SCEdit: Efficient and Controllable Image Diffusion Generation via Skip Connection Editing

Zeyinzi Jiang   **Chaojie Mao**   **Yulin Pan**   **Zhen Han**   **Jingfeng Zhang**

Alibaba Group

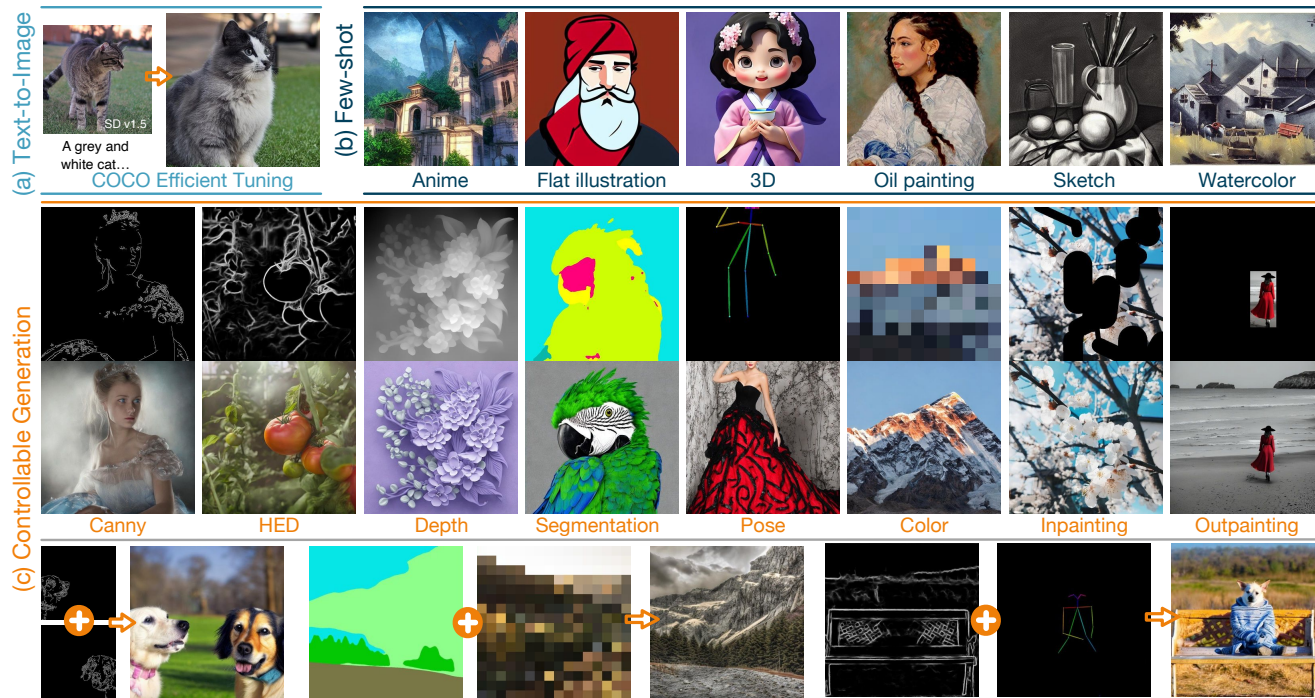{zeyinzi.jzyz, chaojie.mcj, yanwen.pyl, hanzhen.hz, zhangjingfeng.zjf}@alibaba-inc.com

Figure 1. **Images generated by SCEdit.** With a small number of trainable parameters and low memory usage, SCEdit enables efficient fine-tuning on specific datasets (left top) and supports transfer learning with a few samples (right top). Additionally, it adopts various conditions as inputs for efficient controllable generation (middle), while individually learned conditional models combine effortlessly, providing endless compositional possibilities (bottom).

Abstract

*Image diffusion models have been utilized in various tasks, such as text-to-image generation and controllable image synthesis. Recent research has introduced tuning methods that make subtle adjustments to the original models, yielding promising results in specific adaptations of foundational generative diffusion models. Rather than modifying the main backbone of the diffusion model, we delve into the role of skip connection in U-Net and reveal that hierarchical features aggregating long-distance information across encoder and decoder make a significant impact on the content and quality of image generation. Based on the observation, we propose an efficient generative tuning framework, dubbed **SCEdit**, which integrates and edits Skip Connection using a lightweight tuning module named SC-Tuner. Furthermore, the proposed framework allows for straightforward extension to controllable image synthesis by injecting different conditions with Controllable SC-Tuner, simplifying and unifying the network design for multi-condition inputs. Our SCEdit substantially reduces training parameters, memory usage, and computational expense due to its lightweight tuners, with backward propagation only passing to the decoder blocks. Extensive experiments conducted on text-to-image generation and controllable image synthesis tasks demonstrate the superiority of our method in terms of efficiency and performance. Project page: https://scedit.github.io/.*

1. Introduction

Building upon large-scale pre-trained image diffusion models [1–3, 14, 15, 30], researchers have focused on various downstream tasks and applications, including text-to-image generation [7–9, 27, 31, 32], controllable image synthesis [18, 29, 33, 50] and image editing [4, 26, 37, 41, 48]. However, fully fine-tuning a foundation image diffusion model often proves inefficient or even impractical in most customized scenarios due to the constraints of limited training data and computational resources.

Recently, efficient tuning methods [20, 21, 46] have emerged as a practical solution by introducing additional trainable structures on generative tasks. Nonetheless, many of these popular efficient tuning methods still suffer from substantial resource consumption as the network expands. For instance, LoRA [17] typically adds the trainable low-rank matrices to multi-head attention layers all across the U-Net [36], and backward propagation is conducted throughout the entire backbone, resulting in an accumulation of gradients and increase in memory usage during training, even more than that in fully fine-tuning. To address this issue, we properly design our framework by strategically inserting all the trainable modules into the Skip Connections (SCs) between the encoder and decoder of the U-Net, effectively decoupling the encoder from the backpropagation process and significantly reducing computation and memory requirements, as shown in Fig. 2.

