Dense 3D Gaussian Splatting Initialization for Sparse Image Data

S. Seibt[1*], T. Chang[1*], B. von Rymon Lipinski[1], and M. E. Latoschik[2]

[1] Game Tech Lab, Faculty of Computer Science, Nuremberg Institute of Technology, Nuremberg, Germany
[2] Human-Computer Interaction Group, Institute of Computer Science, University of Wuerzburg, Wuerzburg, Germany

Abstract

This paper presents advancements in novel-view synthesis with 3D Gaussian Splatting (3DGS) using a dense and accurate SfM point cloud initialization approach. We address the challenge of achieving photorealistic renderings from sparse image data, where basic 3DGS training may result in suboptimal convergence, thus leading to visual artifacts. The proposed method enhances precision and density of initially reconstructed point clouds by refining 3D positions and extrapolating additional points, even for difficult image regions, e.g. with repeating patterns and suboptimal visual coverage. Our contributions focus on improving "Dense Feature Matching for Structure-from-Motion" (DFM4SfM) based on a homographic decomposition of the image space to support 3DGS training: First, a grid-based feature detection method is introduced for DFM4SfM to ensure a well-distributed 3D Gaussian initialization uniformly over all depth planes. Second, the SfM feature matching is complemented by a geometric plausibility check, priming the homography estimation and thereby improving the initial placement of 3D Gaussians. Experimental results on the NeRF-LLFF dataset demonstrate that this approach achieves superior qualitative and quantitative results, even for fewer views, and the potential for a significantly accelerated 3DGS training with faster convergence.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction; Point-based models;*** *Rendering;*

1. Introduction

3D Gaussian Splatting (3DGS) [KKLD23] emerged as a promising alternative to Neural Radiance Field (NeRF) techniques for novel-view synthesis by offering computational efficient training and accelerated photorealistic rendering using a splatting-based scene representation. Yet, its effectiveness is contingent on the quality and quantity of the initial Structure-from-Motion (SfM) point cloud. This is a limitation particularly for scenarios with a limited number of images, where sparse and inaccurate point clouds can lead to suboptimal training convergence and result in visual artifacts.

Seibt et al. [SVRLCL23] introduced a dense and accurate feature matching approach (DFM4SfM) to improve conventional SfM: It is based on a homographic decomposition of the image space through iterative rematching, which improves precision and density of the point cloud reconstruction using positional refinement of matched feature points, identification of critical non-planar image areas (e.g. due to parallaxes), and extrapolation of additional matches in regions difficult for one-shot matching. DFM4SfM also comprises a multi-view strategy for feature cluster refinement across different views by leveraging a connectivity graph, thus enhancing pose estimation and 3D reconstruction accuracy. Additionally, a global matching extrapolation method is employed (considering multiple neighboring views) to increase matching recall

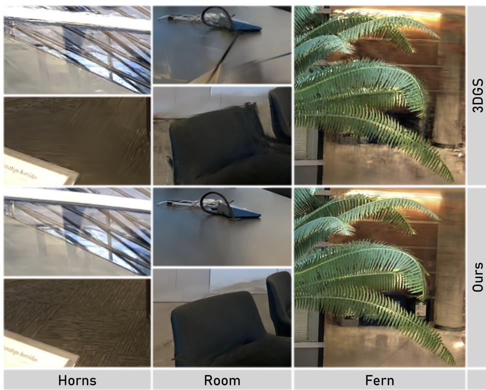

Figure 1: *Visual comparisons on sparse image data from NeRF-LLFF: 3DGS without (top) and with the proposed method (bottom).*

and reconstruct even denser 3D structures. The main contributions of this work are the following DFM4SfM extensions for improving 3DGS rendering: (a) Grid-based feature detection for a well-distributed point cloud generation that captures both foreground and background details and initializes Gaussians across multiple depth planes. This approach assures a more uniform and faster converging splat initialization, also considering image areas with visually less significant features. (b) A multi-homography decomposition strategy supplemented by the estimation of geometrically

* Equal contribution to this work.

