DepthSplat: Connecting Gaussian Splatting and Depth

Appendix

A. Depth Pre-Training for Gaussian Splatting

In the main paper, we have shown that better depth architectures lead to improved view synthesis with Gaussian splatting. We further study the effect of different weight initializations for the depth model. Specifically, we compare three variants: 1) only initializing the monocular feature with Depth Anything V2 [6]; 2) initializing the monocular feature with Depth Anything V2 [6] and the multi-view feature with UniMatch [5]; 3) initializing the full depth model by pre-training the depth model on TartanAir [4]. We can observe from Tab. A.1 that better depth initialization also leads to improved view synthesis results.

Table A.1. **Depth pre-training for Gaussian splatting**. We compare different weight initializations for the depth model when training our full DepthSplat model for view synthesis. Compared to 1) only initializing the monocular feature (mono features) and 2) initializing the monocular and multi-view features (mono & mv features), our pre-trained full depth model (full depth model) achieves the best view synthesis results.

Initialization	PSNR ↑	SSIM ↑	LPIPS ↓
mono features	26.59	0.874	0.1256
mono & mv features	26.76	0.877	0.1234
full depth model	**26.81**	**0.878**	**0.1225**

B. More Visual Results

B.1. Unsupervised Depth Pre-Training with Gaussian Splatting

In Fig. B.1, we show the comparisons of depth estimation results with and without Gaussian splatting pre-training. We observe that pre-training leads to better results in texture-less regions. We hypothesize that pre-training provides regularizations (as also observed in [2]) to the challenging scenarios and accordingly leads to improved performance.

B.2. Cross-Dataset Generalization

In Fig. B.2, we show the cross-dataset generalization results on the DL3DV [3] dataset, which are obtained with the RealEstate10K [7] pre-trained models. Our DepthSplat generalizes more robustly than MVSplat [1] on unseen scenes.

B.3. Visual Comparisons on DL3DV

In Fig. B.3, we compare the visual synthesis results from 4 input views with MVSplat [1]. Our DepthSplat better preserves the scene structures.

B.4. High-Resolution Results

In Fig. B.4 and Fig. B.5, we show the view synthesis and multi-view depth estimation results at 512×960 resolutions. Note that MVSplat [1] runs out-of-memory on such high resolutions with 6 or 12 input views, while our DepthSplat significantly improves the efficiency with two technical components. First, our lowest feature resolution is $1/8$ of the image resolution, while MVSplat uses $1/4$. Second, we use local cross-view attentions unlike the pair-wise global attentions in MVSplat.

C. More Implementation Details

We provide more implementation details on the high-resolution experiments. For high-resolution experiments, we choose our small model which contains a ViT-S monocular branch and a single-scale multi-view branch. We first train our model on RealEstate10K with two input views at 256×448 resolutions for 150K iterations, where the total batch size is 256. We fine-tune the model on the mixed RealEstate10K and DL3DV datasets at 448×768 resolutions for 200K iterations with a total batch size of 64, where the number of input views is randomly sampled from 4 to 10. We use this model to predict results on high resolutions (*e.g.*, 512×960) and different numbers of input views (*e.g.*, 6 and 12), as we show in our project page: haofeixu.github.io/depthsplat.