The SCs bridge the gap between distant blocks in the U-Net architecture, facilitating the integration of information over long distances and alleviating vanishing gradients. Some works [19, 38, 44] have focused on exploring the efficacy of SCs in enhancing the training stability of the U-Net, as well as in improving the generation quality [42]. Inspired, we further investigate the potential of U-Net's SCs in adapting to new scenarios. To gain a deeper insight into each SC within a pre-trained U-Net, we gradually remove SCs and observe the subsequent changes in the value distributions and feature maps across the blocks of the decoder in Fig. 3. We first assess the diversity of information in the final decoder's output by visualizing the distribution of its latent feature values. A higher variance signifies a greater breadth of information, while a variance near zero indicates a significant loss of detail. As we discard an increasing number of SC from half to all, the variance of the final decoder output gradually decreases. In addition, when half of the SCs are removed, there is a marked reduction in the level of detailed structural information within the feature maps of both the 6th and 11th decoder blocks. Eliminating all SCs further intensifies the deterioration of information. These trends further validate the significant impact of SC on the generation of detailed structural information.

In light of this revelation, we propose a simple yet highly efficient approach by **S**kip **C**onnection **Edit**ing for image
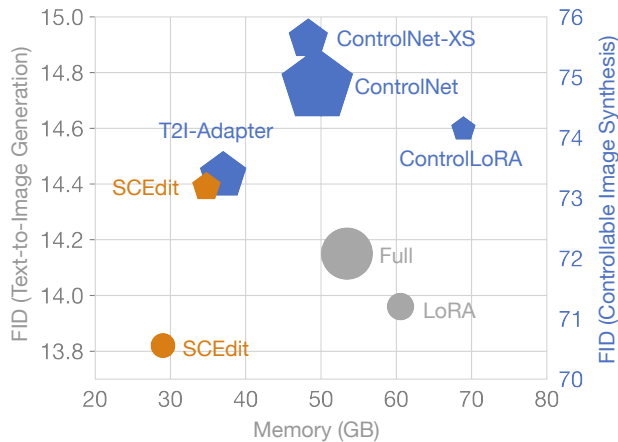

Figure 2. **Performance and efficiency comparison** on both text-to-image generation (circular markers) and controllable image synthesis (pentagonal markers) tasks. The marked area reflects the relative amount of parameters.
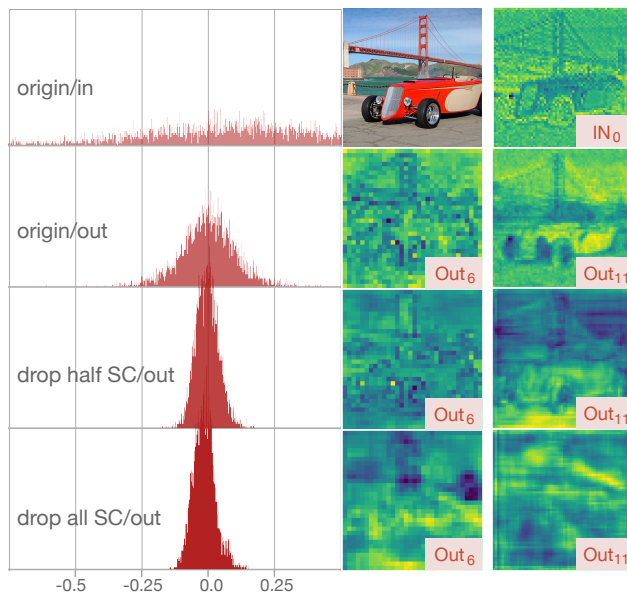

Figure 3. The **output distributions** (left) and **feature maps** (right) of pre-trained U-Net. Removing the skip connections from different layers markedly affects the overall network output.

generation, dubbed `SCEdit`. Specifically, we introduce a lightweight tuning module named `SC-Tuner`, which is designed to edit the latent features within each SC of the pre-trained U-Net for efficient tuning. Furthermore, we extend the capabilities of SC-Tuner to facilitate controllable image synthesis, which can accommodate various input conditions using the Controllable SC-Tuner (`CSC-Tuner`). Our SCEdit framework is adaptable to a broad spectrum of image generation tasks using the proposed tuning modules and by decoupling the encoder blocks in U-Net, it allows for efficient and flexible training, as it enables backward propagation solely through the decoder block.

We evaluate our SCEdit on efficient tuning of text-to-image generation tasks as well as controllable image synthesis tasks. For text-to-image generation tasks, our approach outperforms existing methods on COCO2017 dataset in terms of FID score and qualitative results, while also reducing memory consumption by 52% during the training stage. Additionally, faster transfer and high-quality results are achieved in the few-shot fine-tuning scenario. On the controllable generation tasks, our approach can easily control the results under various conditional inputs and show impressive results, while having lower computational costs than existing methods. It utilizes merely 7.9% of the parameters required by ControlNet and achieves a 30% reduction in memory usage.

2. Related work

Image Diffusion Models [15, 43, 44] have achieved the state-of-the-art performance in sample quality of generative image. Research on text-to-image diffusion models [8, 35] incorporates text latent vectors as conditions. Some works [18, 50] conduct controllable image generation by leveraging various images as conditions, offering personalization, customization, or task-specific image generation. However, the scale and computational cost limit the applications of image diffusion models.

Efficient Tuning [10, 16, 17, 21, 22] has become a popular solution in recent years, attracting significant attention. It allows for easy adaptation of image diffusion models by making light adjustments to the original pre-trained models. LoRA [17] is to use low-rank matrices to learn weight off-sets, which has proven effective in customized image generation scenarios. Other works, such as U-Tuning [20], Res-Tuning [21], and MAM-Adapter [11], propose a unified paradigm for efficient tuning methods, offering more options for tuning pre-trained models. Typically, some tuning modules are integrated into the U-Net architecture of diffusion models, but few works thoroughly analyze the key factors for improving the quality of generative images through tuning modules.

