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

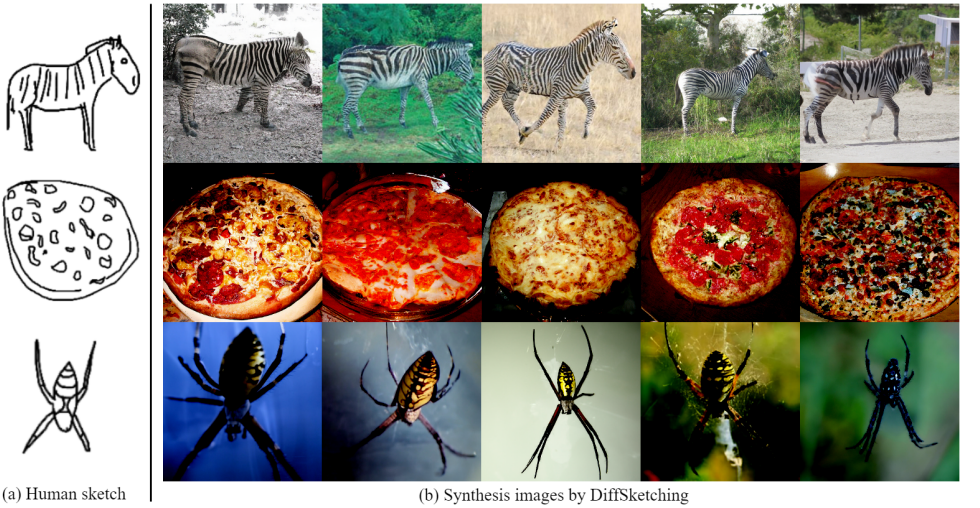
Abstract

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 [13], 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 field [25, 34, 51, 56, 60]. But compared with edges, hand-free sketch is more abstract, imaginative, flexible and challenging. The first work really employ sketch to generate image is SketchyGAN [7], which is an encoder-decoder model and 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 GANs, 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. Prafulla *et al*. [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, 37, 44, 53, 58] have been proposed. Besides image synthesis [7, 38, 39, 42], diffusion models are 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, 30].

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 $\hat{x}_T(\theta)$ and then synthesize the image $\hat{x}_0(\theta)$ from $\hat{x}_T(\theta)$ 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 $\hat{x}_0$ 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 $\beta_1, \ldots, \beta_T$, diffusion models gradually add Gaussian noise ε in step t to get $x_1, \ldots, x_T$:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{1}$$

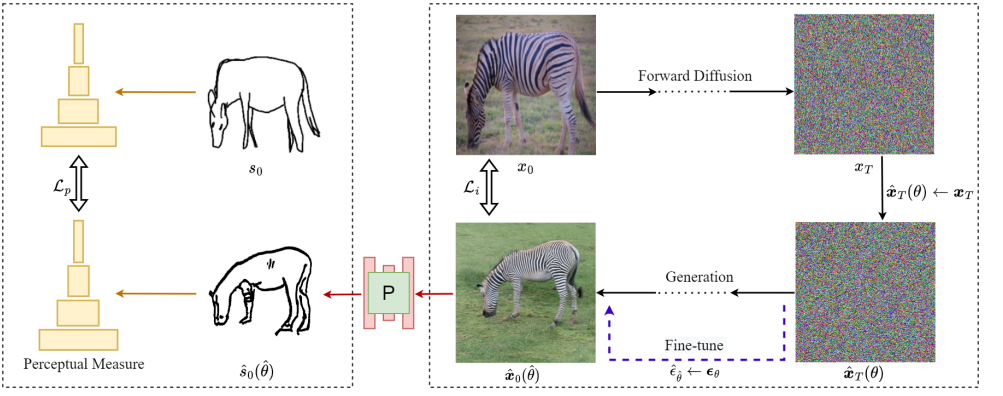


Figure 2: Overview of DiffSketching. Our training Constraints consist two components: (a) $\mathcal{L}_p$: the model $\mathcal{P}$ converts $\hat{x}_0(\hat{\theta})$ to sketch $\hat{s}_0(\hat{\theta})$ and makes perceptual loss with input sketch s_0. (b) $\mathcal{L}_i$: cosine image similarity loss between input image x_0 and generated image $\hat{x}_{t_0}(\theta)$.

