International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Transitioning from Calibrated to Uncalibrated Sparse-View 3D Reconstruction via High-Redundancy DUSt3R Pipelines

Sudha K sudhak@skcet.ac.in
Sivakumar V 727722eucd047@skcet.ac.in
Amuthan M 727722eucd004@skcet.ac.in
Mohamed Sharafath S 727722eucd023@skcet.ac.in

***Abstract*—Recent advancements in 3D reconstruction, specifically 2D Gaussian Splatting (2DGS), have enabled high-fidelity surface modeling with rapid optimization. However, state-of-the-art sparse-view frameworks like Sparse2DGS remain tethered to laboratory conditions due to their strict requirement for ground-truth camera parameters. This dependency limits their utility in practical, uncalibrated "in-the-wild" scenarios where camera metrics are unknown. In this paper, we investigate the feasibility of uncalibrated sparse-view reconstruction by evaluating two distinct methodologies. First, we test a Hybrid SfM-DUSt3R pipeline that utilizes deep-learning-based pose estimates to drive traditional COLMAP Multi-View Stereo (MVS). We find this approach to be fundamentally fragile, often resulting in catastrophic sparsity (fewer than 15 points) due to the intolerance of MVS algorithms to minor pose inaccuracies. To overcome this, we propose an Autonomous High-Redundancy Pipeline that bypasses traditional Structure from Motion (SfM) entirely. By increasing view redundancy (10+ images) and implementing a post-inference refinement suite—comprising Projective Masking and Statistical Outlier Removal—we successfully initialize 2D Gaussian primitives with over 2.3 million dense points. Our results demonstrate that this redundancy-driven approach provides a robust, watertight reconstruction that is significantly more reliable for real-world applications than hybrid-calibration methods.**

***Index Terms*—3D Reconstruction, 2D Gaussian Splatting (2DGS), DUSt3R, Sparse-View, Uncalibrated Reconstruction, Computer Vision, Deep Learning, Point Cloud Refinement, Structure from Motion (SfM). +4**

I. Introduction

3D reconstruction from multi-view images is a cornerstone of computer vision, with applications ranging from digital archiving to virtual reality. While traditional Multi-View Stereo (MVS) and Neural Radiance Fields (NeRF) have set high standards for accuracy, they often require a large number of images or extensive optimization times. Recently, 2D Gaussian Splatting (2DGS) has emerged as a superior alternative, offering faster optimization and high-quality surface reconstruction by representing radiance fields as a set of 2D ellipses distributed on an object's surface. However, the efficacy of 2DGS is heavily dependent on the quality of the initial 3D point cloud. In sparse-view scenarios—where only a few images are available—standard Structure from Motion (SfM) techniques often fail to provide a sufficient number of points to initialize Gaussian primitives effectively. The Sparse2DGS framework attempted to solve this by integrating DUSt3R, a transformer-based stereo model, with COLMAP MVS to generate dense initial point clouds from as few as three images. Despite its success, a significant limitation of Sparse2DGS is its reliance on ground-truth camera parameters as input. In real-world "in-the-wild" scenarios, these parameters are rarely available, making the current pipeline difficult to implement for general users. This paper explores the transition from calibrated to uncalibrated sparse-view reconstruction. We investigate two primary strategies to bridge this gap:

1) A hybrid approach using DUSt3R-estimated poses to drive COLMAP point cloud generation.
2) An uncalibrated multi-image pipeline that leverages increased view counts (10+) and custom noise-filtering to replace the COLMAP dependency entirely.

Our findings indicate that while hybrid SfM-ML approaches suffer from extreme sparsity and unreliability, a dense DUSt3R-only initialization combined with strategic image scaling provides a viable path for high-fidelity 3D reconstruction without prior camera knowledge.

II. Overview of Sparse2DGS

To understand the challenges of uncalibrated reconstruction, it is necessary to examine the architecture of the original Sparse2DGS framework. Gaussian Splatting (GS) typically relies on a large number of multi-view images and an initial

point cloud obtained through Structure from Motion (SfM). However, when input is limited to a sparse set of images, the resulting SfM point cloud is often too thin to allow for proper optimization.