References

[1] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, 2024. 1

[2] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *JMLR*, 2010. 1

[3] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 1

[4] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 1

[5] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 1

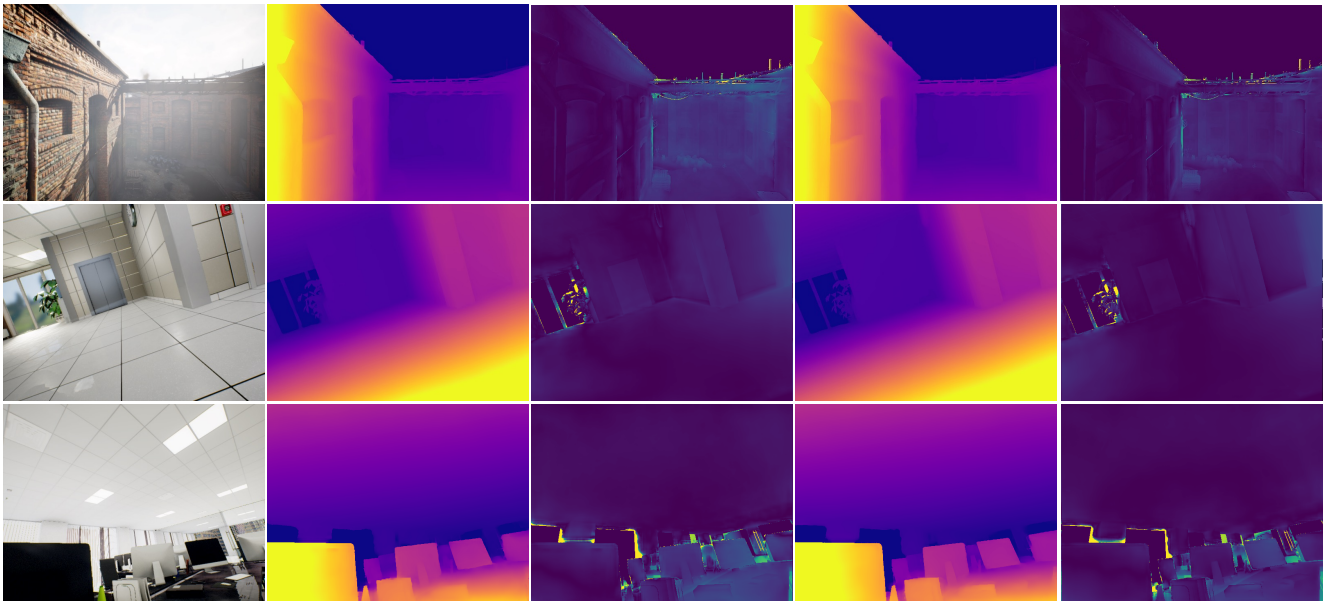

Figure B.1. **Effect of unsupervised depth pre-training**. We observe that the unsupervised pre-training leads to improved performance for texture-less regions.

