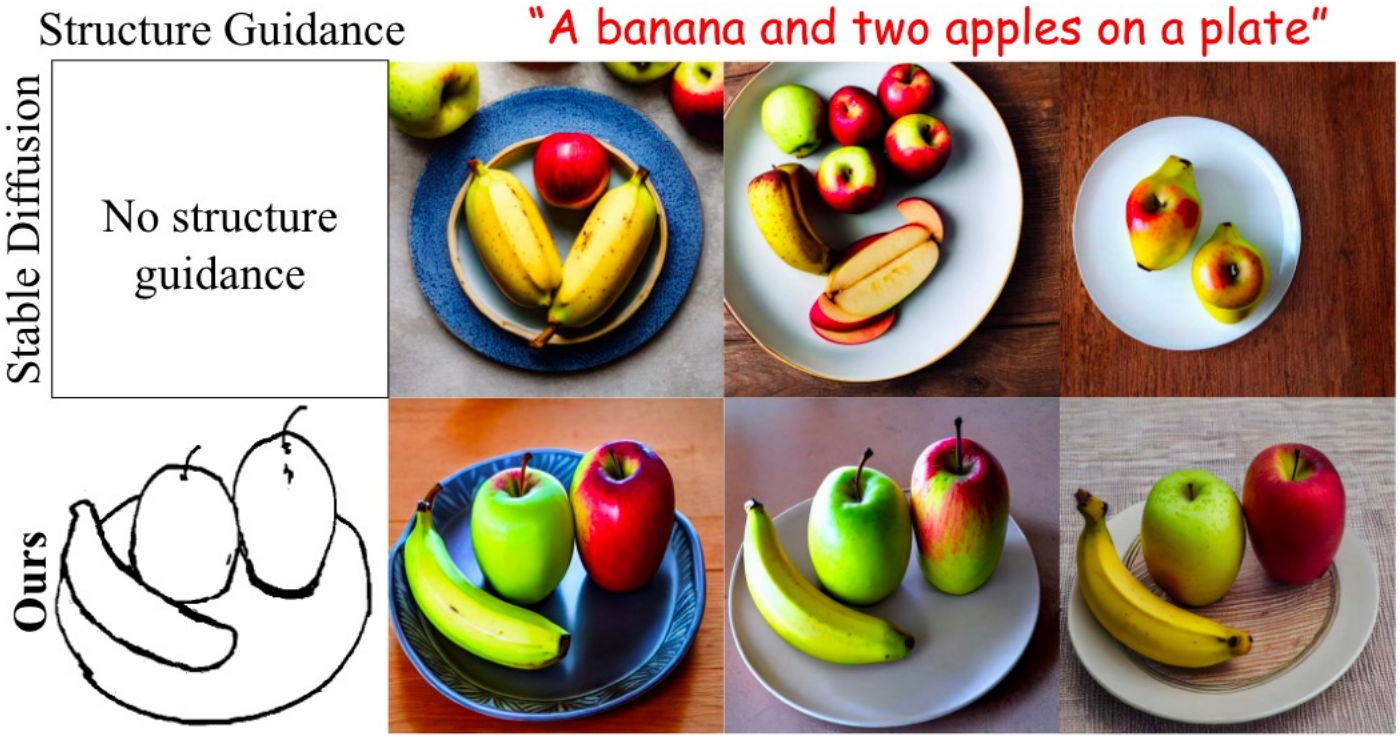
Controllable Text-to-Image Diffusion Models with Additional Conditions

Qi Tianhao
2023/06/27

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

Task Introduction

- 扩散模型能够生成多样化的高质量图像，但是不能对生成图像提供精细化的控制



Methods

- Sketch-Guided Text-to-Image Diffusion Models
- Adding Conditional Control to Text-to-Image Diffusion Models
- T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models