Ho *et al.* [21] used $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$ to represent $\mathbf{x}_t(\mathbf{x}_0, \varepsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$. 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:

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

where $\sigma \in \mathbb{R}^T_{\geq 0}$ is the index of inference distribution family $\mathcal{Q}$, controlling the stochasticity of the forward process.

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

$$x_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} x_t + \left(\sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}} \right) \varepsilon_\theta(x_t, t) \tag{3}$$

The function $\varepsilon_\theta(x_t, t)$ represents the prediction of noise distribution, and θ 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: $\nabla_\theta \left\| \varepsilon - \varepsilon_\theta(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\varepsilon, t) \right\|^2$.

Classifier Guidance Prafulla *et 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 $p_\phi(y|x_t, t)$ on the trained diffusion models, where y is the class label. The predicted noise is defined as:

$$\hat{\varepsilon}_\theta(x_t, t) = \varepsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t} \nabla_{x_t} \log p_\phi(y|x_t) \tag{4}$$

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 define $\mathcal{X}$, $\mathcal{Y}$ as the domains of sketches and images respectively. To bridge the gap between $\mathcal{X}$ and $\mathcal{Y}$, we adopt the pretrained GAN-based network Photo-sketching [32] to translate images into sketches $\mathcal{P} : \hat{x}_0(\hat{\theta}) \rightarrow \hat{s}_0(\hat{\theta})$.

Because sketches have strong space distortion and style variability, classical per-pixel 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 which misguides the model. Therefore, we introduce perceptual metric loss [69] which can express appearance similarity from global semantics to solve this problem. We use a pretrained perceptual sketch feature extractor $F_s(\cdot)$ for feature extraction between s_0 and $\hat{s}_0(\hat{\theta})$. 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:

$$\mathcal{L}_p = \sum_{l=1}^{L} \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (F_s(\hat{s}_0(\hat{\theta}))^l_{hw} - F_s(s_0)^l_{hw})||_2^2 \tag{5}$$

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

3.3 Image Constraint Identity Learning

We observe that only the constraints in the sketch domain will lead to a loss of too many elements in the original image and an increase in generation uncertainty. Because the sketch domain provides less information than the image domain. To solve this problem, we propose 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 $\hat{x}_0(\hat{\theta})$ in an attempt to minimize the cosine distance of the generated image from the input image:

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

$$\mathcal{L}_i = \frac{\mathbf{F_i}(\mathbf{x_0}) \cdot \mathbf{F_i}(\hat{\mathbf{x}}_\mathbf{0}(\hat{\theta}))}{\|\mathbf{F_i}(\mathbf{x_0})\| \|\mathbf{F_i}(\hat{\mathbf{x}}_\mathbf{0}(\hat{\theta}))\|} \tag{6}$$

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. λ is a super parameter for balancing diversity and identity. Our training objective is as follows:

$$\mathcal{L} = \lambda \mathcal{L}_i(x_0, \hat{x}_0(\hat{\theta}))) + (1-\lambda)\mathcal{L}_p(s_0, \hat{s}_0(\hat{\theta})) \tag{7}$$

3.4 Class-Guidance Fine-tuning Reverse Process

According to Eq. 3, the backward generation process is denoising from $\hat{x}_T(\theta)$ to $\hat{x}_0(\theta)$. 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 classifier, 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 classifier. In terms of fine-tuning, our model learns to be self-supervised subject to the constraint of Eq. 7. Once the diffusion model has been fine-tuned, any input sketches can be processed into images, as shown in Fig. 1. More details on the fine-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 fine-tuning stage, we take 12.5K images from Sketchy [45] dataset with each image corresponding to $5 \sim 10$ pieces of sketches.

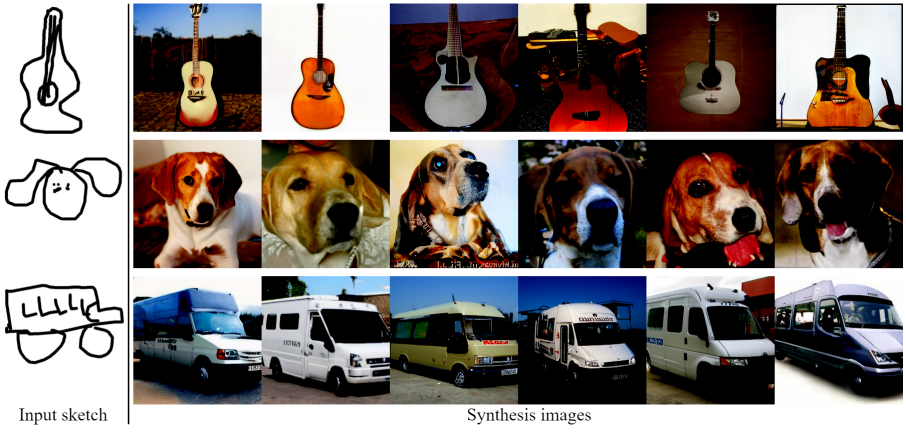


Figure 5: Qualitative results testing on Quickdraw.

