

PI3DGS: ROBUST JOINT OPTIMIZATION OF CAMERA POSES AND 3DGS FROM UNCALIBRATED IMAGES

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent advancements in neural rendering for 3D reconstruction have focused on constructing representations directly from uncalibrated RGB images, bypassing the need for Structure-from-Motion (SfM) preprocessing. A primary challenge in this domain is the joint optimization of scene geometry and camera parameters, a task fraught with inherent ambiguities. Although 3D Gaussian Splatting (3DGS) has achieved photorealistic reconstruction quality, its discrete, point-based representation complicates this joint optimization process. To address these challenges, we propose a robust, SfM-free framework that leverages pre-trained 3D feed-forward models within a coarse-to-fine alignment pipeline. Our method introduces Pi3 for scene initialization and proceeds with the joint training of geometry and camera poses. To enhance the stability of camera pose optimization, we employ 3D and 2D filters to regularize the gradients from signal alignment. Furthermore, we incorporate a geometric regularization based on image matching to provide global constraints for camera pose refinement, which significantly improves both reconstruction quality and pose estimation accuracy. Our method achieves competitive performance in novel view synthesis and camera pose estimation, demonstrating its robustness across diverse datasets.

1 INTRODUCTION

Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) provide a practical framework for learning a neural 3D radiance representation from collections of 2D RGB images, and have motivated a variety of alternative scene representations. NeRF accepts 3D sample points and view directions as input and predicts radiance and density via a neural network, which are then integrated by volume rendering. This implicit representation is a principal contributor to the high computational cost of NeRF-like methods. Recently, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has emerged as an explicit alternative: it represents a scene as a set of 3D Gaussian primitives and leverages conventional rasterization pipelines on modern GPUs, yielding substantially faster training and rendering.

Most reconstruction methods, including NeRF and 3DGS, require accurate camera intrinsics and extrinsics for the input images. These parameters are commonly obtained through Structure from Motion (SfM) tools such as COLMAP (Schonberger & Frahm, 2016). However, classical SfM pipelines have limitations: they depend on sufficient image overlap to establish multiview correspondences and can fail in challenging scenarios. Recent large multitask models for 3D vision, such as MAST3r (Duisterhof et al., 2025) and VGGT (Wang et al., 2025a), demonstrate improved precision and robustness across diverse conditions. Motivated by these advances, we adopt Pi3 (Wang et al., 2025c), a fully permutation-equivariant architecture, to replace SfM as a preprocessing module.

Directly using the outputs of learned large models introduces additional challenges. Network-based approaches can infer 2D–3D relations at a high level, but predicting precise numerical camera parameters (e.g., view matrices) remains difficult in the absence of strict geometric constraints (Wang et al., 2023a). Inaccurate camera parameters often lead to geometric and photometric degradations—such as misalignment, skew, or distortion—particularly in scenes with large depth ranges (e.g., outdoor environments or long image sequences). To address this, we propose Pi3DGS, a joint training framework that reconstructs 3DGS geometry while simultaneously refining camera poses, starting from the imperfect camera estimates produced by Pi3.

Adapting joint optimization strategies from NeRF to 3DGS entails specific difficulties. First, 3DGS employs an explicit representation composed of Gaussian ellipsoids whose attributes are optimized and whose population is incrementally densified to recover higher-frequency detail. This densification process splits or duplicates primitives when they exceed a threshold. Because 3DGS is prone to overfitting, primitives created under supervision of inaccurate early-stage cameras can become trapped in local minima, and there is no inherent mechanism to correct such primitives later in training.

To mitigate early densification artifacts, and inspired by Mip-Splatting (Yu et al., 2024), we introduce a gradient-smoothing strategy. The key idea is to temporarily scale or blur the Gaussians so that each primitive influences more pixels during early training, thereby smoothing gradients and emulating the coarse-to-fine behavior observed in AbsGS (Ye et al., 2024). Consistent with analyses in TensorRF (Chen et al., 2024), this attenuation of high-frequency signals yields more stable camera alignment and reduces the tendency to converge to poor local minima.