U-Net is originally introduced for diffusion models in DDPM [15], and it achieves tremendous success in the field of generative tasks. Recently, there has been a lot of attention focused on the effectiveness of skip connections in U-Net. To address the issue of doubling signal variance caused by skip connections and alleviate the oscillations of U-Net, some methods [38, 44] propose rescaling all skip connections by $\frac{1}{\sqrt{2}}$. ScaleLong [19] provides further theoretical analysis on why scaling the coefficients of skip connections can help improve the training stability of U-Net and using constant scaling or learnable scaling for better stability. FreeU [42] recognizes that the main backbone of U-Net plays a primary role in denoising, while the skip connections introduce high-frequency features into the decoder

module. By leveraging the strengths of both components, the denoising capability of U-Net can be enhanced. Inspired by previous works, we aim to analyze the roles of skip connections in the image generation process and propose a novel and effective framework incorporating tuning modules by editing skip connections.

3. Method

3.1. Preliminaries

Diffusion models [15] are a family of probabilistic generative models that aim at sampling high-fidelity images from Gaussian noise. It combines two fundamental processes: a diffusion process and a denoising process. In the diffusion process, it gradually decreases the signal-to-noise ratio of an image via a T-step Markov chain, following prescribed noise level schedules $[\beta_1, \beta_2, ..., \beta_T]$. At each step t, the noisy intermediate variable x_t is constructed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \quad (1)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. The denoising process reverses the above diffusion process by estimating the noise ϵ via a parameterized neural network, which is trained by minimizing the l_2 loss between estimated noise ϵ_θ and real noise ϵ:

$$\ell_{\text{simple}}^t(\theta) = \mathbb{E}_{x_0, t, \epsilon} \|\epsilon_\theta(x_t, t) - \epsilon\|_2^2. \quad (2)$$

U-Net [36] is already a widely adopted architecture in pixel-corresponded generative tasks, such as image restoration, medical segmentation, and the aforementioned image generation. Specifically, it first encodes the input into multi-scale features through multiple cascaded encoder blocks:

$$x_{i+1} = \mathcal{F}_i(x_i), \quad 0 \leq i \leq N - 1, \quad (3)$$

where i denotes the index number of the encoder layer, $\mathcal{F}_i$ denotes the i-th encoder block operation, N is the total number of encoder layers, x_{i+1} denotes the output of i-th encoder layer and x_0 means the original image input.

Subsequently, the decoder gradually decodes the feature via establishing a skip connection with corresponding encoder layer outputs, which complements the high-frequency information:

$$g_{j+1} = \mathcal{G}_j([x_{N-j}; g_j]), \quad 0 \leq j \leq N - 1, \quad (4)$$

where $[\cdot; \cdot]$ represents the concatenation operation, j denotes the index number of the decoder layer, $\mathcal{G}_j$ denotes the j-th decoder block operation, N is the total number of decoder layers which equals that of encoder layers, g_{j+1} denotes the output of j-th decoder layer, g_0 is the last output block before decoder input.
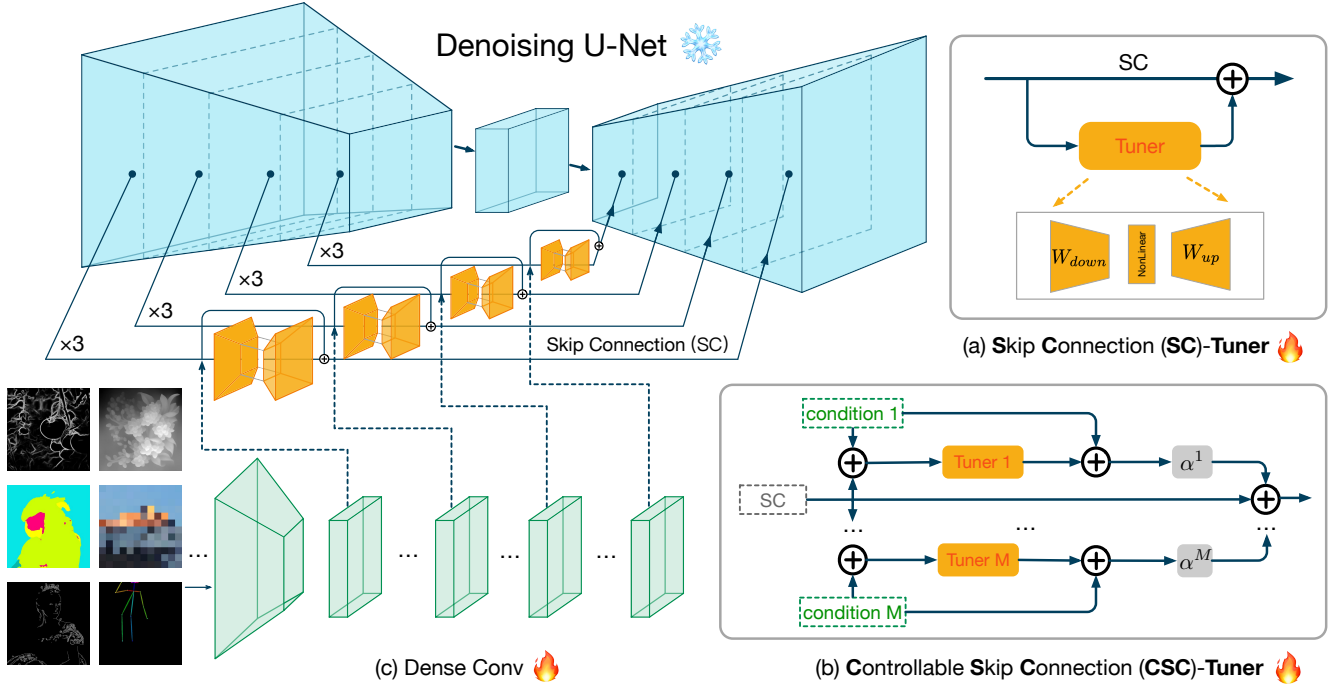
Figure 4. **Illustration of SCEdit framework.** Our method achieves efficient tuning by editing the features on skip connections which is proved to have rich structural information. Leveraging (a) SC-Tuner for text-to-image generation tuning, and controllable image synthesis can be achieved with the assistance of (b) CSC-Tuner and (c) Cascade Dense Convolution.