Quantitative evaluation We measure our model sample quality based on Fréchet Inception Distance [20] (FID), Inception Score [1] (IS) , Precision and Recall metrics [29]. FID measures the distribution similarity between real images and generated images by comparing the mean and variance of image features. IS calculates the classification entropy of the generated image distribution. The Precision measures fidelity that the model samples are close to the data samples in VGG feature space [47], and the Recall measures diversity that the data samples are close to the model samples in VGG feature space.

Human study We conduct human study to judge the synthetic quality by comparing the output of different baseline methods with our method. Given an input sketch and the output of different methods, participants are asked to select the image that best conforms to the characteristics of sketch in the output. A total of 10 viewers were recruited for this test. We randomly selected 500 samples which were randomly displayed. And the percentage of each selected method was counted.

4.2 Comparison

Baselines To the best of our knowledge, this is the first time diffusion models have been used for sketch-based image synthesis and most of the previous works are based on GANs. We choose 3 baseline models and to be fair, all methods are tested on Sketchy evaluation dataset. (i) USPS [35] is an unsupervised GAN model consisting of two steps, translating the sketch shape into a gray-scale image and enriching it into a color image. It proposes an attention module to deal with abstraction and style variations which can improve the quality and realism of generation. (ii) MUNIT [24] is a general unsupervised multimodal image translation framework. MUNIT decomposes the image into a content code and a style code. It recombines the content code with the randomly sampled style code. And the model learns both codes at the same time. (iii) Sketch-YOG [52] pretrains a GAN-based generation model, utilizing cross domain adversarial learning and image space regulation to fine-tune.

Benchmarking and Qualitative results (i) Our model outperforms other baselines in almost all metrics listed in Table 1, indicating that we can restore images with more diversity and high fidelity. The higher human study score shows that our synthetic results are more in line with human intuitive judgments. (ii) Unlike USPS and MUNIT, our model does not

Table 1: Quantitative result. The best value is highlighted in black.

Method	FID ↓	IS ↑	Precision ↑	Recall ↑	Human ↑
USPS	48.73	23.74	0.42	0.38	26.45%
MUNIT	56.50	28.99	0.34	0.51	20.23%
Sketch-YOG	19.94	48.94	**0.70**	0.53	18.85%
Ours	**6.46**	**89.91**	0.68	**0.56**	**34.47**%
Ours(w/o $\mathcal{L}_p$)	7.22	83.43	0.33	0.39	N/A
Ours(w/o $\mathcal{L}_i$)	11.78	63.09	0.40	0.44	N/A
Ours(Quickdraw)	6.65	87.42	0.67	0.49	N/A

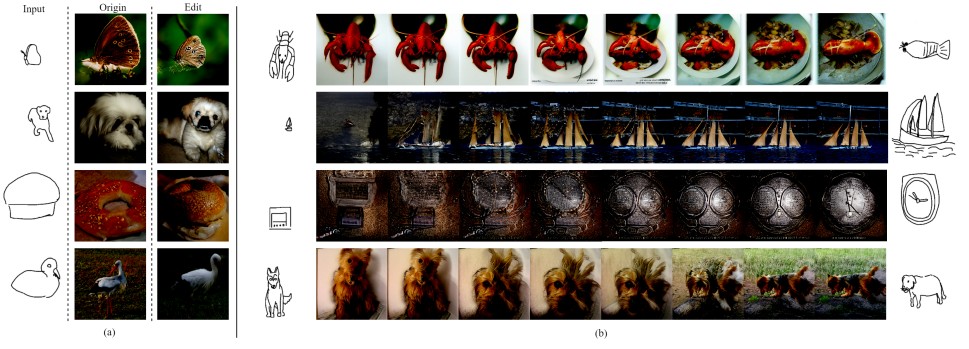


Figure 6: (a) Image editing. (b) Condition Interpolation.