Sketch-Guided Text-to-Image Diffusion Models

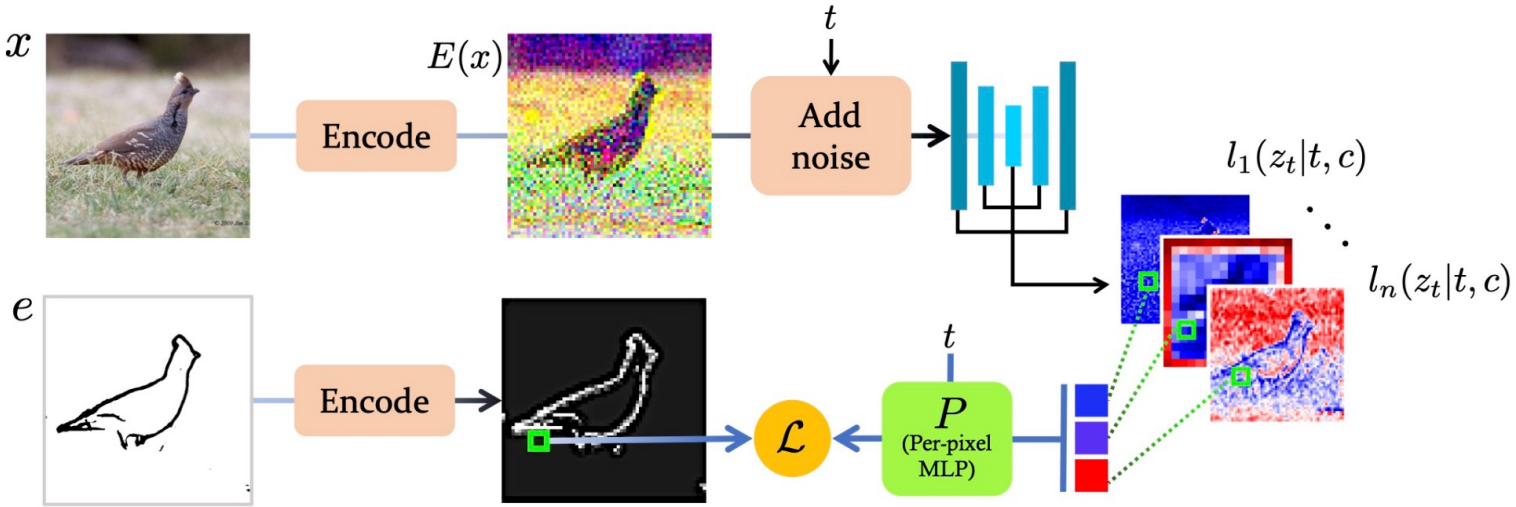
Motivation

结合用户提供的sketch作为辅助信息，训练一个per-pixel MLP预测edge map，并通过梯度指导生成过程朝着接近给定sketch的方向进行

Sketch-Guided Text-to-Image Diffusion Models

- Training

1. 抽取去噪过程中Unet的多个中间层特征，resize到输入尺寸后通道级联
2. 级联特征图通过per-pixel MLP预测edge map，与ground-truth计算损失



$$z_t = \alpha_t \cdot E(x) + \mu_t \cdot \xi \qquad \mathbf{F}(w|c,t) = [l_1(w|c,t), \ldots, l_n(w|c,t)]$$

$$\mathcal{L} = \mathop{\mathbb{E}}_{(x,e,c)\sim\mathcal{D}} \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}([0,1]) \\ \xi\sim\mathcal{N}(0,1)}} \sum_{i,j} \|P(\mathbf{F}(z_t|c,t)_{i,j}, t) - E(e)_{ij}\|^2$$

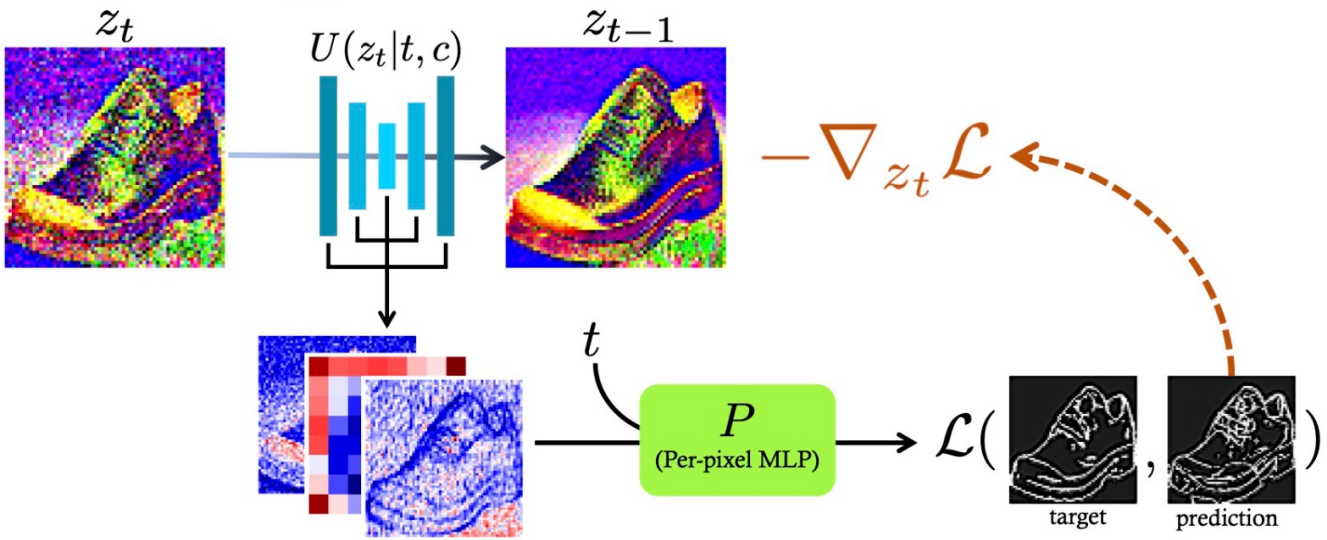
Sketch-Guided Text-to-Image Diffusion Models