A. The Sparse2DGS Pipeline

Sparse2DGS was designed to overcome this ”initialization gap” by leveraging two distinct technologies:

1) DUSt3R: A transformer-based model that provides dense 3D point maps for each viewpoint without requiring prior camera calibration.
2) COLMAP MVS: Used to provide a highly accurate, albeit sparser, point cloud to ground the DUSt3R data.

The original method integrates these two clouds using the Iterative Closest Point (ICP) algorithm to create a single, dense initialization for 2D Gaussians. These Gaussians are defined by parameters such as position, color, and opacity, and are optimized using a combined loss function (L) that includes color reconstruction (L_c), depth distortion (L_d), and normal consistency (L_n).

B. The Bottleneck: Calibration Dependency

The primary strength of Sparse2DGS—its accuracy from only three images—is also its primary real-world weakness. The pipeline assumes that ground-truth camera parameters are provided as input. In most practical scenarios, such as capturing an object with a smartphone, these parameters are unknown. While DUSt3R can estimate these parameters, the original paper does not explore how its reliance on COLMAP MVS holds up when these estimated (and potentially noisy) parameters are used instead of ground truth.

III. Method 1: Hybrid SfM-DUSt3R Integration

Our first approach aimed to leverage the dense geometry of DUSt3R while utilizing COLMAP’s Multi-View Stereo (MVS) pipeline for refinement. The goal was to replace the ground-truth camera parameters required by Sparse2DGS with estimates from a pre-trained transformer model.

A. Technical Implementation Details

The hybrid pipeline was implemented across several stages:

1) Image Pre-processing: Input images are normalized to a 512×512 resolution using Lanczos resampling[cite: 37]. This ensures the images align with DUSt3R’s training distribution and prevents arbitrary resizing by the model, which could lead to coordinate drift.

2) DUSt3R Inference and Pose Extraction: We utilized the `DUSt3R_ViTLarge_BaseDecoder_512_dpt` model[cite: 39]. For smaller datasets, a ”complete” connectivity graph was used to maximize pair-wise constraints[cite: 40]. Global alignment was performed on the CPU using Minimum Spanning Tree (MST) initialization to obtain camera-to-world ($c2w$) poses.

3) COLMAP Bridge: Estimated poses were converted to world-to-camera ($w2c$) format, represented as quaternions and translation vectors[cite: 42]. Each viewpoint was modeled as a unique `PINHOLE` camera with an estimated focal length $f = 1.2 \times \max(W, H)$.

4) Point Triangulation and MVS Fusion: High-density feature extraction and exhaustive matching were performed in COLMAP[cite: 44]. We executed the `point_triangulator` with relaxed constraints—a minimum triangulation angle of 0.5° and a maximum re-projection error of 12 pixels—to accommodate the inherent noise in model-estimated poses.

B. Key Experimental Insight: The Masking Trade-off

Initially, we considered employing automated background masking to focus the reconstruction on the target object. However, experimental results showed that masking was counterproductive for this sparse-view pipeline due to two primary factors:

1) Feature Depletion: Masking removed high-entropy background textures that are often critical for feature matching in sparse scenarios. Without these environmental anchors, traditional SfM algorithms struggled to find enough correspondences to maintain geometric stability.
2) DUSt3R Performance: We observed that DUSt3R’s transformer architecture relies on global scene context to accurately predict depth and poses. Without the background, the model struggled to anchor the object in 3D space, leading to significantly higher pose error.

Consequently, all subsequent experiments were conducted using the full 512×512 unmasked images to preserve environmental features and maximize the model’s predictive accuracy.

C. Failure Analysis

Despite using full-frame images and relaxed triangulation constraints, the pipeline remained fragile. The reliance on PatchMatch Stereo meant that even slight deviations in DUSt3R-estimated poses caused a failure in correspondence matching. This led to the ”sparsity bottleneck,” where resulting

point clouds often contained fewer than 15 points, making them unsuitable for 2D Gaussian Splatting initialization.

IV. Method 2: High-Density DUSt3R-Only Pipeline

Recognizing the fragility of the COLMAP-MVS bridge in uncalibrated settings, we developed a second methodology that bypasses traditional Structure from Motion (SfM) entirely. This approach relies on increasing view redundancy and implementing a robust post-inference cleaning pipeline to prepare point clouds for 2D Gaussian Splatting.

