Optimizing 3D Scene Reconstruction: Integrating DUSt3R with 3D Gaussian Splatting

Authors: Tirth Patel, Daniel Allex

1 Abstract

Multi-view stereo reconstruction has traditionally depended on known camera parameters and carefully calibrated systems to achieve high-quality three-dimensional models. In this work, we present Dust3r Splatting, a method that leverages DUSt3R's intrinsic-free multi-view reconstruction pipeline and integrates it with the efficiency and real-time rendering capabilities of Gaussian Splatting. By decoupling reconstruction from explicit camera calibration and employing a pointmap regression paradigm, Dust3r Splatting supports monocular and binocular depth estimation and camera pose inference under a single unified framework. We show how DUSt3R's outputs can be seamlessly transformed into standard sparse representations acceptable to Gaussian Splatting pipelines, enabling state-of-the-art real-time scene rendering without the need for camera intrinsics. Experimental evaluation reveals that our method handles previously unseen views with reasonable robustness compared to traditional methods relying on known camera intrinsics and structured pipelines such as COLMAP. While a comprehensive quantitative analysis remains future work, our initial results highlight the promise of this unified approach for simplifying and accelerating multi-view reconstruction and novel-view synthesis.

2 Introduction

Multi-view stereo reconstruction involves recovering the underlying three-dimensional structure of a scene from a collection of images. Classical approaches rely on known camera parameters and a carefully orchestrated pipeline, including camera calibration, pose estimation, and multi-view dense reconstruction. However, the demand for flexible and scalable 3D vision solutions has driven research toward methods that do not depend on explicit camera intrinsics or meticulously controlled conditions. Such a shift would enable rapid scene understanding from arbitrary image collections, making applications like robotics navigation, augmented reality, and 3D modeling from internet photo sets more accessible.

Recent advances in learning-based and neural implicit reconstruction methods have substantially improved the quality and robustness of novel-view synthesis, depth estimation, and camera pose inference. Techniques like Neural Radiance Fields (NeRF) demonstrate remarkable results in generating photorealistic novel views given known camera parameters. However, these methods often incur heavy computational costs and do not intrinsically solve for unknown camera intrinsics. On the other hand, DUSt3R introduces a transformer-based pipeline that treats pairwise reconstructions as pointmap regressions, bypassing the need for explicit projective camera models. By providing depth maps, pixel correspondences, and even camera poses without requiring calibrated intrinsics, DUSt3R simplifies the geometry pipeline and can unify monocular and binocular methods under a single framework.

Meanwhile, Gaussian Splatting has emerged as a method that capitalizes on efficient point-based scene representations. By positioning and shaping 3D Gaussians and employing fast visibility-aware rendering

algorithms, Gaussian Splatting can achieve high-quality, real-time novel-view synthesis without the heavy inference and memory costs associated with neural radiance fields. However, Gaussian Splatting pipelines traditionally rely on established camera intrinsics and poses to initialize their scene representations – factors that can be challenging in less controlled scenarios.

In this paper, we propose Dust3r Splatting, a method that marries the advantages of DUSt3R's intrinsic-free reconstruction approach with the speed and quality of Gaussian Splatting. By converting DUSt3R's outputs into a COLMAP-like format through a lightweight intermediate step, we provide Gaussian Splatting with the necessary initialization without relying on known camera intrinsics. Our experiments suggest that this integrated approach not only streamlines the pipeline but also better generalizes to unseen viewpoints. The contributions of this work are threefold: (1) We introduce a unified framework that couples intrinsic-free multi-view reconstruction with state-of-the-art point-based scene rendering. (2) We show how to transform DUSt3R's outputs into a format suitable for Gaussian Splatting, enabling scene rendering without explicit camera intrinsics. (3) We provide initial qualitative comparisons indicating improved robustness and highlight avenues for future quantitative analysis.

3 Related Work

3.1 Neural Radiance Fields (NeRF):

NeRF [3] pioneered the use of continuous, implicit neural representations for novel-view synthesis. By encoding a scene as a function mapping 3D coordinates and viewing directions to radiance and density, NeRF enables photorealistic rendering of new viewpoints from sparse image collections. While NeRF and its subsequent extensions produce impressive results, they rely on known camera parameters and often require lengthy training times.

3.2 Neural Volumes and Related Implicit Representations:

Neural Volumes [2] introduced a voxel-based representation learned directly from images. By performing differentiable ray-marching through the learned volumes, it brought forward an end-to-end framework that can handle dynamic scenes. Although Neural Volumes can deal with changes in appearance and geometry, it also requires known camera parameters and remains computationally expensive.

3.3 DUSt3R and Transformer-Based Reconstructions:

DUSt3R [4] offers a fundamental shift by not relying on explicit camera intrinsics. Leveraging transformer architectures and pre-trained models for feature extraction, DUSt3R treats pairwise reconstruction as a pointmap regression task. By doing so, it inherently unifies monocular and binocular 3D estimation and can provide relative poses and dense reconstructions from arbitrary image collections. This approach significantly reduces the complexity of traditional geometric pipelines, which often depend on well-established epipolar geometry or multi-view constraints that assume known intrinsics.