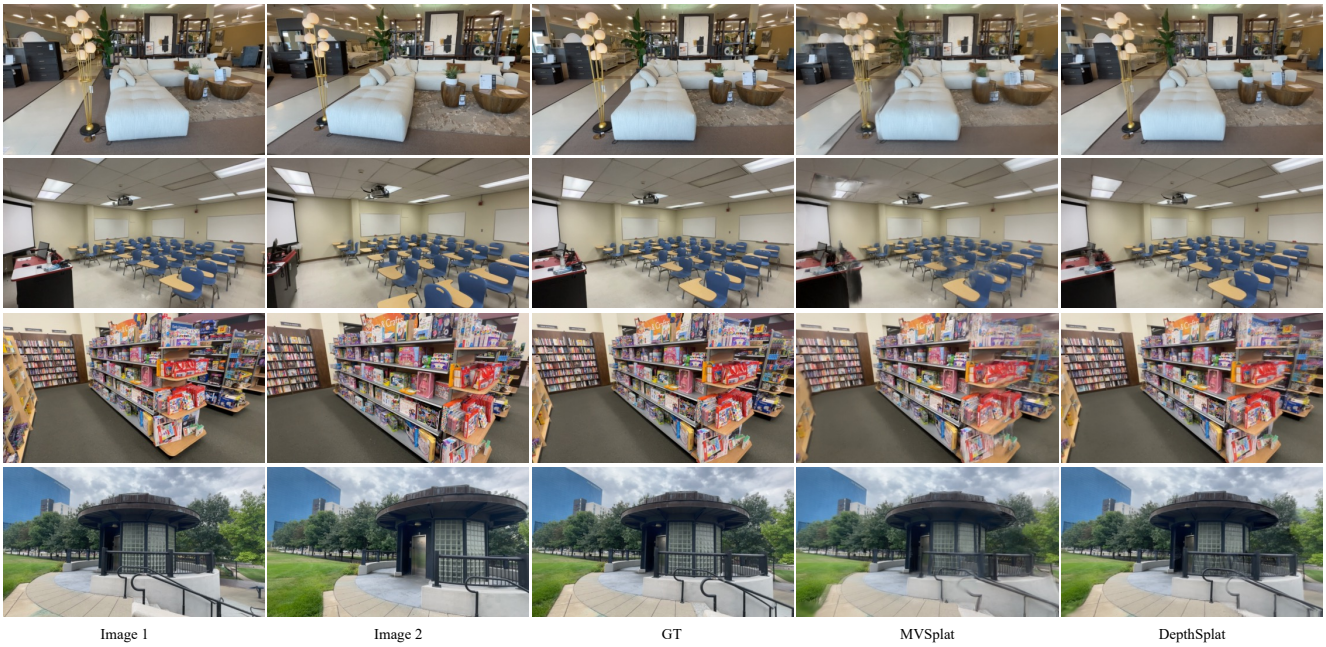

Figure B.2. **Generalization from RealEstate10K to DL3DV**.

[6] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. 2024. 1

[7] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM TOG*, 2018. 1

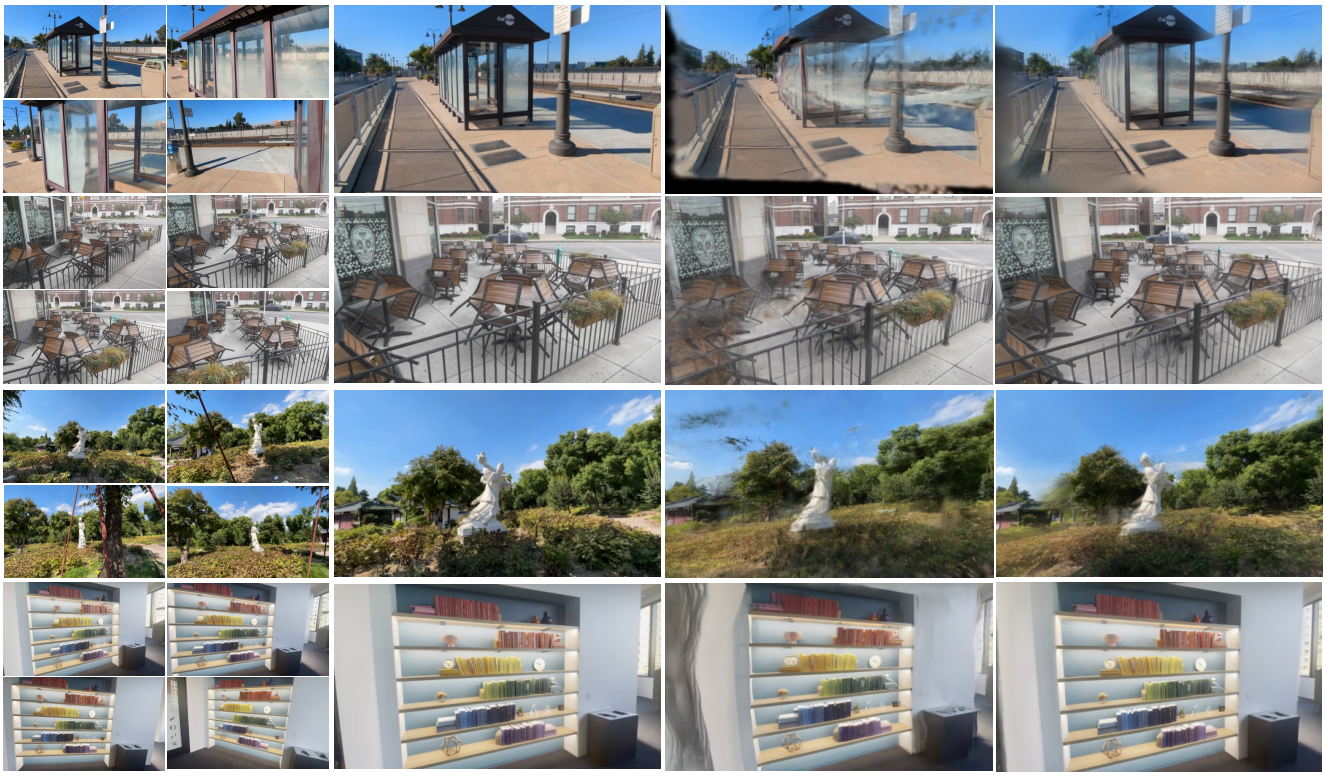

Figure B.3. **View synthesis from 4 input views on DL3DV**.

