Adaptively-Realistic Image Generation from Stroke and Sketch with Diffusion Model

Shin-I Cheng*[1], Yu-Jie Chen*[1], Wei-Chen Chiu[1], Hung-Yu Tseng[2], and Hsin-Ying Lee[3]

[1]National Chiao Tung University, Taiwan, [2]Meta, [3]Snap Inc.

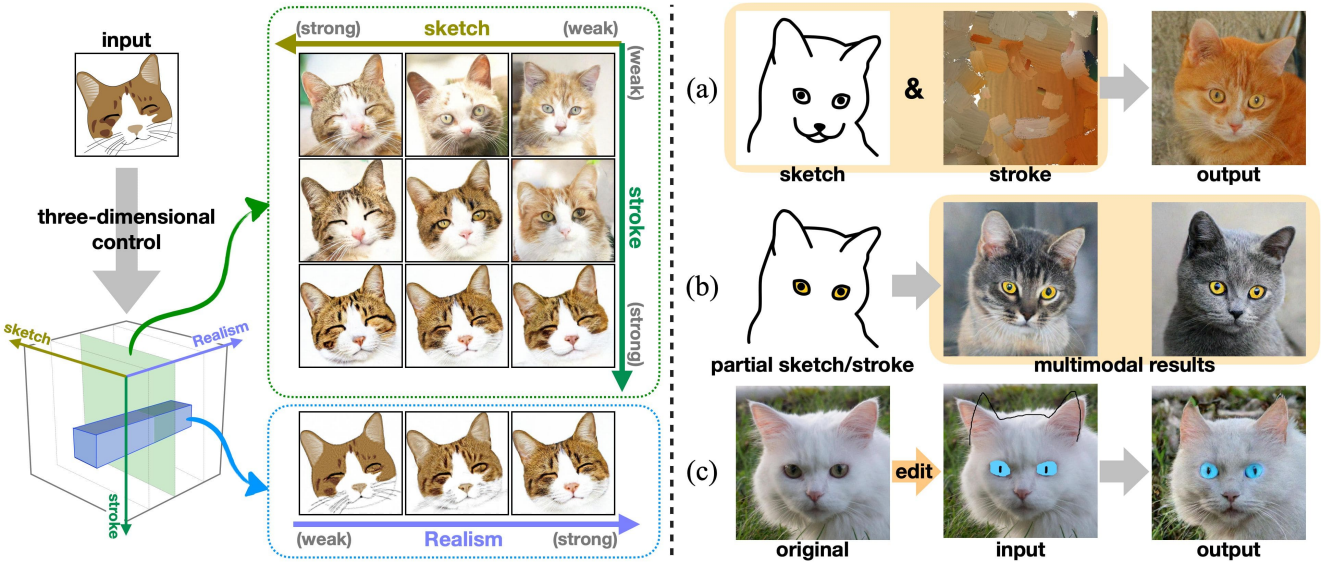


Figure 1: **Three-dimension controls of image generation from stroke and sketch.** (*left*) Our proposed model is able provide three-dimension controls over image synthesis from stroke and sketch. Given sketch and stroke as input, we can control the scales of faithfulness for the synthesized output with respect to the sketch and stroke, as well as the degree of its realism. (*right*) (a) Given sketch and strokes, we perform sketch/stroke-to-image translation. (b) We generate multimodal results with partial sketch/strokes as input. (c) Sketch/strokes conditioned local editing.

Abstract

Generating images from hand-drawings is a crucial and fundamental task in content creation. The translation is difficult as there exist infinite possibilities and the different users usually expect different outcomes. Therefore, we propose a unified framework supporting a three-dimensional control over the image synthesis from sketches and strokes based on diffusion models. Users can not only decide the level of faithfulness to the input strokes and sketches, but also the degree of realism, as the user inputs are usually not consistent with the real images. Qualitative and quantitative experiments demonstrate that our framework achieves state-of-the-art performance while providing flexibility in generating customized images with control over shape, color, and realism. Moreover, our method unleashes applications such as editing on real images, generation with partial sketches and strokes, and multi-domain multi-modal synthesis.

1. Introduction

Sketches and strokes are abstract depictions of objects and scenes. They represent different abstract illustrations that people have in mind and thus serve as important communication mediums. Conceivably, image synthesis from hand-drawn inputs can bridge human creations with reality, unleashing potential applications and assistance toward the content creation process.

*Equal contribution.
Project page: https://cyj407.github.io/DiSS/

Image generation from sketches and strokes is difficult as the translation is ill-defined and multimodal. For each sketch and stroke, different users will expect different outputs under different circumstances in terms of how faithful the results should be to the given inputs. Initially, the problem is formulated as image-to-image translation [21, 1, 15, 23, 21] with the help of generative adversarial networks (GANs) [6]. This stream of works is usually task-dependent, requiring different models for various tasks (e.g. separate models for stroke-to-image and sketch-to-image translations). Moreover, they lack flexibility and controllability in terms of the degree of faithfulness. Recently, diffusion models [18, 5] shed light on the tasks with high-quality image synthesis and stable training procedures. Variants of diffusion models are proposed to handle conditions in different forms, such as category [5, 9], reference image [3], and stroke-based painting [17]. Therefore, utilizing diffusion models, we would like to explore the possibility of a unified framework that can consider all factors of interest, including contour, colors, consistency, and realism.

