ControlNeXt: Powerful and Efficient Control for Image and Video Generation

作者

一作

Bohao Peng

Email: bhpeng22@cse.cuhk.edu.hk

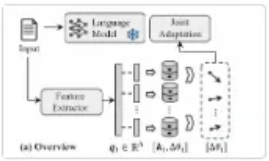
Hong Kong

the Chinese University of Hong Kong

Email

Google Scholar

Publications

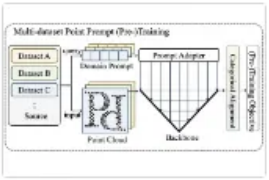


Scalable language model with generalized continual learning

Bohao Peng, Zhuotao Tian, Shu Liu, Mingchang Yang, Jiaya Jia

*nternational Conference on Learning Representations (**ICLR**) 2024*

[Paper] [Code] [Bib]

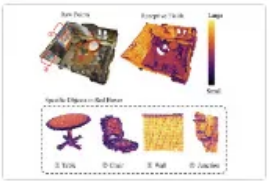


Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training

Xiaoyang Wu, Zhuotao Tian, Xin Wen, **Bohao Peng**, Xihui Liu, Kaicheng Yu, Hengshuang Zhao

*Computer Vision and Pattern Recognition (**CVPR**) 2024*

[Paper] [Code] [Bib]



OA-CNNs: Omni-Adaptive Sparse CNNs for 3D Semantic Segmentation

Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, Jiaya Jia

*Computer Vision and Pattern Recognition (**CVPR**) 2024*

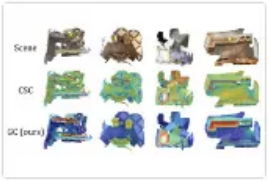
[Paper] [Code] [Bib]

Prompt Highlighter: Interactive Control for Multi-Modal LLMs

Yuechen Zhang, Shengju Qian, **Bohao Peng**, Shu Liu, Jiaya Jia

*Computer Vision and Pattern Recognition (**CVPR**) 2024*

[Paper] [Code] [Bib]



GroupContrast: Semantic-aware Self-supervised Representation Learning for 3D Understanding

Chengyao Wang, Li Jiang, Xiaoyang Wu, Zhuotao Tian, **Bohao Peng**, Hengshuang Zhao, Jiaya Jia

*Computer Vision and Pattern Recognition (**CVPR**) 2024*

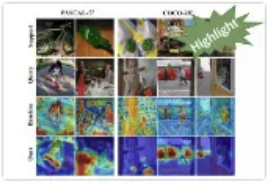
[Paper] [Code] [Bib]

LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, **Bohao Peng**, Shu Liu, Jiaya Jia

Arxiv preprint, 2024

[Paper] [Bib]



Hierarchical dense correlation distillation for few-shot segmentation

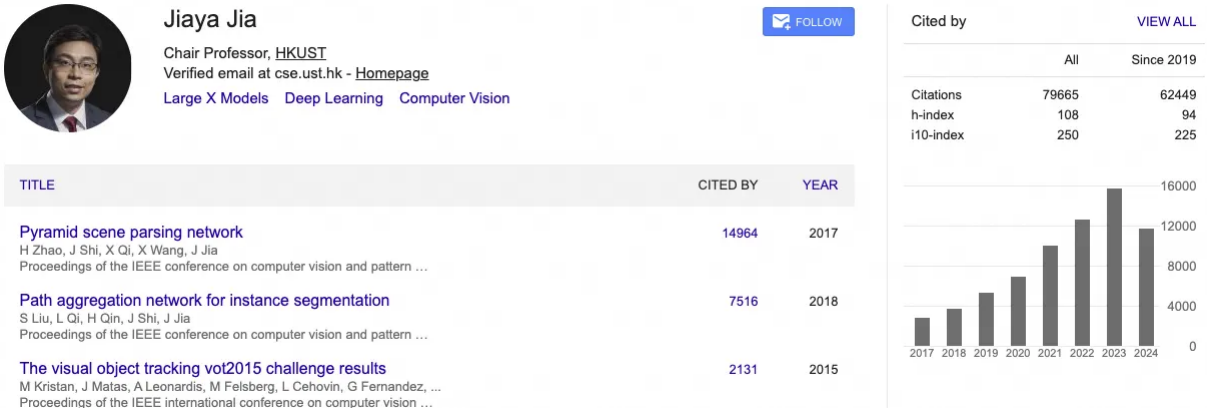
Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, Jiaya Jia

