

Free360: Layered Gaussian Splatting for Unbounded 360-Degree View Synthesis from Extremely Sparse and Unposed Views - Supplementary Material

Chong Bao^{1§} Xiyu Zhang¹ Zehao Yu³ Jiale Shi¹ Guofeng Zhang¹
Songyou Peng² Zhaopeng Cui^{1†}

¹State Key Lab of CAD&CG, Zhejiang University ²ETH Zürich

³University of Tübingen, Tübingen AI Center

In this supplementary material, we first present detailed implementation aspects in Section A. More experimental details are shown in Section B. We show more comparisons in the Sec. C. Additionally, we include a short video summarizing the method with video results, and an offline webpage for interactive visualization of our whole results and comparisons.

A. Implementation Details

For the fine-grained front and back layer masks, we annotate the maximum depth of the front layer in monocular depth [10] of each view. The pixels are selected into the front layer if their depth is smaller than the annotated maximum depth. We build Free360 upon the 2DGS [3] framework. We follow the version implemented in the StableNormal [11]. We use default settings in dense stereo reconstruction models [7, 9] and use the filtered point cloud by predicted confidence map. We transform the world origin to the center of the scene, which is determined by the center depth of the first image. Besides, we rescale the cameras to fit within a sphere of radius 2.

In reconstruction bootstrap optimization, we downsample the point cloud of the front layer before initializing its Gaussian primitives. We initialize the front layer’s Gaussian primitives using its point cloud and train for 10,000 iterations based on the loss defined in Eq. (1). We enable densification from the 166-th iteration to 5000-th iteration.

In the iterative fusion of reconstruction and generation, we define the unknown cameras in two ways. First, we interpolate the poses between input sparse views in the cubic spline interpolator. Second way is to define a target camera pose by jittering the position of an input camera pose while orienting its rotation to face the world origin, and interpolate the poses between the target pose and closet input pose. We empirically define 300 to 400 unknown camera poses in total from these two ways. We use ViewCrafter [12] to generate

the 25 frames each time with the resolution 1024×576 .

In uncertainty-aware training, we set the maximum L1 difference β between conditions and generations as 0.2. We train the Gaussian primitives of the front layer and back layer using Eq. (3) in 10000 iterations without densification. All experiments are conducted on an NVIDIA RTX 6000 GPU.

B. Experimental Details

B.1. Dataset

We double the official downsample factor used by 3DGS [4] in Mip-NerF 360 [1]. For Tanks and Temples [6], we use the processed data from PixelGS [13] and downsample the images by factor 2. To evaluate the metrics between rendered images and ground-truth images, we follow the InstantSplat [2] to align the estimated poses from stereo reconstruction model [7, 9] to the ground-truth poses. Initially, a coarse alignment is obtained through rigid point registration between the estimated and ground-truth camera positions at the training viewpoints. Subsequently, for each rendered image, we fix the Gaussian primitives, and a test-time optimization is performed on the camera pose by minimizing the L1 difference between the rendered and ground-truth images. This optimization is executed for 500 iterations using the Adam optimizer [5], with a learning rate of 0.0003 for position and 0.0001 for rotation.

B.2. Baselines

All compared methods use the same camera poses and point clouds from dense stereo reconstruction [7, 9]. Since ZeroNVS [8] and ViewCrafter [12] are agnostic to the reconstruction backbone, we adopt the same 2DGS backbone [3, 11] for both methods to ensure a fair and rigorous comparison. In geometry evaluation, the same 2DGS [3, 11] backbone is used as the baseline. For low-overlapped sparse views of the unbounded scene, ViewCrafter [12] needs to iteratively generate novel views within a small region, utilize Dust3R on generated views to recover the scene’s point cloud, and subsequently repeat to generate the next portion

[†]Corresponding author.

[§]The work was partially done when visiting ETHZ.



Figure A. **Comparison on Mip-NeRF 360° [1] Dataset on the 3-View Setting.** We qualitatively compare rendering quality with FSGS* [14], InstantSplat [2], ZeroNVS* [8], ViewCrafter [12] given 3 input views.

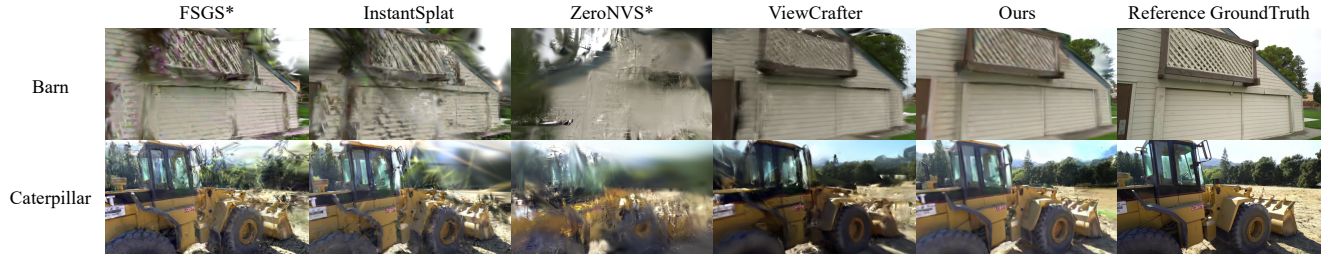


Figure B. **Comparison on Tanks and Temples [6] Dataset on the 4-View Setting.** We qualitatively compare rendering quality with FSGS* [14], InstantSplat [2], ZeroNVS* [8], ViewCrafter [12] given 4 views.

