

3R-GS: Best Practice in Optimizing Camera Poses Along with 3DGS

Zhisheng Huang¹ Peng Wang² Jingdong Zhang¹ Yuan Liu³ Xin Li¹
 Wenping Wang¹

¹Texas A&M University, ²Hong Kong University, ³Hong Kong University of Science and Technology

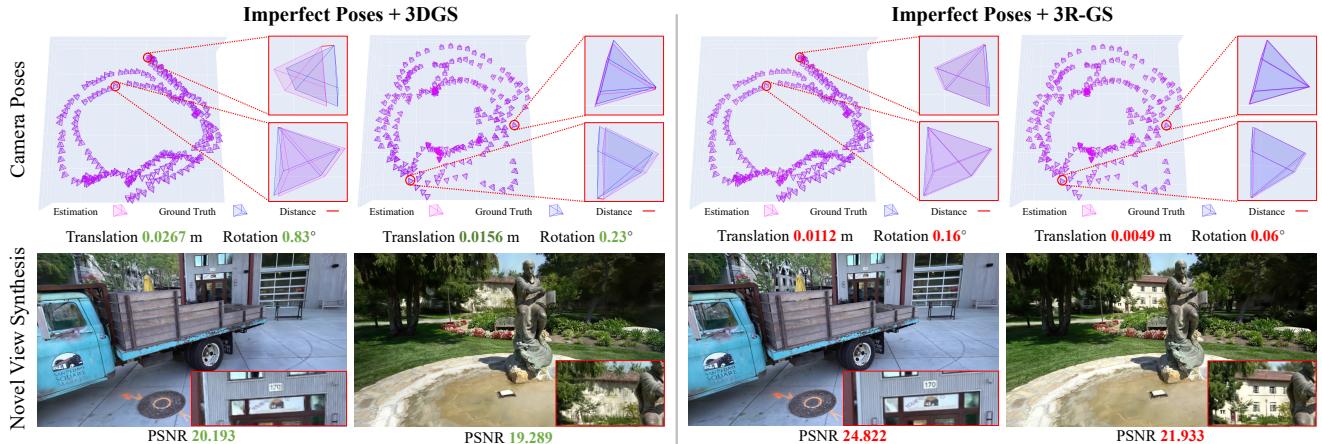


Figure 1. We propose 3R-GS, a robust method for reconstructing high-quality 3D Gaussians and poses from the MAS3R’s imperfect output cameras. Our method outperforms simply joint camera pose optimization along with 3DGS in a large margin.

Abstract

3D Gaussian Splatting (3DGS) has revolutionized neural rendering with its efficiency and quality, but like many novel view synthesis methods, it heavily depends on accurate camera poses from Structure-from-Motion (SfM) systems. Although recent SfM pipelines have made impressive progress, questions remain about how to further improve both their robust performance in challenging conditions (e.g., textureless scenes) and the precision of camera parameter estimation simultaneously. We present 3R-GS, a 3D Gaussian Splatting framework that bridges this gap by jointly optimizing 3D Gaussians and camera parameters from large reconstruction priors MAS3R-SfM. We note that naively performing joint 3D Gaussian and camera optimization faces two challenges: the sensitivity to the quality of SfM initialization, and its limited capacity for global optimization, leading to suboptimal reconstruction results. Our 3R-GS, overcomes these issues by incorporating optimized practices, enabling robust scene reconstruction even with imperfect camera registration. Extensive experiments demonstrate that 3R-GS delivers high-quality

novel view synthesis and precise camera pose estimation while remaining computationally efficient. Project page: <https://zsh523.github.io/3R-GS/>

1. Introduction

Building a 3D representation from 2D images has been a long-standing research challenge over recent decades. Recently, methods such as Neural Radiance Fields (NeRF) [32] and 3D Gaussian Splatting (3DGS) [22] have emerged as powerful approaches for 3D scene representation, particularly for novel view synthesis. These NeRF- and 3DGS-based methods require accurate camera parameters to correctly establish the 3D-2D projection relationship, a task that typically relies on structure-from-motion (SfM) techniques (e.g., COLMAP) [38]. However, SfM processes are often time-consuming and sometimes are not robust for scenes or objects with featureless regions, such as indoor environments — making them less reliable.

Recent advances in learning-based, feed-forward dense reconstruction methods (e.g., DUS3R [45] and MAST3R [12]) have demonstrated the significant potential of large models for inferring 3D structures from uncalibrated images. [12, 13, 51]. They are more robust than

traditional SfM pipelines, especially under challenging conditions such as pure rotational motions, textureless regions, and sparse view scenarios. Despite these improvements, the estimated camera poses of these 3R-based methods still lack perfect accuracy due either to limitations in the feed-forward paradigm or to oversimplified global optimization strategies. Consequently, these imperfections can degrade the performance of subsequent 3DGS training, which requires camera poses with pixel-level accuracy.

In this paper, we introduce 3R-GS, a robust method for reconstructing high-quality 3D Gaussian representations from imperfect outputs of the MASt3R camera. Our method builds on the idea of simultaneously learning camera poses and 3D Gaussian representations. However, directly applying a joint learning approach for 3DGS on imperfectly registered camera frames poses several challenges.

The first challenge is **sensitivity to initialization**. Training 3DGS requires advanced engineering heuristics, such as split/clone and opacity resetting, which require extensive hyperparameter tuning. When camera poses are imperfect, joint optimization can easily get trapped in local minima. This issue becomes more severe with recent feed-forward dense reconstruction methods, such as DUSt3R [45]. These methods often generate point clouds with low accuracy in background regions due to high depth ambiguity present in the training data.

The second challenge is **inefficient pose optimization**. Unlike NeRF-like methods, which render using ray marching with pixel-level precision, 3DGS lacks built-in mechanisms for efficiently optimizing multiple cameras in a single training step. Instead, 3DGS uses a differentiable rasterizer that, in each training step, transforms all points into the same Normalized Device Coordinates (NDC) space, then projects, sorts, and renders them onto the image space, producing full-image level rendering. While this mechanism renders a large batch of pixels at once, all pixels originate from a single camera, meaning that the optimization affects only one camera per training step. In contrast, NeRF-like methods allow more efficient camera optimization since each training step can incorporate pixels rendered from multiple cameras.