Inference

1.级联特征图通过per-pixel MLP预测edge map，与ground-truth计算损失

2.借鉴classifier guidance的思想，用损失梯度指导生成过程

3.由于去噪过程的末端不会影响生成图像的几何布局，故仅在t=T,…,S>1使用edge guidance，其中S=0.5T



$$\mathcal{L}(\tilde{E}, E(e)) = \|\tilde{E} - E(e)\|^2$$

$$\tilde{z}_{t-1} = z_{t-1} - \alpha \cdot \nabla_{z_t}\mathcal{L}$$

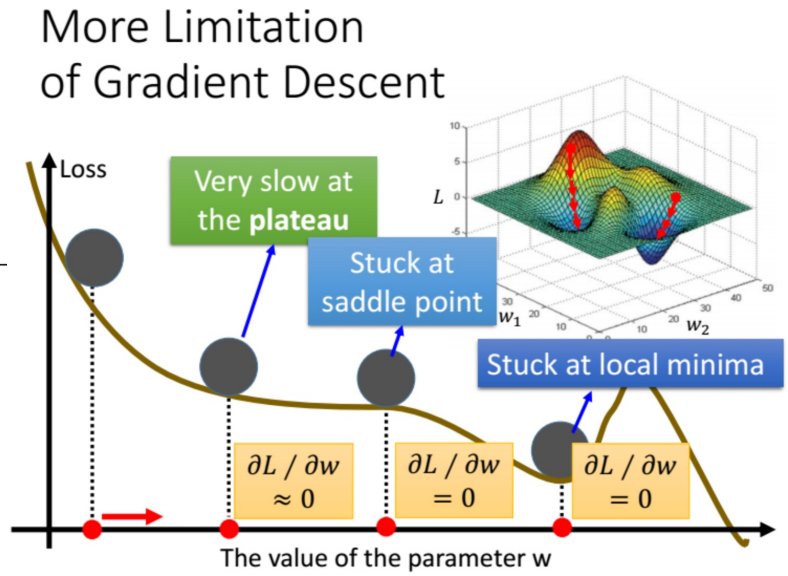
$$\alpha = \frac{\|z_t - z_{t-1}\|_2}{\|\nabla_{z_t}\mathcal{L}\|_2} \cdot \beta$$

Inference

2. 借鉴classifier guidance的思想，用损失梯度指导生成过程

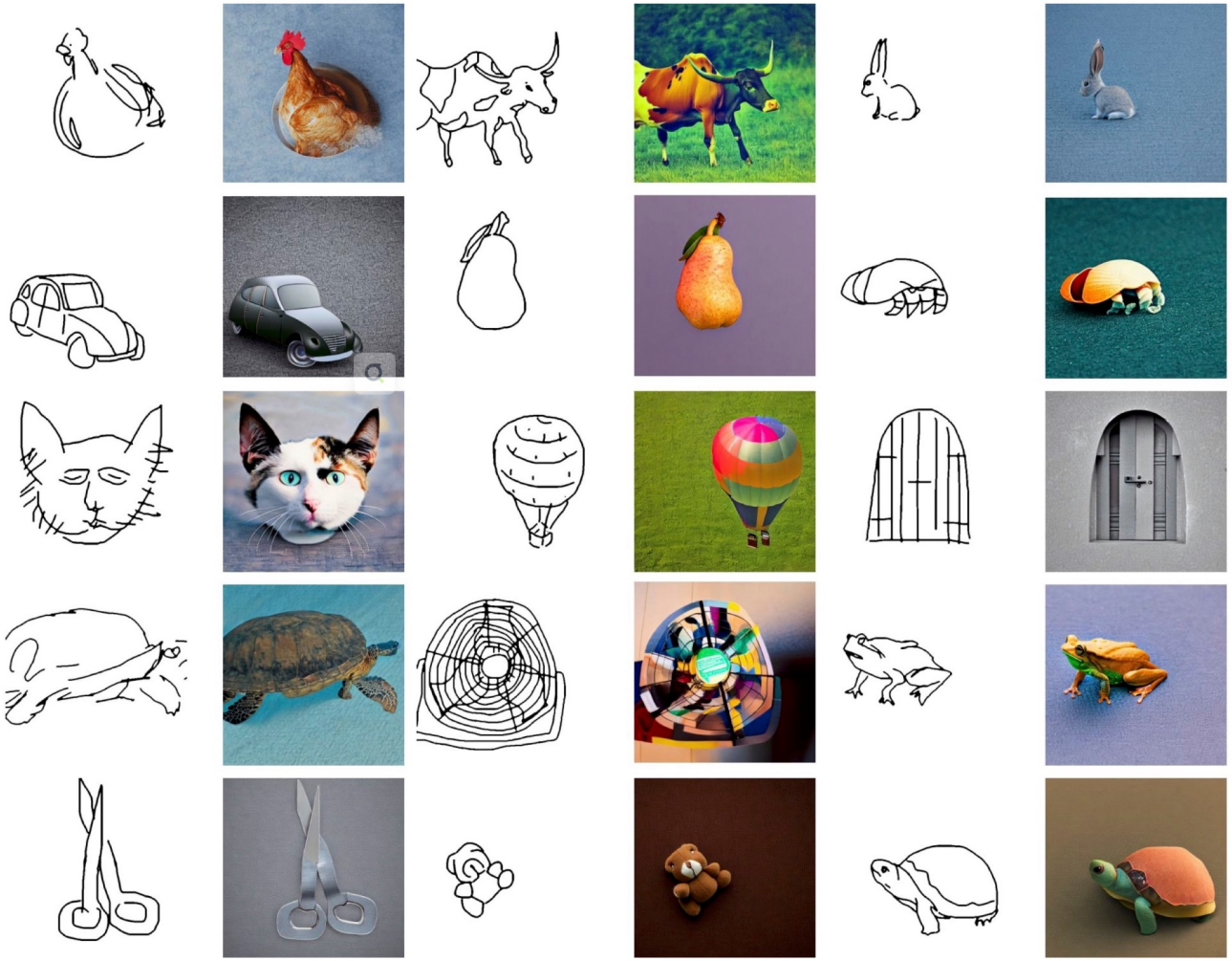
Algorithm 2 Classifier guided DDIM sampling, given a diffusion model $\epsilon_\theta(x_t)$, classifier $p_\phi(y|x_t)$, and gradient scale s.