need large-scale sketch image datasets. Due to the lack of such large datasets, many GAN-based sketch-to-image models can only synthesize a few categories such as shoes and chairs. We compared qualitative results on shoes, shown in Fig. 3. (iii) USPS and MUNIT generate image shapes that strictly match the input sketches and they focus on the generation of color, texture and shading. Whereas Sketch-YOG and our method give the model more imagination in terms of external contours, in particular we are able to generate more complex backgrounds, resulting in a higher IS score. (iv) The Precision is slightly lower than that of Sketch-YOG, indicating that our model is slightly less sensitive to the distribution of real data. More comparison results and analysis can be found in supplementary material.

4.3 Sketch-Based Image Synthesis

Fine-grained sketching controls image synthesis As shown in Fig. 4(a), when sketching a dog, the user' s simple strokes could not be identified as German Shepherd, Briard, Swiss Mountain Dog or any other categories. The trained classifiers enable users to specify the categories they want to generate in Fig. 4(b). More interestingly, when we specify categories that are not related to the original input sketch, we can still synthesize images that are similar in style and shape to the sketch, as displayed in Fig. 4(c).

Test on real human sketches To prove our model is more practical than other methods, we conduct tests on hand-drawn sketches. Quickdraw [17] collects real hand-painted sketches, which is simple and distorted. Qualitative and quantitative results are shown in Fig. 5 and Table 1, respectively. Since we calculate the global perception loss of sketch rather than the one-to-one pairing of local details, our model can still generate the user's ideal results from realistic and distorted input sketches with only slightly reduced quantitative indicators.

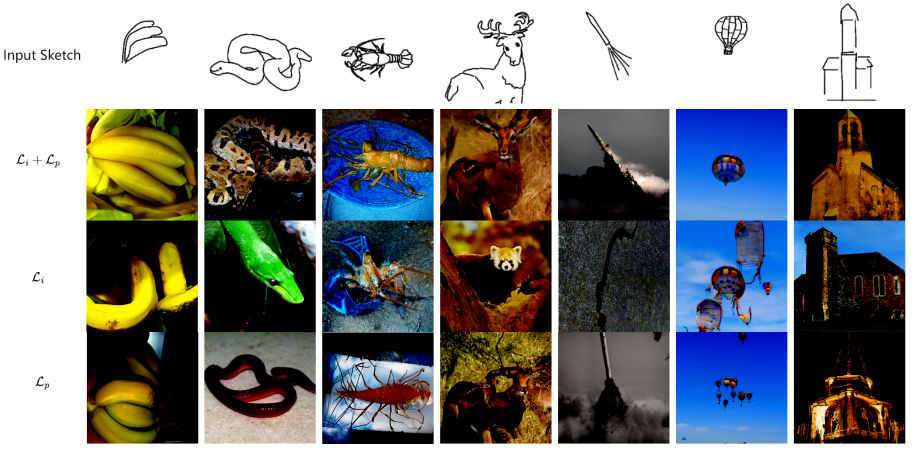


Figure 7: Ablation study of identity loss $\mathcal{L}_i$ and perceptual loss $\mathcal{L}_p$.

4.4 Applications

Image editing Our model can edit the original images under the guidance of input sketches to obtain new images. We use the Attentional block in Luhman *et al*. [36] to input the origin image information, retaining the basic texture and color information of the picture. And we modify the posture and shape features according to the user's input sketch to synthesize a new image. The results are shown in Fig. 6(a). Details about this method are provided in supplementary material.

Condition Interpolation Because of the consistency of DDIM, we use spherical linear [46] to combine different initial latent variables $x_T^{(0)}$ and $x_T^{(1)}$ to get a new $x_T^{(\alpha)}$.

$$x_T^{(\alpha)} = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} x_T^{(0)} + \frac{\sin(\alpha\theta)}{\sin(\theta)} x_T^{(1)} \tag{8}$$

where $\theta = \arccos\left(\frac{(x_T^{(0)})^\top x_T^{(1)}}{||x_T^{(0)}|||x_T^{(1)}||}\right)$, $\alpha \sim (0, \pi/2)$. We linearly extract eight α values, and show the results in Fig. 6(b). The left and right sketches are different inputs. Between them are reconstructed interpolation results in latent space. More results of interpolation method are shown in supplementary material.

