Generalizable Visual Localization for Gaussian Splatting Scene Representations

Supplementary Material

Table 1. **Gaussian splatting PSNR scores.** We provide the PSNR scores for our trained GS models across each scene in the Cambridge Landmarks [4] and 7-Scenes [3, 7] datasets.

Cambridge Landmarks - Outdoor

Kings	Hospital	Shop	StMary	Average
20.8	16.9	22.0	21.7	20.4

7-Scenes - Indoor

Chess	Fire	Heads	Office	Pump.	Kitchen	Stairs	Average
28.9	28.8	30.7	28.9	30.0	24.4	30.4	28.9

Below, we provide additional details about our implementation and examples of challenging scenes from the ScanNet++ dataset in Figure 1. As shown, the scenes are large and diverse, including extensive texture-less areas, which demonstrates the generalization ability of our cross-scene model.

Our code is still a work in progress. We will publish it after finishing to clean and refactor it for easy use. All of the data we used is publicly available. We will also release our pre-trained models.

1. Gaussain Splatting Implementation Details

We use the pre-built COLMAP reconstructions from [1] for the 7scenes dataset and the reconstructions provided in HLoc toolbox [6] for the Cambridge landmarks dataset. We train all the scenes using the vanilla 3DGS [5], for 30k iterations using the default parameters, in Table 1 we report the per-scene PSNR scores for our trained models on the training images. Notably, the rendering quality of outdoor scenes is inferior compared to indoor scenes, which might explain the degradation in our pose estimation accuracy.

Handling challenges in outdoor scenes. To effectively train a 3DGS model for outdoor scene reconstruction, we focus on reconstructing static elements such as buildings, fences, and signs. This approach addresses real-world challenges like varying lighting conditions, dynamic objects, and distant regions. To mitigate these issues, we use a pre-trained semantic segmentation model [2] to mask out sky regions and moving objects, including pedestrians and vehicles. These elements, which constitute only a small portion of the captured images, are excluded from the loss function during training, resulting in more accurate scene reconstruction. For this purpose, we utilize pre-computed segmentation maps provided by [9], generated using the method described in [2].

References

[1] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6218–6228, 2021. 1

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[3] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179. IEEE, 2013. 1

[4] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 1

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1

[6] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1

[7] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 1

[8] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2

[9] Qunjie Zhou, Maxim Maximov, Or Litany, and Laura Leal-Taixé. The nerfect match: Exploring nerf features for visual localization. *arXiv preprint arXiv:2403.09577*, 2024. 1

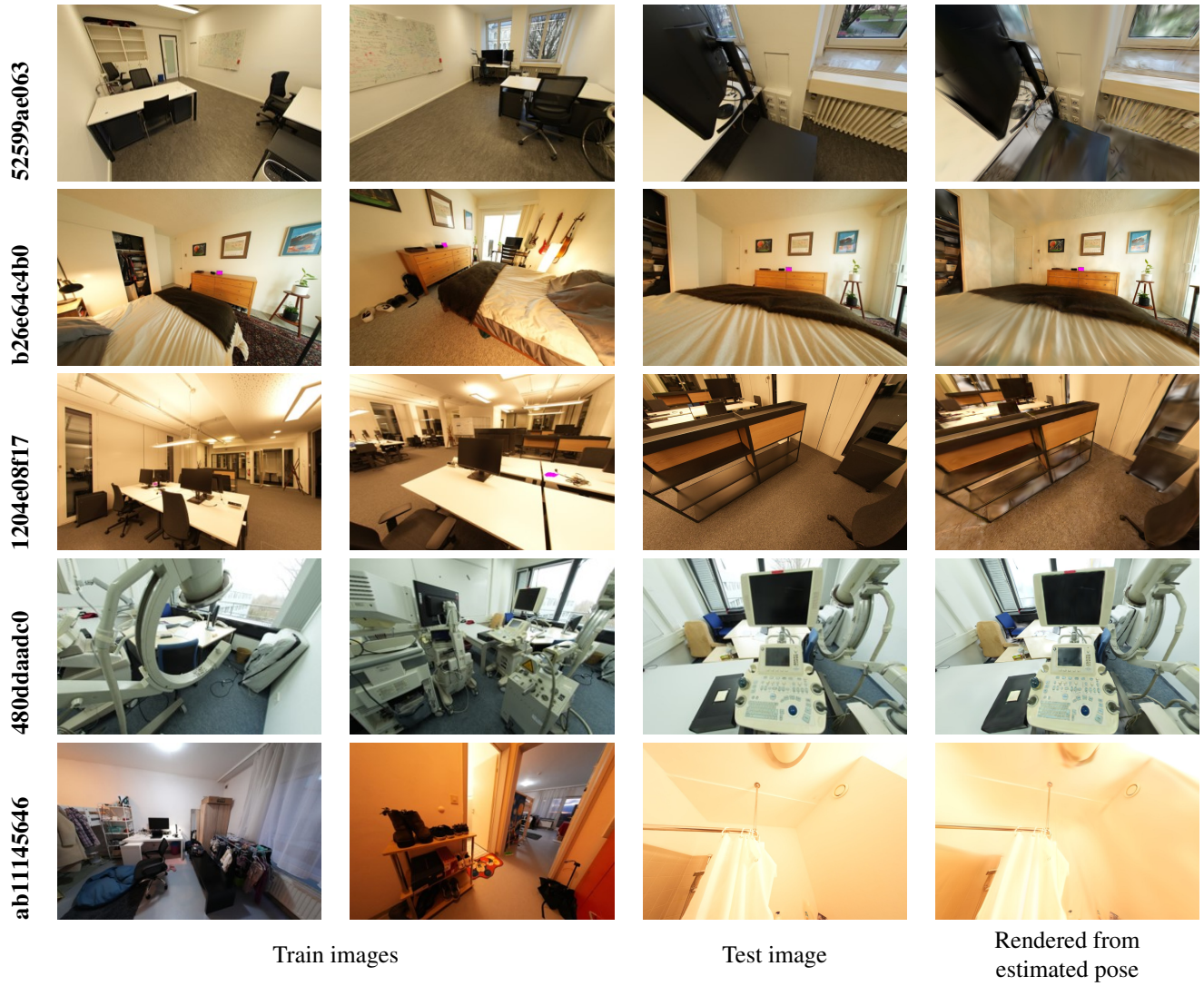

Figure 1. **ScanNet++.** Examples of qualitative results obtained by our cross-scene model on diverse scenes of ScanNet++ [8]. From left to right, two images from the training images, a test query image, and the rendered image from our model pose prediction.