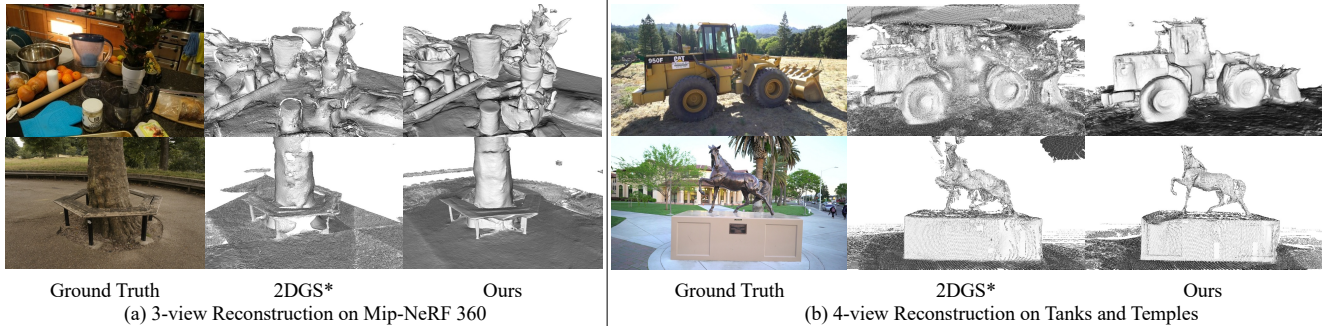


Figure C. **Comparison on 3D Surface Reconstruction.** We qualitatively compare surface reconstruction with 2DGS [3] on (a) Counter (top) and Treehill (down) from Mip-NeRF 360° [1] dataset and (b) Caterpillar (top) and Horse (down) from Tanks and Temples dataset [6].

of the scene conditioned previously generated image and Dust3R [9] point cloud. However, Dust3R’s feed-forward nature struggles with inconsistencies in generated images, leading to inaccurate depths that degrade subsequent generations. In contrast, our iterative fusion framework integrates an uncertainty-aware GS optimization after each iteration to refine the generative error promptly. The optimized 3D-consistent GS rendering is used to condition subsequent generations for consistent multi-view generation guiding the next GS optimization. ViewCrafter [12] and ZeroNVS [8]

use the same group of unknown cameras to generate novel views as our method.

C. More Experiments

C.1. Novel View Synthesis

We show more rendering comparisons on the Mip-NeRF360 [1] and Tanks and Temples [6], as illustrated in Fig. A and Fig. B respectively. FSGS [14] and InstantSplat [2] exhibit severe distortion and needle-like Gaus-

Settings	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
w/o \mathcal{L}_d & \mathcal{L}_n	16.723	0.310	0.505
w/o \mathcal{L}_d	16.799	0.314	0.499
w/o \mathcal{L}_n	16.825	0.313	0.498
Ours	16.95	0.321	0.480

Table A. We perform ablation studies on the geometric priors in our method.

sian artifacts in the rendering results. ZeroNVS [8] fails in synthesizing clear novel views due to limited consistency from generative prior. ViewCrafter [12] cannot present a detailed and consistent rendering of the scene. Instead, our method shows the crisp rendering and complete structure of the scene.

C.2. Geometry Evaluation

We show more geometry evaluation on the Mip-NeRF360 [1] and Tanks and Temples [6], as illustrated in Fig. C. Given the sparse views, the geometry from 2DGS has many missing areas and distorted surfaces, such as the holes in Treehill, the missing legs in Horse, and the floater at the top of Caterpillar. In contrast, our method not only produces a complete and smooth geometry of the scene but also detailed structures. We show the F1-score precision and recall curves of 2DGS and our method in Fig. D. Since the extremely sparse-view surface reconstruction is ambiguous and error-prone, few points lie within the official error threshold in Tanks and Temples [6]. To facilitate a clearer comparison, we increase the error threshold by a factor of 10 (represented by the black dotted line in Fig. D).

C.3. Ablation on geometry prior

As shown in Tab. A, we show ablation of geometry priors \mathcal{L}_d and \mathcal{L}_n in Eq. 1, 2 on *Bicycle* and *Garden* (3 views). The geometry prior improves the rendering quality but is not the key to sparse-view reconstruction.