*Computer Vision and Pattern Recognition (**CVPR**) 2023*

***Highlight** (2.4% acceptance rate)*

[Paper] [Code] [Bib]

导师



Jiaya Jia

FOLLOW

Chair Professor, HKUST
Verified email at cse.ust.hk - Homepage
Large X Models Deep Learning Computer Vision

TITLE	CITED BY	YEAR
Pyramid scene parsing network H Zhao, J Shi, X Qi, X Wang, J Jia Proceedings of the IEEE conference on computer vision and pattern …	14964	2017
Path aggregation network for instance segmentation S Liu, L Qi, H Qin, J Shi, J Jia Proceedings of the IEEE conference on computer vision and pattern …	7516	2018
The visual object tracking vot2015 challenge results M Kristan, J Matas, A Leonardis, M Felsberg, L Cehovin, G Fernandez, ... Proceedings of the IEEE international conference on computer vision …	2131	2015

Cited by VIEW ALL

	All	Since 2019
Citations	79665	62449
h-index	108	94
i10-index	250	225

研究背景

可控文生图，业界主流是Adapter型，需要训练，即插即用，主要分为两类：1.Attention注入 2.Add注入

1.Attention注入

IP-Apdater

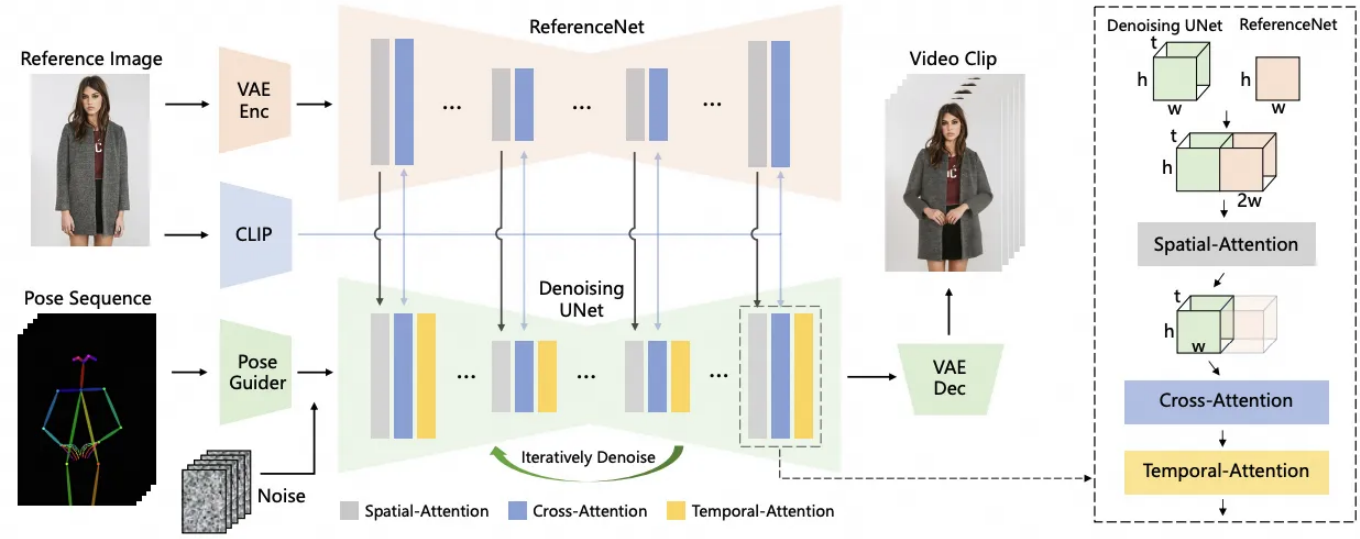
feature embeddings 通过Cross-Attention

https://github.com/tencent-ailab/IP-Adapter

ReferenceNet (Animate Anyone)