A second challenge is the relative paucity of cross-view constraints during camera optimization. NeRF-based methods typically sample rays from multiple images per batch, enabling simultaneous optimization across many cameras and enforcing multi-view consistency. In contrast, 3DGS uses a differentiable rasterization pipeline in which all primitives are transformed, clipped, rasterized, and blended to produce pixel colors; this pipeline limits the number of views that can be jointly optimized per step and, absent explicit geometric coupling, makes cameras susceptible to inconsistent local groupings.

To provide stronger geometric supervision, and drawing on PoRF (Bian et al.), we integrate an image-matching-based geometric constraint into the pose refinement process. This constraint supplements the rasterization loss, stabilizes camera optimization (in particular during opacity resetting), and improves final pose accuracy.

Our main contributions can be summarized as follows.

- We present Pi3DGS, a pipeline for reconstructing 3D Gaussian Splatting scenes from RGB images alone, obviating the need for accurate SfM precomputation. The pipeline combines a learned permutation-equivariant initialization (Pi3) with a joint optimization that constructs the 3DGS representation while simultaneously refining camera intrinsics and extrinsics.
- We introduce a filtering strategy based on gradient smoothing to prevent premature and erroneous densification of Gaussian primitives. By enlarging the effective receptive field of primitives during early training and attenuating high-frequency signals, this strategy mitigates unstable gradients and reduces convergence to poor local minima, thereby enhancing training stability.
- We incorporate image-matching geometric constraints into the pose-refinement stage to enforce robust multi-view consistency. This regularization complements the rasterization loss, stabilizes camera optimization (particularly during opacity resetting), and improves final pose accuracy and scene fidelity.

2 RELATED WORKS

2.1 NOVEL VIEW SYNTHESIS

Novel view synthesis is a subfield of 3D reconstruction that focuses on producing high-quality images from viewpoints not present in the input image set. This topic is closely related to inverse differentiable rendering and other recent trends in neural rendering. Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) marked a milestone by demonstrating the effectiveness of implicit neural fields for photorealistic novel view synthesis. Subsequent approaches based on neural signed distance functions (SDFs) replace radiance-based volumetric representations with surface-centric formulations (Wang et al., 2021; 2023b; Li et al., 2023), which can be more readily converted to classical 3D representations such as triangle meshes. Other works organize scene representations on structured voxel grids to improve efficiency (Deng & Tartaglione, 2023; Takikawa et al., 2021; Reiser et al., 2021).

Although MLP-based methods have achieved impressive visual fidelity, they face inherent challenges in representing high-frequency signals (e.g., complex textures and fine geometric detail) and often require substantial training time. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) addresses some of these limitations by using an explicit representation that models geometry with 3D Gaussian primitives and encodes appearance via spherical harmonics (Atkinson & Han, 2012). By leveraging conventional rasterization pipelines, 3DGS attains excellent visual quality while enabling real-time rendering performance. Consequently, 3DGS has attracted considerable attention and has been applied to diverse tasks, including autonomous driving (Zhou et al., 2024), deformable human modeling (Moreau et al., 2024), and embodied intelligence training (Wang et al., 2025b).

Despite these advances, most novel view synthesis methods still depend on accurate camera poses to construct reliable scene representations; obtaining such poses remains a critical practical bottleneck in many real-world scenarios.

2.2 JOINT REFINEMENT ON CAMERA AND SCENE

The feasibility of jointly optimizing scene representation and camera poses was first demonstrated by BARF (Lin et al., 2021), which trains a neural radiance field from RGB images given only inaccurate initial camera poses. TensorRF (Chen et al., 2024) introduces Gaussian kernel filtering of 2D supervision to smooth the optimization landscape for neural radiance fields, thus mitigating local minima in joint camera-scene optimization. Subsequent work has incorporated additional priors and geometric cues to stabilize pose estimation: Nope-NeRF (Bian et al., 2023) uses depth maps as anchors; SC-NeRF (Jeong et al., 2021) and PoRF (Bian et al.) integrate image matching to refine camera poses.

Recent methods based on 3D Gaussian Splatting (3DGS) aim to address these limitations by leveraging rasterization-friendly primitives. CF-3DGS (Fu et al., 2024) incrementally adds Gaussians and cameras from a sequence of consecutive frames, but fixes registered cameras in later stages, which can lead to accumulated registration errors; moreover, when its COLMAP-free template heuristics fail, CF-3DGS requires precomputed camera intrinsics. InstantSplat (Fan et al., 2024) employs DUST3r (Wang et al., 2024) for initialization but does not scale well to long image sequences. 3R-GS (Huang et al., 2025) proposes an alternative pipeline (see the original reference for details). BAD-Gaussians (Zhao et al., 2024) specifically target scenes affected by motion blur by assigning and optimizing multiple virtual cameras per view to model blur trajectories. KeyGS (Chang et al., 2025) adapts the smoothing strategy from TensorRF to 3DGS through Mip-Splatting (Yu et al., 2024), but continues to rely on the conventional structure-from-motion for initial scene setup.