In this paper, we introduce DiSS, a **DI**ffusion-based framework that generates images from **S**ketches and **S**trokes while enabling a three-dimensional (contour, color, realism) control over the degree of consistency to the input. First, unlike previous works using either black-white sketches or stroke paintings, we propose to handle both factors simultaneously, which is not trivial because it often comes with a trade-off between faithfulness to shape and to color. To provide disentangled control for the consistency of sketch and stroke, we adopt classifier-free guidance [9] to support two-dimensional control. Upon disentangling the shape and color information, we can customize the generative process and separately adjust the sampling results depending on users' demands. However, the input strokes and sketches from general users are often inconsistent with the distribution of real images. Therefore, we propose the third control factor, the realism scale, to realize a trade-off between consistency and realism. Specifically, we apply iterative latent variable refinement [3] and utilize a low-pass filter to adjust the coarse-to-fine features of the referred drawings.

With three-dimensional control, the proposed DiSS provides flexible editability, as shown in Fig. 1. Users can decide to what extent the faithfulness should be to the input sketch and strokes, and to what degree the results are close to real images. DiSS naturally unleashes several applications. First, multi-modal multi-domain translation (Fig. 4) can generate diverse results in multiple domains guided only by sketches and strokes without explicit labels. Second, multi-conditioned local editing (Fig. 5(a)) enables users to edit existing images by simply drawing contours and colors. Third, region-sensitive stroke-to-images (Fig. 5(b)) supports inputs that are not fully colored and provide variations on the blank regions.

We evaluate the proposed framework quantitatively and qualitatively on the three-dimension controllability. We measure the realism and perceptual quality with Fréchet inception distance (FID) [7], LPIPS [24], and subjective study on the AFHQ [4], Oxford Flowers [19], and Landscapes-HQ [22] datasets. Qualitatively, we present the diverse image synthesis conditioned on different kinds of drawings and demonstrate the adjustment of the three scales.

We summarize our contributions as follows: We present a unified framework of adaptively-realistic image generation from stroke and sketch that encodes the condition of the given stroke and sketch with the classifier-free guidance mechanism and adjusts the degree of realism with a latent variable refinement technique. The proposed framework enables a three-dimensional control over image synthesis with flexibility and controllability over shape, color, and realism of the generation, given the input stroke and sketch. Moreover, our proposed work unleashes several interesting applications: multi-conditioned local editing, region-sensitive stroke-to-image, and multi-domain sketch-to-image.

2. Related Works

2.1. Image Generation from Hand-drawings

Sketch-to-Image (S2I) generation aims at learning the mapping and eliminating the domain gap between hand-sketches and real images, which is usually modeled as an image-to-image translation task. Early works on image-to-image translation [10, 25, 12, 26, 2, 20] learn to map machine-generated edge maps or segmentation maps, where the distributions are quite different from real-world hand-drawn sketches ans scribbles, to real images. In turn, Scribbler [21] and SketchyGAN [1] are among the pioneering works to specifically tackle the translation upon sketch inputs. However, the training procedure requires datasets composed of various paired sketch-image data, which is not only hard to collect but also potentially limits the resultant translation model to tackle the general misalignment between sketches and images other than the one being seen in training set. More recently, efforts are made to address the S2I task via unsupervised and self-supervised learning, where the gray-scale version of photos serve as an auxiliary intermediate representation [15] or an autoencoder is adopted to learn disentangled style and content factors [14]. However, these GAN-based methods suffer from unstable training and quality. Moreover, hand-drawings can be decomposed into sketches that model contours and strokes that model colors, yet the GAN-based models are usually task-specific, where different pipelines are adopted for different settings. Finally, it is difficult for GAN-based models to control the level of faithfulness to the input, which is a crucial property of S2I. Therefore, inspired by the recent success of diffusion models [8, 5], we propose a diffusion-based framework supporting flexible controls over the level

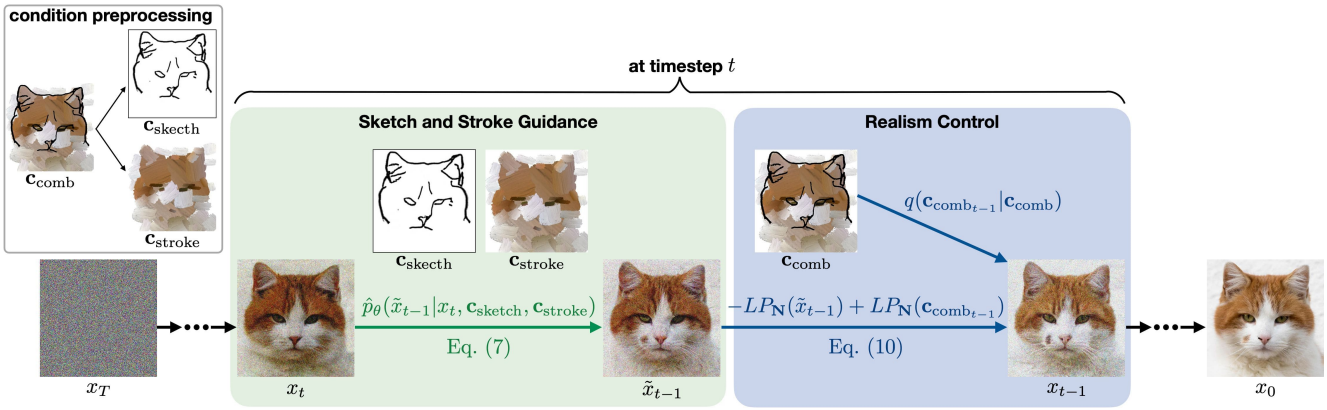


Figure 2: **Conditional denoising process .** At each time-step t, our proposed pipeline first performs classifier-free guidance with $\mathbf{c}_{\text{sketch}}$ and $\mathbf{c}_{\text{stroke}}$, which are extracted from a single input of colorful drawing $\mathbf{c}_{\text{comb}}$, and then controls the fidelity/realism by refining x_{t-1} with the input $\mathbf{c}_{\text{comb}}$, in which such realism control is realized by iterative latent variable refinement.

of faithfulness to shapes, colors, and realism.