latent feature通过multi-stage的Self-Attention注入

https://arxiv.org/pdf/2311.17117

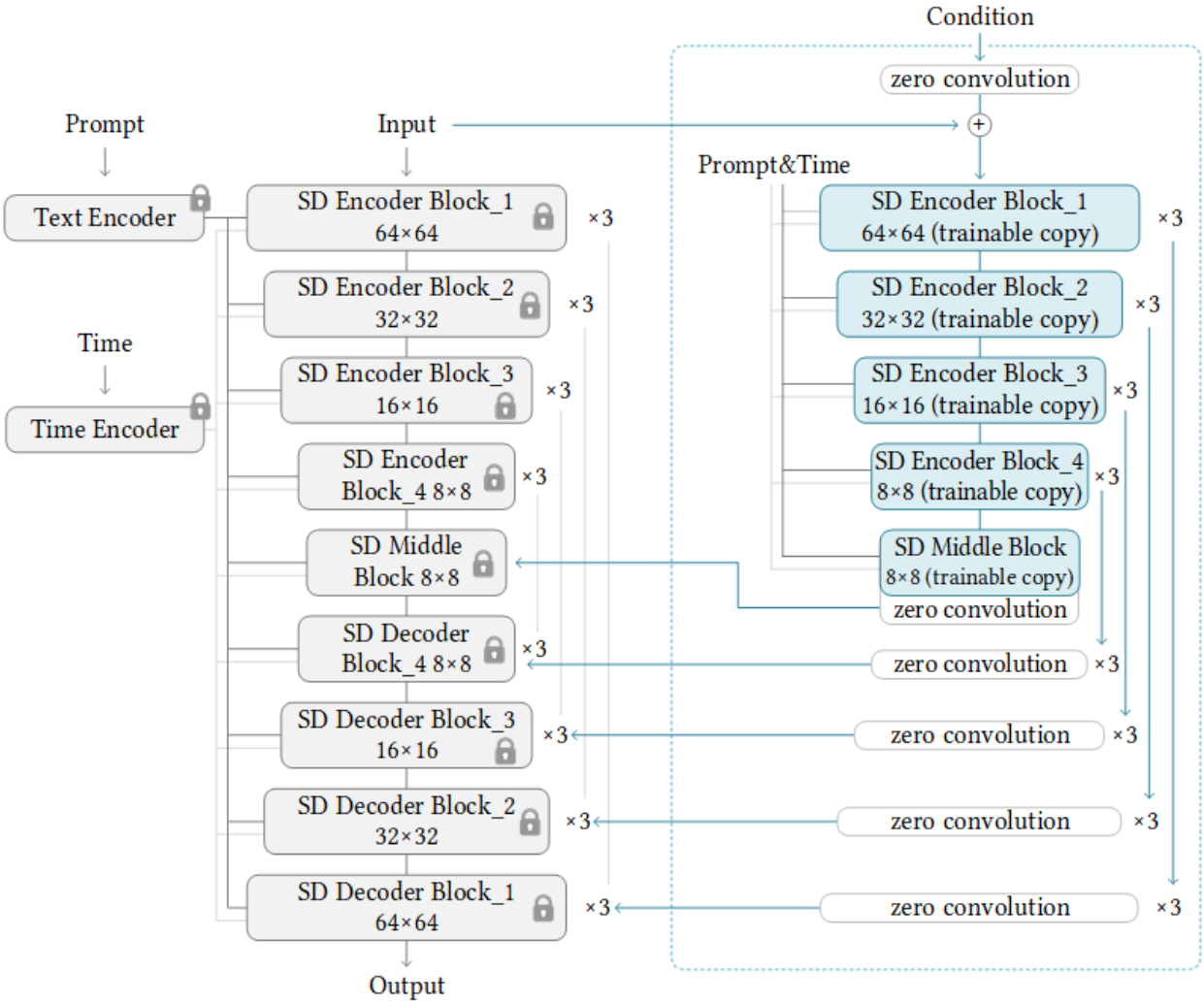


2.Add注入

ControlNet

复制一份Unet的一半参数作为编码器，输出结果经零卷积+multi-stage Add到Unet的上采样层

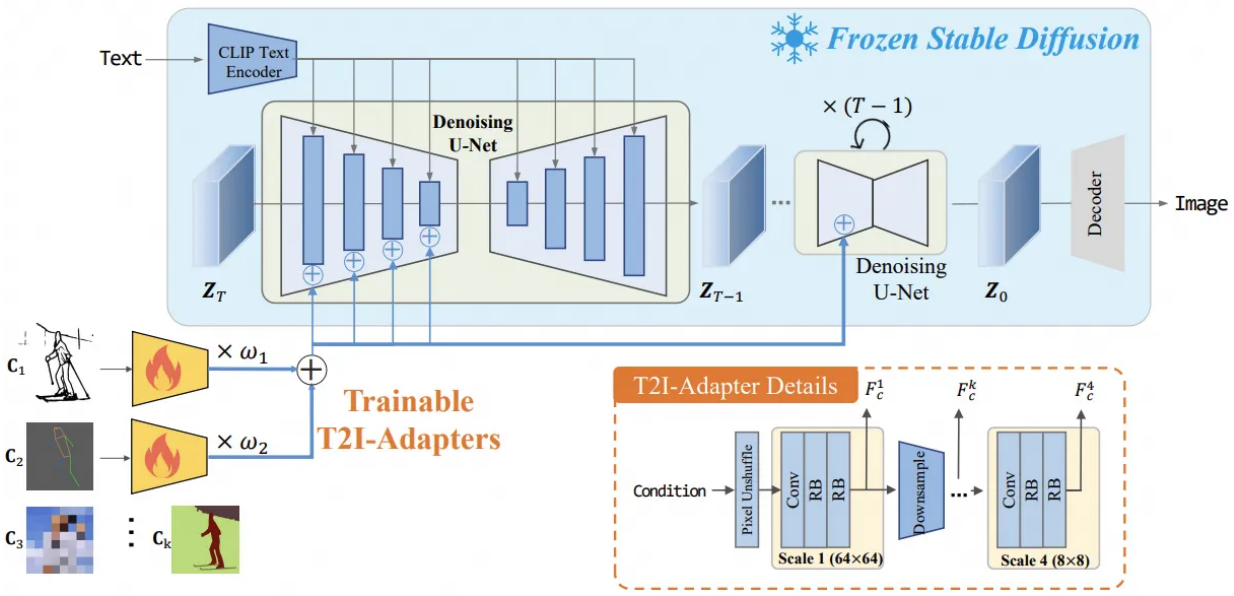
https://github.com/lllyasviel/ControlNet



T2I-Adapter

更轻量化的卷积层，multi-stage Add到Unet的下采样层

https://arxiv.org/pdf/2302.08453



研究动机

1.Zero Cov收敛慢。