3.2. Tuner modules

We present **S**kip **C**onnection **Tuner**, termed `SC-Tuner`, a method designed to directly edit the latent features within skip connections. As illustrated in Fig. 4a, the SC-Tuner is composed of a tuning operation, referred to as Tuner OP, and a residual connection. The j-th SC-Tuner takes x_{N-j} as input and produces the sum of x_{N-j} and the output of Tuner OP with x_{N-j} as input. This process can be mathematically formulated as follows:

$$O_j^{SC}(x_{N-j}) = \mathcal{T}_j(x_{N-j}) + x_{N-j}, \qquad (5)$$

where $O_j^{SC}(x_{N-j})$ denotes the j-th SC-Tuner module with x_{N-j} as input, $\mathcal{T}_j$ denotes the Tuner OP of j-th SC-Tuner module.

The efficient tuning paradigm [21] has demonstrated comparable effectiveness across various Tuner OPs, including LoRA OP, Adapter OP, and Prefix OP, as well as their versatile combinations. In these independent OPs, we adapt the form of an Adapter OP, as it has been proven to be the simplest yet relatively effective method. The Tuner OP $\mathcal{T}_j$ can be defined as follows:

$$\mathcal{T}_j(x_{N-j}) = \mathbf{W}_j^{up}\phi(\mathbf{W}_j^{down}x_{N-j}), \qquad (6)$$

where $\mathbf{W}^{up}$ and $\mathbf{W}^{down}$ are up and down tunable projection matrices, respectively, ϕ is a GELU [13] activation

function. Formally, we can employ tuners of various types and scales to modify the features of skip connections.

Furthermore, The SC-Tuner can be easily adapted to support controllable image synthesis by incorporating the conditions and x_{N-j} as input, where conditions accommodate both single and multiple conditions. In the case of multi-condition controllable generation, we assign weights to different condition branches and combine them with original skip connections. This modified structure is depicted in Fig. 4b, named **C**ontrollable **SC-Tuner** or `CSC-Tuner`. We extend Eq. (5) by injecting extra conditional information which can be formulated as follows:

$$O_j^{CSC}(x_{N-j}, C_j) = \sum_{m=1}^{M}\alpha^m(\mathcal{T}_j^m(x_{N-j} + c_j^m) + c_j^m) \\ + x_{N-j}, \qquad (7)$$

where $O_j^{CSC}(x_{N-j}, C_j)$ denotes the j-th CSC-Tuner module with x_{N-j} and M conditions $C_j = \{c_j^0, \ldots, c_j^M\}$ as inputs, c_j^m is the j-th hint block's output of m-th condition features, α^m is the weight of different independent condition embeddings, and $\sum_{m=1}^{M}\alpha^m = 1$. In practice, we can engage in joint training with multi-conditions or perform inference under combined conditions directly after completing single-condition training. The control conditions include but are not limited to the canny edge, depth, hed boundary, semantic segmentation, pose keypoint, color, and masked image.

3.3. SCEdit framework

We introduce **SCEdit**, a framework designed for efficient **S**kip **C**onnection **Edit**ing in image generation that utilizes the SC-Tuner and CSC-Tuner. All the skip connections and conditions are fed into the SC-Tuner and CSC-Tuner, with the outputs subsequently concatenated to the original feature maps and input into the corresponding decoder blocks. We depict this process as follows:

$$g_{j+1} = \mathcal{G}_j([\text{O}_j(x_{N-j}, C_j); g_j]), \qquad (8)$$

where $\text{O}_j(x_{N-j}, C_j)$ denotes the j-th SC-Tuner or CSC-Tuner modules. By editing the original x_{N-j}, we can adapt it to different tasks.

SCEdit can facilitate flexibility and efficiency across both text-to-image generation and controllable image synthesis tasks. As illustrated in Fig. 4, the SC-Tuner is applied to the text-to-image generation task and integrated into all skip connections within the pre-trained U-Net architecture. The CSC-Tuner is employed for controllable image synthesis, where the input of corresponding condition features is encoded via a cascaded convolution network, as shown in Fig. 4c, consisting of a multi-layer hint block and several dense modules for feeding into the skip connection branch, where each level contains zero convolution layers and SiLU [13] activation functions, ensuring compatibility with the dimensions of skip connection.

4. Experiments

4.1. Experimental setups

Evaluation. We evaluate the flexibility and efficiency of our SCEdit, mainly through the text-to-image generation and controllable image synthesis tasks. We analyze its quantitative metrics such as trainable parameters, memory consumption, training speed, and FID [40] evaluation metrics to assess its performance and efficiency. Additionally, we consider qualitative evaluations of the quality and fidelity of the generated images.

Datasets. In the text-to-image generation task, we use the COCO2017 [24] dataset for training and testing, which consists of 118k images and 591k training prompts. Furthermore, we employ customized style datasets [28] with a limited samples to further validate the effectiveness of our approach. In controllable image synthesis tasks, for each kind of condition data, we utilize a filtered version of the LAION artistic dataset [39] that includes approximately 600k images that have been removed of duplicates, low-resolution images, those with a risk for harm, and those that are of low quality.