2.2. Diffusion Models

Diffusion models are flourishing in recent years as a powerful family of generative models with diversity, training stability, and easy scalability, which GANs commonly lack. Fundamentally, diffusion models fulfill the sampling from a target distribution by reversing a progressive noise diffusion process, in which the process is defined as a Markov chain of diffusion steps for adding noise to the data. In addition to providing competitive or even superior capability on unconditional image generation [8, 18] in comparison to GANs, diffusion models also make significant progress on various tasks of conditional generation. Given a target class, [5] proposes a classifier-guidance mechanism that adopts a pretrained classifier to provide gradients as guidance toward generating images of the target class. More recently, *classifier-free diffusion guidance*[9] introduces a technique which jointly trains a conditional and an unconditional diffusion model without any pretrained classifier. Other than directly modifying the network of an unconditional diffusion model for conditional generation, ILVR [3] instead proposes to iteratively introduce the condition into the generative process via refining the intermediate latent images with a noisy reference image at each time-step during sampling. Therefore, ILVR is able to sample high-quality images while controlling the amount of high-level semantics being inherited from the given reference images. As the nature of diffusion models for adopting a progressive denoising process, the generation/synthesis via sampling can start from a noisy input (similar to the intermediate stage of sampling) instead of always beginning from random noise. SDEdit [17] hence realizes stroke-based image synthesis by starting the sampling from a stroke input with noise injected, in which the generative model used in SDEdit is built upon stochastic differential equations where its mechanism is quite similar to diffusion models (e.g. sampling via iterative denoising). In this work, we exploit both the techniques of classifier-free diffusion guidance and ILVR into our diffusion-based framework of image generation for fulfilling a three-dimensional control on the synthesized images in terms of their realism and the consistency with respect to the stroke and sketch conditions.

3. Method

As motivated above, our proposed framework, named DiSS, aims to perform image generation conditioned on the input of stroke and sketches with three-dimensional control over the faithfulness to the conditions and the realism of the synthesized output. In the following we sequentially describe our proposed method, starting from the preliminaries for diffusion models (Section 3.1) and the modifications we make for realizing the conditional generation and the discussion for the sketch and stroke guidance (enabled by the technique of classifier-free diffusion guidance, Section 3.2), and the control over realism (achieved by the technique of iterative latent variable refinement, Section 3.3).

3.1. Preliminaries

Denoising diffusion probabilistic models (DDPM) [8, 18] are a class of generative models (and the diffusion models that our proposed framework is based on) which adopt a denoising process to formulate the mapping from a simple distribution (e.g., isotropic Gaussian) to the target distribution. The forward diffusion process gradually adds noises to the data sampled from the target distribution, while the backward denoising process attempts to learn the reverse mapping. Both processes are modeled as Markov chains. Here we briefly introduce the process following the formulations and notations in [18].

Given a sample from the target data distribution $x_0 \sim q(x_0)$, the forward diffusion path of DDPM is a Markov

chain produced by gradually adding Gaussian noise to x_0 with total T steps:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $t \sim [1, T]$ and β_1,...,β_T is a fixed variance schedule with $\beta_t \in (0, 1)$. Sampling x_t at an arbitrary timestep t can be expressed in a closed form:

$$\begin{aligned} &q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \\ &x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \end{aligned} \quad (2)$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$. Consequently, x_t can be viewed as a linear combination of the original data x_0 and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The true posterior $q(x_{t-1}|x_t)$ can be well approximated by a diagonal Gaussian when the magnitude of noise β_t added at each step is small enough. Moreover, x_T is nearly an isotropic Gaussian $\mathcal{N}(0, \mathbf{I})$ when T is large enough. These behaviors facilitate a generative (denoising) process learning, the reverse of the forward path, to approximate the true posterior $q(x_{t-1}|x_t)$. Specifically, DDPM adopts a deep neural network (typically U-Net is adopted) to predict the mean and the covariance of x_{t-1} given x_t as input and the generative process is expressed as parameterized Gaussian transitions:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (3)$$

Ho *et al*. [8] propose to predict the noise $\epsilon_\theta(x_t, t)$ instead and derives $\mu_\theta(x_t, t)$ using Bayes's theorem:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)). \quad (4)$$

To perform the learning of the denoising process, we first generate sample $x_t \sim q(x_t|x_0)$ by adding Gaussian noise ϵ to x_0(i.e. Eq. 2), then train a model $\epsilon_\theta(x_t, t)$ to predict the added noise using a standard MSE loss:

$$L_{\text{simple}} := E_{t\sim[1,T], x_0\sim q(x_0), \epsilon\sim\mathcal{N}(0,\mathbf{I})}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2]. \quad (5)$$