3 METHOD

As a classic task in computer vision, taking a sequence of plain 2D RGB images $\{I_i\}_{i=1}^N, I_i \in \mathbb{R}^{H \times W \times 3}$ as input, our goal is to reconstruct the 3D geometry using 3D Gaussian splatting and learn the extrinsic and intrinsic parameters of the cameras for the corresponding images simultaneously.

3.1 PRELIMINARY: 3D GAUSSIAN SPLATTING

Kerbl et al. (2023) proposed a novel representation for 3D geometry that can be rendered with photorealistic quality and high performance by leveraging the conventional rasterization pipeline. The primitive is an ellipsoid formulated with a 3D Gaussian function:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$

Each 3D Gaussian is parameterized by a mean position $\boldsymbol{\mu}$ and a covariance matrix $\Sigma \in \mathbb{R}^{3 \times 3}$. The ellipsoid defined by Equation 1 is equivalent to a sphere that has been scaled along the coordinate axes and then rotated. Consequently, the covariance can be expressed as a composition of a scaling matrix $S \in \mathbb{R}^{3 \times 3}$ (typically diagonal) and a rotation $R \in \mathbb{R}^{3 \times 3}$, e.g. $\Sigma = RSS^T R^T$, which ensures that Σ is positive semi-definite.

To render a 3D Gaussian splat (3DGS), it is projected onto the screen (canvas). Given the camera pose represented by a rigid-body transform $W \in \text{SE}(3)$ and camera intrinsics $K \in \mathbb{R}^{3 \times 4}$, the mean

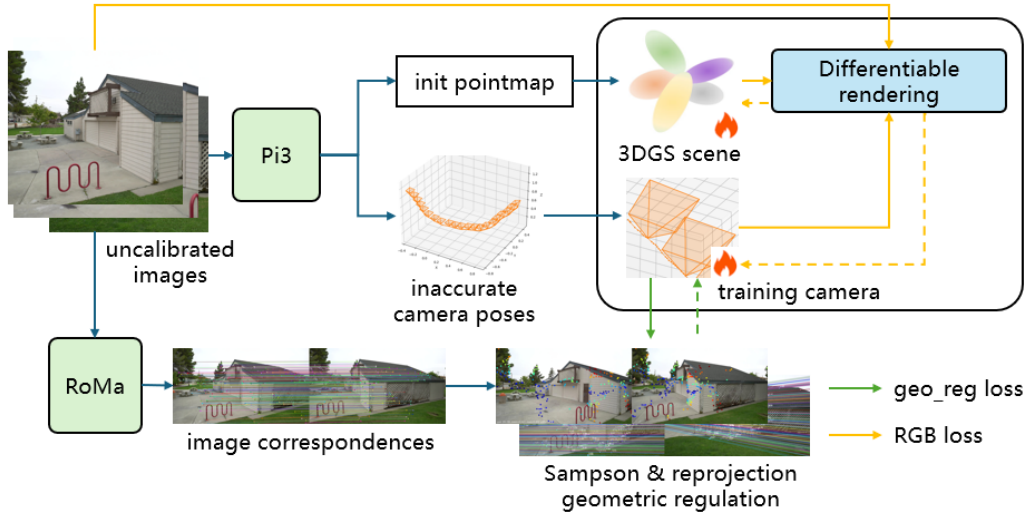


Figure 1: Overview of our Pi3GS pipeline. The geometry and the cameras are optimized simultaneously.

μ is projected to image coordinates x_j^{2D} . The projected covariance is approximated by propagating the 3D covariance via J , the Jacobian of the camera projection matrix (Zwicker et al., 2001): $\Sigma^{2D} = J W \Sigma W^T J^T$, where J denotes the Jacobian of the projection at the Gaussian center. This approximation yields a 2D Gaussian in image space and enables rasterization without explicit 3D sampling, making the representation compatible with standard GPU rasterization pipelines.

