Generative Sparse-View Gaussian Splatting

Hanyang Kong Xingyi Yang Xinchao Wang*
National University of Singapore
{hanyang.k, xyang}@u.nus.edu, xinchao@nus.edu.sg

1. Implementation Details

We build our pipeline based on the official 3DGS [2] and SpacetimeGS [4] for sparse-view 3D and 4D scene reconstruction, respectively. We train the overall pipeline with $10,000$ iterations for all datasets. We apply pre-trained DPT [7] to estimate monocular depth for all training and pseudo views. The initial point clouds for each scene are reconstructed by Structure-from-Motion (SfM) with the sparse training views. We apply Stable Diffusion XL (SDXL) as the base stable diffusion model and the depth adapter [6] as the depth-conditioned module. The parameters of SDXL and the depth adapter are frozen during the training procedure.

2. More Experiment Results

2.1. Effects of Training Views

Table 1. **Quantitative results on the effects of training views.** We conduct experiments on the LLFF [5] dataset with different numbers of training views. Our GS-GS outperforms other baselines across all experimental settings. PSNR scores are reported.

	3 Views	6 Views	9 Views
Mip-NeRF	16.11	22.91	24.88
3DGS	17.43	22.87	24.65
DietNeRF	14.94	21.75	24.28
RegNeRF	19.08	23.10	24.86
FreeNeRF	19.63	23.73	25.13
SparseNeRF	19.86	23.64	24.97
DNGaussian	19.12	22.18	23.17
FSGS	20.31	23.64	25.89
Ours	24.82	26.87	28.12

We evaluated our GS-GS method on the LLFF [5] dataset using 3, 6, and 9 training views, as shown in Tab. 1, and found that it consistently outperforms all baseline methods across all settings in terms of PSNR. With only 3 views, our method achieves a PSNR of 24.82, significantly higher than the second-best method, FSGS, which attains 20.31, marking an improvement of over 4.5 dB. This substantial gain demonstrates our method's superior ability to reconstruct high-quality scenes from extremely sparse data. As the number of views increases to 6, our method's PSNR rises to 26.87, surpassing FSGS's 23.64 by more than 3 dB, indicating effective utilization of additional views. At 9 views, our method achieves a PSNR of 28.12, exceeding the next best method, FSGS, which records 25.89, by over 2 dB. These consistent performance improvements highlight the effectiveness of our hallucination strategy, which generates reliable images from pseudo-novel views to enhance supervision during optimization. By boosting the training of the 3D/4D Gaussian Splatting (GS) model, our approach leads to reconstructions with richer structural details and mitigates the challenges posed by limited training views, underscoring the robustness and effectiveness of our method in sparse-view scene reconstruction.

2.2. Effects of Geometry-aware Diffusion Fine-tuning

In this section, we demonstrate the effectiveness of the feature-level geometry correspondence of the diffusion model. DIFT [8] has demonstrated the efficiency of matching pixel locations in two images by the pre-trained diffusion model. We take the horn scene on the LLFF [5] dataset as an example. Given a rendered image at pesudo view with low quality, we select four different points on the rendered image, and query the corresponding pixel from the ground truth image at the pseudo view and the ground truth image at another camera view. The visualized results are shown in Fig. 1. Though the rendered images are low-quality with artifacts, the diffusion feature could establish semantic correspondence across images from various camera views.

2.3. More Visualization Results

In this part, we illustrate more qualitative results on the LLFF dataset [5], the MipNeRF-360 dataset [1], and the Neural Video dataset [3]. The results of the first three datasets (static scenes) are illustrated in Fig. 2, and the visualization on the Neural Video dataset (dynamic scenes) is shown in the video supplementary material.

*Corresponding author.