For $\Sigma_\theta(x_t, t)$, Nichol *et al*. [18] presents an effective learning strategy as an improved version of DDPM with fewer steps needed and applies an additional loss term L_{vlb} (the details are shown in [18]) that interpolates between the upper and lower bounds for the fixed covariance proposed by the original DDPM. The overall hybrid objective that we adopt is:

$$L_{\text{hybrid}} := L_{\text{simple}} + L_{\text{vlb}}. \quad (6)$$

3.2. Sketch- and Stroke-Guided Diffusion Model

To generate images based on the given sketches and strokes, our proposed method concatenates the sketch condition $\mathbf{c}_{\text{sketch}}$ and the stroke condition $\mathbf{c}_{\text{stroke}}$ along with x_t as input for the U-Net model (which is responsible for posterior prediction). The modified parameterized Gaussian transition for conditioning generation is then represented as:

$$\begin{aligned} &\hat{p}_\theta(\tilde{x}_{t-1}|x_t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}}) \\ &:= \mathcal{N}(\tilde{x}_{t-1}; \mu_\theta(x_t, t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}}), \Sigma_\theta(x_t, t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}})). \end{aligned} \quad (7)$$

In practice, as following [8], the conditioning denoising process learns the noise prediction with additional sketch and stroke information, denoted as $\hat{\epsilon}_\theta(x_t, t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}})$:

$$\hat{L}_{\text{simple}} := E_{t,x_0,\epsilon}[\|\epsilon - \hat{\epsilon}_\theta(x_t, t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}})\|^2]. \quad (8)$$

To separately control the guidance level of the sketch and stroke conditions, we leverage classifier-free guidance [9] and modify it for two-dimensional guidance. In practice, we adopt a two-stage training strategy. First, we train the model with complete sketches and strokes as conditions. Then we fine-tune the model by randomly replacing 30% of each condition with an image filled with gray pixels, denoted as $\emptyset$, for unconditional representation. During sampling, the ratio between the degree of faithfulness to strokes and sketches is controlled through the following linear combination with two guidance scales s_{sketch} and s_{stroke}:

$$\begin{aligned} &\hat{\epsilon}_\theta(x_t, t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}}) = \hat{\epsilon}_\theta(x_t, t, \emptyset, \emptyset) \\ &+ s_{\text{sketch}}(\hat{\epsilon}_\theta(x_t, t, \mathbf{c}_{\text{sketch}}, \emptyset) - \hat{\epsilon}_\theta(x_t, t, \emptyset, \emptyset)) \\ &+ s_{\text{stroke}}(\hat{\epsilon}_\theta(x_t, t, \emptyset, \mathbf{c}_{\text{stroke}}) - \hat{\epsilon}_\theta(x_t, t, \emptyset, \emptyset)). \end{aligned} \quad (9)$$

With this formulation, our model supports multi-guidance on a single diffusion model.

3.3. Realism Control

In reality, sketches and strokes provided by users are usually inconsistent to the real images. Therefore, it is essential to provide the control over how faithful the output should be to the inputs. In other words, how realistic the output should be. We then provide realism control in addition to the two-dimensional classifier-free guidance with sketch and stroke information. We apply iterative latent variable refinement [3] to refine each intermediate transition in the generative process with a downsampled reference image. The proposed realism control allows additional trade-off between consistency to the provided strokes/sketches and the distance to target data distribution (i.e. real images). Let LP represents a linear low pass filtering operation which performs downsampling to a transformed size $\mathbf{N}$ and upsampling back. Given a realism scale $s_{\text{realism}} \sim [0, 1]$ as an indication of the transformed size $\mathbf{N}$ and a reference image combining sketch and stroke information $\mathbf{c}_{\text{comb}}$ of size $\mathbf{m} * \mathbf{m}$, the realism adjustment during the conditioning generative process at timestep t can be expressed as:

$$\begin{aligned} &\tilde{x}_{t-1} \sim \hat{p}_\theta(\tilde{x}_{t-1}|x_t, \mathbf{c}_{\text{sketch}}, \mathbf{c}_{\text{stroke}}), \\ &x_{t-1} := \tilde{x}_{t-1} - LP_{\mathbf{N}}(\tilde{x}_{t-1}) + LP_{\text{N}}(\mathbf{c}_{\text{comb}t-1}), \\ &\text{in which } \mathbf{N} = -s_{\text{realism}}(\mathbf{m}/8 - 1) + (\mathbf{m}/8) + k \end{aligned} \quad (10)$$

where $\mathbf{c}_{\text{comb}t-1} \sim q(\mathbf{c}_{\text{comb}t-1}|\mathbf{c}_{\text{comb}_0})$ with $\mathbf{c}_{\text{comb}_0} = \mathbf{c}_{\text{comb}}$ showing that $\mathbf{c}_{\text{comb}t-1}$ is sampled following Eq. 2 as gradually injecting noise into $\mathbf{c}_{\text{comb}}$ by $t-1$ steps. In details, as x_{t-1} can be seen as combining the high-frequency contents of $\tilde{x}_{t-1}$ (produced by $\tilde{x}_{t-1} - LP_{\mathbf{N}}(\tilde{x}_{t-1})$) with the low-frequency contents of the corrupted reference $\mathbf{c}_{\text{comb}t-1}$, the