To address the first issue, we propose adopting 3DGS-MCMC [24] to enhance robustness against imperfect initialization. We view 3D Gaussians as MCMC samples drawn from a distribution that accurately represents the scene. Through state transitions, the Gaussian primitives are relocated, helping them escape local minima and improving convergence. This reduces the method’s reliance on high-quality point cloud and camera pose initialization. In addition, with 3DGS-MCMC, we eliminate the need for heuristic densification and pruning strategies in 3DGS, removing the burden of hyperparameter fine-tuning.

To address the second issue, drawing inspiration from

PoRF [2] and ACE0 [4], we improve camera pose optimization by modeling correlations between camera poses using a multilayer perceptron (MLP). Specifically, we jointly train a globally shared MLP alongside per-camera embeddings to refine camera poses. Additionally, to further enhance camera pose optimization, we incorporate an epipolar distance loss as a geometric constraint for refining camera poses. This approach directly leverages pairwise correspondences image matching to optimize camera poses. By leveraging all available pairwise correspondences, we can better optimize camera poses using more direct geometric supervising signals. To the best of our knowledge, our approach is the first to apply MLP pose modeling and epipolar loss to tackle the unique challenges of joint 3DGS and camera pose optimization.

In summary, our contributions are:

1. We propose 3R-GS, a robust method for reconstructing high-quality 3D Gaussians and poses from the MASt3R’s imperfect output cameras.
2. Identifying two main challenges in bundle-adjusting 3DGS, we propose an effective solution that combines 3DGS-MCMC, an MLP-based pose refiner, and an epipolar distance loss to address these issues.
3. Our experiments demonstrate the superior performance of 3R-GS in both novel view synthesis and camera pose estimation.

2. Related Work

2.1. Camera Pose Estimation from Images

Estimating camera poses robustly and accurately using only RGB images is a long-standing and fundamental challenge. Depending on whether the images are captured in an ordered or unordered sequence, methods such as Structure-from-Motion (SfM) or Simultaneous Localization and Mapping (SLAM) can be employed to achieve precise pose estimation.

In this paper, we primarily focus on unordered settings, as exemplified by structure-from-motion methods. Traditional SfM method, such as COLMAP [38], is a long pipeline compromising several stages like feature matching, camera registration, and bundle adjustment, is complex and not robust to some challenging cases like texture-less regions and pure rotations. Recent learning-based sparse feature extraction and matching methods [28, 34, 37, 42] have sought to improve traditional feature matching, yet their robustness still leaves room for improvement. Unlike complex pipelines, recent methods—such as DUSt3R [45] and its variants [12, 43, 44] have demonstrated the power of directly predicting 3D structures using large models. However, despite their robustness, these approaches lack pixel-level accuracy, which leads to suboptimal downstream reconstruction results.

2.2. Novel View Synthesis (NVS)

Novel view synthesis, as its name suggests, aims to generate images from unseen viewpoints, leveraging input images. This capability is pivotal in applications like virtual reality, telepresence, etc. In recent years, Neural Radiance Fields (NeRF) [32] and 3D Gaussian Splatting (3DGS) [21] have greatly improved the quality of novel view synthesis, achieving photo-realistic results. In particular, 3DGS-like methods have been at the forefront of NVS research recently due to their clear, explicit representation and real-time rendering capabilities. Several variants of 3DGS have been proposed, each targeting a specific aspect of the problem—for example, approaches for large-scale reconstruction [23, 29, 29, 35, 50, 62], feed-forward models [30, 40, 49, 58], surface reconstruction [17, 55–57], and methods for handling reflective objects [20, 52, 60]. However, these NVS methods often depend on dense, accurate camera poses obtained through SfM pipelines. When the input camera poses are inaccurate, misalignments and artifacts can occur in the synthesized views.

2.3. Joint NVS and Pose Estimation

To address the challenge of reliance on the known camera poses as described above, recent methods have integrated pose optimization, designing end-to-end frameworks for joint camera pose estimation and NVS. For example, Guo et al. [16] proposed a two-stage network that synthesizes novel views directly from a 6-DoF camera pose, decoupling geometric mapping and texture rendering to enhance robustness against variable operating conditions.

Early works [3, 6, 8, 19, 27, 41, 47, 54] on NeRF try to eliminate such requirement. Among them, NeRFmm [47] demonstrates the joint optimization of camera parameters and NeRF parameters through an empirical, two-stage pipeline. BARF [27], in contrast, introduces a single coarse-to fine registration on coordinate-based scene representation. GARF [8] and [48] employ special activation functions, alleviating issues with high-frequency positional encoding and systematic sub-optimality in NeRFmm respectively. NoPe-NeRF [3] adopts additional single view depth estimation to provide strong geometry cues. SPARF [41], SC-NeRF [19] and PoRF [2] incorporate image correspondence in the joint optimization. Though impressive results have been achieved, these methods are limited to either forward-facing scenes or short video clips with simple trajectories. Moreover, NeRF representation makes these methods slow to converge.

Recent studies have shifted focus from NeRF to 3DGS, as it enables real-time rendering, improved rendering quality and faster training speed. In the scope of joint NVS and pose estimation, CF-3DGS [15] and [39] assume a sequential video frame inputs and processes frames in a sequential manner, progressively training the 3DGS. InstantSplat

[14] leverages DUST3R [45] for camera pose initialization, but limited to very few images. ZeroGS [7] utilizes a pre-trained model as neural scene representation, enabling training 3DGS from hundreds of unposed and unordered images. However, they also features progressive training and need a two-stage strategy for convergence. BAD-Gaussian [61] and [10] also consider camera optimization in training 3DGS, but focus on addressing motion blur.

Our work is largely inspired by pioneering works on NeRF. For instance, the MLP pose refiner has been introduced by PoRF and ACE0 [4], and epipolar distance loss function has also been utilized by SC-NeRF and PoRF. However, different from them, we use these approaches to effectively handle unique problems in bundle adjusting 3DGS. And compared progressive methods in 3DGS, our methods only need to modify the standard 3DGS training pipeline slightly with negligible overhead and can be applied to both short video clips and full video sequences.