DOI: 10.2312/egp.20241038

plausible feature matches, enhancing homography estimation for wide-baseline image pairs with complex visual structures or multiple depths. This novel strategy minimizes sparse and potentially incorrect splat initializations of traditional SfM approaches, typically caused by fundamental matrix degeneration in such input images.

Related Work To enable 3DGS for scenes with sparse image data, [ZFJW23] proposed a proximity-based and depth-supported re-distribution of Gaussians to prevent overfitting and visual artifacts. Since their method depends on an additional unpooling process and requires a pre-trained monocular depth estimation model, it is not able to accelerate training convergence.

2. Methods

To support a uniform feature detection – and thus a well-distributed point cloud for 3DGS – the image is initially split into an $w \times h$ grid. For each grid cell, a coarse-to-fine detection approach is employed: Given an approximate maximum feature count f_{max}, the corresponding detection threshold is adjusted iteratively until $\frac{f_{max}}{w*h}$ is reached. If the number of detected features saturates during iteration, then this sub-process terminates early for that grid cell (typical for visually homogeneous image areas, like the sky). Based on our experiments, we propose a 16×16 grid for full HD images, as higher resolutions yielded no notable improvements w.r.t. feature detection and uniformity, respectively. Furthermore, DFM4SfM generates a high-density point cloud by successively rematching detected features over multiple iterations. To reduce the risk of mismatches in later iterations due to increasing matching distances, the basic feature descriptor distances are complemented with positional encodings. These are used to derive geometric relationships of matches w.r.t. previous rematching iterations (with potentially lower matching distances) in order to "stabilize" feature correspondences as follows: In each rematching iteration n, let p_s be an unmatched feature point in the source image, and (N_s, N_t) all the associated Delaunay-adjacent feature points that have already matched. Let $P_t = \{p_t^1, p_t^2, \ldots\}$ be a list of kNN matches for $k = n$ w.r.t. p_s in the target image, ordered by their basic feature descriptor distances. Now, every point in P_t that is located within the convex hull of N_t is then ranked higher (w.r.t. to the sorting order) than points outside of the hull, since geometric proximity of adjacent matching candidates correlates with visual correspondence. Then, all "top-ranked" pairs, i.e. $p_i^1 = (p_{i,s}, p_{i,t}^1)$, are used for DFM4SfM's homographic estimation to compute a consensus set C. Finally, each resulting pair that is detected as an outlier, i.e. $p_i^1 \notin C$, is replaced by the second-best consenting pair, thus with the index $\min\{t \in [2..k] \mid p_i^t \in C\}$, if available. Using this geometric plausibility check for DFM4SfM's homographic decomposition increases the number of matching inliers, resulting in a more robust multi-plane recovery in 3D space, and thus a denser and preciser point cloud for 3DGS initialization.

3. Experiments

Benchmarks were performed for the NeRF-LLFF dataset [MSOC*19] on an Intel i9 14900KF CPU, 64GB RAM, and a GeForce RTX 4060 GPU. Ca. 10.000 keypoints were detected per image, while COLMAP was used for the default 3D reconstruction. Its built-in feature matching was fine-tuned and used for comparative evaluation with the DFM4SfM-based method. On average, using DFM4SfM takes two times longer, but reconstructs 213% more 3D points. In Table 1, quantitative results for 30.000 3DGS training iterations are shown. Table 2 shows further results for the *Fern* NeRF-LLFF scene with varying numbers of training images and iterations. Using DFM4SfM results in a significantly increased performance across all metrics (on average: SSIM +11.7%, PSNR +15.0%, LPIPS −35.0%), even for considerably sparser configurations, e.g. only nine images or 1.000 iterations (resulting also in significantly reduced training times). Without the proposed 3DGS enhancements, performance increases with "pure" DFM4SfM are less (SSIM +7.8%, PSNR +12.1%, LPIPS −20.0%).

