Sparfels: Fast Reconstruction from Sparse Unposed Imagery
– Supplementary material –

Shubhendu Jena*, Amine Ouasfi*, Mae Younes, Adnane Boukhayma
Inria, Univ. Rennes, CNRS, IRISA

1. Evaluation of Novel View Synthesis and Camera pose estimation

In this section, we provide qualitative and quantitative comparisons on the Tanks and Temples [11], MipNeRF360 [2] datasets and MVImgNet [19] datasets for both novel view synthesis and camera pose estimation metrics. Tab 1 presents quantitative results for these experiments on MipNeRF360 [2] and MVImgNet [19] with qualitative results presented in Fig 1 and Fig 2. For Tanks and Temples [11], quantitative and qualitative results are reported in Tab 2 and Fig 3 respectively. We observe that NoPe-NeRF [3] and NeRF-mm [16] suffer markedly in their novel view performance and camera pose estimation metrics. Being implicit, volumetric rendering methods, they also suffer from slow training and inference times. CF-3DGS [8] also encounters artifacts when rendering from novel viewpoints, stemming from its complex optimization pipeline and erroneous pose estimations. InstantSplat [5, 6] variants provide good performance, but still lag behind our method in most metrics, particularly in the challenging 3-view setting. For the Tanks and Temples comparison in Tab 2, Fig 3, we also outperform SPARF [14] by a sizeable margin on all metrics, while requiring order of magnitudes less training and inference time, since it takes around 10 hours to train on a single scene and needs more than a minute to render a single image during inference, owing to its volumetric rendering framework. Our method significantly outperforms all baselines on various datasets in terms of SSIM, LPIPS (novel view synthesis metrics) and ATE (camera pose estimation metric), demonstrating its robustness to complex scenes with challenging lighting conditions.

Method	SSIM (MVImgNet)			LPIPS (MVImgNet)			ATE (MVImgNet) ↓			MipNeRF360 (12 Training Views)			
	3-view	6-view	12-view	3-view	6-view	12-view	3-view	6-view	12-view	SSIM	PSNR	LPIPS	ATE ↓
NoPe-NeRF [3]	0.4326	0.4329	0.4686	0.6168	0.6614	0.6257	0.2780	0.1740	0.1493	0.3580	16.16	0.6867	0.2374
CF-3DGS [8]	0.3414	0.3544	0.3655	0.4520	0.4326	0.4492	0.1593	0.1981	0.1243	0.2443	13.17	0.6098	0.2263
NeRF-mm [16]	0.3752	0.3685	0.3718	0.6421	0.6252	0.6020	0.2721	0.2376	0.1529	0.2003	11.53	0.7238	0.2401
Instantsplat-S [5, 6]	0.5489	0.6835	0.7050	0.3941	0.2980	0.3033	**0.0184**	0.0259	0.0165	0.4647	17.68	0.5027	0.2161
Instantsplat-XL [5, 6]	0.5628	0.6933	0.7321	0.3688	0.2611	0.2421	**0.0184**	0.0259	**0.0164**	0.4398	17.23	0.4486	0.2162
Ours	**0.8313**	**0.8801**	**0.9008**	**0.2215**	**0.1658**	**0.1410**	0.0273	**0.0244**	0.0172	**0.8168**	**26.21**	**0.2199**	**0.2067**

Table 1. NVS performance comparison of different methods on MVImgNet and MipNeRF360

*Equal contribution.

Method	SSIM↑			LPIPS↓			ATE↓		
	3-view	6-view	12-view	3-view	6-view	12-view	3-view	6-view	12-view
COLMAP + 3DGS [10]	0.3755	0.5917	0.7163	0.5130	0.3433	0.2505	-	-	-
COLMAP + FSGS [21]	0.5701	0.7752	0.8479	0.3465	0.1927	0.1477	-	-	-
NoPe-NeRF [3]	0.4570	0.5067	0.6096	0.6168	0.5780	0.5067	0.2828	0.1431	0.1029
CF-3DGS [8]	0.4066	0.4690	0.5077	0.4520	0.4219	0.4189	0.1937	0.1572	0.1031
NeRF-mm [16]	0.4019	0.4308	0.4677	0.6421	0.6252	0.6020	0.2721	0.2329	0.1529
SPARF [14]	0.5751	0.6731	0.5708	0.4021	0.3275	0.4310	0.0568	0.0554	0.0385
Instantsplat-S [6]	0.7624	0.8300	0.8413	0.1844	0.1579	0.1654	0.0191	0.0172	0.0110
Instantsplat-XL [6]	0.7615	0.8453	0.8785	0.1634	**0.1173**	0.1068	0.0189	**0.0164**	0.0101
Ours	**0.8752**	**0.9020**	**0.9180**	**0.1623**	0.1283	**0.1050**	**0.0150**	0.0174	**0.0078**

Table 2. Performance comparison of different methods across SSIM, LPIPS, and ATE metrics for 3-view, 6-view, and 12-view settings on the Tanks and Temples dataset.