Figure 1. **The effectiveness of the feature-level geometry correspondence of diffusion models.** We select four red source points on the low-quality rendered image at pseudo-view 1 and identify the corresponding pixel-level keypoints in the ground truth image at view 1 and in another ground truth image at view 2 by the pre-trained diffusion model.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 1

[2] 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

[3] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 1

[4] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8508–8520, 2024. 1

[5] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 1, 3

[6] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[7] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1

[8] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1

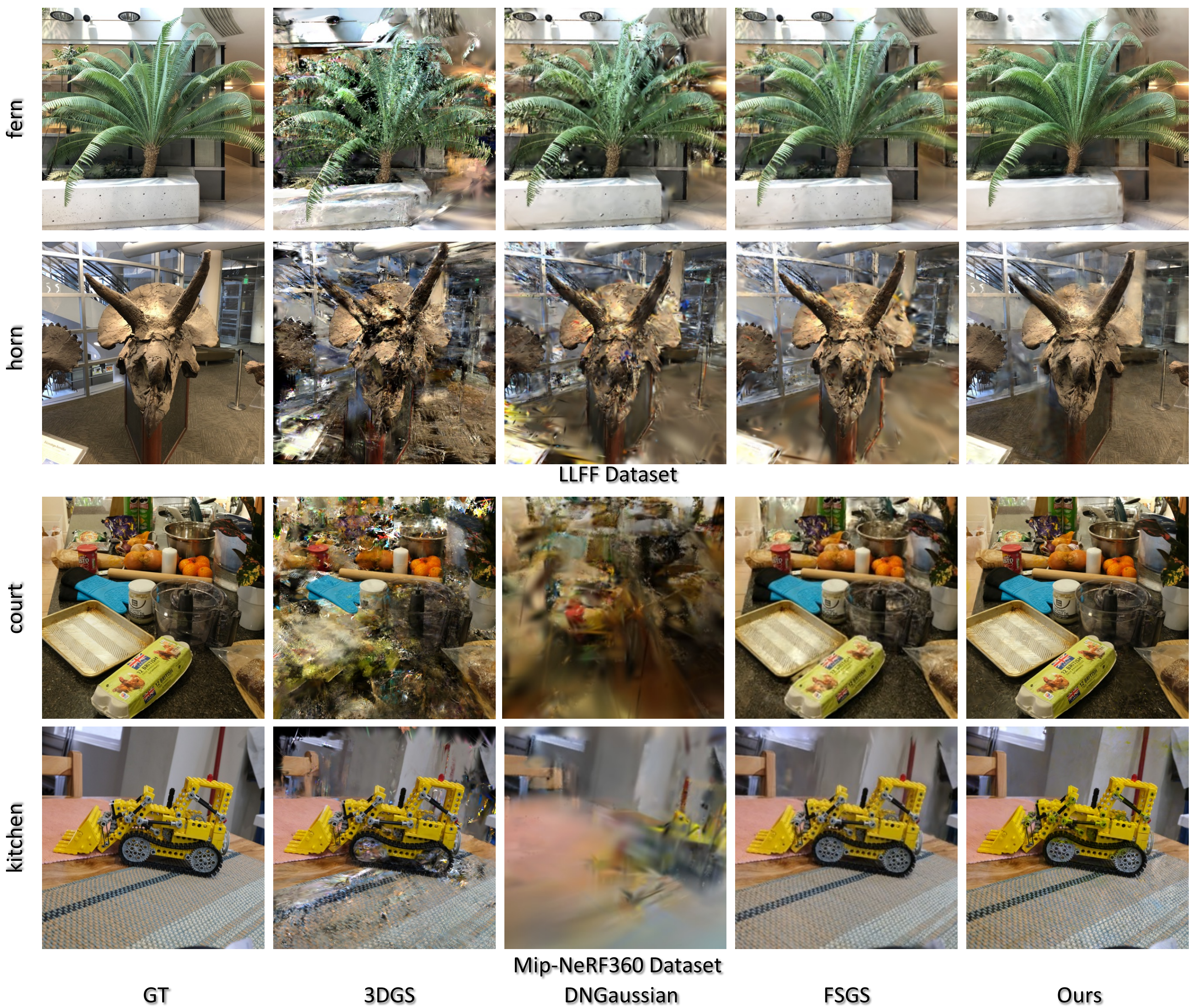

Figure 2. **The visualization results on the LLFF [5] and Mip-NeRF360 datasets.** Our method produces detailed foreground geometry and renders high-quality novel views with sparse camera views.