3. Method

3.1. Overview

Given a set of images captured without known poses in challenging scenes, our goal is to reconstruct both high-quality 3D Gaussian representations and accurate camera poses. Existing 3D Gaussian reconstruction methods rely heavily on accurate camera poses—typically obtained from traditional structure-from-motion techniques (e.g., COLMAP [5]) as input, and often struggle in scenarios such as textureless indoor environments.

To this end, we build on recent techniques that incorporate large reconstruction priors [12, 45], and specifically, we employ MASt3R-SfM [12] to robustly estimate camera poses.

While MASt3R-SfM outperforms traditional SfM methods such as COLMAP [38] in terms of robustness under various conditions, its estimated camera poses remain imperfect due to a lack of pixel-level accuracy, posing challenges for the downstream 3DGS reconstruction. Our 3R-GS, a joint 3DGS and camera poses learning framework from MASt3R-SfM, aims to address the above issue. However, naively optimizing the imperfect camera poses during 3DGS training leads to only limited improvements, introducing two challenges - sensitivity to initialization and inefficient pose optimization, as described in the introduction.

To address these challenges, we introduce: (1) a robust pose refinement strategy leveraging Markov Chain Monte Carlo (Section 3.2), (2) a global camera correlation model using an MLP-based refiner (Section 3.3), and (3) a rendering-free geometric constraint based on epipolar loss (Section 3.4).

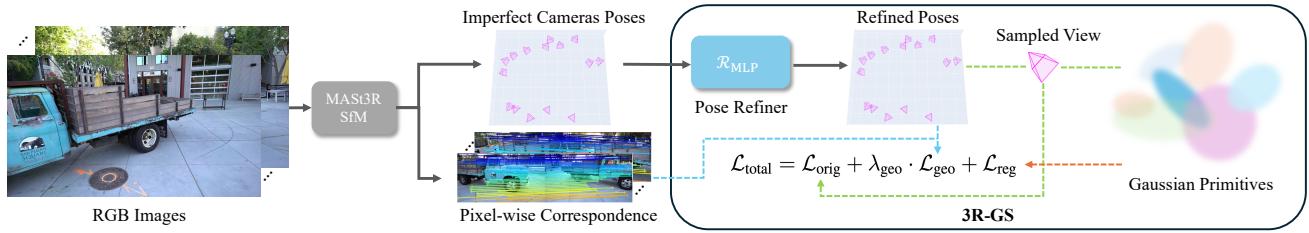


Figure 2. Overview of the 3R-GS pipeline. The pipeline jointly refines camera poses and 3D Gaussian parameters.

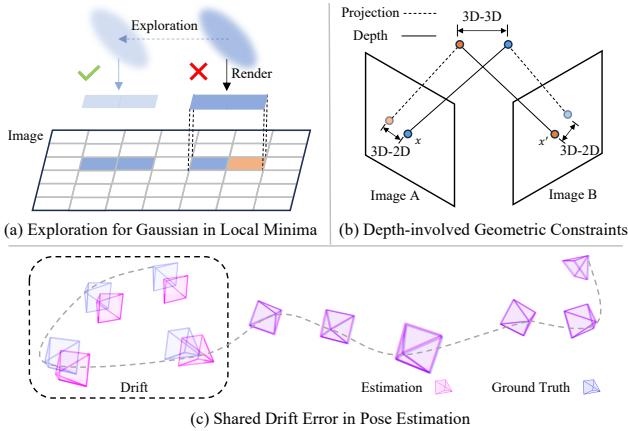


Figure 3. Motivations for 3R-GS; see Sec. 3.2, 3.4, and 3.3.

3.2. MCMC-based Pose Optimization

Motivation: Vanilla 3DGS optimization is highly sensitive to initialization because Gaussian primitives have limited adaptability and rely on accurate initial point clouds [24, 36]. For example, as shown in Fig. 3(a), if a Gaussian primitive is initially placed slightly away from its ideal position due to imperfect initialization, it may struggle to correct itself. This happens because the photometric rendering loss only provides gradients within a small local region, making it hard for the primitive to escape local optima and reach the correct position. As a result, poor initialization can lead to suboptimal convergence and degraded scene reconstruction.

Furthermore, adaptive density control in 3DGS depends on gradient magnitude-based thresholds, which require manual tuning or adjustments to the densification strategy [11] when introducing new training objectives. This reliance not only complicates optimization but also poses additional challenges when jointly optimizing camera poses with 3DGS.

Solution: We adopt 3DGS-MCMC [24], which improves robustness to initialization by reformulating 3D Gaussian Splatting as Markov Chain Monte Carlo (MCMC) sampling. This approach interprets training as sampling from a distribution $p(\mathcal{G})$ that assigns high probability to collections of Gaussians faithfully reconstructing training images. It reveals that standard 3DGS optimization resembles Stochastic Gradient Langevin Dynamics (SGLD) updates:

$$\mathcal{G} \leftarrow \mathcal{G} + a \cdot \nabla_{\mathcal{G}} \log p(\mathcal{G}) + b \cdot \eta$$

where η is exploration noise, and parameters a and b balance convergence and exploration. With this noise, ex-

ploration in Fig. 3(a) can be achieved. Moreover, 3DGS-MCMC removes the need for heuristic-based densification and pruning by replacing them with principled state transitions. We also incorporate their regularizer to promote parsimonious use of Gaussians.

With 3DGS-MCMC, we achieve robust joint optimization of camera poses and 3DGS, which addresses the “sensitivity to initialization” issue. In the following sections, we introduce two techniques to tackle the “inefficient pose optimization” challenge.

3.3. MLP-Based Global Pose Refinement

Motivation: In practice, multiple cameras often share common drift errors — while their relative poses may be correct, they collectively deviate from the ground truth with a shared rotation and translation error as shown in Fig. 3(c). However, directly optimizing individual camera poses treats them as independent, which can distort originally correct local relative poses and make the optimization more prone to local minima due to the inherent non-convexity of the problem [33].