3.4 Gaussian Splatting:

Gaussian Splatting [1] proposes a novel scene representation using anisotropic 3D Gaussians. Starting from sparse camera-calibrated points, it optimizes both geometry and rendering in a process that supports real-time, high-resolution (1080p, ≥100 fps) novel-view synthesis. The method refines anisotropic Gaussian distributions during training to achieve compact and accurate scene representations. While Gaussian Splatting excels at speed and quality, it presupposes initial camera parameters for the scene representation, making it less suitable in scenarios where camera intrinsics are unknown.

3.5 InstantSplat and Intrinsic-Free Approaches:

InstantSplat [6] aims to rapidly generate point-based scene representations for sparse-view novel-view synthesis without the traditional Structure-from-Motion (SfM) overhead. Although it speeds up the scene-to-representation pipeline, InstantSplat still depends on initial geometric cues or partial intrinsic knowledge, making it not fully intrinsic-free.

3.6 CroCo and Cross-View Complements:

Methods like CroCo [5] introduce cross-view completion tasks to improve geometric reasoning. CroCo leverages masked image modeling across multiple viewpoints, pushing the network to learn spatial correspondences and not just semantics. Notably, it can be used to improve binocular tasks and helps with pose estimation.

3.7 Summary:

While NeRF and other implicit methods produce high-quality renderings, they depend on known camera parameters or offline calibration. DUSt3R breaks away from this dependency, providing a foundation for intrinsic-free reconstruction. Gaussian Splatting provides efficient, high-quality rendering, but previously required known camera intrinsics. Dust3r Splatting bridges this gap by combining DUSt3R's intrinsic-free pipeline with Gaussian Splatting's real-time rendering capabilities, resulting in a unified approach that does not rely on pre-calibrated camera parameters.

4 Methods

Our method, Dust3r Splatting, integrates DUSt3R's intrinsic-free reconstruction capabilities with the efficient Gaussian Splatting rendering pipeline. By doing so, we remove the requirement of having known camera intrinsics – a typical bottleneck in multi-view stereo – while still achieving high-quality novel-view synthesis.

4.1 Overview

DUSt3R starts by processing image pairs through a transformer-based architecture to regress pointmaps that encapsulate depth, pixel correspondences, and relative camera poses. Instead of modeling the geometry strictly through known projective camera models, DUSt3R relies on learned features and multi-view constraints to infer depth and pose directly from the images. This leads to a unified setup where monocular and binocular 3D estimation share the same formulation, simplifying the pipeline and allowing it to handle arbitrary image sets with minimal assumptions.

We take DUSt3R's outputs – essentially point correspondences and estimated camera positions – and transform these results into a format interpretable by downstream pipelines. Using pycolmap, a Python interface to the well-established COLMAP library, we convert DUSt3R outputs into a structure akin to COLMAP reconstructions. Although DUSt3R does not require known intrinsics, we populate the COLMAP-style data structure with approximate intrinsics or identity placeholders. This step creates a bridge: while the underlying reconstruction does not depend on intrinsics, the final representation now matches what Gaussian Splatting expects.

4.2 Intuition for Improved Performance

The key intuition is that DUSt3R's intrinsic-free approach yields robust pairwise relationships between images, which translates into stable scene geometries and camera poses. Traditional pipelines that rely on known camera models can become brittle or fail entirely if the intrinsic parameters are imperfect or if the images come from diverse, uncalibrated sources. DUSt3R's learned approach sidesteps these issues, providing a stable geometric initialization. When fed into Gaussian Splatting, this stable initialization supports the optimization of anisotropic Gaussians, resulting in accurate, compact representations. Thus, the pipeline becomes more versatile and can handle previously unseen views better. In effect, we bring the robustness and generality of DUSt3R's front-end into the efficient rendering domain of Gaussian Splatting.

4.3 Detailed Model and Algorithm Description

Step 1: DUSt3R Reconstruction:
Given a set of images, DUSt3R predicts point clouds and camera intrinsics. Through the transformer-based architecture and pretrained encoders, DUSt3R effectively infers these relationships without prior knowledge of camera intrinsics.

Step 2: Global Alignment and Pose Fusion:
From DUSt3R's pairwise models, we employ a global alignment strategy to merge these partial reconstructions into a common reference frame. DUSt3R's outputs include relative pose information, allowing us to establish a consistent global coordinate system. The merged point cloud and camera positions are then represented in a format that mimics the sparse reconstruction output of a traditional SfM pipeline.

Step 3: Conversion to COLMAP Format via pycolmap:
We then use pycolmap, a Python interface for COLMAP's data structures, to store the merged reconstruction. Even though true camera intrinsics are unknown, we assign nominal placeholders or use a standard pinhole model with generic parameters. Since Gaussian Splatting only needs an initialization, this step provides it a familiar interface, bridging the gap between DUSt3R's intrinsic-free reconstruction and Gaussian Splatting's expected input.