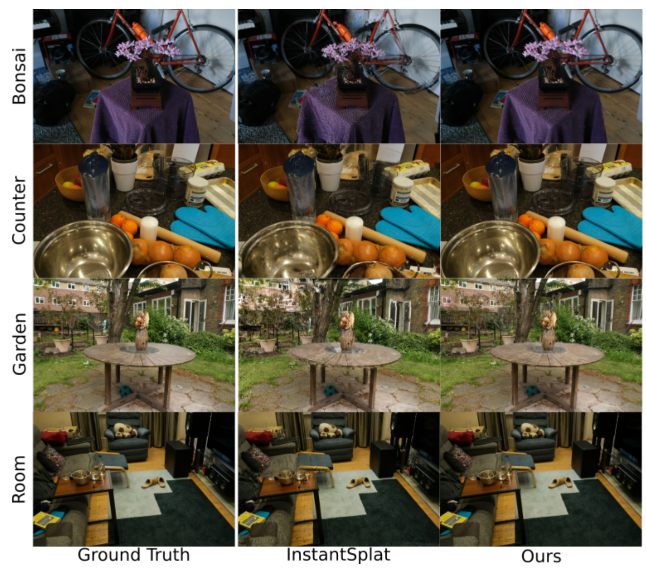

Figure 1. **Qualitative comparison of novel view synthesis on MipNeRF360 dataset from 12 input images.**

2. Additional qualitative comparison on 3D reconstruction

We also provide a qualitative comparison in Fig. 4 to SpaRP [17] on the DTU [1] dataset, a recent method that leverages 2D diffusion models for efficient 3D reconstruction and pose estimation from unposed sparse-view images. For comparison using 3 input images, our method achieves mesh reconstructions with greater fidelity to the input images, as seen in the comparisons. We also provide video

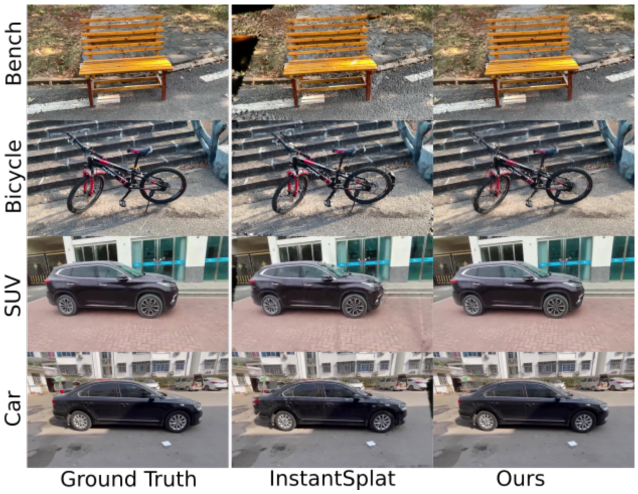

Figure 2. **Qualitative comparison of novel view synthesis on MVImgNet dataset from 3 input images.**

Figure 3. **Qualitative comparison of novel view synthesis on Tanks and Temples dataset from 3 input images.**

results depicting our reconstructions and novel view results on the DTU [1] and BlendedMVS [18] datasets and their comparison to other methods.

3. Color variance plot

We plot the average color variance over optimization iterations (Fig. 5) for models w/ and w/o variance loss. Models with the loss activated effectively maintain lower color variance consistently, which aligns with our goal of encouraging stable, low-uncertainty renderings. This supports the effectiveness of the proposed loss in guiding convergence toward robust geometry.

4. Alternative priors

This example (Fig. 6) demonstrates that initializing our framework with VGGT [15], a recent state-of-the-art feed-forward method that avoids the global optimization step of MASt3R [12] also produces successful results. This highlights the modularity of our approach and its compatibility with different geometric priors.