The appearance of each Gaussian is represented using spherical harmonics: $c_i = SH_i(\mathbf{v})$, where $\mathbf{v} \in \mathbb{R}^3$ is the normalized view direction. Parameterized by a low-dimensional feature vector, the spherical harmonics basis compactly encodes low-frequency, view-dependent color variation.

The final pixel color is obtained by compositing the contributions of individual Gaussians in a front-to-back order. Denoting the opacity contribution of the i -th Gaussian by $\alpha_i = o_i G_i(x)$ (where o_i accounts for per-splat opacity), the color is

$$\mathbf{C} = \sum_{i=1}^N c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

which corresponds to standard alpha compositing applied to the 2D Gaussian splats.

3.2 Pi3DGS PIPELINE

Our objective is to reconstruct a high-fidelity 3D geometry and shading (3DGS) representation from a collection of uncalibrated multi-view RGB images. Conventional 3DGS reconstruction typically requires accurate camera intrinsics and poses, which restricts its applicability for many downstream tasks. Pi3 (Wang et al., 2025c) is an image-matching model pre-trained on extensive 3D datasets and thus embodies a strong 3D prior. Given a sequence of 2D images, Pi3 predicts an initial 3D point cloud and estimates camera intrinsics via reprojection error minimization. Leveraging a large neural model, Pi3 can produce satisfactory initial point clouds and camera poses in many cases. However, in the absence of explicit regularization on camera poses, these initializations may contain non-negligible noise, particularly for outdoor scenes with large depth ranges.

Because Pi3 does not provide dense pixel correspondences, its initialization alone is insufficient to establish a robust geometric constraint. We therefore employ RoMa (Edstedt et al., 2024), an efficient and robust image-matching model, to extract pairwise correspondences across the image sequence and thereby strengthen the geometric supervision.

Camera modeling. We represent each camera pose subject to refinement as the composition of an initial pose and a trainable corrective transformation:

$$T_i = \Delta T_i \hat{T}_i, \quad (3)$$

where $\hat{T}_i$ denotes the Pi3 initialization and ΔT_i is a learnable refinement. To reduce coupling between rotation and translation during optimization, we parameterize ΔT_i by decoupling its rotation and translation components: the rotation is represented in the Lie algebra $\mathfrak{so}(3)$, while the translation is represented in $\mathbb{R}^3$ and multiplied by a learnable scalar s . The scalar s enhances the robustness of translation optimization across scenes with varying absolute scales.

Initialization. For each input image $I_i \in \{I_i\}_{i=1}^N$, Pi3 produces a local point map $\mathbf{X}_i^N \in \mathbb{R}^{H \times W \times 3}$ and an approximate camera pose $\hat{T}_i = [R_i \mid \mathbf{t}_i]$. The intrinsic matrix K_i is not provided and must be estimated. Given the approximate extrinsics $\hat{T}_i$, we estimate K_i by minimizing the reprojection error between the 3D points and their observed image projections.

Let $\mathbf{x}_j^{3D} \in \mathbf{X}_i^N$ be a 3D point and $\mathbf{x}_j^{2D} \in \mathbb{R}^2$ its observed image location. In homogeneous coordinates the projection relation is

$$\bar{\mathbf{x}}_j^{2D} \sim K_i [R_i \mid \mathbf{t}_i] \bar{\mathbf{x}}_j^{3D}, \quad (4)$$

where $\bar{\mathbf{x}}_j^{2D}$ and $\bar{\mathbf{x}}_j^{3D}$ denote homogeneous coordinates, K_i is the intrinsic matrix, and $[R_i \mid \mathbf{t}_i]$ are the extrinsics from Pi3. The intrinsics are recovered by minimizing the reprojection error:

$$K_i = \arg \min_K \sum_j \|\pi(K [R_i \mid \mathbf{t}_i] \bar{\mathbf{x}}_j^{3D}) - \mathbf{x}_j^{2D}\|^2, \quad (5)$$

where $\pi(\cdot)$ denotes perspective division. This problem can be linearized and solved in a least-squares sense to obtain an initial estimate of K_i . If all cameras are assumed to share identical intrinsics, the shared intrinsic matrix K is estimated jointly over the entire sequence.