Input: class label y, gradient scale s
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t} \log p_\phi(y|x_t)$
$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}$
end for
return x_0



Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in Neural Information Processing Systems* 34 (2021): 8780-8794.

Sketch-Guided Text-to-Image Diffusion Models



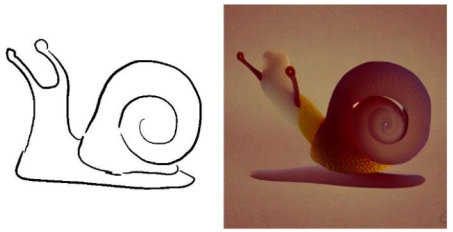
Limitation:

1.对用户提供sketch的笔触风格敏感

2.初始化影响生成结果，难以处理复杂且混乱的sketch

3.梯度引导不考虑text prompt且不存在任何交互，不能很好地贴合语义

failed initialization



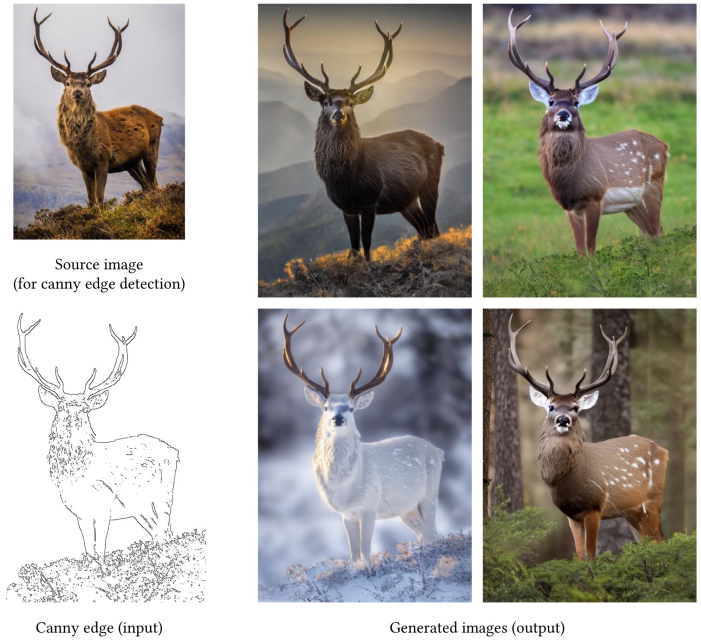
failed scene comprehension

ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang and Maneesh Agrawala
Stanford University

- 2021 苏州大学本科毕业
- 2021-2022 香港中文大学科研助理
- 2022至今 Standford University
- Style2Paints 17.1k star
- ControlNet 27.3k star