Solution: We introduce an MLP-based global pose refiner, which learns to predict pose corrections $\Delta \mathbf{T}_i$ from a latent camera representation:

$$\Delta \mathbf{T}_i = \mathcal{R}_{\text{MLP}}(\mathbf{z}_i), \quad (1)$$

where \mathbf{z}_i is a learnable camera embedding jointly optimized with the MLP refiner. The corrections consist of translation ($\Delta \mathbf{t}_i \in \mathbb{R}^3$) and rotation ($\Delta \mathbf{r}_i \in \mathbb{R}^6$) components. The MLP is initialized with a zero-mean prior to ensure stable refinement. This formulation captures global pose relationships by using a shared MLP across all views, enabling more accurate camera adjustment. In practice, it achieves significantly better results than directly optimizing individual camera poses.

3.4. Rendering-Free Geometric Constraint

Motivation: In addition to the factor previously mentioned that contributes to inefficient camera optimization, another issue is that relying solely on rendering loss lacks direct geometric supervision for camera poses. A straightforward approach to imposing direct geometric supervision is to use correspondence-based geometric losses. We note that MAST3R-SfM provides matching correspondences, which can be potentially used for our geometric optimization.

Specifically, MAST3R-SfM constructs a sparse scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $I \in \mathcal{V}$ represents an image, and each edge $e = (n, m) \in \mathcal{E}$ denotes an undirected connection between two likely overlapping images I^n and I^m . Based on the graph, MAST3R-SfM computes correspondence matches $\mathcal{M}^{n,m}$.

To utilize the correspondences, common choices include the 3D-2D projection loss [12, 41] and 3D-3D loss [5, 12], both relying on depth as shown in Fig. 3(b). The 3D-3D loss computes distances between corresponding points in an image pair by back-projecting them to 3D space using depth and camera parameters; and 3D-2D loss re-projects these 3D points onto the image plane to compute 2D distances to their correspondences. These methods usually require multiple image pairs to simulate global bundle adjustment, ensuring more consistent gradients. However, integrating these optimization targets into 3DGS training presents significant challenges, primarily because 3DGS employs per-view depth sorting for rendering (both RGB and depth) and requires tens of thousands of iterations for training. This computational constraint severely limits the number of views that can be processed in each step. Incorporating additional views substantially increases training time and memory consumption to prohibitive levels. Consequently, only a subset of image pairs can be considered in each step when applying the aforementioned geometric constraints, preventing the enforcement of a truly global objective and ultimately leading to suboptimal results.

Solution: We propose a rendering-free global geometric constraint based on epipolar distances. Given image correspondences $\mathcal{M}^{n,m}$ from MAST3R-SfM, we define the loss as:

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{E}|} \sum_{(n,m) \in \mathcal{E}} \frac{1}{|\mathcal{M}^{n,m}|} \sum_{(x_i, x'_i) \in \mathcal{M}^{n,m}} \text{conf}_i \cdot d(x_i, x'_i) \quad (2)$$

where conf_i is confidence provided by MAST3R for correspondence (x, x') , and $d(x, x')$ is the symmetric epipolar distance computed from the fundamental matrix F , which is derived from the camera poses and intrinsics. Unlike PoRF [2], we consider the correspondences from all image pairs $(n, m) \in \mathcal{E}$ in MAST3R-SfM during each training iteration. This enables a more globally informed joint optimization of camera pose. While MAST3R-SfM can provide thousands of correspondences for each image pair, we empirically find that only a few hundred are necessary, so we subsample the correspondences uniformly.

3.5. Final Training Objective

Our complete training objective integrates the original 3D Gaussian Splatting rendering losses with additional regularization terms from 3DGS-MCMC [24], along with our geometric constraints \mathcal{L}_{geo} in Eq. 2.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{orig}} + \lambda_{\text{geo}} \cdot \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{reg}} \quad (3)$$

The original 3DGS training loss $\mathcal{L}_{\text{orig}}$ follows [22]:

$$\mathcal{L}_{\text{orig}} = (1 - \lambda_{\text{D-SSIM}}) \cdot \mathcal{L}_1 + \lambda_{\text{D-SSIM}} \cdot \mathcal{L}_{\text{D-SSIM}} \quad (4)$$

where \mathcal{L}_1 measures L1 color error and $\mathcal{L}_{\text{D-SSIM}}$ measures structural similarity, with $\lambda_{\text{D-SSIM}} = 0.2$. The regularization term \mathcal{L}_{reg} in 3DGS-MCMC promotes efficient use of Gaussians by encouraging fewer Gaussians:

$$\mathcal{L}_{\text{reg}} = \lambda_o \cdot \sum_i |o_i|_1 + \lambda_{\Sigma} \cdot \sum_{ij} |\sqrt{\text{eig}_j(\Sigma_i)}|_1 \quad (5)$$

where $\text{eig}_j(\cdot)$ denotes the j -th eigenvalue of the covariance matrix.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on three widely used real-world datasets: Tanks and Temples [25], Mip-NeRF360 [1], and DTU [18], selecting four representative scenes from each. The Mip-NeRF360 dataset contains indoor and outdoor scenes captured with cameras distributed evenly along 360-degree trajectories, with each scene comprising approximately 100–300 images. Tanks and Temples follows a similar setup in terms of camera poses and scene scale but exhibits greater variations in illumination and appearance. In contrast, DTU focuses on object-level indoor scenes captured under controlled lighting, with each sequence containing 49 or 64 images and precise ground truth poses. These datasets are widely used in the 3D Gaussian Splatting literature [17, 22]. Following prior work, we adopt the same evaluation protocol and training view resolution for each dataset.

Metrics. Following BARF [27] and CF-3DGS [15], we evaluate both Novel View Synthesis (NVS) and camera pose registration. For camera pose evaluation, we report the average rotation error and the Root Mean Square Error (RMSE) of the Absolute Trajectory Error (ATE) [31] (in meters) on the training views. To account for similarity transformations, we align the optimized training poses with the ground truth using Procrustes analysis on camera locations, following prior work [27]. For NVS, we report PSNR, SSIM [46], and LPIPS [59]. Since NVS requires test view poses, we perform test-time rendering optimization to obtain optimal test poses, consistent with previous approaches [3, 15, 27].