Figure B.4. **View synthesis at** 512×960 **resolutions from 6 input views**.

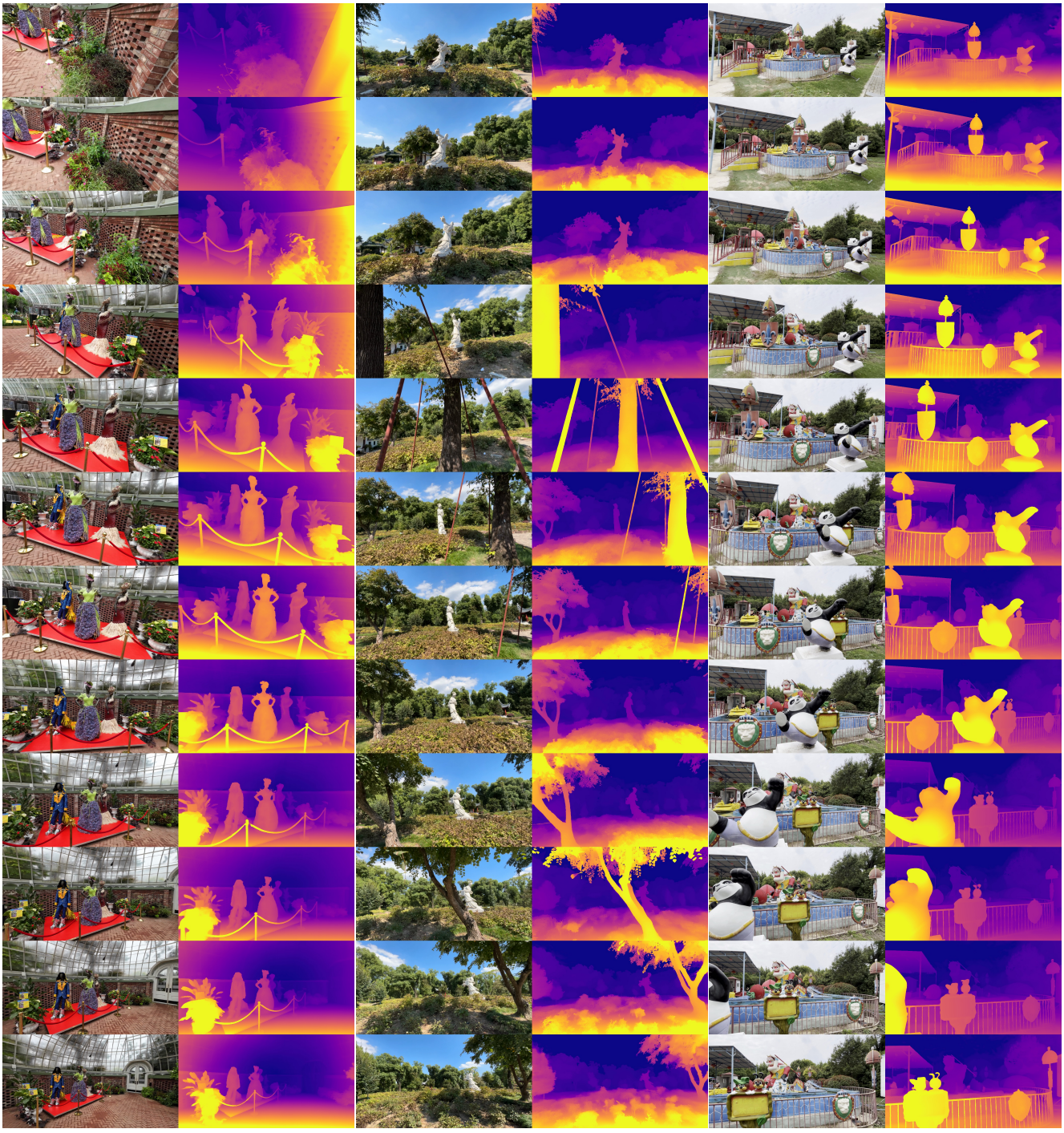

Figure B.5. **Depth predictions on DL3DV with 12 input views**. The image resolutions are 512×960.