3.3 ALIGNMENT WITH SMOOTHED GRADIENT

3D Gaussian blurring can be formulated as a smooth operator acting on the scale component of a 3DGS. Concretely,

$$G_i(\mathbf{x}) = \sqrt{\frac{|\Sigma|}{|\Sigma + \epsilon_i^2 I|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\Sigma + \epsilon_i^2 I)^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \quad (6)$$

This operation increases the effective covariance of each Gaussian and attenuates its peak amplitude, thereby simulating a spatial spread of the underlying density.

To ensure consistent blurring of projected Gaussians in image space, we set the blur magnitude ϵ_i proportionally as

$$\epsilon_i = \epsilon_{\text{base}} \cdot d_i, \quad (7)$$

where d_i denotes the depth of the Gaussian center in camera coordinates (i.e., its z coordinate). This depth-proportional scaling balances sampling density across the image plane and applies a scale-appropriate blur to each Gaussian, yielding more stable and consistent training dynamics.

For 2D supervision, rendered images are smoothed by convolution with a 2D Gaussian kernel. By scheduling the intensity of both the 3D and 2D blurring terms, we realize a coarse-to-fine alignment strategy: stronger blurring is applied early to suppress high-frequency noise, and the blur is gradually reduced to handle finer signals. Following the analysis of TensorRF (Chen et al., 2024), this controlled blurring of high-frequency content promotes robust and stable camera-pose alignment and regularizes the frequency spectrum of the 3DGS primitives.

3.4 IMAGE MATCHING BASED GEOMETRIC REGULATION

Although Pi3 has provided a strong prior on various 3D information including point cloud, camera poses. However, according to our observation, several problems arise in supplying these initializations directly to the 3DGS reconstruction pipeline. Pi3 does not output correspondence like other



Figure 2: Effect of camera perturbations in the absence of geometric regularization. From left to right: ground truth, rendered result, and error map. From top to bottom: (a) an outlier camera that fails to align the ground-truth and rendered views; (b) artifacts resulting from misaligned cameras; (c) results produced when our geometric constraint is applied. Zoom in for finer details.

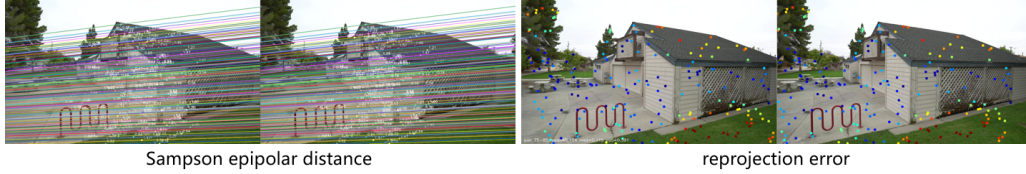


Figure 3: Visualization of epipolar and reprojection geometry. When camera motion exhibits a small baseline, the epipolar lines become nearly parallel, which reduces the effectiveness of the epipolar constraint (left). Therefore, we report the reprojection error as a more direct measure (right). Colors map error magnitude from blue (low) to red (high).

large 3D models, we exploit RoMa (Edstedt et al., 2024), a robust but accurate pre-trained image matching model to acquire pairwise correspondences.

Based on the correspondences, we implement our framework with both the Sampson epipolar distance and reprojection error to maintain the relationship between the cameras. This improves the global stability for the cameras in searching for blurred phrases. Specifically, the former is a robust and symmetric version of the epipolar distance. The latter is a more sensitive regulation based on image matching, which is more effective for low-baseline scenarios.

$$\mathcal{L}_{\text{match}} = \frac{\lambda_{\text{epi}}}{\sum_{i=1}^N c_i} \sum_{i=1}^N c_i \cdot d_{\text{epi}}(\mathbf{x}_i, \mathbf{x}'_i) + \frac{\lambda_{\text{reproj}}}{\sum_{i=1}^N c_i} \sum_{i=1}^N c_i \cdot d_{\text{reproj}}(\mathbf{x}_i, \mathbf{x}'_i) \quad (8)$$

Here, $d_{\text{epi}}(\mathbf{x}_i, \mathbf{x}'_i)$ and $d_{\text{reproj}}(\mathbf{x}_i, \mathbf{x}'_i)$ represent the symmetric Sampson distance and reprojection error of correspondence $(\mathbf{x}_i, \mathbf{x}'_i)$ in pixel coordinates.