Implementation details. Our method is implemented in PyTorch, building upon the 3D Gaussian Splatting framework gsplat [53]. For all experiments, we employ consistent weighting factors: $\lambda_{\text{D-SSIM}} = 0.2$, $\lambda_o = 0.01$, $\lambda_{\Sigma} = 0.01$,

Scenes	3DGS			Spann3R			ZeroGS			CF-3DGS			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Truck	20.91	0.723	0.181	10.67	0.398	0.863	-	-	-	-	-	-	24.82	0.860	0.121
Ignatius	18.96	0.665	0.249	13.32	0.298	0.589	21.95	0.665	0.234	-	-	-	21.93	0.778	0.198
Caterpillar	19.29	0.539	0.349	12.57	0.348	0.720	-	-	-	12.96	0.340	0.616	23.37	0.773	0.235
Meetingroom	22.78	0.784	0.239	11.87	0.462	0.834	-	-	-	-	-	-	25.93	0.867	0.177
garden	24.85	0.729	0.126	18.13	0.281	0.485	25.47	0.839	0.107	-	-	-	26.44	0.82	0.131
counter	27.57	0.862	0.209	15.02	0.537	0.632	26.87	0.873	0.124	-	-	-	28.80	0.897	0.157
bicycle	17.52	0.303	0.567	16.09	0.256	0.634	23.10	0.707	0.201	-	-	-	24.89	0.727	0.252
room	30.66	0.899	0.204	14.06	0.563	0.709	-	-	-	-	-	-	31.82	0.924	0.154
scan69	26.37	0.865	0.134	15.76	0.447	0.565	-	-	-	18.09	0.554	0.521	26.62	0.868	0.112
scan83	28.36	0.882	0.172	20.45	0.759	0.321	-	-	-	12.81	0.572	0.546	28.44	0.881	0.117
scan106	32.74	0.923	0.109	20.30	0.664	0.379	-	-	-	18.00	0.550	0.530	34.35	0.936	0.066
scan110	31.46	0.905	0.142	21.58	0.752	0.323	-	-	-	18.87	0.644	0.482	32.63	0.931	0.074

Table 1. Quantitative comparison of novel view synthesis. (-) denotes unreported results for ZeroGS and failed scenes for CF-3DGS.

Scenes	3DGS		Spann3R		ZeroGS		CF-3DGS		Ours	
	Rotation(°)↓	ATE(m)↓								
Truck	0.83	0.027	51.69	3.156	-	-	-	-	0.16	0.011
Ignatius	0.23	0.016	5.87	0.391	0.03	0.002	-	-	0.06	0.005
Caterpillar	1.41	0.402	10.95	0.695	-	-	82.50	3.743	0.32	0.020
Meetingroom	0.75	0.052	26.14	2.021	-	-	-	-	0.24	0.023
garden	0.19	0.003	2.08	0.147	0.03	0.002	-	-	0.03	0.002
counter	0.25	0.011	4.08	0.332	0.03	0.002	-	-	0.05	0.003
bicycle	1.07	0.034	11.11	1.516	0.04	0.005	-	-	0.09	0.013
room	0.27	0.016	8.46	0.908	-	-	-	-	0.13	0.012
scan69	0.23	0.006	5.10	0.158	-	-	47.95	0.955	0.1	0.003
scan83	0.26	0.007	3.04	0.184	-	-	155.34	1.286	0.19	0.005
scan106	0.13	0.004	4.09	0.121	-	-	46.40	0.902	0.11	0.003
scan110	0.48	0.007	3.00	0.129	-	-	66.78	0.983	0.13	0.004

Table 2. Quantitative comparison of camera pose registration. (-) indicates unreported results for ZeroGS and failed scenes for CF-3DGS.

and $\lambda_{\text{geo}} = 2$. We find that the epipolar geometric constraint \mathcal{L}_{geo} plays a crucial role in the early training phase of 3DGS, helping to establish correct geometry, its influence becomes less critical in later stages. Thus we decay λ_{geo} to 0 after 3,000 iterations. This strategy is consistently applied across all our experiments. Our training procedure follows the standard 3DGS pipeline [17, 22], where each iteration samples and renders a single training view. The key distinction is that we additionally compute \mathcal{L}_{geo} at each step and propagate camera pose gradients through our pose refiner and associated camera latent codes. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB memory.

4.2. Comparison with Previous Methods

Results on full video sequence. We evaluate our method against four state-of-the-art baselines: 3DGS [22], Spann3R [43], ZeroGS [7], and CF-3DGS [15]. The evaluation is conducted on three standard datasets: Tanks and Temples [25], Mip-NeRF360 [1], and DTU [18]. For 3DGS, we utilize the gsplat implementation [53]. To ensure fair comparison, we configure 3DGS to use the same camera poses obtained from MASt3R-SfM and enable camera pose optimization during training, as supported by gsplat [53]. Since Spann3R employs a different scene representation, we use its estimated camera poses for 3DGS training with camera optimization enabled, maintaining consistency with our experimental setup. For ZeroGS, which lacks publicly available code at the time of writing, we report results directly from their paper. CF-3DGS experiments are conducted using their official implementation across all three datasets.

Table 1 and Fig. 4 present the novel view synthesis results. Our approach demonstrates superior performance across all three datasets, significantly outperforming 3DGS, which uses identical MASt3R-SfM camera poses. While ZeroGS shows slightly lower performance compared to our method, CF-3DGS consistently fails on Mip-NeRF360 and Tanks and Temples datasets. This failure can be attributed to early camera tracking loss during their progressive training pipeline, particularly evident in scenes with large camera motion.

For camera registration (Table 2 and Fig. 5), our method significantly outperforms 3DGS with camera pose optimization, while showing comparable results to ZeroGS. The marginal difference (0.02° in rotation error and 0.003 m in ATE on average) is negligible, especially considering our superior novel view synthesis results. Moreover, while ZeroGS employs a complex two-stage training strategy with progressive image registration similar to classical incremental SfM, our approach achieves competitive results through a simpler process that enhances standard 3DGS training using SfM outputs, introducing minimal computational overhead.

Results on short video clips. We also evaluate our method on short video clips from the Tanks and Temples dataset, following the experimental protocol of CF-3DGS[15] and using their preprocessed data. For comparison, we select state-of-the-art baselines including BARF [27], SC-NeRF [19], and Nope-NeRF [3].