Step 4: Gaussian Splatting Initialization:
Gaussian Splatting begins from a set of 3D points and corresponding camera parameters. In our pipeline, these come from the created COLMAP output. Gaussian Splatting then converts these points into 3D

Gaussians, optimizing their shape and position to accurately represent the scene. Because DUSt3R has already established stable geometry and relative poses, Gaussian Splatting's optimization converges efficiently.

Step 5: Real-Time Novel-View Synthesis:
With the optimized Gaussian representation, we can generate novel views at real-time rates. The resulting pipeline is completely free from any earlier camera calibration, making it more robust to diverse and unstructured image collections.

5 Experiments

Our experiments aim to validate Dust3r Splatting on diverse scenes and to highlight its ability to handle previously unseen views. We use the same datasets commonly employed by SfM and novel-view synthesis pipelines, comparing Dust3r Splatting with baseline methods that rely on COLMAP or known camera intrinsics.

5.1 Testbed and Experimental Questions

We set up experiments on indoor and outdoor scenes with varying complexity. The key questions are:
(1) Does Dust3r Splatting achieve competitive or improved performance in reconstructing scene geometry compared to baseline pipelines that require known intrinsics?
(2) Can Dust3r Splatting handle previously unseen viewpoints with similar accuracy to methods that rely on calibrated cameras?
(3) Does the integration of DUSt3R and Gaussian Splatting preserve the real-time rendering capability and visual quality achieved by Gaussian Splatting?

5.2 Experimental Details

We begin by taking multiple images of a scene with no initial knowledge of camera intrinsics. DUSt3R processes these images to yield relative poses and pointmaps. Next, we use pycolmap to generate a COLMAP-like reconstruction – albeit with arbitrary intrinsics. Finally, we feed this reconstruction into Gaussian Splatting. The baseline comparison involves using a standard SfM pipeline (e.g., COLMAP with known intrinsics) to provide input for Gaussian Splatting. We then evaluate the rendering quality of novel views, both visually and through qualitative metrics such as feature consistency, sharpness, and scene completeness.

5.3 Results

Initial qualitative results indicate that Dust3r Splatting can produce visually comparable reconstructions and novel-view renderings to those generated via standard SfM pipelines. In particular, scenes reconstructed without known intrinsics remain stable and show minimal distortion, attesting to DUSt3R's intrinsic-free robustness. When rendering previously unseen viewpoints, Dust3r Splatting maintains coherent scene geometry and visually pleasing results, only including more noise than baseline methods.

Dust3r Splatting

COLMAP + Gaussian Splatting

Due to the complexity of establishing fully quantitative benchmarks without intrinsics, our current work focuses on qualitative comparisons and stability analyses. Future work could emphasize controlled quantitative evaluations, including metrics such as depth accuracy, reprojection error, and perceptual image quality measures.

5.4 Discussion

While our experiments show promise, they are preliminary and mainly qualitative. A lack of extensive quantitative metrics leaves open the question of how Dust3r Splatting compares rigorously across various datasets and conditions. Nevertheless, the improvement in handling unseen views and maintaining stable geometry without prior calibration suggests significant practical advantages. Future experiments could include more rigorous metrics, comparisons against a broader set of methods, and exploration of different types of scenes.

6 Conclusion and Future Work

This paper introduces Dust3r Splatting, a unified approach to dense multi-view reconstruction that combines DUSt3R's intrinsic-free pipeline with the state-of-the-art Gaussian Splatting rendering technique. By bridging these methods through a COLMAP-like intermediate format, we achieve high-quality, real-time novel-view synthesis from arbitrary image collections without requiring camera intrinsics. The initial qualitative results suggest that Dust3r Splatting handles previously unseen views more robustly and simplifies the complex geometric pipelines that commonly hinder multi-view stereo.

Looking ahead, future work will involve extensive quantitative evaluations, including metrics like depth accuracy, camera pose estimation error, and rendering quality under controlled conditions. We will also consider more challenging datasets, integrate advanced regularization strategies, and investigate the

potential of Dust3r Splatting for dynamic scenes. Ultimately, we believe our approach can provide a more accessible and robust solution to 3D reconstruction, empowering a wide range of applications in 3D vision, robotics, and augmented reality.

7 References

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[2] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural Volumes: Learning Dynamic Renderable Volumes from Images," *arXiv preprint arXiv:1906.07751*, June 18, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1906.07751

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *European Conference on Computer Vision (ECCV) 2020 (oral)*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2003.08934

[4] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D Vision Made Easy," *arXiv preprint arXiv:2312.14132*, 2023. [Online]. Available: https://arxiv.org/abs/2312.14132

[5] Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., & Revaud, J. (2022). CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. *Advances in Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=wZEfHUM5ri

[6] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, Yue Wang. InstantSplat: Sparse-view SfM-free Gaussian Splatting in Seconds**,** https://ar5iv.org/pdf/2403.20309, March 2024.