionally, we present applications of our method in image editing and interpolation.

1 Introduction

Free-hand sketch is an intuitive way for human beings to express the real world, while imagining from any given sketch to colored realistic images is a desirable ability for intelligent machines. A high quality sketch-to-image synthesis model can help design animation, games and other works. However, sketch contains far less information than image due to its simplicity, abstraction and inaccuracy. The cross-domain synthesis lacks important information such as color, shadow and texture. And the way that people do hand drawing is space distorted and imperfect, which makes this task very difficult.

Early sketch based image synthesis methods [5, 6, 12] are based on image retrieval which do not have real generation ability. In recent years, with the rise of GAN [15], a large number of solutions have been proposed [7, 24, 35, 52], but most of these methods rely on large sketch-image pairing datasets, which are very precious and hard to obtain. Sketchy [45] is the largest sketch-image pairing dataset including 125 categories at present. But each category only contains 50 images, which is far from enough for deep generative models. In



(a) Human sketch (b) Synthesis images by DiffSketching

Figure 1: Diffsketching synthesis (b) are a large number of images from (a) one or more real human sketches. Shape, pose, texture and other features of the sketch can be faithfully preserved.

addition, almost no research has been attempted on the complex and variable ImageNet [9] dataset.

Diffusion models rapidly become popular and beat GANs in some key indicators [10], bringing new creativity to research the generative models. Inspired by this, we propose DiffSketching, a sketch-guided image synthesis method through diffusion models. The input image is converted into latent noise by forward diffusion process. With the guidance of sketches, adjust the score function to invert to new images. This process does not need to rely on large sketch-image pairing datasets, and can beat the GAN-based method on qualitative comparisons and human evaluation results.

Our goal is that users only need to input a sketch, and our model can generate many corresponding images. There are three main challenges in using diffusion model to complete this task. (i) Existing diffusion models generate data in a single domain, so we need an appropriate guidance method for cross-domain generation, and an appropriate method to measure data distribution in two different domains. (ii) Unlike edge maps extracted from images, sketches and corresponding images are more inconsistent in space and geometry, so it is difficult to measure the cross domain matching of sketch image. (iii) The sketch entered by the user contains little information and often has ambiguity (*e.g.* drawing a dog, it is difficult to tell whether it is a German Shepherd or a Briard). We need to introduce more information to eliminate such ambiguity.

In order to solve the above challenges, our work makes the following major contributions. (i) We propose a model that can synthesize sketch-faithful, and photo-realistic images from a single sketch (Fig. 1), performing better on benchmarks than GAN-based models. (ii) We can guide the generation process more finely and eliminate the singularity and uncertainty of input sketches. (iii) We prove that our method is capable of editing images and conducting image interpolation conditioned on sketch.

2 Related works

Sketch Based Image Synthesis There are numerous research on edge based image synthesis, which belongs to image translation [25, 34, 51, 56, 60]. But compared with edge, hand-free sketch is more abstract, imaginative, flexible and challenging. The first work realize employ sketch to generate image is SketchyGAN [7], which is an encoder-decoder model adopts a two stage strategy for shape and appearance completion based on the paired sketch image data. Subsequently, there has been many works on automatically synthesizing natural images [13, 14, 35, 52] and human portraits [4, 33, 55]. Most of them are based on GAN require adversarial training which often suffers from unstableness and mode collapse.

As a mirror task, image-to-sketch work has also been extensively studied [32, 40, 49]. Photo-sketching [32] trains an image-conditioned contour generator for multiple diverse outputs, achieving the state-of-the-art (SOTA) performance in boundary detection and contour rendering. This method does not generate edge graph, but uses antagonistic training to make the generated result closer to the ground truth hand-drawn sketch. So we use it as a cross domain converter between sketches and images at the stage of measuring perceptual loss.