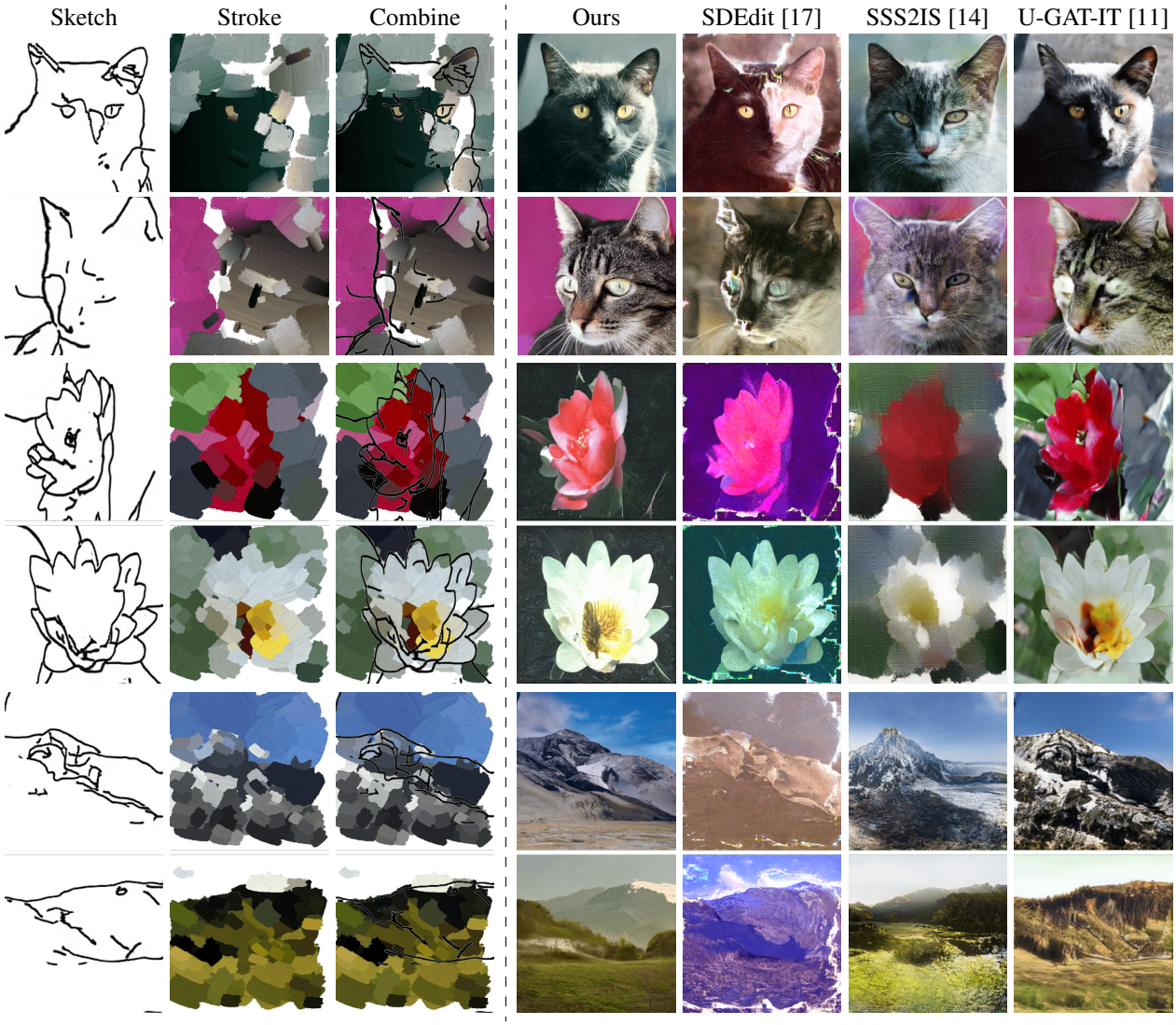


Figure 3: **Qualitative comparisons.** We present results from different approaches on the (*top two rows*) AFHQ, (*middle two rows*) Oxford Flower, and (*bottom two rows*) Landscapes datasets. U-GAT-IT [11], as an image-to-image translation method, takes as input the combination of sketches and strokes (third column). SDEdit [17], SSS2IS [14] and our model take the contour and color as separate inputs (the leftmost two columns).

downsampled size $\mathbf{N}$ controlled by s_{realism} determines how faithful the output should be to the reference (on the other hand, the tendency of the synthesized output towards the target distribution). The detailed discussion and explanation for the computation of $\mathbf{N}$ is provided in the supplementary materials.

The overall three-dimensional control of our proposed framework is illustrated in Figure 2, in which it is realized via the combination of the sketch- and stroke-guidance with the realism control.

4. Experiments

We conduct extensive qualitative and quantitative experiments to validate the effectiveness of the proposed DiSS method on the task of image generation from stroke and sketch. First, we compare our approach with several recent state-of-the-art frameworks, and demonstrate the three-dimensional control (contour, color, realism) over the generation process. Second, we show two applications: multi-conditioned local editing and region-sensitive stroke-to-image generation. Finally, we discuss the trade-off and the interaction between the three controllable dimensions.

Datasets. We conduct experiments using the AFHQ [4], Landscapes [22] and Oxford Flower [19] datasets. We use Photo-sketching [13] to generate the black sketches, and the stylized neural painting [27] as well as the paint transformer [16] model to synthesize the colored strokes for all the datasets. We provide more data preparation details in