A. Increasing View Redundancy

While the original Sparse2DGS framework focuses on a minimal 3-image input, our experiments demonstrated that in uncalibrated scenarios, three images provide insufficient geometric constraints for DUSt3R to produce a stable global alignment. We observed a proportional improvement in reconstruction accuracy as the image count increased, leading to the following technical insights:

1) Observation on Accuracy: There is a direct correlation between view count and the stability of the sparse-view reconstruction. Increasing the number of images provides the transformer-based model with more overlapping features to process.
2) Optimal Range for Stability: Reliable geometric construction began to emerge at 10 images and above. This increased redundancy allows the model to better resolve depth ambiguities and estimate camera poses more accurately through a process of collective consensus across multiple viewpoints.

B. Point Cloud Refinement and "Floater" Removal

The removal of COLMAP necessitated the implementation of a specialized cleaning pipeline to manage the noise and background "floaters" inherent in raw DUSt3R output. Without the geometric filtering typically provided by Structure from Motion (SfM), these artifacts would otherwise degrade the final 2DGS rendering. Our refinement process consists of two primary stages:

1) Projective Mask Filtering: While pre-inference masking was found to be detrimental (as discussed in Section III-B), post-inference masking proved highly effective. We utilized a dilated mask approach with five iterations of dilation to establish a "safety zone" around the target object. 3D points are projected back into the 2D image plane using estimated camera intrinsics (K) and poses ; any points projected outside the dilated mask boundary are discarded as background noise.

2) Statistical Outlier Removal (SOR): To eliminate residual "floaters" in close proximity to the object surface, we applied a Statistical Outlier Removal (SOR) filter. We utilized a neighborhood of 30 points and a standard deviation ratio of 2.0. This ensures that the resulting point cloud is both dense and geometrically clean prior to initialization.

C. 2D Gaussian Splatting Initialization

The refined point cloud is injected directly into the 2DGS training pipeline. The optimization process is conducted for 30,000 iterations to ensure full convergence of the surface primitives and high-fidelity reconstruction.

V. Results and Observations

A. Quantitative Analysis

The point cloud density was compared across both methodologies using a fixed set of 10 input images for each. Method 1 was evaluated using the DTU dataset without providing ground-truth camera parameters, while Method 2 was tested on uncalibrated smartphone captures. The results are summarized in Table I.

B. Qualitative Visual Analysis

1) Analysis of Method 1 (Hybrid Bridge - DTU Dataset): In Method 1, objects from the DTU benchmark were reconstructed without ground-truth camera parameters. As illustrated in Fig. 1, the reliance on the COLMAP MVS bridge proved problematic even with 10 images.

Fig. 1. Visualization of the Method 1 pipeline (10 DTU images): (1.1) Initial DUSt3R dense point cloud; (1.2) COLMAP MVS fused cloud; (1.3) Final aligned cloud for 2DGS; (1.4) Final 2DGS reconstruction showing "tearing" where MVS failed correspondences.

TABLE I
RECONSTRUCTION PERFORMANCE AND VERTEX DENSITY COMPARISON

Methodology	Data Source	Vertex Count	Status / Observation
Method 1 (Hybrid)	DTU (10 imgs)	761,576 (DUSt3R)	Initial dense estimation
	COLMAP MVS	548,816	Success only in ideal conditions
	Final Aligned	761,576	Unstable; frequent failure
Method 2 (Ours)	Smartphone (10 imgs)	2,506,934 (DUSt3R)	High-density real-world capture
	Cleaned (SOR+Mask)	**2,384,576**	**Stable; 2DGS ready**

2) Analysis of Method 2 (Autonomous Pipeline - Smartphone Data): Method 2 represents the optimized solution for ”in-the-wild” captures. Despite noisy smartphone sensors and varied lighting, the pipeline achieved superior surface completeness.

