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

Supplementary Material

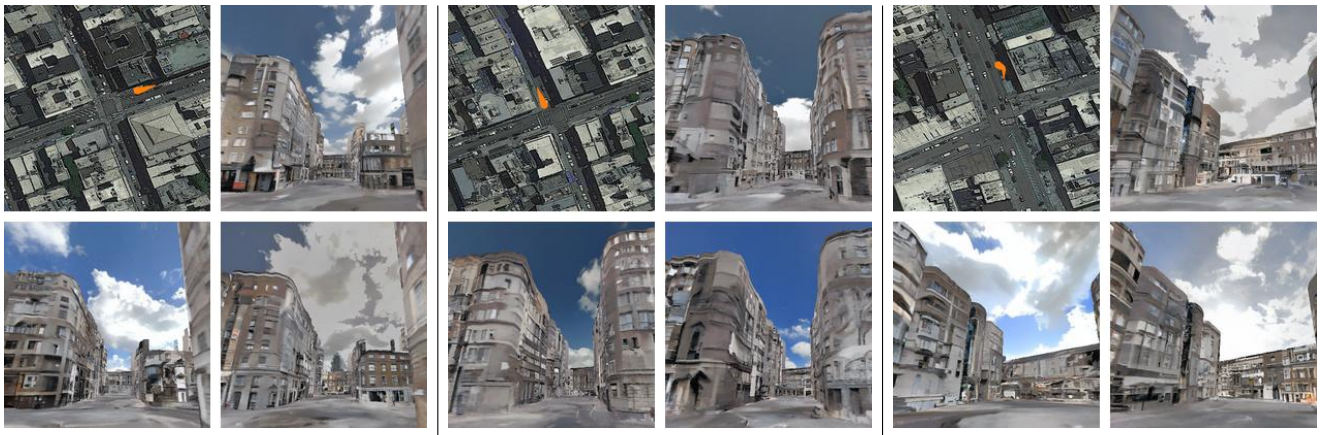

Figure 7. **Denoising procedure visualization with three different diversities**. We align the denoising processes of background and foreground for better visual effects. For each scene, we also provide its satellite top view in the upper-left corner with current locations. Please use **Adobe Reader / KDE Okular** to see **animations**. Use the scroll to drag frames while the mouse is over the video. Leave the current page and back to replay.

In this supplementary material, we provide further details (Sec. 6), present additional experiment results (Sec. 7), and discuss the limitations of our method along with potential future directions (Sec. 8).

6. Additional details

OmniCity [18] video poses. All the videos in Fig. 1, Fig. 6, and Fig. 7 use trajectories along the right side of the road. The ground-view poses are at a height level of 2m, 15° pitch facing the sky and zero roll, while bird-view poses are at a height level of 10m, 15° pitch facing the road and zero roll.
Data augmentation. Since the number of scenes (∼5k) of the HoliCity [45] dataset is not sufficient enough, we add the following data augmentations during training. We randomly flip the whole scene along the two horizontal axes, while we also randomly rotate the scene around the vertical axis. In addition, we perform Gamma correction on the point cloud RGB ground truth, with a random γ value between 0.8 and 1.25.
Inference timing. The generation of the 3D sparse diffusion models takes approximately 10 minutes for the whole denoising process of a single scene. For the neural rendering phase, the inference takes around 1.6s per frame with a resolution of 512×512.
Other related baselines. We did not include experiments with Vid2Vid [38], WC-Vid2Vid [25] and Sat2Density [31] due to the following reasons. Vid2Vid [38] has already exhibited deficiencies in temporal consistency, as observed in Sat2Vid [19]. WC-Vid2Vid [25], on the other hand, necessitates semantic video input, which exceeds the scope of our problem setting. Furthermore, Sat2Density [31] requires for precise satellite-ground image correspondence, which is unavailable in the HoliCity [45] dataset.