Baselines. According to task categories, the evaluation tasks can be characterized into two groups: **(i)** the text-to-image tuning includes fully fine-tuning and LoRA [17] tuning strategies; **(ii)** the controllable image synthesis focuses

| SD v1.5 | Full | LoRA | SCEdit |

[An Amtrak train sitting on the train track]

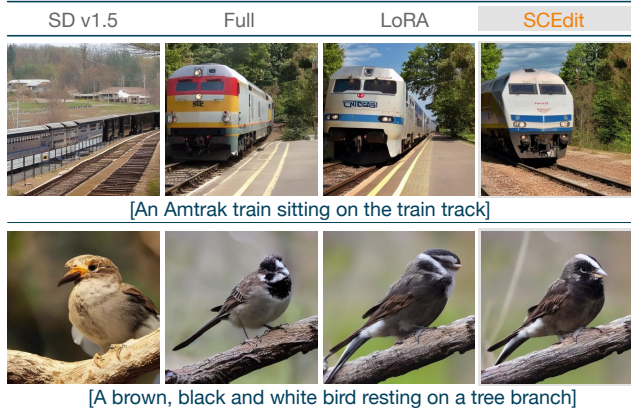[A brown, black and white bird resting on a tree branch]

Figure 5. **Qualitative comparison** of the original SD v1.5, existing tuning strategies, and our SCEdit using the same prompts.

Table 1. **Comparison of FID and efficiency on text-to-image generation** using COCO2017 dataset. For the FID score, we follow the default settings of SD v1.5 and randomly sample 10k image-text pair. In terms of efficiency, we compare the aspects of parameter-, memory-, and training time efficiency. We compare the performance of LoRA and our method under two different parameter settings.

Method	FID↓	Params	Mem.	Time
SD v1.5 [1]	15.48	-	-	-
Full	14.15	859.52M	53.46G	×1.0
LoRA/r=64 [17]	13.96	23.94M	60.57G	×1.24
LoRA/r=6 [17]	15.12	2.24M	59.94G	×1.20
SCEdit	**13.82**	19.68M	29.02G	×0.78
SCEdit$_{10}$	13.99	**1.98**M	**28.06**G	**×0.77**

on methods including additional information as input, such as ControlNet [50], T2I-Adapter [29], ControlLoRA [12], and ControlNet-XS [49]. Due to the distinct task characteristics of these two categories, we apply the SC-Tuner module for the former and CSC-Tuner for the latter.

Implementation details. For all experiments, we perform efficient fine-tuning based on the Stable Diffusion (SD) pre-trained model, where SD v1.5 [1] is used for the text-to-image task and SD v2.1 [2] for the conditional task, and the input image sizes for training are set to 512×512. We utilize the AdamW [25] optimizer with a fixed learning rate of 5e-5. Unless otherwise specified, models are trained for 100k steps.

Following ControlNet, we use different input conditions, *e.g.*, edge map [5, 47], depth map [34], segmentation map [23], and body key points [6], in addition to the color map condition in T2I-Adapter [29] and the mask-generation strategy presented in LaMa [45].

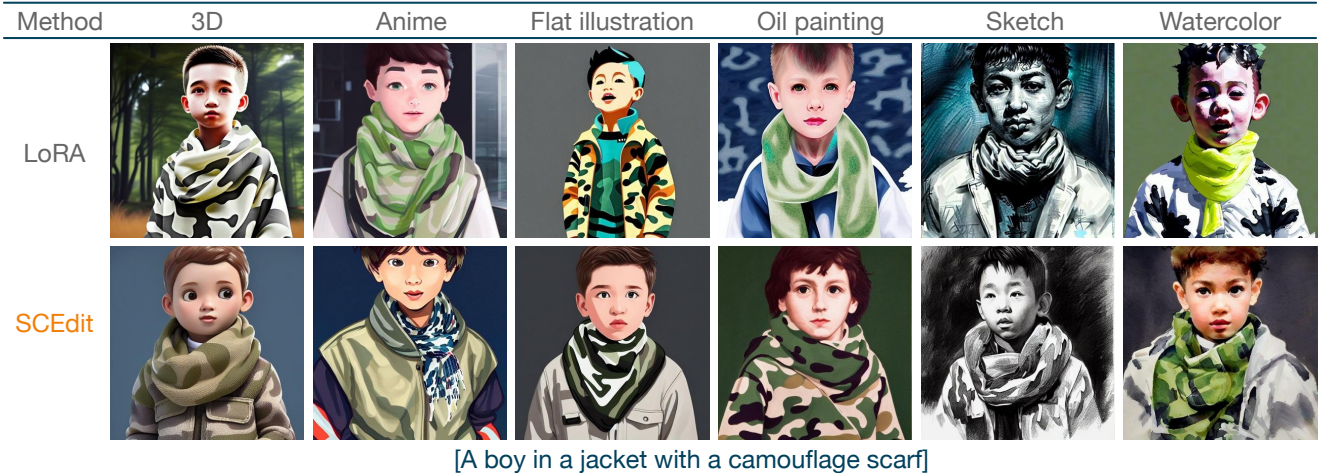| Method	3D	Anime	Flat illustration	Oil painting	Sketch	Watercolor

[A boy in a jacket with a camouflage scarf]

Figure 6. **Few-shot transfer learning comparison** on various custom-stylized datasets. Compared to LoRA tuning, SCEdit achieves more precise learning of style characteristics and generates images of superior quality.

4.2. Text-to-image generation

Performance and efficiency comparison. To evaluate the transferability of our method in downstream tasks, we fine-tuning COCO2017 with pre-trained U-Net and compare our method with other training strategies in terms of qualitative and quantitative results.

The qualitative results can be seen in Fig. 5, the left-most column is the zero-shot result of the original model, while the fine-tuned model obtained semantic comprehension capabilities on downstream tasks. Compared to existing strategies, our method exhibits fewer artificial artifacts in the generated images and higher visual quality. For instance, in the second row, the generated bird has more realistic details in the head while maintaining semantic comprehension.

In addition, as shown in Tab. 1, compared to the fully fine-tuning baseline, our SCEdit achieves 0.33 performance gain in FID score while using only 2.29% of the parameters and reducing nearly 22% training time. It is worth noted that at lower parameters compared to LoRA/r=64 strategies with rank 64, our method is lower in both FID, while the training memory can be reduced by 52.1% and the training time of LoRA is $1.6\times$ longer than ours. When we further reduce the hidden dimensions of the tuner by $10\times$, denoted as $SCEdit_{10}$, which corresponds to a reduction of parameters by the same factor compared to LoRA/r=6, our FID score shows a significant decreased by 1.13, while also maintaining a clear advantage in terms of memory usage and training time.