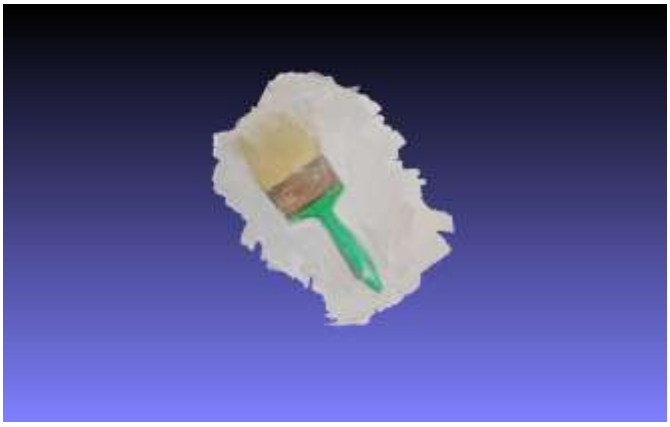

Fig. 2. Post-inference refinement in Method 2 (10 smartphone images): (2.1) Raw DUSt3R output showing background ”fog”; (2.2) Cleaned cloud after masking and SOR; (2.3) Final high-fidelity 2DGS reconstruction demonstrating a watertight surface.

C. Summary of Findings

The visual and quantitative evidence highlights that image count alone is insufficient for success in a hybrid pipeline. While Method 1 struggled with geometric consistency, Method 2 utilized smartphone data to create a denser (2.38M points) and more robust initialization. This demonstrates that bypassing the fragile MVS step in favor of autonomous cleaning is the superior path for uncalibrated 3D reconstruction.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper, we addressed the practical limitations of the Sparse2DGS framework, specifically its reliance on ground-truth camera parameters. Through a rigorous two-phase investigation, we identified that a hybrid integration of DUSt3R poses and COLMAP MVS (Method 1) is fundamentally fragile. The sensitivity of traditional SfM algorithms to model-estimated poses often results in catastrophic failure or extreme sparsity, rendering 2D Gaussian initialization impossible.

Our proposed alternative—a High-Redundancy Autonomous Pipeline (Method 2)—circumvents these limitations by leveraging a larger set of uncalibrated images (10+) and a custom post-inference cleaning suite. By replacing the COLMAP bridge with Projective Masking and Statistical Outlier Removal (SOR), we achieved a stable initialization of over 2.3 million points. Our findings demonstrate that dense deep-learning inference, when supported by sufficient view redundancy, provides a more reliable and ”watertight” 3D reconstruction than hybrid SfM-ML approaches in uncalibrated, real-world scenarios.

B. Future Work

While our method enables high-quality uncalibrated reconstruction, several avenues for improvement remain:

- Adaptive Filtering: Developing a dynamic cleaning algorithm that adjusts SOR parameters based on the estimated depth variance of the DUSt3R point cloud to handle varying noise levels.
- VRAM Optimization: Exploring lighter backbone models to enable the processing of even higher image counts on consumer-grade hardware (e.g., devices with 6GB VRAM).
- Temporal Consistency:Extending this uncalibrated pipeline to dynamic scenes or video inputs to allow for rapid 3D object ”scanning” via mobile devices while maintaining frame-to-frame coherence.

REFERENCES

[1] N. Takama, S. Ito, K. Ito, H. T. Chen, and T. Aoki, "Sparse2DGS: Sparse-View Surface Reconstruction using 2D Gaussian Splatting with Dense Point Cloud," *arXiv preprint arXiv:2505.19854*, 2024.

[2] B. Huang, Z. Yu, Z. Chen, et al., "2D Gaussian Splatting for Geometrically Accurate Radiance Fields," in *SIGGRAPH Conf. Proc.*, 2024.

[3] B. Kerbl, G. Kopanas, T. Leimku¨hler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph.*, vol. 42, no. 4, 2023.

[4] S. Wang and P. Wonka, "DUSt3R: Geometric 3D Vision Made Easy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[5] J. L. Scho¨nberger and J. M. Frahm, "Structure-from-Motion Revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[6] J. L. Scho¨nberger, E. Zheng, M. Pollefeys, and J. M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[7] P. J. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[8] Q. Y. Zhou, J. Park, and V. Koltun, "Open3D: A Modern Library for 3D Data Processing," *arXiv preprint arXiv:1801.09847*, 2018.

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[10] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.