LooseControl: Lifting ControlNet for Generalized Depth Conditioning

Supplementary Material

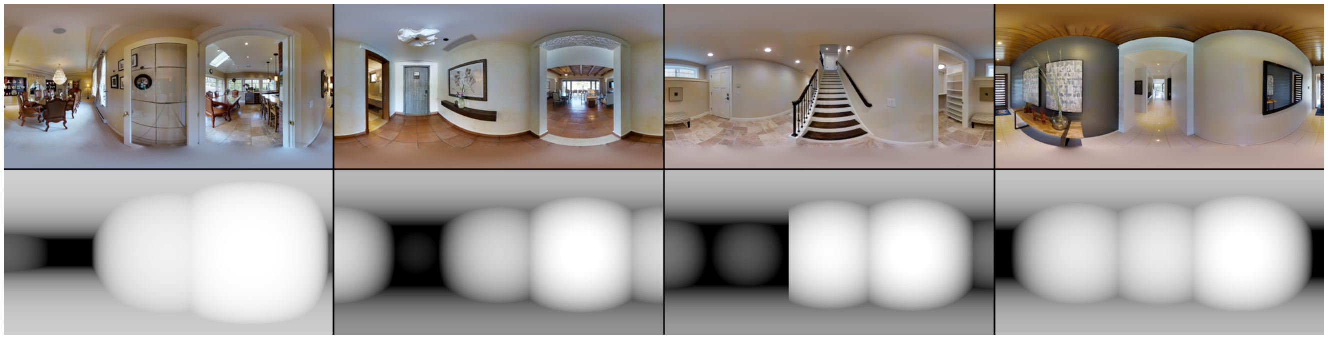

Figure 1. Upper-bound violations in Matterport3D [7]. *Top:* RGB panorama images. *Bottom:* Depth rendered using the layout labels from the dataset.

A. Implementation details

For all our experiments, we use Stable Diffusion v1.5 [12] and the corresponding ControlNet [13] checkpoint *"lllyasviel/control_v11f1p_sd15_depth"* hosted on HugginFace [1]. We use the PyTorch [10] framework and the diffusers [3] library as the framework for diffusion models. We use ZoeDepth [6] and SAM [9] for extracting depth maps and segmentation maps respectively. For SAM, we use `min_mask_area=1e4`. We make use of Pytorch3D [11] and Open3D [14] in our 3D framework and PyTorch3D for rendering the depth maps for obtaining the proxy depths for fine-tuning. For backprojection, we randomly choose an FOV in the range of 43 and 57 degrees. For both Scene Boundary Control and 3D Box Control, we use LoRA rank $r = 8$ and fine-tune only LoRA layers for 200 steps with a learning rate of 0.0001, and batch size of 12 with Adam [8] optimizer. We use `controlnet_conditioning_scale=1.0` and LoRA scale factor $\gamma = 1.2$ as default. We built our 3D editor user interface using Gradio [5] and BabylonJS [2].

B. Issues with Room Layout Datasets

As discussed in the main paper, a possible alternative for extracting the scene boundary for the preparation of the dataset for Scene Boundary Control could be room layout datasets. However, we note that these datasets often contain ambiguous layout labels that directly violate our tight upper-bound condition required to implement scene boundary control. We provide some examples of these violations for the popular MatterPort3D dataset as the representative in Fig. 1

C. User study additional details

We created an anonymous user study as a web form using streamlit [4]. Users were presented with 'Two-alternative forced choice' (2-AFC) with two options as images generated by our baseline (ControlNet) and our result and asked to respond with their preference for the given text prompt and condition image. The options were anonymized and did not indicate the name of the method. Each user was asked to respond to 10 randomized questions in total (5 for Scene Boundary Control, and 5 for 3D box control). The order of options was also randomized. As mentioned in the main paper, over 95% of responses were in favor of our method.

D. Additional Results

We provide additional results and interactive examples for controlled generations and editing in the attached HTML web page.

References

[1] https://huggingface.co/. 1

[2] https://www.babylonjs.com/. 1

[3] https://huggingface.co/docs/diffusers/index. 1

[4] https://streamlit.io/. 1

[5] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 1

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song,

Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[11] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 1

[12] 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*, pages 10684–10695, 2022. 1

[13] 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*, pages 3836–3847, 2023. 1

[14] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 1