Additionally, regularly resetting the opacity is a common scheme to reduce artifacts by removing wrongly generated floater Gaussians. This operation produces a large fluctuation to the RGB gradient, causing unsteadiness or even mis-convergence to the cameras.

3.5 TOTAL TRAINING LOSS

The overall training objective comprises three terms: the 3DGS rendering loss, a geometric regularization term based on image matching, and a scale regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{scale.reg}}. \quad (9)$$

The photometric rendering loss follows standard practice and combines an L1 term with a structural similarity term:

$$\mathcal{L}_{\text{rgb}} = (1 - \lambda_{\text{SSIM}}) \mathcal{L}_1 + \lambda_{\text{SSIM}} \mathcal{L}_{\text{SSIM}}, \quad (10)$$

where $\lambda_{\text{SSIM}} = 0.2$ balances the two components. To prevent degenerate or ill-conditioned Gaussian primitives, we introduce a scale regularization term that constrains the aspect ratio of each Gaussian’s principal axes:

$$\mathcal{L}_{\text{scale.reg}} = \frac{1}{N} \sum_{i=1}^N \left[\max \left(\frac{\max(S_i)}{\min(S_i)}, r \right) - r \right]_+, \quad (11)$$

where $S_i \in \mathbb{R}^d$ is the vector of scale parameters (the lengths of the principal axes) for the i -th Gaussian, $r > 1$ denotes the allowable maximum aspect ratio, and $[\cdot]_+ = \max(\cdot, 0)$ ensures the term is nonnegative.

Table 1: Novel view evaluation on Tanks and Temples dataset. The best results are highlighted in bold.

Method	Ours			KeyGS			CF-3DGS			NoPe-NeRF		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Chruch	31.61	0.95	0.04	30.62	0.92	0.06	30.23	0.93	0.11	25.17	0.73	0.39
Barn	33.61	0.94	0.03	34.25	0.95	0.04	31.23	0.90	0.10	26.35	0.69	0.44
Museum	35.85	0.97	0.01	33.46	0.94	0.03	29.91	0.91	0.11	26.77	0.76	0.35
Family	34.54	0.97	0.02	33.05	0.95	0.04	31.27	0.94	0.07	26.01	0.74	0.41
Horse	35.02	0.97	0.02	33.65	0.96	0.03	33.94	0.96	0.05	27.64	0.84	0.26
Ballroom	36.63	0.98	0.01	33.70	0.95	0.02	32.47	0.96	0.07	25.33	0.72	0.38
Francis	35.07	0.96	0.05	34.45	0.93	0.08	32.72	0.91	0.14	29.48	0.80	0.38
Ignatius	33.29	0.96	0.03	30.85	0.92	0.06	28.43	0.90	0.09	23.96	0.61	0.47
Mean	34.45	0.96	0.03	33.00	0.94	0.05	31.28	0.93	0.09	26.34	0.74	0.39

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We evaluate our method on two datasets. The Tanks and Temples dataset (Knapitsch et al., 2017) contains large-scale reconstruction scenarios covering both indoor and outdoor environments. CO3Dv2 (Reizenstein et al., 2021) comprises numerous multi-view captures of common objects. We compare against CF-3DGS (Fu et al., 2024), NoPe-NeRF (Bian et al., 2023), and keyGS (Chang et al., 2025).

Evaluation addresses both novel-view synthesis and camera-pose estimation; an ablation study quantifies the contribution of individual components. For novel view synthesis we report PSNR, SSIM, and LPIPS as measures of photometric and perceptual quality. Camera pose accuracy is assessed using translational and rotational relative pose error (RPE_t and RPE_r) as well as the absolute trajectory error (ATE).

Implementation Details. Our implementation is built on Nerfstudio (Tancik et al., 2023) with the gsplat extension (Ye et al., 2025), which provides a flexible framework for constructing a custom 3DGS reconstruction pipeline. To jointly optimize the 3DGS representation and camera poses, we

tune several hyperparameters. For camera optimization we employ exponential decay schedules, using base learning rates of 1×10^{-5} and 5×10^{-5} for rotation and translation, respectively, to balance their relative update magnitudes.

