Sat2Scene: 3D Urban Scene Generation from Satellite Images with Diffusion

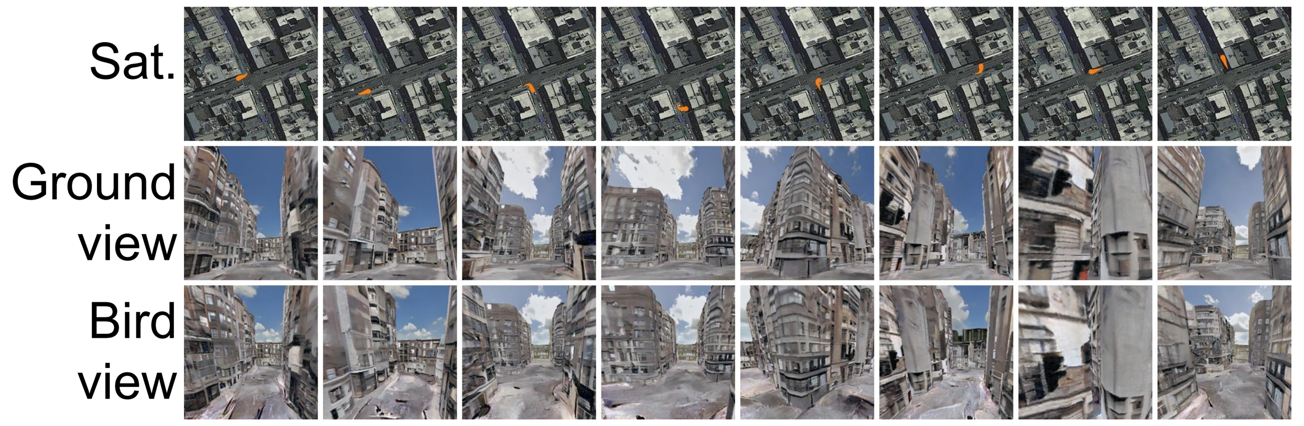
ETH zürich · UTokyo · 浙江大学 ZHEJIANG UNIVERSITY · Microsoft · UNIVERSITY OF AMSTERDAM

Zuoyue Li Zhenqiang Li Zhaopeng Cui Marc Pollefeys Martin R. Oswald

CVPR SEATTLE, WA JUNE 17-21, 2024

1 Introduction

Task Generate 3D urban scene on a given or predicted geometry and render arbitrary 2D views with robust consistency



Why 3D generation?

- Consistency naturally holds
- Do not need preset trajectory

Why diffusion models instead of GANs?