	Scene	SSIM ↑		PSNR ↑		LPIPS ↓		Time (mm:ss) ↓	
		3DGS	Ours	3DGS	Ours	3DGS	Ours	3DGS	Ours
NeRF-LLFF	Fern	0.68	**0.83**	21.16	**24.40**	0.25	**0.16**	36:58	**34:47**
	Flower	0.71	**0.91**	22.06	**29.36**	0.26	**0.09**	29:47	**19:29**
	Fortress	0.87	**0.91**	27.84	**30.97**	0.13	**0.09**	30:30	**28:47**
	Horns	0.80	**0.92**	24.21	**28.41**	0.18	**0.10**	35:44	**29:36**
	Leaves	0.64	**0.70**	18.68	**20.31**	0.31	**0.26**	35:20	**33:43**
	Orchids	0.62	**0.74**	18.72	**20.92**	0.22	**0.16**	37:56	**34:01**
	Room	0.91	**0.96**	27.70	**31.76**	0.12	**0.08**	19:24	**17:23**
	Trex	0.89	**0.93**	24.73	**26.82**	0.14	**0.09**	30:59	**24:05**
Mean		0.77	**0.86**	23.14	**26.62**	0.20	**0.13**	32:05	**27:44**

Table 1: *Results for 3D Gaussian Splatting (3DGS) and 3DGS initialized with DFM4SfM (Ours) on NeRF-LLFF [MSOC*19].*

	N	Iters	SSIM ↑		PSNR ↑		LPIPS ↓		Time (mm:ss) ↓	
			3DGS	Ours	3DGS	Ours	3DGS	Ours	3DGS	Ours
Fern (NeRF-LLFF)	16	30000	0.68	**0.83**	21.16	**24.40**	0.25	**0.16**	36:58	**34:47**
	12	30000	0.54	**0.75**	18.13	**22.60**	0.32	**0.21**	27:16	**21:55**
	9	30000	0.52	**0.69**	17.58	**20.33**	0.33	**0.25**	25:07	**20:35**
	6	30000	0.48	**0.65**	16.60	**18.86**	0.38	**0.27**	21:23	**19:19**
	3	30000	0.33	**0.47**	13.38	**14.75**	0.52	**0.44**	19:37	**18:34**
	16	*15000*	0.69	**0.82**	21.34	**24.17**	0.24	**0.16**	15:54	**15:43**
	16	*5000*	0.72	**0.84**	22.07	**24.50**	0.24	**0.17**	**03:43**	04:15
	16	*1000*	0.57	**0.76**	19.59	**23.66**	0.51	**0.26**	**00:37**	00:45
	16	*500*	0.50	**0.66**	17.68	**21.68**	0.61	**0.41**	**00:19**	00:25
	16	*100*	0.49	**0.53**	15.91	**19.01**	0.59	**0.57**	**00:04**	00:06

Table 2: *Results with varying number of images (N) and 3DGS training iterations (Iters) for the scene Fern [MSOC*19]. Underlined: Ours surpassing 3DGS's best with lower N and Iters.*

4. Conclusion

We propose a well-distributed feature detection and geometrically primed feature matching method for a DFM4SfM-optimized 3D Gaussian Splatting initialization. It yields superior qualitative and quantitative rendering results for scenes with sparse image data, indicating potential for significantly accelerated 3DGS training.

References

[KKLD23] Kerbl B., Kopanas G., Leimkuehler T., Drettakis G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. on Graph.* (2023). 1

[MSOC*19] Mildenhall B., Srinivasan P. P., Ortiz-Cayon R., Kalantari N. K., Ramamoorthi R., Ng R., Kar A.: Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. on Graph.* (2019). 2

[SVRLCL23] Seibt S., Von Rymon Lipinski B., Chang T., Latoschik M. E.: Dfm4sfm - dense feature matching for structure from motion. In *IEEE Inter. Conf. on Img. Proc. Wksp.* (2023). 1

[ZFJW23] Zhu Z., Fan Z., Jiang Y., Wang Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting, 2023. 2