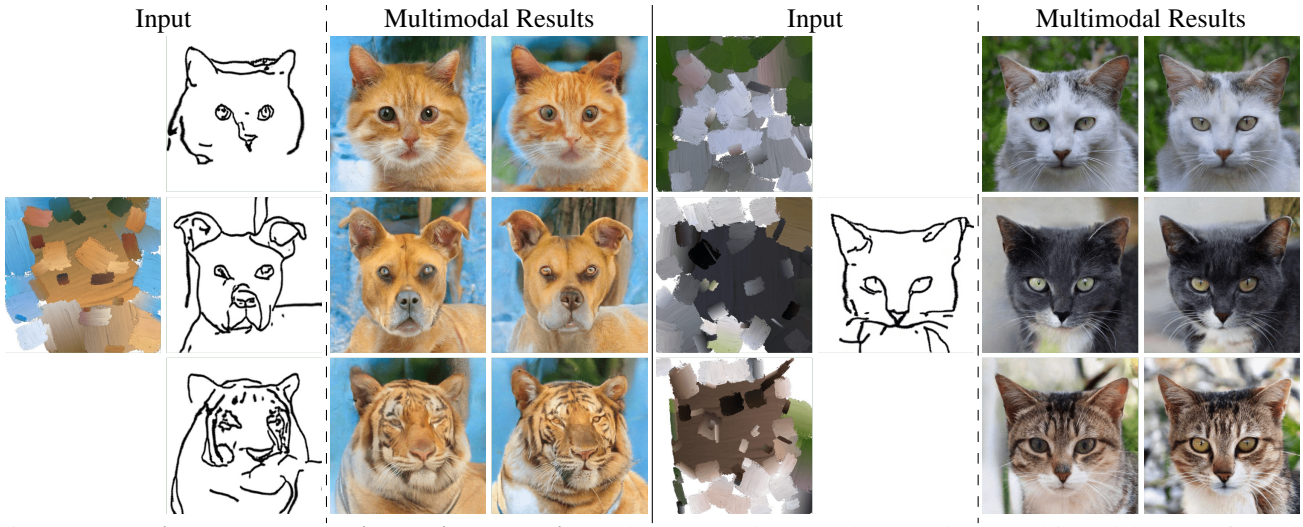


Figure 4: **Multi-modal and multi-domain generation.** The proposed approach 1) produces multi-modal results from the same set of input data, 2) understands the implicit class information from the input sketch image (as shown in the left hand side of the figure), and 3) is robust to *un-aligned* sketch stroke input data (note the input sketch and stroke are extracted from different source images in this example).

the supplementary document.

Compared methods. We compare our method with three recent state-of-the-art frameworks on image generation via the stroke and sketch task:

- **U-GAT-IT** [11] is a recent image-to-image translation approach. To leverage U-GAT-IT, we overlay the black sketches and the colored stroke to form the drawing image, which is considered to belong to the source domain. The corresponding photo-realistic image is then treated as the target domain image.
- **SSS2IS** [14] is a self-supervised GAN-based scheme that takes as input a black sketch and a style image. We retrain the model by replacing the input style images with a colored stroke image, and computing the regression loss between the real image and the autoencoder output.
- **SDEdit** [17] is a diffusion-based algorithm for the stroke-to-image generation task. To involve the sketch guidance, we retrain the model to take the sketch as the conditional signal by concatenating the sketch image with the original input of the U-Net network.

4.1. Qualitative Evaluation

Adaptively-realistic image generation from sketch and stroke . We present the qualitative comparisons between the proposed DiSS and other methods in Figure 3. Compared to the other frameworks, the proposed DiSS approach produces more realistic results on the object-level (cats and flowers) and scene-level (landscapes) datasets. Moreover,

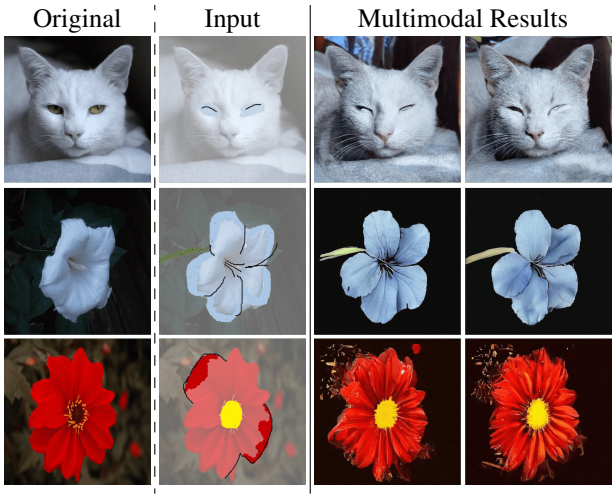
[†]Source from https://thenounproject.com/icon/cat-975212/
[‡]Source from https://freesvg.org/kocka
[§]Source from https://free-vectors.net/nature/green-field-vector

Table 1: **Quantitative comparisons.** We use the FID (↓) metric to measure the generated image quality, and the LPIPS (↓) score to evaluate the consistency between the synthesized images and the input sketches.

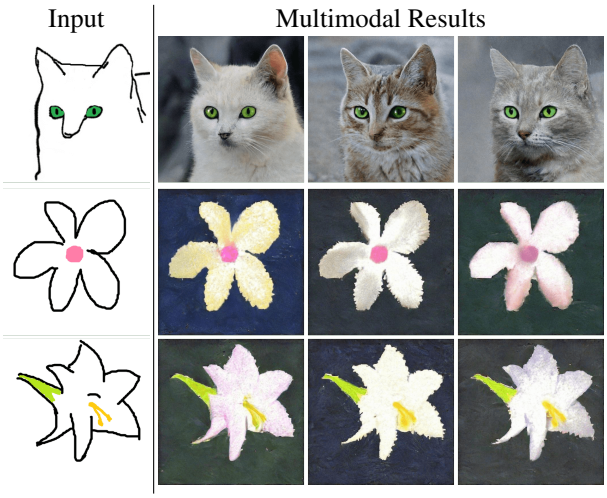
	AFHQ-cat		**Flowers**		**LHQ**	
	FID	LPIPS	FID	LPIPS	FID	LPIPS
U-GAT-IT	24.75	0.185	**74.27**	0.207	**36.93**	0.188
SSS2IS	85.48	0.23	275.24	0.227	62.25	0.143
SDEdit	30.55	0.178	138.97	0.196	84.67	0.15
Ours	**15.27**	**0.148**	83.12	**0.125**	38.83	**0.117**

the images generated by our scheme faithfully correspond to the input contour and color information. It is also noteworthy that our method is robust to different levels of details provided by the contour image. For example, in the second row of Figure 3, the proposed DiSS still synthesizes photo-realistic result even the contour image does not indicate the eye position of the cat. Finally, we demonstrate the variation produced by changing the three controllable scales in Figure 1 (sketch/stroke), Figure 7 (realism), and Figure 8 (sketch/stroke).