7. Additional experiment results

Visualization of the denoising processes on the exemplary scenes are presented in Fig. 7. In the initial seconds of the videos, the transformation of texture from complete noise to meaningful visual patterns is evident. Larger noise patches are observed in the sky background, and the denoising process exhibits stability due to the utilization of LDMs [32].
Multi-style generation examples are also shown in Fig. 7. For each scene, we present three different diversities which are denoised from different noise seeds. It is clear that our model can generate different styles of texture for the same geometry. We noticed that no matter how the style varies, the number of floors of the generated building facade remains similar (approximately 3m per floor) as long as the building height is constant. Also, the style of the ground floor is generally different from that of the higher floors, which is also consistent with real-world buildings.
Supplementary experiments including additional qualitative baseline comparisons and a qualitative ablation study, are presented in Fig. 8 and Fig. 9, respectively. These figures serve as extensions of Fig. 4 and Fig. 5.
The inferior consistency of MVDiffusion [36] compared to its original paper is evident. In addition to the overlap ratio mentioned in Sec. 4.3, we believe this may also be

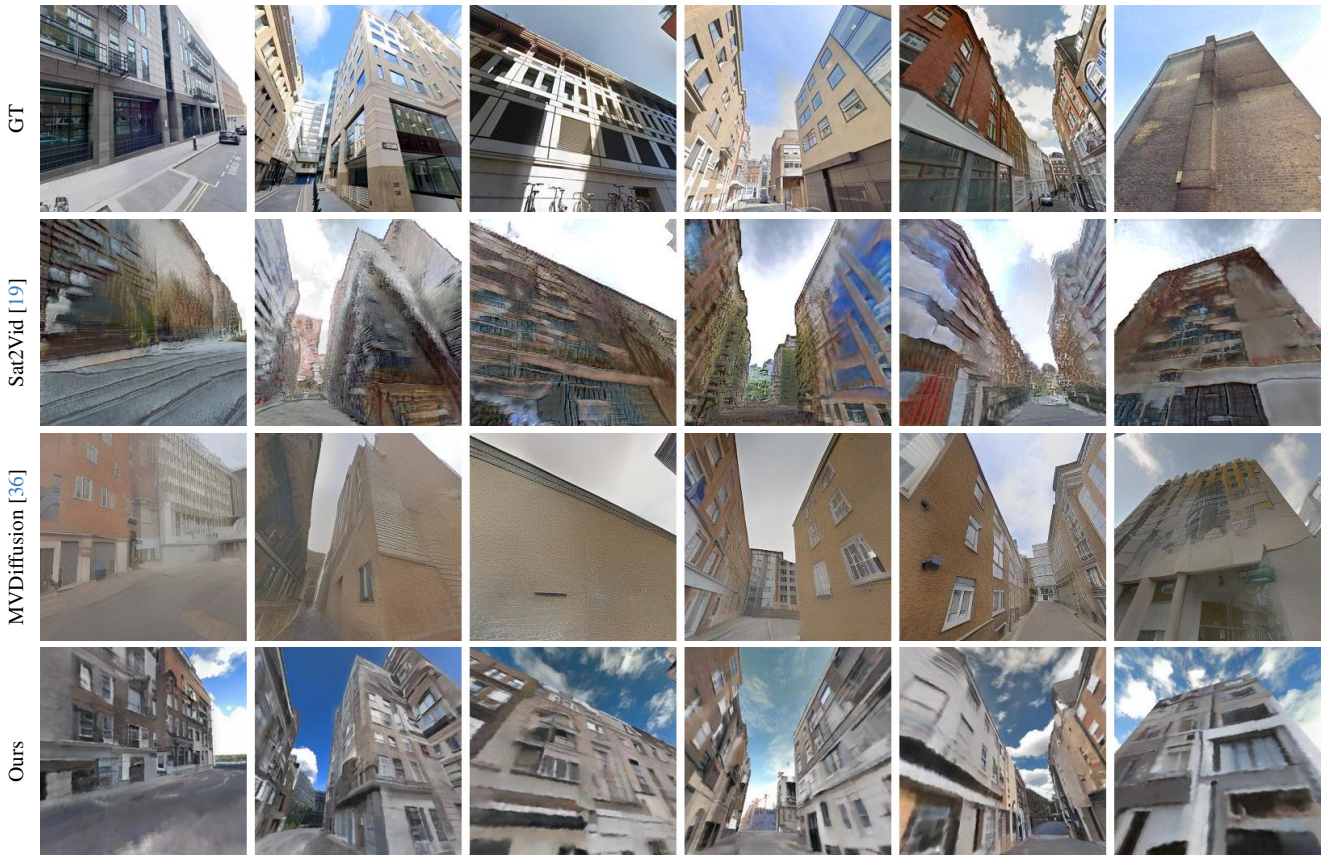

Figure 8. **Additional qualitative baseline comparison on the HoliCity [45] dataset.** Our method produces higher-quality videos with better temporal consistency compared to the baselines. Please use **Adobe Reader / KDE Okular** to see **animations**.

attributed to factors such as text prompt and depth detail. MVDiffusion [36] utilizes long and diversified text prompts to provide the network with rich scene information. To ensure a fair comparison with our approach and other baselines, we use a universal text prompt, which may increase the difficulty of learning color information or detailed texture. As for the depth detail, the indoor depths in MVDiffusion [36] exhibit high quality, even allowing for the recognition and inference of object boundaries. In contrast, the outdoor depths in HoliCity [45] appear more "flat" and lack high-frequency details such as windows and facade decorations. Thus, learning the mapping from geometry to appearance in the HoliCity dataset becomes more challenging compared to the indoor scenarios in MVDiffusion [36].