- Better performance
- Stability during training

2 Related Work

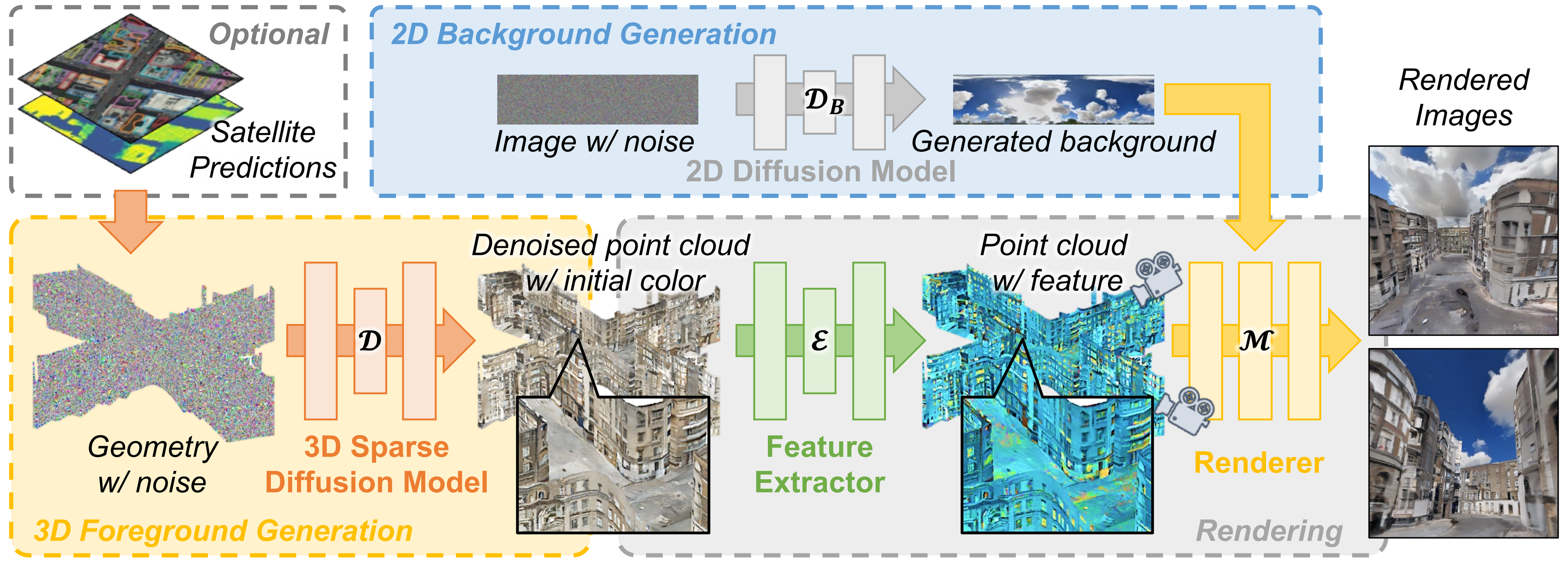
Foundation work

- Diffusion models
- Point-NeRF
- Minkowski Engine

Baselines w/ different generative models

- **Sat2Vid**: 3D GAN-based method
- **InfiniCity**: 2D GAN-based method
- **MVDiffusion**: 2D diffusion-model-based method

3 Method



4 Experiment

Baseline comparison

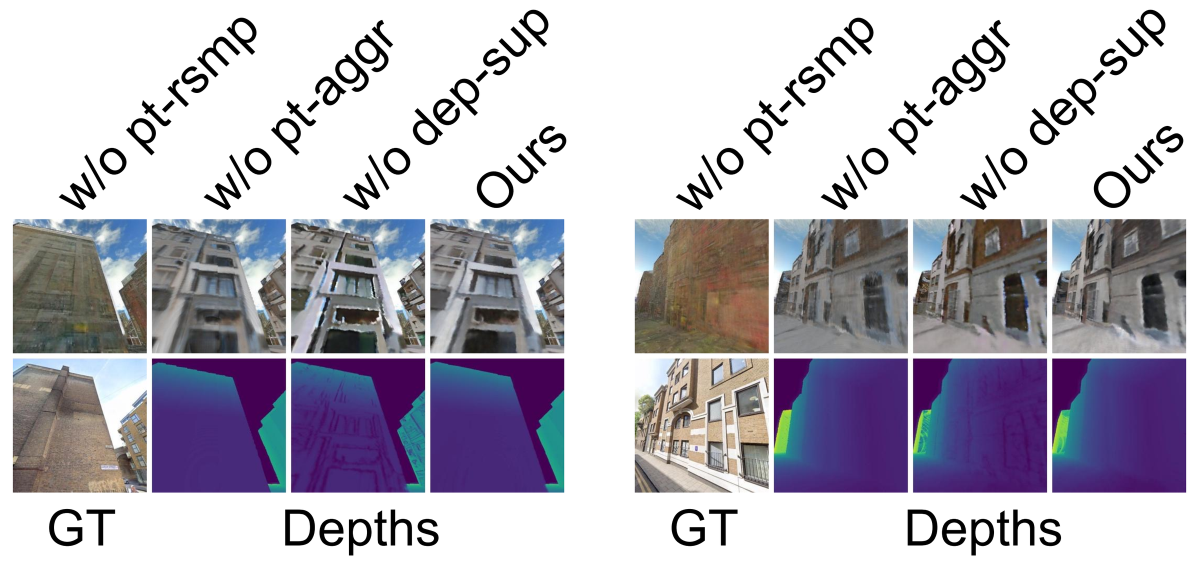
- HoliCity dataset
- GT geometry
- Various metrics