Figure 3. ControlNeXt achieves significantly faster training convergence and data fitting. It can learn to fit the conditional controls with fewer training steps, which also significantly alleviates the *sudden convergence* problem.

2.ControlNet参数量大，推理慢。

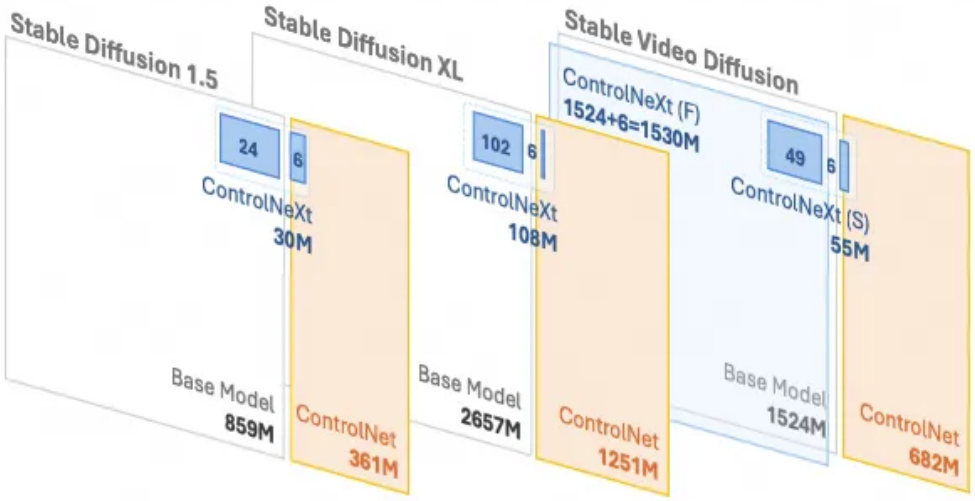


Figure 4. Parameter efficiency of ControlNeXt. We present the number of learnable parameters with various base models.