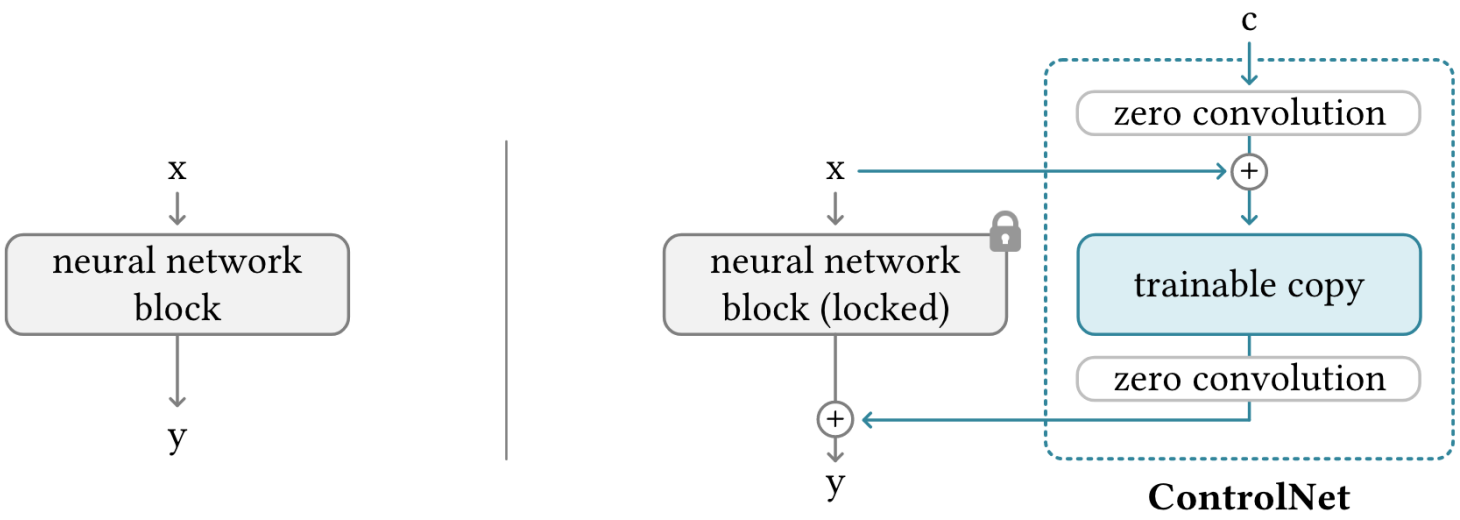
ControlNet

Motivation

1.冻结原文生图模型参数，以保留在数十亿数据上训练得到的文生图能力

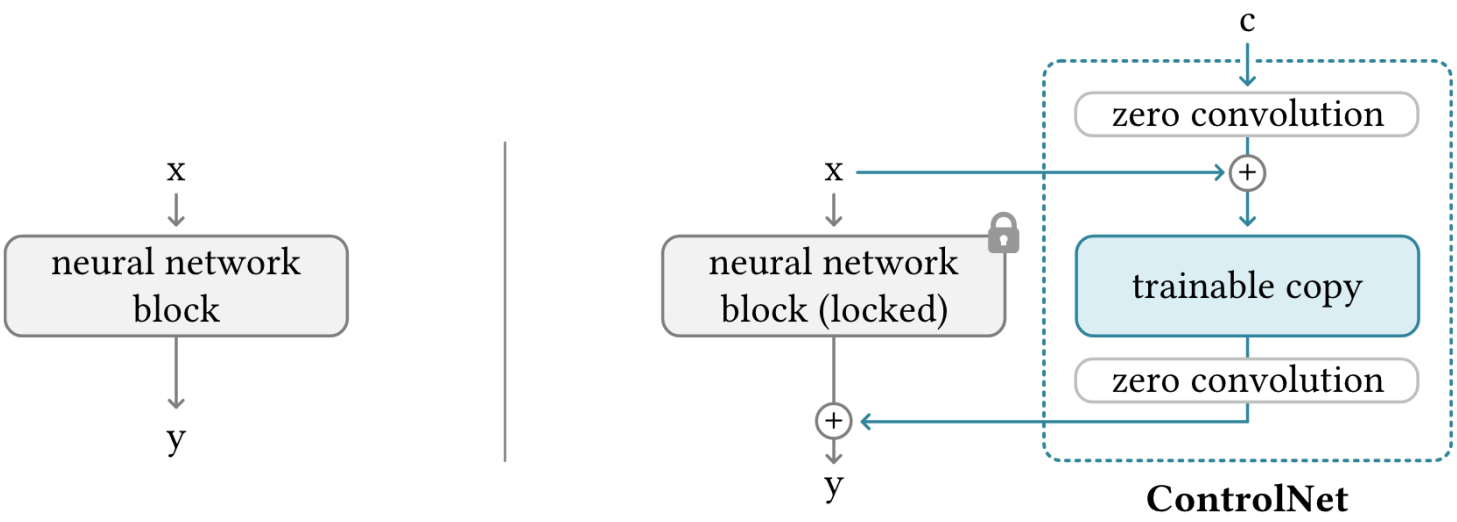
2.引入额外分支，以实现额外条件的控制作用

3.zero convolution保证模型在优化前与原模型的容量、功能和结果保持一致