Diffusion Models Recently, many works on iterative generative models [3], such as denoising diffusion probabilistic models (DDPM) [21], score-based generative model [50] can produce samples comparable to those of GANs. Denoising diffusion implicit models (DDIM) [48] exert fewer sampling steps to obtain higher quality samples. Pratul [10] achieves the SOTA performance in image synthesis by improving DDIM architecture. Because diffusion models do not need adversarial training, they fundamentally solve the mode collapse problem of GANs.

However, a significant drawback of diffusion models is that it simulates many time steps of Markov chain to generate samples. Beyond DDIM, many acceleration methods [28, 44, 53, 58] have been proposed. Besides image synthesis [2, 38, 39, 42], diffusion models widely used in various fields, such as image-to-image translation [8, 31, 54], text-to-image translation [16, 26, 41, 43], video generation [18, 22, 57] and audio generation [23, 27, 3

3 Diffusion Models for Sketch-Guidance Image Generation

The overview of our proposed DiffSketching's framework is shown in Fig. 2. The input image x_0 is converted into latent noise x_T through the forward diffusion process. We clone x_T to $\hat{x}_T(q)$ and then synthesize the image $\hat{x}_0(q)$ from $\hat{x}_T(q)$ via a reverse generation process which is achieved through a fine-tuning process.

3.1 Background

Forward Diffusion Process Diffusion models slowly inject noise into the original data to destroy the initial data distribution. During the reverse generation process, the probability distribution of the desired data x_0 is obtained by learning to predict the noise and denoising. For the distribution of each training data $x_0 \sim q_{data}(x_0)$, through a variance schedule $b_1, \dots, b_T$, diffusion models gradually add Gaussian noise in step t to get $x_1, \dots, x_T$:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

Figure 2: Overview of DiffSketching. Our training Constraints consist two components: (a) L_p: the model P converts $\hat{x_0}(\hat{q})$ to sketch $\hat{s_0}(\hat{q})$ and makes perceptual loss with input sketch s_0. (b) L_i: cosine image similarity loss between input image and generated image $x_{t_0}(q)$.

$\sqrt{1-\bar{a}_t}e$, where $e \sim N(0,I)$. Song et al. [48] proposed DDIM that changed forward Markov process to Non-Markov process by using variable information. It becomes an implicit probabilistic model:

$$q_s(x_t|x_{t-1};x_0) = \frac{q_s(x_{t-1}|x_t;x_0)\,q_s(x_t|x_0)}{q_s(x_{t-1}|x_0)} \tag{2}$$

where $s \in R^{T}{}_{0}$ is the index of inference distribution family Q, controlling the stochasticity of the forward process.

Reverse Generation Process In the reverse generation process $q_q(x_t)$, diffusion models allow for different reverse samples to be generated by varying the variance of noise. establishes the mapping relationship from latent to image and conducts denoising from x_t to get x_{t-1}:

$$x_{t-1} = \sqrt{\frac{\bar{a}_{t-1}}{a_t}}x_t + \left(\sqrt{1-\bar{a}_{t-1}} - \sqrt{\frac{\bar{a}_{t-1}(1-a_t)}{a_t}}\right) e_q(x_t;t) \tag{3}$$

The function $e_q(x_t;t)$ represents the prediction of noise distribution, and q denotes the learnable parameter. Training process randomly samples the image with noise in time step t, and adopts simple mean squared error loss to make predicted noise closer to true noise $\tilde{N}_q \quad e_q(\sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}e;t)^2$.

Classifier Guidance Prafullaet al. [10] adopted classifier to guide the generation of images which does not need additional training. This method directly generates the desired image through the gradient guidance of the trained external classifier $p_f(y|x_t;t)$ on the trained diffusion models, where y is the class label. The predicted noise is defined as:

$$\hat{e}_q(x_t;t) = e_q(x_t;t) - \sqrt{1-\bar{a}_t}\,\tilde{N}_{x_t}\log p_f(y|x_t) \tag{4}$$

During sampling, the sampling center of the expected noisy image x_t is guided by the classifier to the condition that the predicted noise is as close as possible to the true noise and can guide the reverse diffusion direction to the desired category.

