Uni3R: Unified 3D Reconstruction and Semantic Understanding via Generalizable Gaussian Splatting from Unposed Multi-View Images

Supplementary Material

Table 1. **Out-of-distribution performance comparison.** Our method shows superior performance when zero-shot evaluation on DTU and ScanNet++ using the model solely trained on RE10k.

Method	DTU		ScanNet++	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
pixelSplat	11.551	0.633	18.434	0.277
MVSplat	13.929	0.385	17.125	0.297
NoPoSplat	17.899	0.279	22.136	0.232
Ours	**18.256**	**0.266**	**22.221**	**0.227**

Table 2. Ablation Study for confidence mask ratio (top-K) on the ScanNet dataset under 2-views setup on source views.

Top-k ratio mask	mIoU↑	Acc.↑	rel↓	τ ↑	PSNR↑	SSIM↑	LPIPS↓
w/o geo. loss	53.88	82.18	5.81	47.99	24.24	0.850	0.108
w/ ratio 100%	53.86	82.46	3.92	60.75	24.24	**0.851**	0.108
w/ ratio 90%	**54.03**	**82.55**	**3.87**	**61.37**	**24.35**	**0.851**	**0.107**
w/ ratio 70%	54.00	82.47	4.55	55.28	24.17	0.848	0.111

1. Appendix

1.1. Results on the DTU and ScanNet++ dataset

To evaluate the cross-domain generalization of Uni3R, we follow NoPoSplat [5]: training on RE10K [8] dataset and testing on DTU [3] and ScanNet++[6] dataset. As shown in Tab. 1, Uni3R consistently outperforms all baseline methods on the benchmarks.

1.2. More Ablation Study on confidence parameter setting in geometry-guided loss

To validate the effectiveness of our confidence mask in geometry-guided loss, we conduct an ablation study by varying the top-K ratio used for supervision. As shown in Table 2, applying a 90% confidence mask yields the best performance in mIoU, depth accuracy, and rendering quality, demonstrating that filtering out low-confidence regions improves overall performance.

Futhermore, the geo. loss from the point map is an essential stability anchor for our unified tasks. In Fig. 1, training without this constraint under complex setups (e.g., 4-view) leads to model collapse due to the high degree of freedom in Gaussian optimization. Furthermore, Tab. 2 shows in 2-view, the geometry loss significantly improves geometric (47.99 → 61.37) while simultaneously improving mIoU (53.88 → 54.03). We believe that observing performance improvements across three distinct tasks using only a geometric loss provides a non-trivial insight for the field.

Figure 1. **Model training w/ and w/o geo. loss on 4 views.**

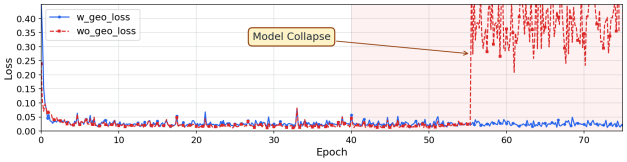

Table 3. **Comparison of our method against per-scene optimized methods.**

Method	8 views		16 views	
	rel↓	τ↑	rel↓	τ↑
Feature-3DGS [7]	17.28	13.31	23.71	10.57
Ours	**4.46**	**56.88**	**5.88**	**42.88**

1.3. Depth Evaluation under Multi-View Settings

For fair comparison, we follow LSM [2] and adopt Absolute Relative Error (rel) and Inlier Ratio (τ) with a threshold of 1.03 for per-scene depth evaluation. This setting is consistently used throughout the paper.

As shown in Tab. 3, Uni3R outperforms the per-scene optimized method on depth estimation under both 8-view and 16-view settings. Notably, our method achieves better depth evaluation performance in one feed-forward.

1.4. Training and Evaluation Details

As described in our main paper, we trained our model on three datasets including ScanNet [1], RE10k [8] and ACID [4].

For model training on ACID [4] and RE10K [8] dataset, we progressively train 2, 4 and 8 view model. For 2-view training on ACID [4] and RE10K [8], we follow NoPoSplat [5]. For 4-view training on RE10K, we initialize the model from the 2-view checkpoint and train it on 8×H100 GPUs with a learning rate of 4e-5 for 40,000 iterations, using a batch size of 4 per GPU. For 8-view training, we further initialize from the 4-view checkpoint and train under the same settings, with a batch size of 1 per GPU.

For the ScanNet [1] dataset, we train Uni3R under 2-view, 8-view, and 16-view settings. For the 2-view setup, we follow the LSM [2]. For the 8-view training, we initial-

ize from the 2-view checkpoint and train the model with a learning rate of 5e-5, with a 5 epochs warmup and 50 total epochs. The batch size is set to 4 per GPU. For the 16-view training, we also initialize from the 2-view checkpoint, with all settings identical to the 8-view setup except for the batch size 2 per GPU.

Additionally, for our arbitrary-view model in the main paper, we uniformly sample 2, 4, and 8 input views from the ScanNet [1] dataset and train the model using a batch size of 1 per GPU. The training is performed with a learning rate of 1e-4, including a 10-epoch warm-up and 100 total epochs. As demonstrated in the main paper, our arbitrary-view model achieves consistently comparable performance across different numbers of input views.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2432–2443, 2017. 1, 2

[2] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, and Marco Pavone. Large spatial model: End-to-end unposed images to semantic 3d. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1

[3] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 406–413. IEEE Computer Society, 2014. 1

[4] Andrew Liu, Ameesh Makadia, Richard Tucker, Noah Snavely, Varun Jampani, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14438–14447. IEEE, 2021. 1

[5] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025. 1

[6] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12–22. IEEE, 2023. 1

[7] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21676–21685, 2024. 1

[8] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1