As shown in Table 3 and Table 4, our approach demonstrates substantial improvements over existing methods in both novel view synthesis and camera pose estimation. For

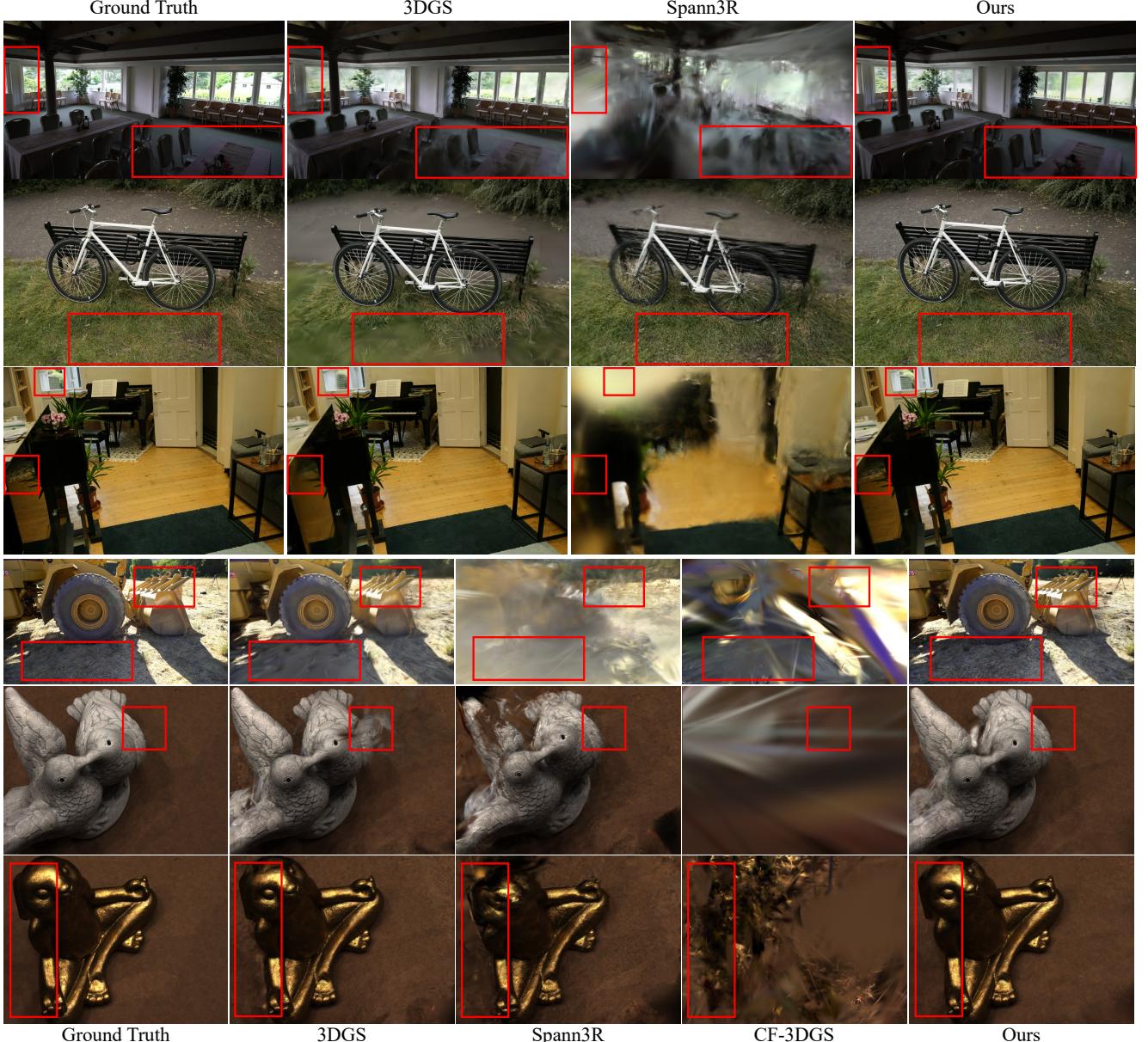


Figure 4. Results for novel view synthesis. We omit the failure scenes for CF-3DGS and unreported results for ZeroGS.

Scenes	BARF			SC-NeRF			Nope-NeRF			CF-3DGS			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Barn	25.28	0.64	0.48	23.26	0.62	0.51	26.35	0.69	0.44	31.23	0.93	0.11	36.95	0.97	0.01
Museum	23.58	0.61	0.55	24.94	0.69	0.45	26.77	0.76	0.35	29.91	0.91	0.11	35.78	0.97	0.01
Ballroom	20.66	0.50	0.60	22.64	0.61	0.48	25.33	0.72	0.38	32.47	0.96	0.07	34.56	0.97	0.01
Ignatius	21.78	0.47	0.60	23.00	0.55	0.53	23.96	0.61	0.47	28.43	0.90	0.09	31.16	0.93	0.03

Table 3. Quantitative comparison of novel view synthesis on short video clips.

fair comparison, we follow the evaluation protocol from prior work, where the Absolute Trajectory Error (ATE) is scaled by a factor of 100, and camera pose alignment with ground truth is performed using both translation and rotation. All quantitative results are obtained using the official evaluation code provided by [15].

Comparison with COLMAP. While both the Tanks and Temples dataset and MipNeRF360 use COLMAP-generated camera poses as ground truth due to its gener-

ally high accuracy, we identify specific scenarios where COLMAP faces challenges. To demonstrate this, we evaluate several challenging scenes from the ScanNet [9] dataset, with results presented in Table 5. Our method demonstrates superior performance compared to COLMAP on these scenes, highlighting the limitations of traditional structure-from-motion approaches in challenging scenarios.