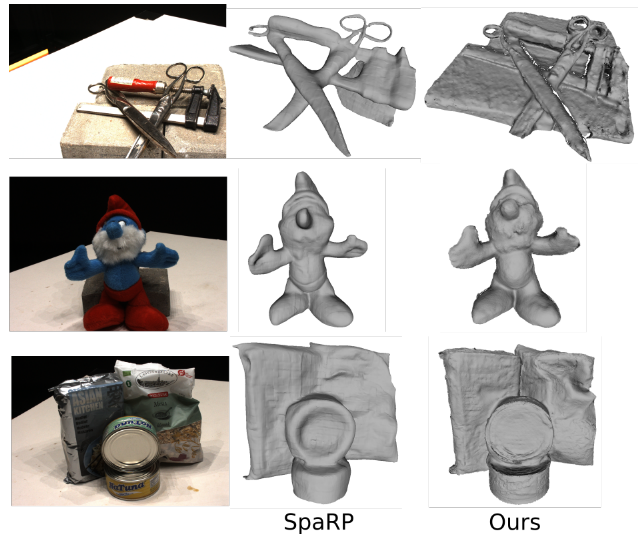

Figure 4. **Qualitative comparison with SpaRP [17] on DTU dataset from 3 input images.**

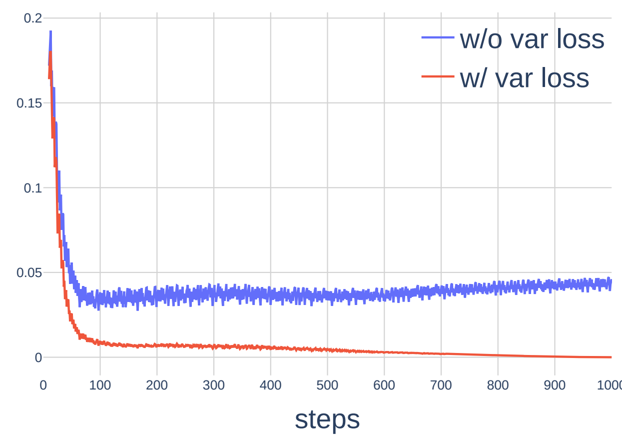

Figure 5. **Color variance.** The variance loss keeps color variance lower, encouraging stable, robust convergence.

5. Variance loss motivation

Our goal is to hedge against epistemic uncertainty in geometry estimation inherent to the unposed surface reconstruction problem. Under sparse-view supervision, the rendering objective admits many geometries that fit the training images but generalizes poorly (see Sec.3 in [20]). This issue is exacerbated when camera poses are optimized during training, introducing additional noise into the supervision. In Gaussian Splatting, scene geometry is encoded through splat parameters defining the 3D density. Among the many plausible geometries, we seek to bias the model toward those that remain predictive even under small per-

Figure 6. **Alternative prior.** VGGT initialization shows successful results (3 input images), demonstrating our framework's modularity.

turbations, *i.e*. robust solutions less sensitive to noisy supervision signals. To formalize this, we minimize the worst-case deviation in rendered color under perturbations to the geometric density field (Eq. 8), yielding a variance regularization loss (Eq. 10) that penalizes color variance along rays. From a learning-theoretic perspective, this can be interpreted as seeking flat minima [9] in the space of densities, an idea supported by arguments from both statistical and deep learning viewpoints, and shown to be effective across a range of machine learning applications [4, 7, 9, 13]. Empirically, this leads to more stable and consistent reconstructions from sparse views (Fig. 5, Tab. 4).

6. Prior failures

The example below illustrates a typical scenario where MASt3R's [12] feed-forward geometry prediction struggles: reconstruction from only 6 images without known camera poses. Challenging regions such as texture-less surfaces, highly reflective materials (e.g., glass doors and shiny faucets), and thin structures often lead to noisy or incomplete results. In contrast, our method recovers plausible geometry in these cases thanks to robust test-time optimization, which refines both the pose and the reconstructed shape despite imperfect initializations.

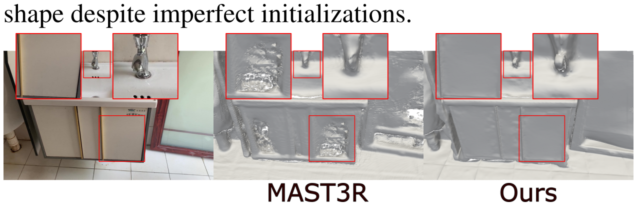

Figure 7. **Robustness to MASt3R failure (3 input images).** Our method recovers geometry where MASt3R struggles.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 1, 2

[2] 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

[3] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 1

[4] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019. 3

[5] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2(3):4, 2024. 1

[6] Zhiwen Fan, Kairun Wen, Wenyan Cong, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Sparse-view sfm-free gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024. 1

[7] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 3

[8] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. *arXiv preprint arXiv:2312.07504*, 2023. 1

[9] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 3

[10] 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

[11] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 1

[12] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 3

[13] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 3

[14] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 1

[15] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[16] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. 2021. 1

[17] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pages 143–163. Springer, 2024. 1, 2

[18] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 2

[19] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 1

[20] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[21] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00451*, 2023. 1