4.2 RESULTS AND COMPARISONS

Novel view synthesis. We first evaluate on a Tanks and Temples version preprocessed by CF-3DGS (Fu et al., 2024), which provides densely sampled frames extracted from short video clips. As reported in Table 1, our approach achieves superior visual-quality metrics in the majority of cases compared to the baselines.

On the CO3Dv2 dataset, we further evaluate the effectiveness of our approach under large camera motions. This dataset provides frames that capture objects over camera trajectories exceeding 180 degrees in most cases. As reported in Table 2, our method demonstrates robust performance and maintains competitive visual quality for novel view synthesis under these challenging conditions.

Table 2: Novel view evaluation on CO3Dv2 dataset. The best results are highlighted in bold.

Method	Ours			KeyGS			CF-3DGS		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Apple	33.98	0.94	0.07	33.53	0.94	0.07	29.69	0.89	0.29
Bench	31.75	0.92	0.10	26.35	0.73	0.30	26.21	0.73	0.32
Hydrant	28.66	0.91	0.07	25.33	0.80	0.15	22.14	0.64	0.34
SkateBoard	34.51	0.94	0.11	32.74	0.93	0.16	27.24	0.85	0.30
Teddybear	33.89	0.94	0.09	32.67	0.93	0.09	27.75	0.86	0.20
Mean	32.56	0.93	0.09	30.12	0.87	0.15	26.61	0.79	0.29

Camera pose estimation. We include KeyGS for reference only, since it is initialized with COLMAP results, which serves as ground-truth camera poses in our evaluation. As reported in Table 3, our method achieves the best performance in estimating the translation component of camera poses, which explains the superior visual quality observed in the novel view synthesis.

Table 3: Camera poses estimation on Tanks and Temples dataset. The best results are highlighted in bold except “KeyGS*” method as it is initialize with COLMAP sequence mode.

Method	ours			Nope-NeRF			CF-3DGS			KeyGS*		
	RPE $\downarrow$	RPE $\downarrow$	ATE $\downarrow$	RPE $\downarrow$	RPE $\downarrow$	ATE	RPE $\downarrow$	RPE $\downarrow$	ATE	RPE $\downarrow$	RPE $\downarrow$	ATE
Church	0.007	0.034	0.000	0.008	0.018	0.002	0.034	0.008	0.008	0.006	0.013	0.000
Barn	0.007	0.036	0.000	0.034	0.034	0.003	0.046	0.032	0.004	0.008	0.016	0.001
Museum	0.024	0.039	0.001	0.052	0.215	0.005	0.207	0.202	0.020	0.025	0.025	0.002
Family	0.012	0.033	0.000	0.022	0.024	0.002	0.047	0.015	0.001	0.012	0.012	0.000
Horse	0.082	0.037	0.001	0.112	0.057	0.003	0.179	0.017	0.003	0.078	0.002	0.001
Ballroom	0.015	0.033	0.000	0.037	0.024	0.003	0.041	0.018	0.002	0.015	0.014	0.000
Francis	0.008	0.036	0.001	0.029	0.154	0.006	0.057	0.009	0.005	0.007	0.016	0.001
Ignatius	0.008	0.033	0.001	0.033	0.032	0.005	0.026	0.005	0.002	0.001	0.010	0.001
mean	0.020	0.035	0.000	0.0409	0.0698	0.0036	0.080	0.038	0.006	0.019	0.014	0.001

4.3 ABLATION STUDY

In this section, we highlight the contributions of the key components and strategies of our proposed approach. All experiments were conducted on the Tanks and Temples dataset. As summarized in Table 4, gradient smoothing (GS) and geometry regularization (Geo-Reg) are evaluated both



Figure 4: Comparison of image-matching approaches for correspondence extraction. For scenes with larger inter-view camera motion, we employ RoMa (Edstedt et al., 2024) to obtain more reliable correspondences.

individually and jointly. The results show that each component yields measurable improvements in visual quality on its own and that applying both components together produces further gains across the evaluated quantitative metrics.

Table 4: Ablation study of model components.