论文方法

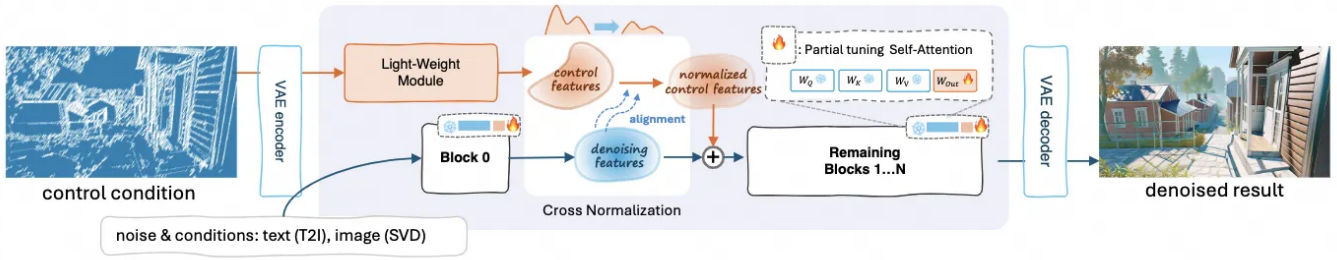


Figure 2. Training pipeline of ControlNeXt. We explore a more remarkable parameter-efficient framework than directly adopting a trainable copy.

1.轻量的控制网络：几层卷积块，其控制特征只Add到UNet的Middle Block输出上（Single-Stage Add）

2.Cross Normalization：用主干网络feature的mean和var去norm 控制网络的输出，再Add

We represent the feature maps processed from the main denoising branch and the control transferring branch as $\boldsymbol{x}_d$ and $\boldsymbol{x}_c$, respectively, where $\boldsymbol{x}_m, \boldsymbol{x}_c \in \mathbb{R}^{h \times w \times c}$. The key of Cross Normalization is to use the mean and variance calculated from the main branch $\boldsymbol{x}_m$ to normalize the control features $\boldsymbol{x}_c$, ensuring their alignment. First, calculate the mean and variance of the denoising features,

$$\boldsymbol{\mu}_m = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{m,i} \ , \tag{8}$$

$$\boldsymbol{\sigma}_m^2 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_{m,i} - \boldsymbol{\mu}_m)^2 \ . \tag{9}$$

Then, we normalize the control features using the mean and variance of the denoising features,

$$\hat{\boldsymbol{x}}_c = \frac{\boldsymbol{x}_c - \boldsymbol{\mu}_m}{\sqrt{\boldsymbol{\sigma}_m^2 + \epsilon}} * \gamma, \tag{10}$$

where ϵ is a small constant added for numerical stability and γ is a parameter that allows the model to scale the normalized value.

代码中的实现不一致：

```
scale = mid_block_additional_residual['scale']
mid_block_additional_residual = mid_block_additional_residual['out']
mid_block_additional_residual=nn.functional.adaptive_avg_pool2d(mid_block_additional_residual, sample.shape[-2:])
mid_block_additional_residual = mid_block_additional_residual.to(sample)
mean_latents, std_latents = torch.mean(sample, dim=(1, 2, 3), keepdim=True), torch.std(sample, dim=(1, 2, 3), keepdim=True)
mean_control, std_control = torch.mean(mid_block_additional_residual, dim=(1, 2, 3), keepdim=True), torch.std(mid_block_additional_residual, dim=(1, 2, 3), keepdim=True)
mid_block_additional_residual = (mid_block_additional_residual - mean_control) * (std_latents / (std_control + 1e-12)) + mean_latents
sample = sample + mid_block_additional_residual * scale
```

3.Finetune极少部分Base Model参数：根据上图看是Attention层的 W_out

实验

1.参数量

Model	Method	Parameters (M) Total	Learnable
SD1.5	ControlNet	1,220	361
	ControlNeXt$_{(\mathbf{Our})}$	865	30
	Base model	859	-
SDXL	ControlNet	3,818	1,251
	ControlNeXt$_{(\mathbf{Our})}$	2,573	108
	Base model	2,567	-
SVD	ControlNet	2,206	682
	ControlNeXt-S$_{(\mathbf{Our})}$	1,530	55
	ControlNeXt-F$_{(\mathbf{Our})}$	1,530	1,530
	Base model	1,524	-

2.推理速度

Method	Inference Time (s) SD1.5	SDXL	SVD	Δ
ControlNet	0.31	1.01	1.73	$_{+}$ 41.9%
ControlNeXt$_{(\mathbf{Our})}$	0.24	0.82	1.29	$_{+}$ 10.4%
Base model	0.22	0.70	1.23	-

Table 2. Comparison of the inference time with various backbones. Our method adds only minimal latency compared to the pretrained base model.