(a) Before (b) After

ControlNet



(a) Before

(b) After

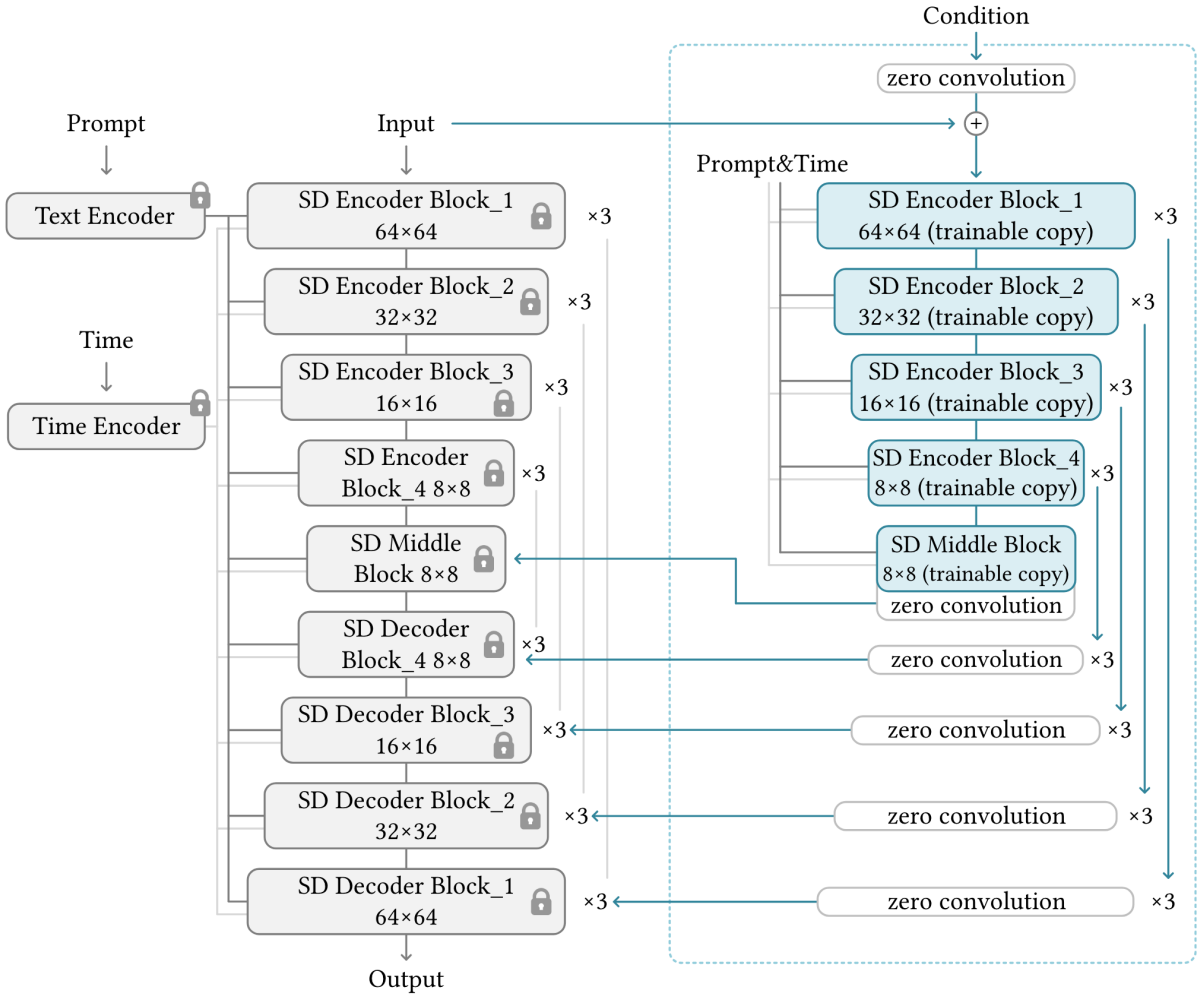
$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \Theta)$$