References

- [1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1, 2, 3
- [2] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds, 2024. 1, 2
- [3] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2
- [4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1
- [5] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1
- [6] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 1, 2, 3
- [7] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1
- [8] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *CVPR*, 2024. 1, 2, 3
- [9] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 2
- [10] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 1
- [11] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics*, 2024. 1
- [12] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 2, 3
- [13] Zheng Zhang, Wenbo Hu, Yixing Lao, Tong He, and Hengshuang Zhao. Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. *arXiv preprint arXiv:2403.15530*, 2024. 1
- [14] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *ECCV*, 2024. 2

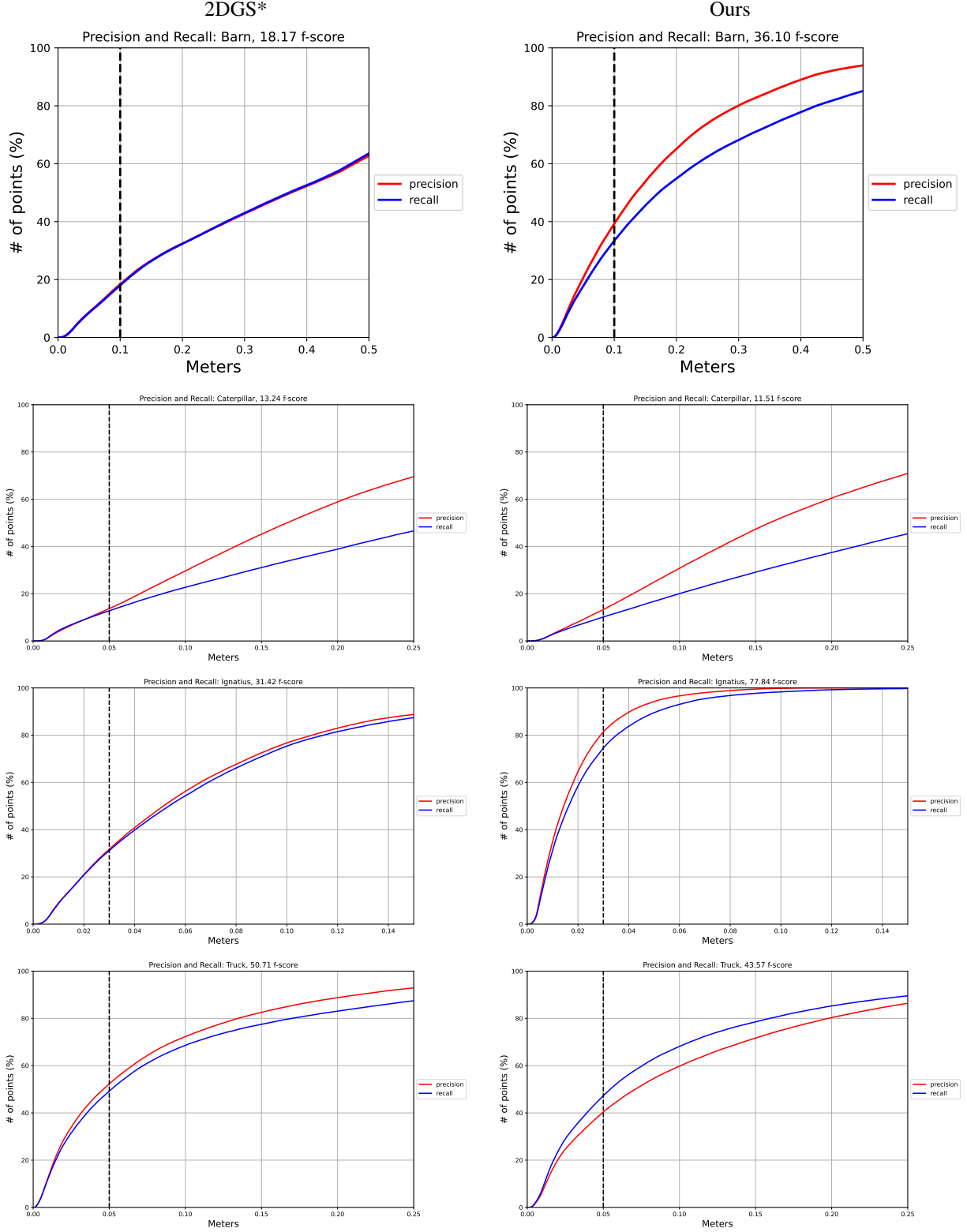


Figure D. **The Precision and Recall Curves of F1-score Comparisons.** We show detailed precision and recall curves of F1-score comparisons between 2DGS and our method on Barn, Caterpillar, Ignatius, and Truck, given 4 views. The black dotted line in each subfigure denotes the error threshold.