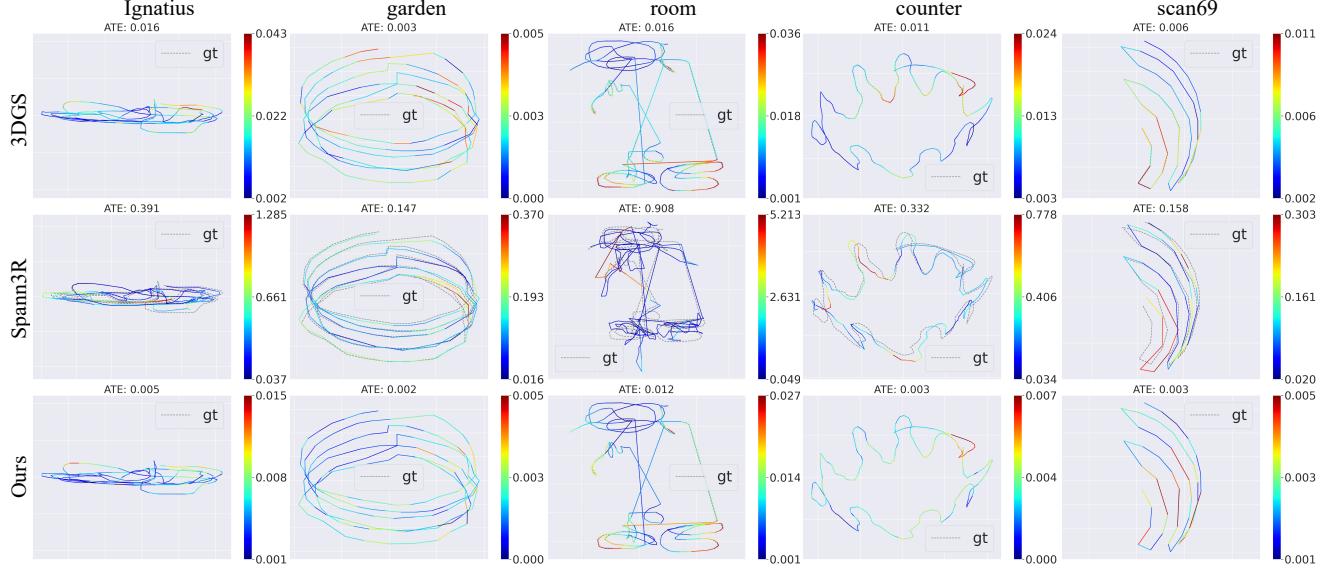


Figure 5. Visualization of camera pose registration for the three best-performing methods. ZeroGS results are unavailable.

scenes	BARF			SC-NeRF			None-NeRF			CF-3DGS			Ours		
	RPE _t ↓	RPE _r ↓	ATE ↓	RPE _t ↓	RPE _r ↓	ATE ↓	RPE _t ↓	RPE _r ↓	ATE ↓	RPE _t ↓	RPE _r ↓	ATE ↓	RPE _t ↓	RPE _r ↓	ATE ↓
Barn	0.314	0.265	0.050	1.317	0.429	0.157	0.046	0.032	0.004	0.034	0.034	0.003	0.009	0.020	0.000
Museum	3.442	1.128	0.263	8.339	1.491	0.316	0.207	0.202	0.020	0.052	0.215	0.005	0.018	0.020	0.000
Ballroom	0.531	0.228	0.018	0.328	0.146	0.012	0.041	0.018	0.002	0.037	0.024	0.003	0.016	0.013	0.000
Ignatius	0.736	0.324	0.029	0.533	0.240	0.085	0.026	0.005	0.002	0.033	0.032	0.005	0.010	0.012	0.001

Table 4. Quantitative comparison of camera pose registration on short video clips.

Scenes	COLMAP			Ours		
	Pose		PSNR↑	Pose		PSNR↑
	R(°)↓	ATE(m)↓	R(°)↓	ATE(m)↓	R(°)↓	PSNR↑
0079.00	3.55	0.014	30.78	2.45	0.014	32.58
0301.00	133.83	0.169	23.63	9.30	0.009	30.11
0418.00	5.03	0.012	29.03	4.34	0.012	31.62

Table 5. Comparison with COLMAP on ScanNet.

3DGS	MCMC	\mathcal{R}_{MLP}	\mathcal{L}_{geo}	Pose		NVS		
				R(°)↓	ATE(m)↓	PSNR↑	SSIM↑	LPIPS↓
✓				0.805	0.124	20.49	0.678	0.255
✓	✓			0.783	0.032	21.94	0.724	0.222
✓	✓	✓		0.353	0.025	23.61	0.803	0.193
✓	✓	✓	✓	0.195	0.015	24.02	0.820	0.183
Geometric Constraints Only				0.430	0.038	21.94	0.731	0.239
Local Geometric Constraints				0.458	0.034	21.97	0.736	0.228

Table 6. Ablation Study on Tanks and Templates.

4.3. Ablation Studies

Component-wise analysis. Our method comprises three key components: 1) 3DGS as MCMC, 2) rendering-free global geometric constraints, and 3) a correlation-modeling global camera pose refiner. To evaluate each component’s contribution, we conduct ablation studies as shown in Table 6. We evaluate on scenes from the Tanks and Temples dataset (Table 1), reporting average metrics for both novel view synthesis and camera pose estimation. The results demonstrate significant improvements from each component. While the baseline 3DGS method includes camera pose optimization during training (provided by gsplat [53]), its effectiveness is limited without our proposed components. Detailed per-scene ablation results for Tables 1 and 2 are provided in the supplementary materials.

Synergistic effects of photometric and geometric losses.

While Table 6 validates the effectiveness of our geometric constraints \mathcal{L}_{geo} , we further investigate the role of photometric rendering in camera pose registration and NVS. We conduct an experiment by detaching camera pose parameters during gaussian splitting, eliminating gradients from the photometric rendering to camera parameters. Additionally, we maintain λ_{geo} constant instead of decaying it to 0 after step 3000 as described in Section 4.1. Results in the second-to-last row of Table 6 demonstrate that the rendering loss in Eq. 3 is crucial for both high-quality NVS and accurate camera pose registration. The optimal performance is achieved through the synergistic combination of rendering loss and geometric constraints.

Advantages of global over Local geometric constraints.

To demonstrate the superiority of global geometric constraints over local alternatives, we compare against a variant that randomly samples each image pair for correspondence at every step, rather than utilizing all pairs $(n, m) \in \mathcal{E}$. The results, shown in the last row of Table 6, indicate that local geometric constraints provide minimal benefit compared to our global approach, validating the effectiveness of \mathcal{L}_{geo} .