8. Limitations and discussion

Although Sat2Scene produces photo-realistic street-view videos with robust temporal consistency and outperforms existing methods, there are still several major limitations.

Large-scale generation. Our model may not handle very large-scale scenes, *e.g.*, a city-scale scene due to potentially limited computation resources. Generating blocks separately can be an alternative to solve the scale problem, but it also introduces the potential problem of texture discontinuity between neighboring blocks.

Computation only on surfaces. Our model only performs on the potentially visible surfaces rather than all the surfaces of the scene to reduce computational efforts. The invisible grounds at the base of the buildings were also removed.

Style diversity. The generated building textures do not hold very well diversity, which can be attributed to several factors: **(1)** The GT geometry in HoliCity [45] dataset is not sufficiently detailed, lacking representations of subtle features such as slightly elevated sidewalks or recessed windows on building facades. Such a limitation can hinder the learning process for mapping geometry to appearance. **(2)** Our approach generates entire scenes simultaneously, necessitating the learning of appearance correlations among different building instances, in addition to individual building's appearance. This may lead the network to pay less attention to diversity. **(3)** 3D networks typically have a smaller network scale (number of layers, channels, *etc.*) than 2D ones due to 3D convolution, which limits their capacity to generate diverse appearances. As potential solu-

tions, one can consider introducing latent diffusion models in 3D sparse space or adopting generation based on individual building instances. Also, incorporating semantics as input conditions may contribute to enhancing diversity.

Road surface. The road surface is not very well generated and there are three potential reasons stemming from the dataset: **(1)** The presence of cars, passengers, and shadows on roads in GT images poses a challenge to the segmentation model in filtering them out, potentially resulting in black spots on the road surface during inference. **(2)** Due to the GT poses often looking a bit up to the sky, a small area around the camera origin tends to lack sufficient supervision signals during training. **(3)** The pixel ratio for buildings ($\sim$43.2%) surpasses that of roads ($\sim$15.8%) in GT images, leading the network to prioritize learning building facades in this unbalanced scenario. A potential solution for addressing (3) is to introduce a balance factor in the loss functions or employ distinct networks dedicated to handling buildings and roads, respectively.

Future direction. Below we discuss potential future development directions. Instead of the current setting which is generating the whole scene, we can further divide the whole geometry into small elements, *e.g.*, building instances, and generate textures individually, which could lead to better diversity. In terms of network architecture, we may also follow LDMs [32] to perform the diffusion models in a latent space, given that there is enough 3D urban scene data to train a good auto-encoder with a sparse setting. Furthermore, we can incorporate satellite image information mainly for road surface generation. Also, our model can be combined with a natural language setting, or can even build a side branch similar to ControlNet [43] on top of the trained one, to further extend the model to a conditional generation, *e.g.*, having semantic information as input.