Method / Metric	FVD↓	$KVD_{\times 100}$↓	FID↓	$KID_{\times 100}$↓	PSNR↑	SSIM↑	LPIPS↓	User study
Sat2Vid	37.06	$4.03^{\pm 0.05}$	137.84	$13.76^{\pm 0.10}$	25.25	0.741	0.252	2.92%
InfiniCity	-	-	108.47	$8.40^{\pm 0.10}$	-	-	-	-
MVDiffusion	22.79	$2.36^{\pm 0.03}$	**50.78**	$\mathbf{4.14}^{\pm 0.07}$	17.56	0.593	0.259	15.62%
Ours	**20.30**	$\mathbf{1.90}^{\pm 0.03}$	71.98	$5.91^{\pm 0.06}$	**31.54**	**0.956**	**0.237**	**81.46%**

Ablation study

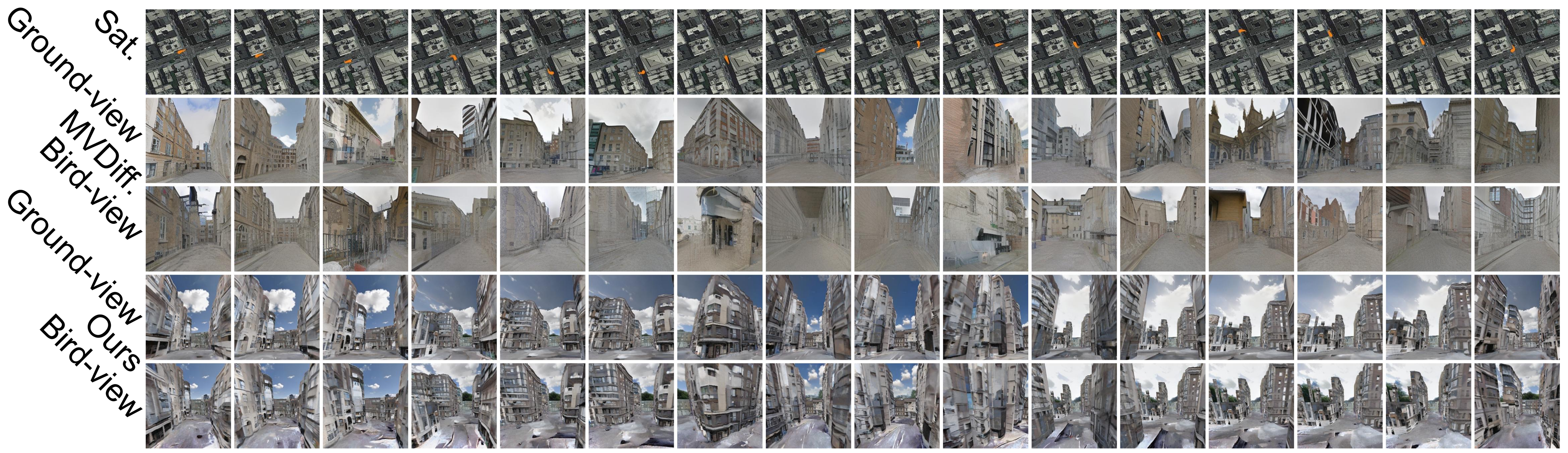
- w/o point resampling
- w/o point aggregation
- w/o depth supervision

Variant / Metric	FID ↓	$KID_{\times 100}$ ↓	Dep. RMSE
w/o pt-rsmp	131.38	$12.66^{\pm 0.12}$	-
w/o pt-aggr	85.58	$7.79^{\pm 0.08}$	3.22
w/o dep-sup	80.40	$7.22^{\pm 0.08}$	3.44
Ours	**71.98**	$\mathbf{5.91}^{\pm 0.06}$	**3.07**



Model generalization

OmniCity dataset, long-seq generation on predicted geometry



5 Conclusion

Contributions

- 3D sparse diffusion models
- Integrated with neural rendering
- Photorealism & robust consistency
- Large-scale 3D scene generation

Future directions

- 3D sparse latent diffusion models
- Advanced scene representation
- Conditional generation

arXiv

Code

Page