Few-shot transfer learning. In the image generation community, few-shot learning is a practical technique that enables users to train a personalized model with just a small subset of data. Our experiment involved performing few-

Table 2. **Comparison of FID and efficiency on controllable image synthesis** using LAION dataset within the canny edge condition. "k" denotes the convolution kernel size of conditional model, and a larger size performs better on FID score, albeit with a moderate increase in parameters.

Method	FID↓	Params.	Mem.	Time
ControlNet [50]	74.86	364.23M	49.51G	$\times 1.0$
T2I-Adapter [29]	73.37	77.37M	37.01G	$\times 0.92$
ControlNet-XS [49]	75.63	55.30M	48.32G	$\times 0.97$
ControlLoRA [12]	74.14	**21.52M**	68.92G	$\times 1.34$
SCEdit/k=1	73.18	28.82M	**34.78G**	$\times$**0.87**
SCEdit/k=3	**71.78**	99.11M	35.28G	$\times$**0.87**

shot transfer learning on a custom-stylized dataset, which included classes 3D, anime, flat illustration, oil painting, sketch, and watercolor, each with only 30 image-text pairs. Moreover, to prevent style leakage, we perform a cleansing of style-related words from the original prompts and incorporated specific trigger words <sce> during training to ensure the reliability of the experiment. In Fig. 6, we provide a comparison of the quality of samples between LoRA and our method, employing the same training setup. The results demonstrate that our method more accurately captures the style, aligning with the distribution of the original training data while maintaining text alignment. For instance, the flat illustration style retained the descriptions of a camouflage scarf, while the sketch style preserved the line-drawn depictions in monochrome.

4.3. Controllable image synthesis

Performance and efficiency comparison. Extensive experiments are conducted to evaluate the effectiveness of our
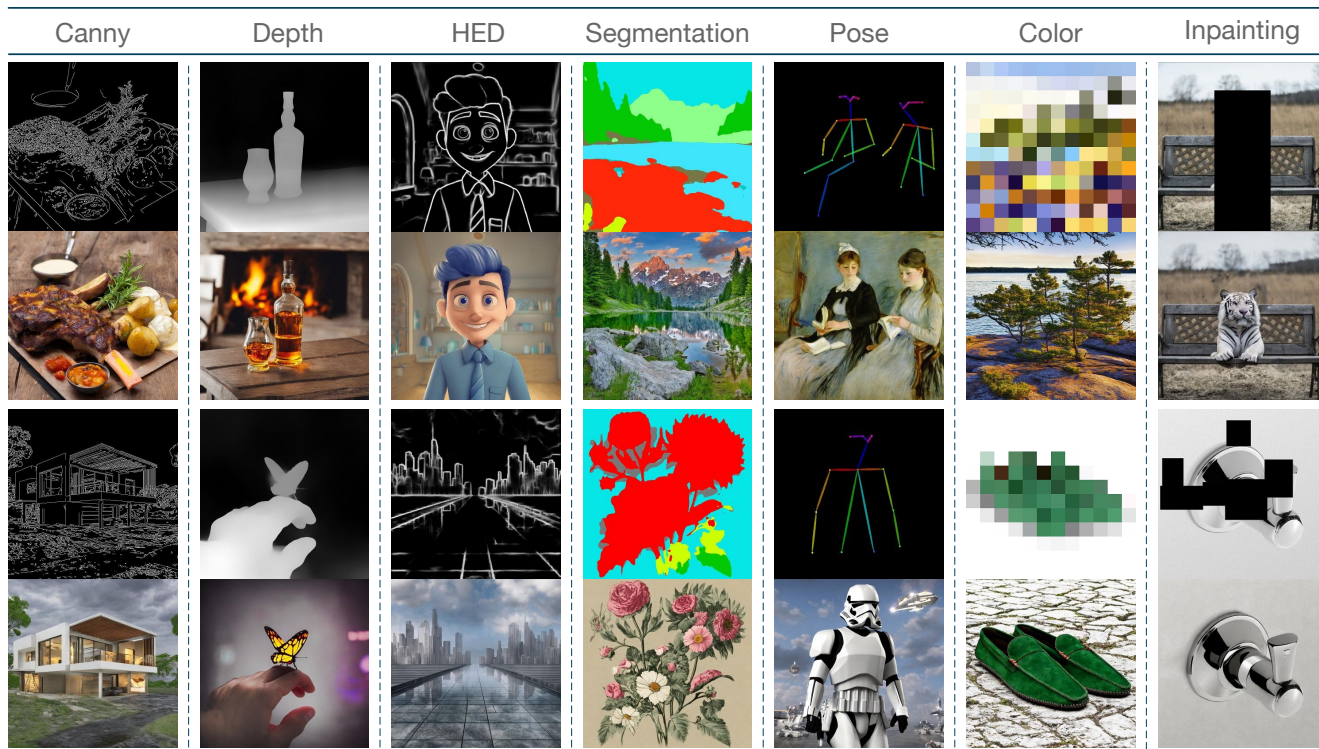
Figure 7. **Controllable image synthesis with various conditions.** The odd-numbered rows represent the input conditions, while the even-numbered rows correspond to the generated results. SCEdit is capable of generating high-quality images precise to input conditions.
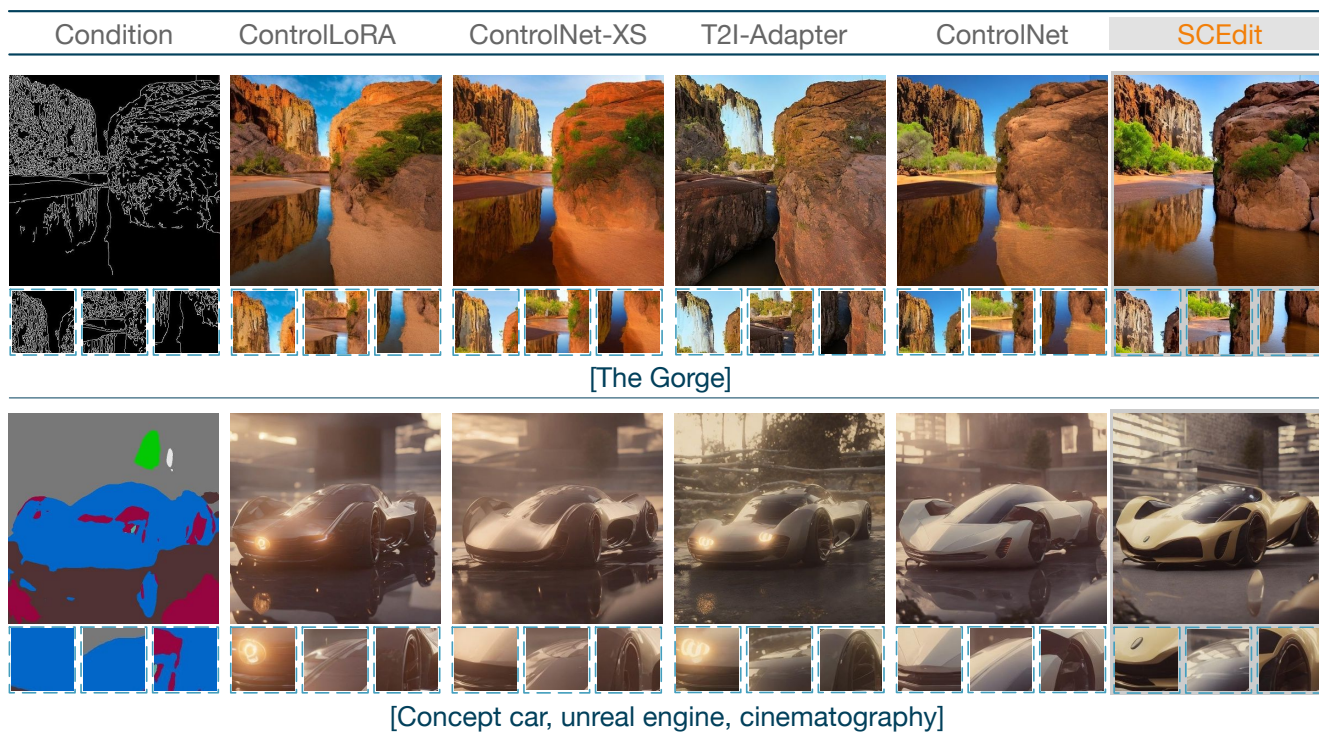