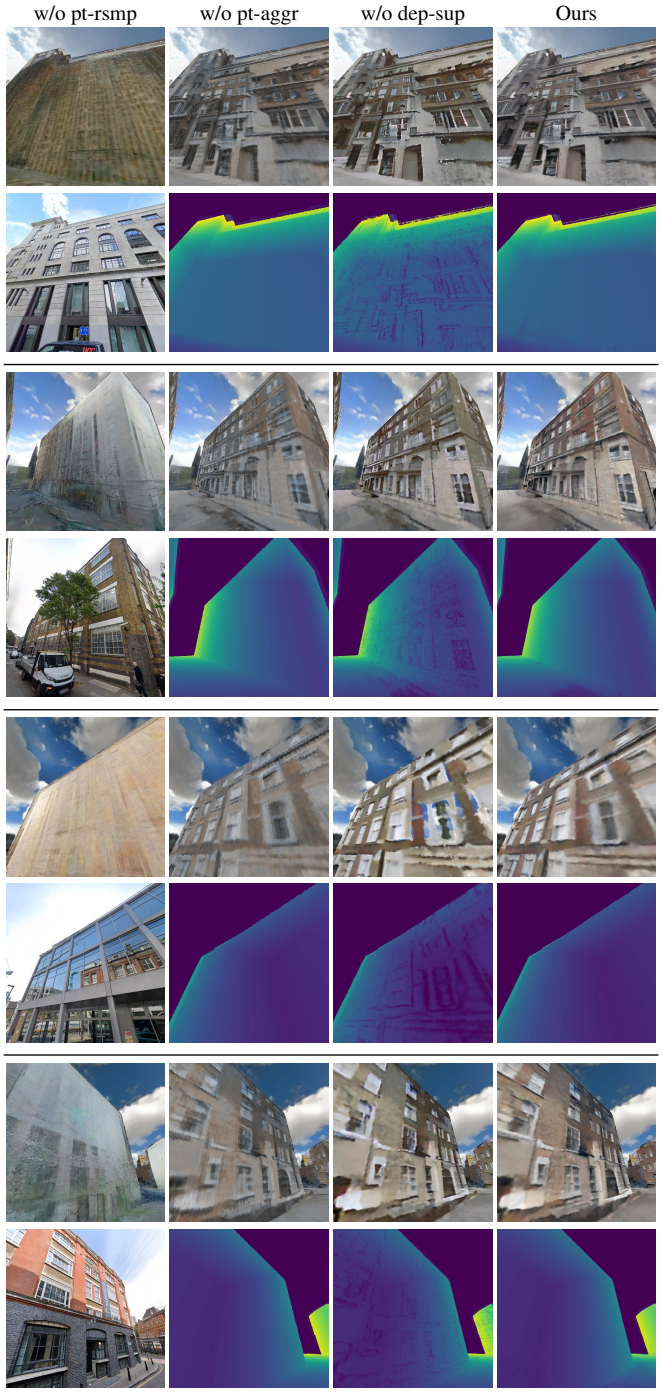

Figure 9. **Additional qualitative ablation study.** We present further qualitative results for various ablations of our method. The rendered images visibly contain more details and the depths are recovered better with our full method. **The second line of each example shows the depth in pseudo colors, except the bottom left ones which are GT images.**

References

[1] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 5

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[3] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 3

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations*, 2023. 5

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 4

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 7

[9] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14313, 2021. 3

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2

[11] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14300–14310, 2023. 2

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[13] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In *ICCV*, 2021. 3, 5

[14] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[15] 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*. Curran Associates, Inc., 2017. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2, 3

[17] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8496–8506, 2023. 2

[18] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17397–17407, 2023. 2, 5, 8, 11

[19] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Rongjun Qin, Marc Pollefeys, and Martin R. Oswald. Sat2vid: Street-view panoramic video synthesis from a single satellite image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12436–12445, 2021. 1, 3, 5, 6, 7, 11, 12

[20] Zuoyue Li, Tianxing Fan, Zhenqiang Li, Zhaopeng Cui, Yoichi Sato, Marc Pollefeys, and Martin R. Oswald. Compnvs: Novel view synthesis with scene completion. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, pages 447–463. Springer, 2022. 3

[21] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22808–22818, 2023. 1, 3, 5, 6

[22] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Thirty-four Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3

[23] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 3

[24] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[25] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020. 3, 11

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *The European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3

[27] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4328–4338, 2023. 2

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. (SIGGRAPH)*, 41(4):102:1–102:15, 2022. 1

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3

[30] 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. 3

[31] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3683–3692, 2023. 1, 3, 11

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 5, 7, 11, 13

[33] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 4

[34] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. In *TPAMI*, 2022. 1

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2, 3

[36] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5, 6, 7, 8, 11, 12

[37] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*, 2019. 5

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3, 11

[39] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5438–5448, 2022. 1, 3

[40] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 2

[41] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Zhengzhe Liu, and Xiaojuan Qi. Texture generation on 3d meshes with point-uv diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4206–4216, 2023. 2

[42] Cem Yuksel. Sample Elimination for Generating Poisson Disk Sample Sets. *Computer Graphics Forum*, 2015. 4

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2, 5, 13

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[45] Yichao Zhou, Jingwei Huang, Xili Dai, Linjie Luo, Zhili Chen, and Yi Ma. HoliCity: A city-scale data platform for learning holistic 3D structures. *arXiv*, 2020. arXiv:2008.03286 [cs.CV]. 2, 5, 6, 7, 11, 12