4.5 Ablations

From Fig. 7, both image identity loss $\mathcal{L}_i$ and sketch perceptual loss $\mathcal{L}_p$ are the keys to the success of our model. (i) As shown in Table 1, the absence of either $\mathcal{L}_p$ or $\mathcal{L}_i$ degrades the quality of the synthesis. (ii) Without introducing $\mathcal{L}_p$, the model is unable to guide the generation process of image. And the synthesis result is not associated with input sketch. (iii) Without introducing $\mathcal{L}_i$, although the model can still recreate the general shape and position of the sketch, a great deal of identity and detail information will be lost, making the generated image feel vague in texture.

5 Conclusion

We propose DiffSketching, the first cross-domain sketch-to-image synthesis method utilizing diffusion models. Our method can be self-supervised when matching inputs, overcoming the large domain gap between sketch and generator's parameter space. We can distinguish sketches of simple lines through the classifier, showing strong content inference ability. And the DiffSketching outperforms GAN-based methods on many key metrics, achieving high-quality and realistic results. We further show the potential for application to other tasks, such as image editing and condition interpolation.

References

[1] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[2] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[3] Yoshua Bengio, Eric Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pages 226–234. PMLR, 2014.

[4] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics (TOG)*, 39(4):72–1, 2020.

[5] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM transactions on graphics (TOG)*, 28(5):1–10, 2009.

[6] Tao Chen, Ping Tan, Li-Qian Ma, Ming-Ming Cheng, Ariel Shamir, and Shi-Min Hu. Poseshop: Human image database construction and personalized content synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 19(5):824–837, 2012.

[7] 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.

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

[12] Mathias Eitz, Ronald Richter, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Photosketcher: interactive sketch-based image synthesis. *IEEE Computer Graphics and Applications*, 31(6):56–66, 2011.

[13] 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.

[14] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1171–1180, 2019.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[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. *arXiv preprint arXiv:2111.14822*, 2021.

[17] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[18] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[23] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022.

[24] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[26] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *arXiv preprint arXiv:2205.15996*, 2022.

[27] Yuma Koizumi, Heiga Zen, Kohei Yatabe, Nanxin Chen, and Michiel Bacchiani. Specgrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping. *arXiv preprint arXiv:2203.16749*, 2022.

[28] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021.

[29] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiang-Yang Li, Tao Qin, et al. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *arXiv preprint arXiv:2205.14807*, 2022.

[31] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Vqbb: Image-to-image translation with vector quantized brownian bridge. *arXiv preprint arXiv:2205.07680*, 2022.

[32] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photosketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1403–1412. IEEE, 2019.

[33] Yuhang Li, Xuejin Chen, Binxin Yang, Zihan Chen, Zhihua Cheng, and Zheng-Jun Zha. Deepfacepencil: Creating face images from freehand sketches. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 991–999, 2020.

[34] Han Lin, Maurice Pagnucco, and Yang Song. Edge guided progressively generative image outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 806–815, 2021.

[35] Runtao Liu, Qian Yu, and Stella X Yu. Unsupervised sketch to photo synthesis. In *European Conference on Computer Vision*, pages 36–52. Springer, 2020.

[36] Troy Luhman and Eric Luhman. Diffusion models for handwriting generation. *arXiv preprint arXiv:2011.06704*, 2020.

[37] Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.

[38] 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.

[39] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

[40] Kaiyue Pang, Da Li, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep factorised inverse-sketching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–52, 2018.

[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*, 2022.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[44] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[45] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35 (4):1–12, 2016.

[46] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[49] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 801–810, 2018.

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[51] Nao Takano and Gita Alaghband. Generator from edges: Reconstruction of facial images. In *International Symposium on Visual Computing*, pages 430–443. Springer, 2020.

[52] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14050–14060, 2021.

[53] Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2021.

[54] Julia Wolleb, Robin Sandkühler, Florentin Bieder, and Philippe C Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models. *arXiv preprint arXiv:2204.02641*, 2022.

[55] Xian Wu, Chen Wang, Hongbo Fu, Ariel Shamir, Song-Hai Zhang, and Shi-Min Hu. Deepportraitdrawing: Generating human body images from freehand sketches. *arXiv preprint arXiv:2205.02070*, 2022.

[56] Shunxin Xu, Dong Liu, and Zhiwei Xiong. E2i: Generative inpainting from edge to image. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1308–1322, 2020.

[57] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022.

[58] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.