Figure 8. **Qualitative comparison** to the state-of-the-art controllable image synthesis methods based on the canny edge and semantic segmentation conditions. The areas in the boxes underneath are enlarged for detailed comparisons.

(a) Using the same canny edge map and textual prompt in combination with different color maps results in the depiction of a mountain across various seasons.
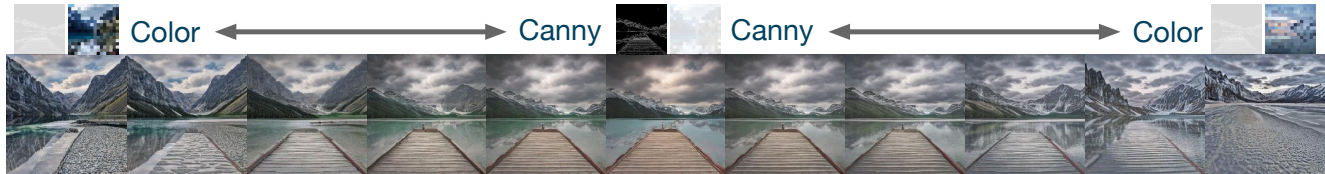
(b) Interpolations within the canny edge map and color maps.

Figure 9. **Composable generation.** Combinations of multi-conditions provide more compositional possibilities.

proposed SCEdit using CSC-Tuner on various conditional generation tasks, *e.g.*, canny edge, depth, hed boundary, semantic segmentation, pose keypoint, color, masked image, and so on. As illustrated in Fig. 7, our method shows superior performance and precise control ability under diverse scenarios and styles control, including real scenes, imagined scenarios, and artistic styles.

We also compare our approach with the state-of-the-art methods that generate images from different conditions. From a qualitative perspective, we present the generated results based on the canny edge map and semantic segmentation conditions with different methods and zoom-in on the detailed parts of the image in Fig. 8. It can be observed that our method achieves better quality in terms of fidelity and realism, such as the preservation of the reflection on the lake surface in the first row and the texture information of the car in the second row. From a quantitative perspective, as seen in Tab. 2, our method utilizes only 7.9% of the number of parameters in ControlNet, resulting in a 30% reduction in memory consumption, and accelerates the training process while also yielding a lower FID score.

Composable generation. In addition to generating control based on a single condition, we also support the input of multi-conditions simultaneously. In Fig. 9, we demonstrate the joint application of separately trained canny edge and color map models and present the results on unpaired condition data for training-free scene-level image translation. By anchoring the controlled subject to a canny edge map and text prompt, and incorporating various color maps, we produce a distinctive seasonal transition effect, as illustrated in Fig. 9a. Additionally, by traversing the embedding space representations of CSC-Tuner between the two conditions, we can blend them to achieve variations. In Fig. 9b, SCEdit
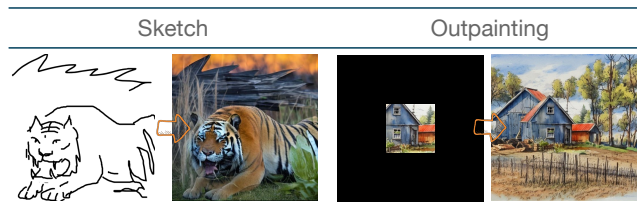

Figure 10. **Control generalization.** Learned edge condition model is capable of sketch-to-image task and inpainting condition model can control outpainting generation.

further provides precise control capabilities, enabling different interpolation effects through balancing among multiple elements.