$$\boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x}; \Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c); \Theta_{z2})$$

$$\begin{cases} \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}) = \boldsymbol{0} \\ \mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c) = \mathcal{F}(\boldsymbol{x}; \Theta_c) = \mathcal{F}(\boldsymbol{x}; \Theta) \\ \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c); \Theta_{z2}) = \mathcal{Z}(\mathcal{F}(\boldsymbol{x}; \Theta_c); \Theta_{z2}) = \boldsymbol{0} \end{cases}$$

$$\boldsymbol{y}_c = \boldsymbol{y}$$

ControlNet



(a) Stable Diffusion (b) ControlNet

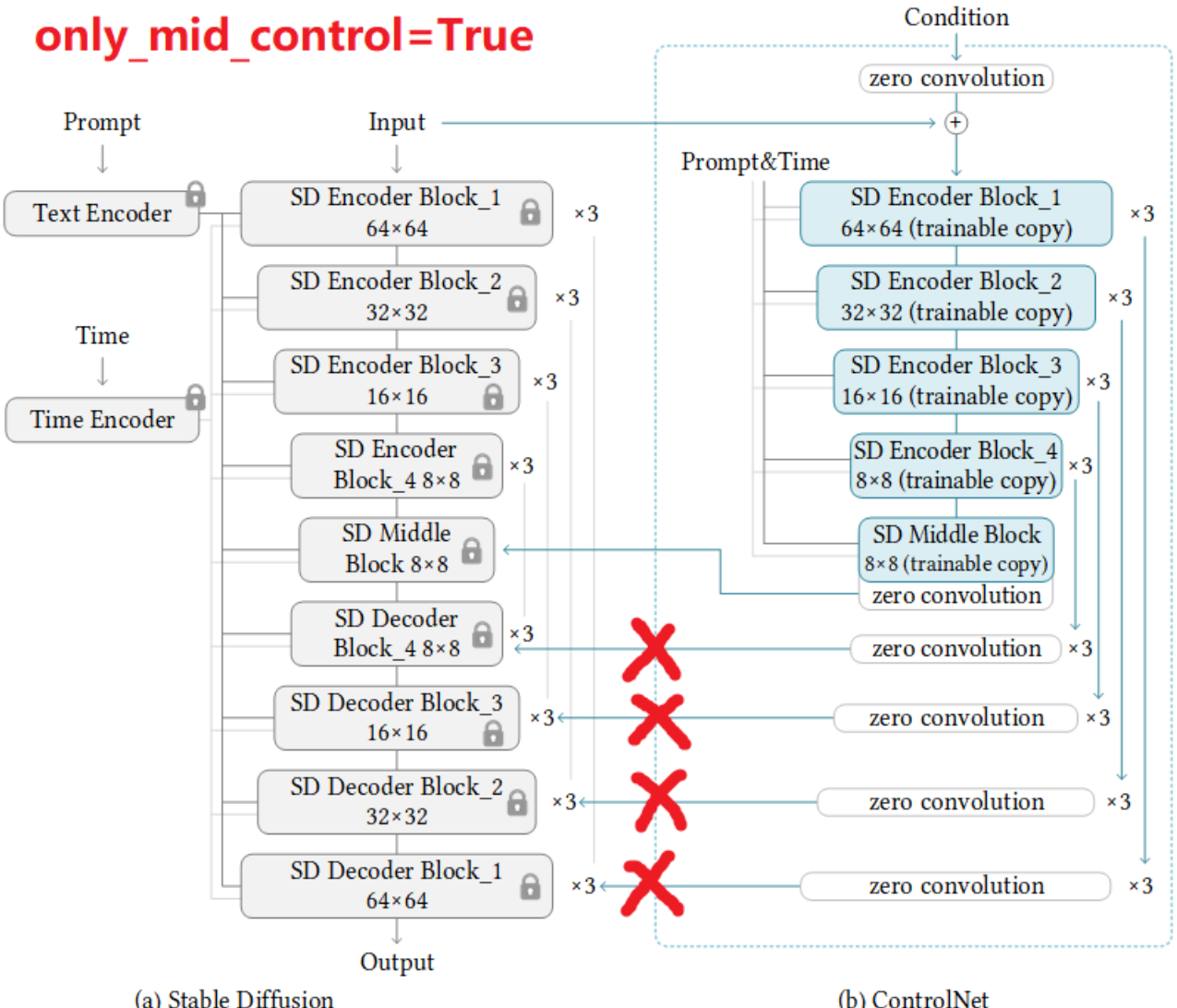
1.Condition通过一个tiny network将图像空间的条件映射到隐空间（8 conv layers in total, 3 conv layers with stride 2）