Figure 3: Qualitative results compared with baselines under the same sketch input.

3.2 Perceptual Diversity Learning

We de ne X, Y as the domains of sketches and images respectively. To bridge the g between X and Y, we adopt the pretrained GAN-based network Photo-sketching [32] t translate images into sketches $\hat{x}_0(\hat{q}) \to \hat{s}_0(\hat{q})$.

Because sketches have strong space distortion and style variability, classical per-p measurement methods such as L_1 Manhattan Distance or L_2 Euclidean Distance [11] will greatly damage the diversity of the generated sketches and enlarge the input defect w misguides the model. Therefore, we introduce perceptual metric loss [59] which can expr appearance similarity from global semantics to solve this problem. We use a pretrai perceptual sketch feature extractor $F_s(\cdot)$ for feature extraction between s_0 and $\hat{s}_0(\hat{q})$. w_l is denoted to scale the activation channel-wise for each layer l. Then we calculate L_2 distance, average over space and sum over channel wise:

$$L_p = \sum_{l=1}^{L} \frac{1}{H_l W_l} \sum_{h,w} \| w_l \quad (F_s(\hat{s}_0(\hat{q}))^l_{hw} \quad F_s(s_0)^l_{hw}) \|_2^2 \tag{5}$$

where $w_l \in R^{C_l}$ and $F_s(\hat{s}_0), F_s(s_0) \in R^{H_l \quad W_l \quad C_l}$.

3.3 Image Constraint Identity Learning

We observe that only the constraints in the sketch domain will lead to a loss of too ma elements in the original image and an increase in generation uncertainty. Because the sk domain provides less information than the image domain. To solve this problem, we prop an image constraint identity loss to compare the input image with the generated one.

We trained ResNet-50 [19] as the image constraint feature extractor F_i to extract features from x_0 and $\hat{x}_0(\hat{q})$ in an attempt to minimize the cosine distance of the generated image fro the input image:

Figure 4: Fine-grained sketch controlling. (a) is input sketch, (b) is a ne-grained category that can be speci ed by users, (c) is a category that independent of input sketch.

$$L_i = \frac{F_i(x_0) \quad F_i(\hat{x}_0(\hat{q}))}{kF_i(x_0)kkF_i(\hat{x}_0(\hat{q}))k} \tag{6}$$

Image constraint identity loss enables the generated results to have more identity information and enhances the robustness of the model. Sketch perceptual loss adds diversity and imagination. l is a super parameter for balancing diversity and identity. Our training objective is as follows:

$$L = l\, L_i(x_0; \hat{x}_0(\hat{q}))) + (1 \quad l)L_p(s_0; \hat{s}_0(\hat{q})) \tag{7}$$

3.4 Class-Guidance Fine-tuning Reverse Process

According to Eq. 3, the backward generation process is denoising from $x_t(q)$ to $\hat{x}_0(q)$. The sketch drawn with a few strokes is too simple and can easily mislead the model to generate inaccurate results. To prevent the generated data distribution from deviating from the category center, we constrain the model by a classi er, via the Eq. 4.

To take full advantage of the image synthesis performance of diffusion models, we pretrain the forward and reverse process of diffusion models with a classi er. In terms of ne-tuning, our model learns to be self-supervised subject to the constraint of Eq. 7. Once the diffusion model has been ne-tuned, any input sketches can be processed into images, as shown in Fig. 1. More details on the ne-tuning procedure and the structure of the model are analyzed in the supplementary materials.

4 Experiments

4.1 Evaluations

Datasets We select ImageNet dataset with 256 256 resolution, including 128K images of 1000 categories, to pretrain class-guidance diffusion models. In the ne-tuning stage we take 12.5K images from Sketchy [45] dataset with each image corresponding to 5 pieces of sketches.