Controllable generalization. We find that additional controllable generation tasks could be attempted on models that had already been trained under specific conditions. In Fig. 10, we achieve impressive results using the edge-based model for the sketch-to-image task and the inpainting-based model for the outpainting task, although there is no adaptation based on non-real conditions such as hand-drawn sketches for the former, and no specialized adjustment of mask generation patterns for outpainting for the latter.

5. Conclusion

We propose SCEdit as an efficient and controllable method for image diffusion generation. We introduce SC-Tuner to edit skip connections and extends it to CSC-Tuner, enabling a diverse range of conditional inputs. Our method achieves high efficiency by exclusively performing backward propagation in the decoupled U-Net decoder. As a lightweight and plug-and-play module, SCEdit supports arbitrary single- and multi-condition generation and demonstrates remarkable superiority in terms of performance.

References

[1] Runway AI. Stable Diffusion v1.5 Model Card, https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022. 2, 5

[2] Stability AI. Stable Diffusion v2-1 Model Card, https://huggingface.co/stabilityai/stable-diffusion-2-1, 2022. 5

[3] Stability AI. Stable Diffusion XL Model Card, https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0, 2022. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning To Follow Image Editing Instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18392–18402, 2023. 2

[5] John Canny. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 679–698, 1986. 5

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186, 2021. 5

[7] Huiwen Chang, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3

[9] Alibaba Cloud. Tongyi Wanxiang, https://tongyi.aliyun.com/wanxiang/, 2023. 2

[10] David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *Int. Conf. Learn. Represent.*, 2016. 3

[11] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a Unified View of Parameter-Efficient Transfer Learning. In *Int. Conf. Learn. Represent.*, 2021. 3

[12] Wu Hecong. ControlLoRA: A Lightweight Neural Network To Control Stable Diffusion Spatial Information, https://github.com/HighCWu/ControlLoRA, 2023. 5, 6

[13] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2023. 4, 5

[14] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Adv. Neural Inform. Process. Syst.* Curran Associates, Inc., 2020. 2, 3

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Int. Conf. Mach. Learn.*, pages 2790–2799, 2019. 3

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Int. Conf. Learn. Represent.*, 2022. 2, 3, 5

[18] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and Controllable Image Synthesis with Composable Conditions. In *Int. Conf. Mach. Learn.*, 2023. 2, 3

[19] Zhongzhan Huang, Pan Zhou, Shuicheng Yan, and Liang Lin. ScaleLong: Towards More Stable Training of Diffusion Model via Scaling Network Long Skip Connection. In *Adv. Neural Inform. Process. Syst.*, 2023. 2, 3

[20] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Yiliang Lv, Deli Zhao, and Jingren Zhou. Rethinking Efficient Tuning Methods from a Unified Perspective. *arXiv preprint arXiv:2303.00690*, 2023. 2, 3

[21] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Ao Ma, Yiliang Lv, Yujun Shen, Deli Zhao, and Jingren Zhou. Res-Tuning: A Flexible and Efficient Tuning Paradigm via Unbinding Tuner from Backbone. In *Adv. Neural Inform. Process. Syst.*, 2023. 2, 3, 4

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Conf. Empirical Methods NLP*, 2021. 3

[23] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning. In *Int. Conf. Learn. Represent.*, 2021. 5

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 5

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Int. Conf. Learn. Represent.*, 2018. 5

[26] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *Int. Conf. Learn. Represent.*, 2021. 2

[27] Midjourney. Midjourney, https://www.midjourney.com/, 2023. 2

[28] ModelScope. Customized Style Dataset Card, https://modelscope.cn/datasets/damo/style_custom_dataset/summary, 2023. 5

[29] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 5, 6

[30] 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*, 2022. 2

[31] OpenAI. DALL·E 2, https://openai.com/dall-e-2, 2022. 2

[32] OpenAI. DALL·E 3, https://openai.com/dall-e-3, 2023. 2

[33] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *Adv. Neural Inform. Process. Syst.*, 2023. 2

[34] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1623–1637, 2022. 5

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 3

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Med. Image Comput. Computer-Assisted Interv.*, 2015. 2, 3

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22500–22510, 2023. 2

[38] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *arXiv preprint arXiv:2104.07636*, 2021. 2, 3

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R. Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Adv. Neural Inform. Process. Syst.*, 2022. 5

[40] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch, https://github.com/mseitzer/pytorch-fid, 2020. 5

[41] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise Image Editing via Recognition and Generation Tasks. *arXiv preprint arXiv:2311.10089*, 2023. 2

[42] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free Lunch in Diffusion U-Net. *arXiv preprint arXiv:2309.11497*, 2023. 2, 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *Int. Conf. Learn. Represent.*, 2021. 3

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *Int. Conf. Learn. Represent.*, 2021. 2, 3

[45] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-Robust Large Mask Inpainting With Fourier Convolutions. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 2149–2159, 2022. 5

[46] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. DiffFit: Unlocking Transferability of Large Diffusion Models via Simple Parameter-efficient Fine-Tuning. In *Int. Conf. Comput. Vis.*, pages 4230–4239, 2023. 2

[47] Saining Xie and Zhuowen Tu. Holistically-Nested Edge Detection. In *Int. Conf. Comput. Vis.*, pages 1395–1403, 2015. 5

[48] Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. CycleNet: Rethinking Cycle Consistency in Text-Guided Diffusion for Image Manipulation. In *Adv. Neural Inform. Process. Syst.*, 2023. 2

[49] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. ControlNet-XS: Designing an Efficient and Effective Architecture for Controlling Text-to-Image Diffusion Models. *arXiv preprint arXiv:2312.06573*, 2023. 5, 6

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023. 2, 3, 5, 6