5. Conclusion

We present 3R-GS, a robust framework for optimizing camera poses and 3D Gaussian representations from imperfect MAST3R-SfM outputs. Experimental results demonstrate our 3R-GS achieves superior performance in both novel view synthesis and camera pose registration compared to prior work. We hope this will benefit 3R-based [26, 45]

methods and their downstream application with 3DGS in community. Possible future work could explore extending our approach to handle dynamic scenes and support real-time applications.

References

- [1] 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. [5](#), [6](#)
- [2] Jia-Wang Bian, Wenjing Bian, Victor Adrian Prisacariu, and Philip Torr. Porf: Pose residual field for accurate neural surface reconstruction. *arXiv preprint arXiv:2310.07449*, 2023. [2](#), [3](#), [5](#)
- [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. [3](#), [5](#), [6](#)
- [4] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Pos-ing of image collections via incremental learning of a relocalizer. In *European Conference on Computer Vision*, pages 421–440. Springer, 2024. [2](#), [3](#)
- [5] Shu Chen, Yang Zhang, Yixin Xu, and Beiji Zou. Structure-aware nerf without posed camera via epipolar constraint. *arXiv preprint arXiv:2210.00183*, 2022. [3](#), [5](#)
- [6] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8264–8273, 2023. [3](#)
- [7] Yu Chen, Rolando Alexandros Potamias, Evangelos Ververas, Jifei Song, Jiankang Deng, and Gim Hee Lee. Zerogs: Training 3d gaussian splatting from unposed images. *arXiv preprint arXiv:2411.15779*, 2024. [3](#), [6](#)
- [8] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *Eu-ropean Conference on Computer Vision*, pages 264–280. Springer, 2022. [3](#)
- [9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. [7](#)
- [10] François Darmon, Lorenzo Porzi, Samuel Rota-Bulò, and Peter Kortschieder. Robust gaussian splatting. *arXiv preprint arXiv:2404.04211*, 2024. [3](#)
- [11] Xiaobiao Du, Yida Wang, and Xin Yu. Mvgs: Multi-view-regulated gaussian splatting for novel view synthesis. *arXiv preprint arXiv:2410.02103*, 2024. [4](#)
- [12] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. [1](#), [2](#), [3](#), [5](#)
- [13] Sven Elflein, Qunjie Zhou, Sérgio Agostinho, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. *arXiv preprint arXiv:2501.14914*, 2025. [1](#)
- [14] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2(3):4, 2024. [3](#)
- [15] 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*, pages 20796–20805, 2024. [3](#), [5](#), [6](#), [7](#)
- [16] Xiang Guo, Bo Li, Yuchao Dai, Tongxin Zhang, and Hui Deng. Novel view synthesis from only a 6-dof camera pose by two-stage networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5028–5035. IEEE, 2021. [3](#)
- [17] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. [3](#), [5](#), [6](#)
- [18] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. [5](#), [6](#)
- [19] 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*, pages 5846–5854, 2021. [3](#), [6](#)
- [20] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussian-shader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. [3](#)
- [21] 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. [3](#)
- [22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [1](#), [5](#), [6](#)
- [23] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. [3](#)
- [24] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *Advances in Neural In-*

- formation Processing Systems*, 37:80965–80986, 2025. 2, 4, 5
- [25] 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. 5, 6
- [26] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 8
- [27] 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*, pages 5741–5751, 2021. 3, 5, 6
- [28] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 2
- [29] Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 3
- [30] Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Namias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiewei Li. Matrix3d: Large photogrammetry model all-in-one. *arXiv preprint arXiv:2502.07685*, 2025. 3
- [31] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 5
- [32] 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. 1, 3
- [33] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, pages 58–77. Springer, 2024. 4
- [34] Guilherme Potje, Felipe Cadar, André Araújo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 2
- [35] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Lining Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *arXiv preprint arXiv:2403.17898*, 2024. 3
- [36] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. Revising densification in gaussian splatting. In *European Conference on Computer Vision*, pages 347–362. Springer, 2024. 4
- [37] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2
- [38] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 3
- [39] Wei Sun, Xiaosong Zhang, Fang Wan, Yanzhao Zhou, Yuan Li, Qixiang Ye, and Jianbin Jiao. Correspondence-guided sfm-free 3d gaussian splatting for nvs. *arXiv preprint arXiv:2408.08723*, 2024. 3
- [40] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3
- [41] 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. 3, 5
- [42] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2
- [43] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 6
- [44] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2
- [45] 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*, pages 20697–20709, 2024. 1, 2, 3, 8
- [46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [47] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv e-prints*, pages arXiv–2102, 2021. 3
- [48] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. *arXiv preprint arXiv:2210.04553*, 2022. 3
- [49] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 3
- [50] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024. 3
- [51] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt

- Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *arXiv preprint arXiv:2501.13928*, 2025.
- [52] Keyang Ye, Qiming Hou, and Kun Zhou. 3d gaussian splatting with deferred reflection. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024.
- [53] 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. *arXiv preprint arXiv:2409.06765*, 2024.
- [54] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inertf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021.
- [55] Mulin Yu, Tao Lu, Lining Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction. *Advances in Neural Information Processing Systems*, 37:129507–129530, 2025.
- [56] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 43(6):1–13, 2024.
- [57] Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiaoxiao Long, and Ping Tan. Rade-gs: Rasterizing depth in gaussian splatting. *arXiv preprint arXiv:2406.01467*, 2024.
- [58] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.
- [59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [60] Youjia Zhang, Anpei Chen, Yumin Wan, Zikai Song, Junqing Yu, Yawei Luo, and Wei Yang. Ref-gs: Directional factorization for 2d gaussian splatting. *arXiv preprint arXiv:2412.00905*, 2024.
- [61] Lingzhe Zhao, Peng Wang, and Peidong Liu. Bad-gaussians: Bundle adjusted deblur gaussian splatting. In *European Conference on Computer Vision*, pages 233–250. Springer, 2024.
- [62] 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*, pages 21634–21643, 2024.