GS	Geo-Reg	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
		33.98	0.93	0.08
✓		34.12	0.94	0.05
	✓	34.24	0.93	0.07
✓	✓	34.45	0.96	0.03

According to our analysis, initialization plays a key role in reconstruction quality. The 3DGS representation is sensitive to initial camera estimates and has limited ability to correct erroneously initialized Gaussian ellipsoids. Table 5 compares several initialization strategies: COLMAP, representing a conventional SfM pipeline, and MASt3r, representing a learning-based dense reconstruction method. The results demonstrate that our proposed initialization outperforms these baselines across the evaluated metrics in terms of both reconstruction accuracy and visual fidelity.

Table 5: Ablation study of initialization strategies. The “MASt3r” initialization employs correspondences extracted by MASt3r.

Method	Geo-Reg	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
COLMAP		31.76	0.91	0.09
MASt3r	✓	32.73	0.93	0.05
Ours		34.12	0.94	0.05
Ours	✓	34.45	0.96	0.03

5 CONCLUSION

3D reconstruction is a fundamental problem in computer vision. We propose Pi3DGS, an efficient yet robust framework that jointly optimizes camera poses and the 3D Gaussian Splatting (3DGS) representation from uncalibrated multi-view images. We investigate the impact of camera pose optimization on 3DGS reconstruction and further introduce two key contributions: (i) leveraging the large-scale 3D model Pi3 to enhance representation quality, and (ii) a gradient-smoothed pipeline with regularization guided by image matching. Experimental results demonstrate that our framework achieves superior performance in both novel view synthesis and pose estimation compared to existing approaches.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical

- report. *arXiv preprint arXiv:2303.08774*, 2023.
- Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.
- Jia-Wang Bian, Wenjing Bian, Victor Adrian Prisacariu, and Philip Torr. Porf: Pose residual field for accurate neural surface reconstruction. In *The Twelfth International Conference on Learning Representations*.
- 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*, pp. 4160–4169, 2023.
- Keng-Wei Chang, Zi-Ming Wang, and Shang-Hong Lai. Keygs: A keyframe-centric gaussian splatting method for monocular image sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1989–1997, 2025.
- Bo-Yu Chen, Wei-Chen Chiu, and Yu-Lun Liu. Improving robustness for joint optimization of camera pose and decomposed low-rank tensorial radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 990–1000, 2024.
- Chenxi Lola Deng and Enzo Tartaglione. Compressing explicit voxel grid representations: fast nerfs become also small. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1236–1245, 2023.
- Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MAST3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *International Conference on 3D Vision 2025*, 2025. URL <https://openreview.net/forum?id=5uw1GRBFoT>.
- Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19790–19800, 2024.
- 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.
- Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20796–20805, June 2024.
- Zhisheng Huang, Peng Wang, Jingdong Zhang, Yuan Liu, Xin Li, and Wenping Wang. 3r-gs: Best practice in optimizing camera poses along with 3dgs. *arXiv preprint arXiv:2504.04294*, 2025.
- Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5846–5854, 2021.
- 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.
- 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.
- Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8456–8465, 2023.
- Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5741–5751, 2021.

- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- Arthur Moreau, Jifei Song, Helisa Dharmo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 788–798, 2024.
- Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14335–14345, 2021.
- Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.
- Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.
- Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11358–11367, 2021.
- Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH ’23*, 2023.
- Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9773–9783, 2023a.
- 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*, pp. 5294–5306, 2025a.
- Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.
- Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.
- Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. Embodiedgen: Towards a generative 3d world engine for embodied intelligence. *arXiv preprint arXiv:2506.10600*, 2025b.
- Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3 : Scalable permutation-equivariant visual geometry learning, 2025c. URL <https://arxiv.org/abs/2507.13347>.
- Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3295–3306, 2023b.

- Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025.
- Zongxin Ye, Wenyu Li, Sidun Liu, Peng Qiao, and Yong Dou. Absgs: Recovering fine details in 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1053–1061, 2024.
- Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19447–19456, June 2024.
- Lingzhe Zhao, Peng Wang, and Peidong Liu. Bad-gaussians: Bundle adjusted deblur gaussian splatting. In *European Conference on Computer Vision*, pp. 233–250. Springer, 2024.
- Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21634–21643, 2024.
- Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS’01.*, pp. 29–538. IEEE, 2001.

A APPENDIX

A.1 USE OF LLMs

With respect to the involvement of large language models (Liu et al., 2024; Achiam et al., 2023) in this research, their use was limited to grammatical review and stylistic polishing to increase formality. They were not employed to generate scientific content, design experiments, analyze data, or interpret results; all substantive intellectual contributions were made by the authors.