3.可视化效果

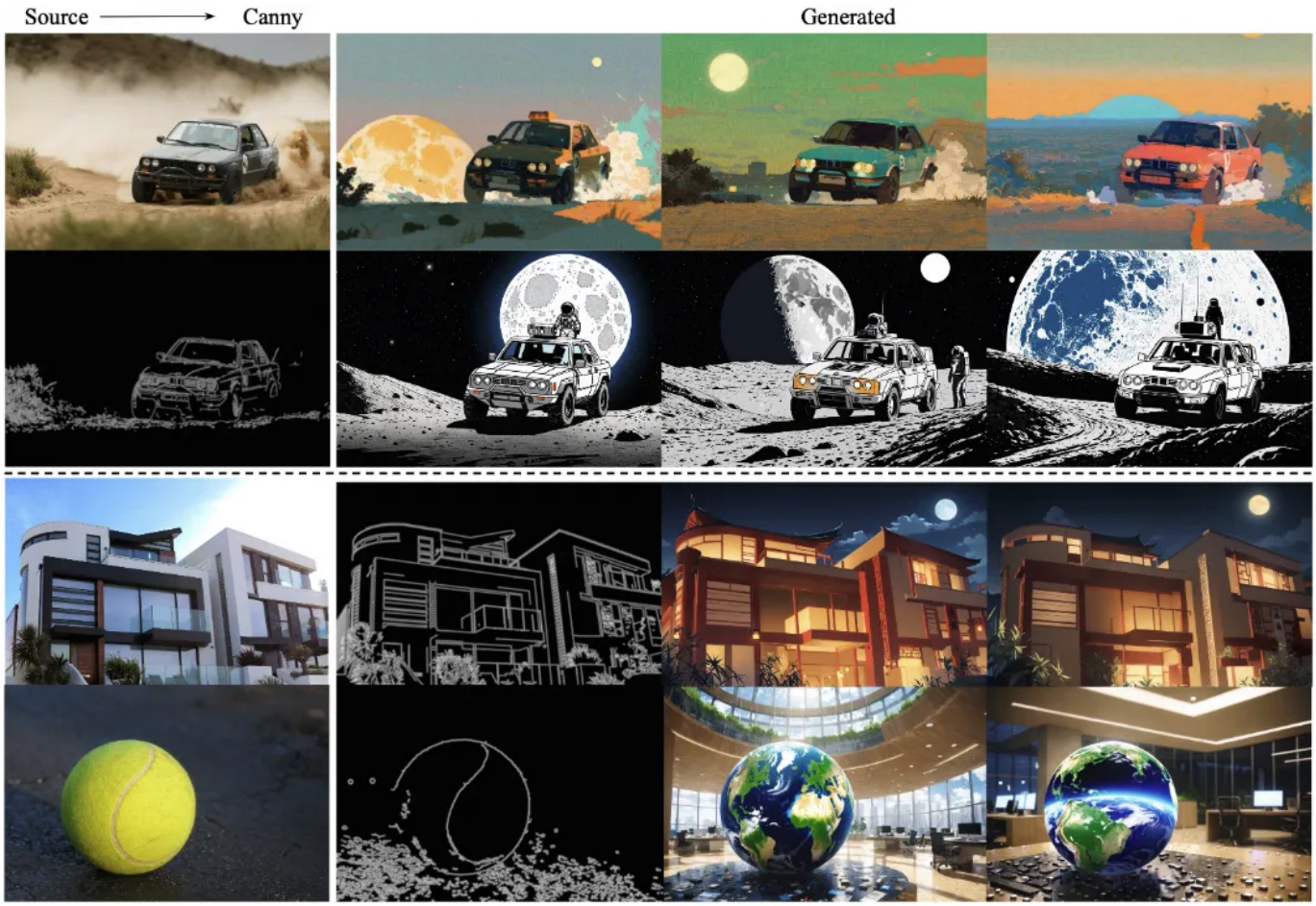


Figure 7. Detailed generation results of the stable diffusion xl are provided. We extract the Canny edges from the input natural image and implement the style transfer using our SDXL mod

4.适配不同的Lora

Figure 8. Our method can serve as a plug-and-play module that adapts to various generation models and LoRA weights, enabling changes in generation style without the need for training.

反思总结

1.论文写作很差，图表中大量错误，实验严重不足（对比ControlNet、T2I-apdater、消融），完全不理解为什么很多公众号推荐

2.确实发现了该领域痛点，例如推理慢，微调占用显存大，但该方法有效性待验证

3.关于Base Model微调 和 Plug-and-Play是否冲突，待验证，模型放得倒挺快

https://github.com/dvlab-research/ControlNeXt/issues/14

4.是否可以有基于特征调制（mul&add）的Adapter网络