Multi-modal multi-domain translation. As the input only contains rough contour and colored stroke information, our DiSS approach is capable of synthesizing multiple (i.e. multimodal) image generation results (based on different initial randomly-drawn noises x_T and stochastic sampling procedure). The results are shown in Figure 4. Note all the images (cats, dogs, and wild animals) are synthesized from the same trained model. This suggests that the proposed model is able to understand the implicit category information from the input sketches. In addition to the



(a) Multi-conditioned local editing.



(b) Region-sensitive stroke-to-image.

Figure 5: **Applications.** (a) By drawing the new contour or color on an existing image, the proposed model enables the mask-free image editing. (b) With the partial colored stoke as the input, the proposed method synthesizes more diverse contents in the non-colored region. Here we use a cat contour[†] and hand-drawing flowers as examples.

diverse generation results, since the input sketch and stroke images are not extracted from the same source image, we demonstrate that our method is also robust to the *un-aligned* sketch-stroke input data.

Applications. The proposed DiSS approach not only offers the three-dimensional control over the generation process, but also enables two interesting applications: multi-conditioned local editing and region-sensitive stroke-to-image generation. Note that we do *not* retrain our model, but only design a specific inference algorithm for these two applications. We provide the details in the supplementary document. First, we present the visual editing results in

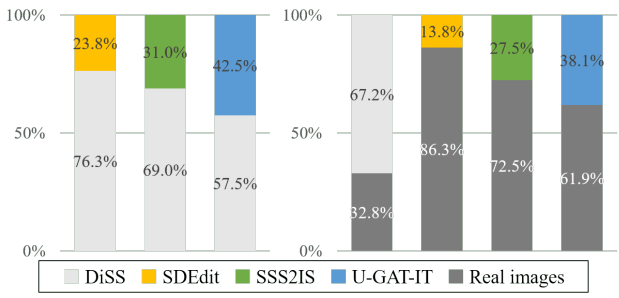


Figure 6: **User preference study.** We conduct the study that asks participants to select results (based on images generated from AFHQ-cat and Landscapes datasets) that are *more realistic*. The number indicates the percentage of preference for that particular pair-wise comparison.

Figure 5 (a). Our model enables flexible local manipulation on an existing image, which refers to both the hand-drawn contour and colored strokes. Secondly, we demonstrate the region-sensitive stroke-to-image generation results in Figure 5 (b). The proposed approach can take partial-sketch as input and produces results that 1) match the appearance in the region of partial-sketch and 2) exhibit multiple plausible contents in the non-colored region.

4.2. Quantitative Evaluation

Image quality and correspondence to input sketch. We use the Fréchet Inception Distance (FID) [7] to measure the realism of the generated images. To evaluate whether the synthesized images correspond to the input sketch, we compute the Learned Perceptual Image Patch Similarity (LPIPS) [24] score on the *sketch* level. Specifically, we calculate the similarity between the input sketch and the sketch inferred from the generated image (via Photo-sketching [13]). Lower FID and LPIPS values indicate better perceptual quality and correspondence, respectively. The quantitative results in Table 1 show that our method performs favorably against other representative approaches.

User preference study. To further understand the visual quality of images generated from sketches and strokes, we conduct a user study (with more than 80 candidates in total) by pairwise comparison. We use the results generated from the AFHQ-cat and Landscapes datasets. Given a randomly-swapped pair of images sampled from real images and images generated from various methods, we ask the participants to choose the image which is *more realistic*. Figure 6 presents the statistics of users' preferences. The results validate the effectiveness of the proposed approach.

Realism vs. correspondence to input guidance. Figure 7 demonstrates the trade-off between the generated image realism and the correspondence between the generated image and the input guidance. We change the realism scale from low (0.0) to high (1.0) in this experiment.