2.随机mask 50%的text prompt，使得controlnet从额外条件中学到更丰富的语义信息

3.Training objective

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, t, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(z_t, t, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f})) \|_2^2 \right]$$

ControlNet with Limited Computing Resources



1.速度1.6x
2.收敛后再设成False finetune一些steps

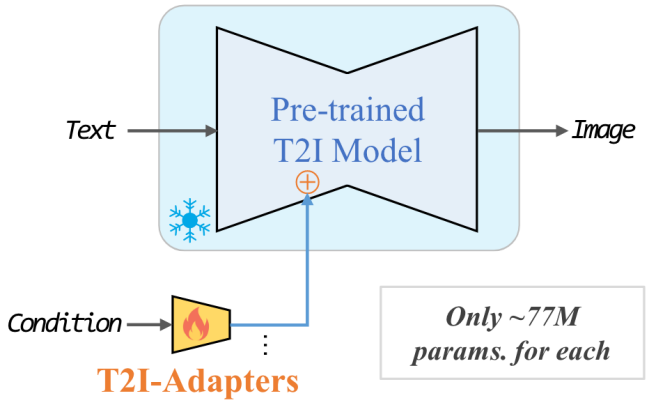
ControlNet

- More Examples

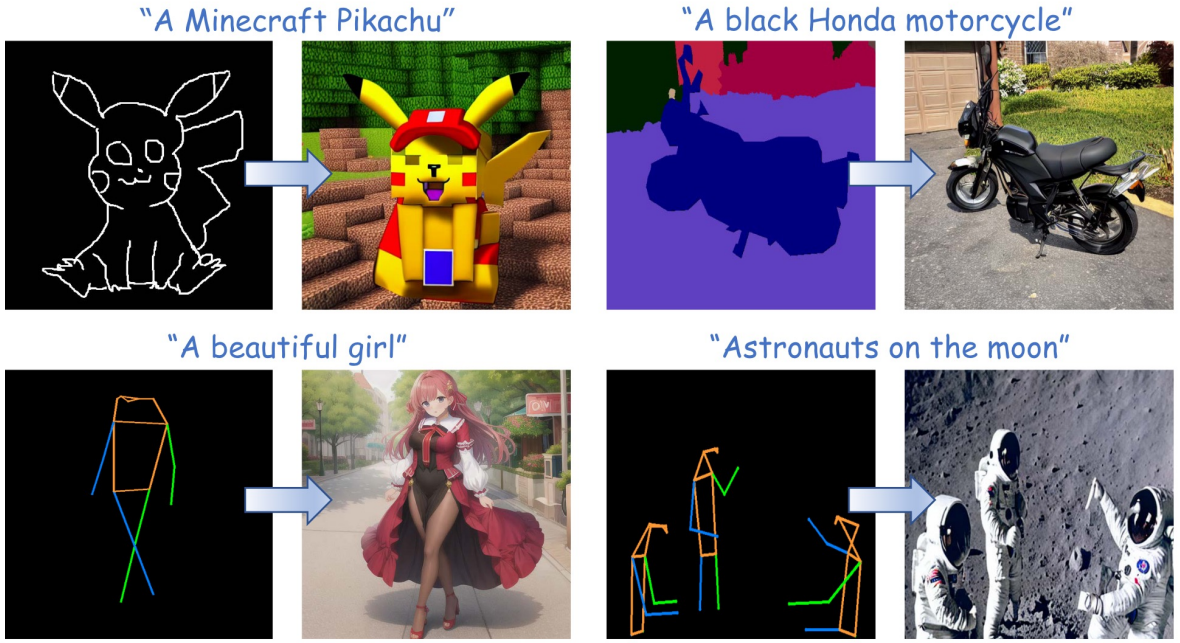
T2I-Adapter

T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models

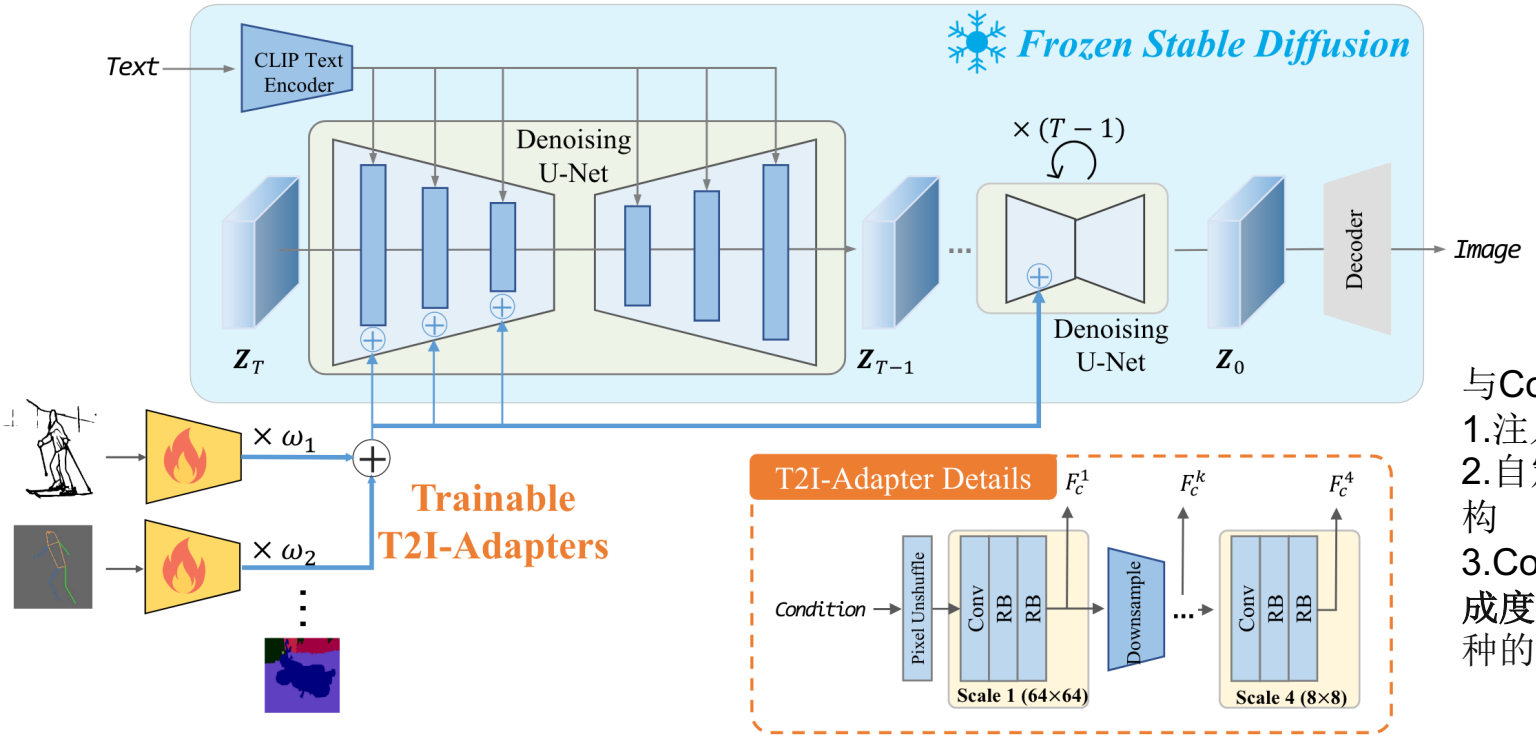
Chong Mou[*1,2] Xintao Wang[†2] Liangbin Xie[*2,3,4] Jian Zhang[†1]
Zhongang Qi[2] Ying Shan[2] Xiaohu Qie[2]
[1]Peking University Shenzhen Graduate School [2]ARC Lab, Tencent PCG [3]University of Macau [4]Shenzhen Institute of Advanced Technology



- ✓ ***Plug-and-play.*** *Not affect original network topology and generation ability*
- ✓ ***Simple and small.*** *~77M parameters and ~300M storage*
- ✓ ***Flexible.*** *Various adapters for different control conditions*
- ✓ ***Composable.*** *Several adapters to achieve multi-condition control*
- ✓ ***Generalizable.*** *Can be directly used on customed models*



T2I-Adapter



与ControlNet的不同：
1.注入Unet encoder
2.自定义的Adapter架构
3.ControlNet可用性完成度更高，支持更多种的condition（9 大类）

One condition: $\mathbf{F}_c = \mathcal{F}_{AD}(\mathbf{C})$

Multiple conditions: $\mathbf{F}_c = \sum_{k \in \mathcal{G}} \omega_k \mathcal{F}_{AD}^k(\mathbf{C}^k)$

$$\hat{\mathbf{F}}_{ldm}^i = \mathbf{F}_{ldm}^i + \hat{\mathbf{F}}_c^i,\ i \in \{1, 2, 3, 4\}$$

https://pytorch.org/docs/stable/generated/torch.nn.PixelUnshuffle.html

T2I-Adapter

- More Examples

Conclusion

- 效果为先，需要有令人眼前一亮的效果或demo展示
- Adapter架构仍是当前研究的热点方向，且其算力开销可以接受
- 直接注入参考图是待研究方向