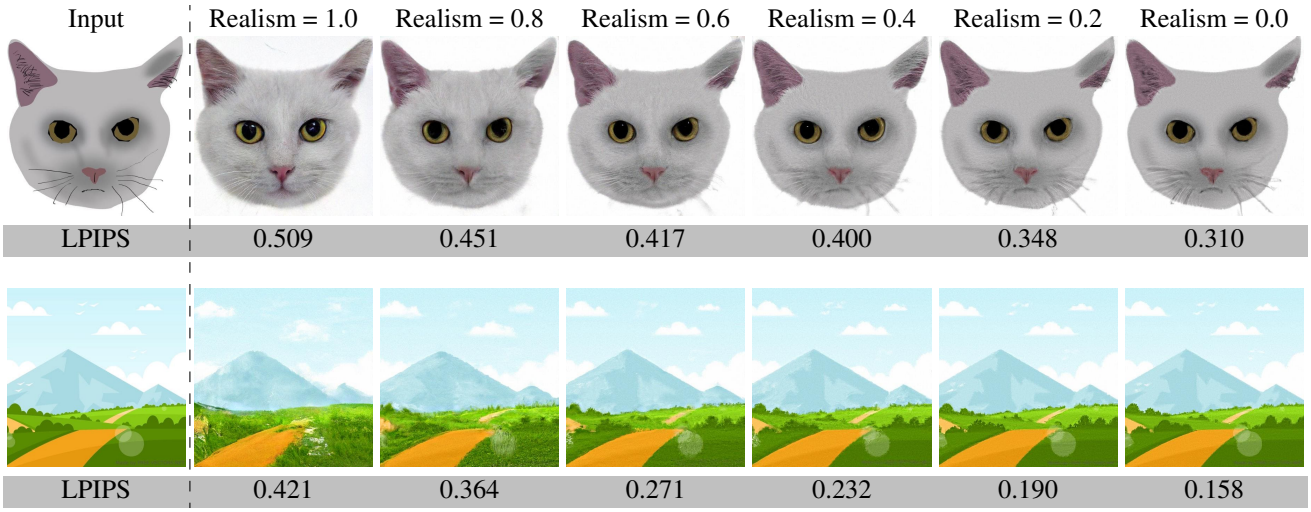


Figure 7: **Trade-off between realism and consistency to image guidance.** We demonstrate the trade-off between the image realism and the correspondence to the input guidance, where the realism scale is varied from low (0.0, *right*) to high (1.0, *left*). We also show the LPIPS scores between the generated image and the input guidance. Both the object-level (a cat drawing[‡]) and scene-level (a landscape painting[§]) input guidance images are used in this experiment.

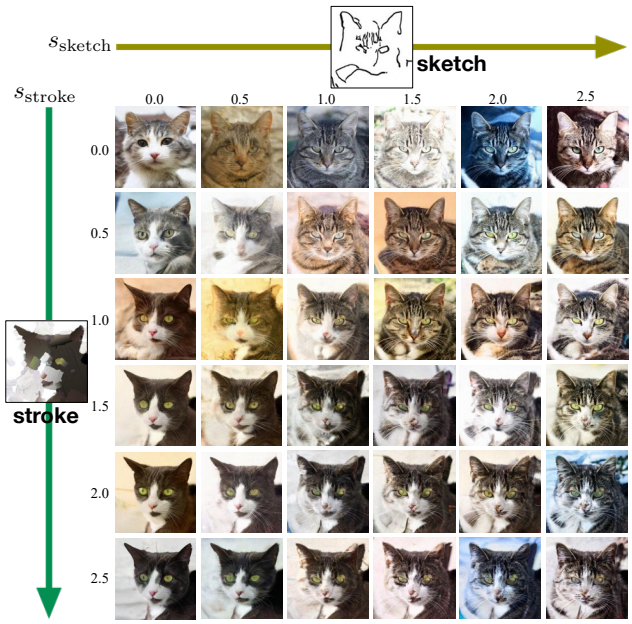


Figure 8: **Qualitative results of using different stroke and sketch scales.** The top-left corner show the results generated without guidance. Stronger scale values lead to results which are more consistent to the input guidance.

Controlling stroke and sketch scales. We conduct an ablation study to understand the impact of using different stroke and sketch scales. Figure 8 shows the qualitative results, and Figure 9 reports the FID scores computed using the AFHQ-cat dataset. The results show that we can obtain the best generated image quality by setting the sketch and stroke scale values in the interval of $[1.5, 2.5]$. More ablation study results about stroke, sketch and realism scales are provided in the supplementary document.

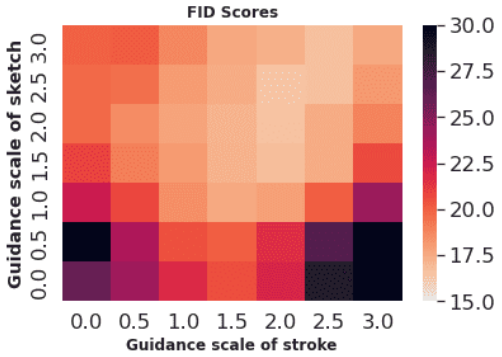


Figure 9: **Impact of various stroke and sketch scales on generated image quality.** We report the FID scores to indicate the generated image quality. The results suggest that setting the stroke and sketch scale values to be in interval of $[1.5, 2.5]$ lead to the best image quality.

5. Conclusion

In this work, we introduce DiSS, a versatile and flexible framework that synthesizes photo-realistic images from the sketch and colored stroke guidance. Our method uses 1) two-directional classifier-free guidance and 2) iterative latent variable refinement to offer the three-dimensional control (sketch, colored stroke, realism) over the image generation process. Extensive experimental results verify the effectiveness of the proposed approach against several representative schemes. Furthermore, we demonstrate that the proposed DiSS framework enables more interesting applications, such as mask-free local editing and region-sensitive stroke-to-image generation.

References

[1] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Yen-Chi Cheng, Hsin-Ying Lee, Min Sun, and Ming-Hsuan Yang. Controllable image synthesis via segvae. In *European Conference on Computer Vision*, 2020.

[3] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[7] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations (ICLR)*, 2020.

[12] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018.

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

[14] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Self-supervised sketch-to-image synthesis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[15] Runtao Liu, Qian Yu, and Stella X Yu. Unsupervised sketch to photo synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[16] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[17] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

[18] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

[19] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008.

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[21] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[23] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own GAN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[26] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017.

[27] Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. Stylized neural painting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.