LongSplat: Robust Unposed 3D Gaussian Splatting for Casual Long Videos

Chin-Yang Lin[1] Cheng Sun[2] Fu-En Yang[2]
Min-Hung Chen[2] Yen-Yu Lin[1] Yu-Lun Liu[1]

[1]National Yang Ming Chiao Tung University [2]NVIDIA Research

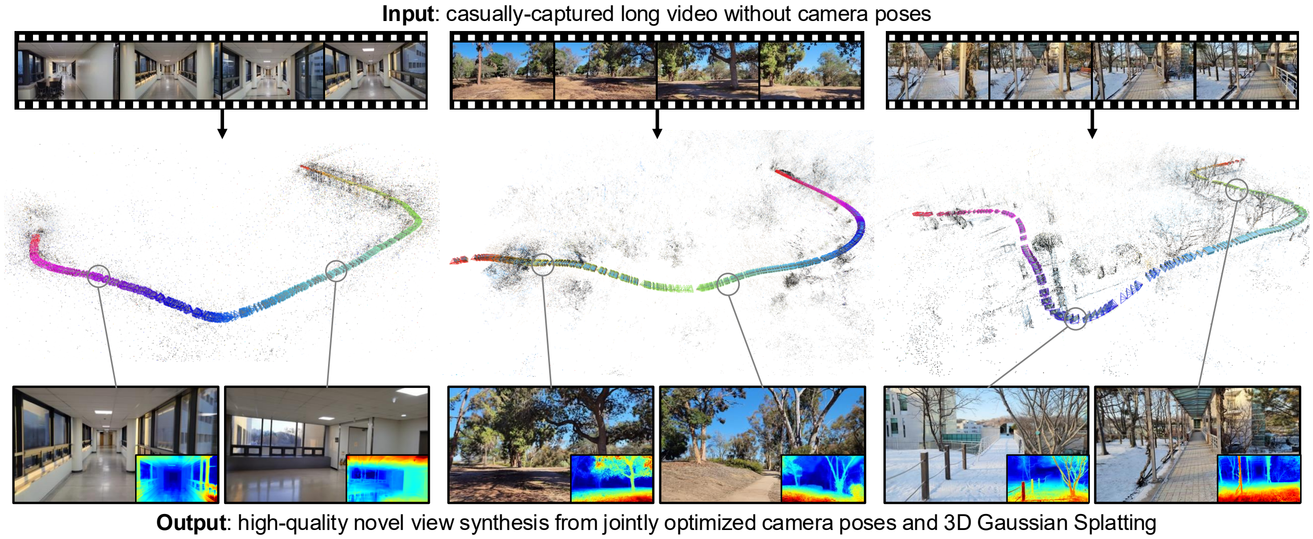

Input: casually-captured long video without camera poses
Output: high-quality novel view synthesis from jointly optimized camera poses and 3D Gaussian Splatting

Figure 1. **LongSplat achieves robust novel view synthesis from casually captured long videos without provided camera poses.** Our approach jointly optimizes camera poses and 3D Gaussian Splatting, producing accurate and visually coherent reconstructions even under challenging conditions.

Abstract

LongSplat addresses critical challenges in novel view synthesis (NVS) from casually captured long videos characterized by irregular camera motion, unknown camera poses, and expansive scenes. Current methods often suffer from pose drift, inaccurate geometry initialization, and severe memory limitations. To address these issues, we introduce LongSplat, a robust unposed 3D Gaussian Splatting framework featuring: (1) Incremental Joint Optimization that concurrently optimizes camera poses and 3D Gaussians to avoid local minima and ensure global consistency; (2) a Pose Estimation Module leveraging learned 3D priors; and (3) an adaptive Octree Anchor Formation mechanism that dynamically adjusts anchor densities, significantly reducing memory usage. Extensive experiments on challenging benchmarks demonstrate that LongSplat achieves state-of-the-art results, substantially improving rendering quality, pose accuracy, and computational efficiency compared to prior approaches. Project page: https://linjohnss.github.io/longsplat/

1. Introduction

High-quality 3D reconstruction and novel view synthesis (NVS) are essential for applications such as virtual reality, augmented reality, virtual tourism, and cultural heritage preservation. They also play a crucial role in video editing tasks like stabilization, visual effects, and digital mapping for real estate or pedestrian-level navigation. With the widespread availability of smartphones and action cameras, casually captured videos have emerged as a significant source of 3D content. Unlike professionally acquired datasets, casual videos present challenging characteristics: irregular camera trajectories, long sequences spanning hundreds or thousands of frames, and the absence of reliable camera poses or precise geometric priors.

Addressing novel view synthesis (NVS) from casually captured videos poses two critical challenges: accurate camera pose estimation over extended trajectories and efficient representation of large-scale scenes. Traditional methods rely on precise poses from Structure-from-Motion (SfM) preprocessing, yet as shown in Fig. 2, pipelines like COLMAP [50] frequently fail in casual settings. COLMAP-free methods, such as CF-3DGS [14], often encounter severe memory constraints, limiting their effectiveness for large-scale scenarios. Foundation models like MASt3R [27] provide fast initial estimates but suffer inaccuracies and drift in long videos, severely affecting reconstruction quality. Similarly, methods like LocalRF [39] struggle with complex camera trajectories, resulting in fragmented reconstructions.

To address these limitations, we introduce **LongSplat**, a robust unposed 3D Gaussian Splatting (3DGS) [22] framework designed specifically for casual long videos. As illustrated in Fig. 1, LongSplat achieves accurate novel view synthesis without relying on provided camera poses. LongSplat departs from traditional pipelines by jointly optimizing camera poses and 3DGS in a unified framework. It integrates a correspondence-guided Pose Estimation Module with 3DGS geometry and photometric refinements to improve pose accuracy, even under large-scale and unstructured camera motion. Furthermore, an adaptive Octree Anchor Formation mechanism dynamically adjusts anchor densities in the 3DGS representation, drastically reducing memory usage while preserving detailed scene structures. These components work together in an incremental joint optimization strategy that avoids local minima and ensures global geometric consistency across extensive sequences.

Extensive experiments on challenging datasets, including Tanks and Temples, Free, and Hike datasets, demonstrate that LongSplat consistently outperforms existing approaches, significantly improving rendering quality and pose accuracy. Compared to conventional methods shown in Fig. 2, LongSplat produces clearer and more coherent reconstructions, effectively addressing pose drift and memory limitations and substantially advancing the state-of-the-art. The main contributions of our work are:

- An incremental joint optimization approach for simultaneous camera pose and 3DGS reconstruction, reducing local minima and ensuring global consistency.
- A robust pose estimation module leveraging learned 3D priors for accurate camera pose estimation.
- An adaptive Octree Anchor Formation strategy that significantly reduces memory usage while preserving reconstruction quality.

2. Related Work

Novel View Synthesis. Novel View Synthesis (NVS) generates new perspectives from captured images, evolving from early pixel interpolation methods [8] to depth-based warping techniques [28] and 3D reconstruction-based rendering [6, 12]. Various representations have been explored, including planes [17, 18], meshes [20, 47, 48], point clouds [67, 74], and multi-plane images [29, 57, 76]. Neural Radiance Fields (NeRF) [40] revolutionized photorealistic rendering, with subsequent improvements in anti-aliasing [2–4, 73], reflectance [1, 58], sparse view training [24, 43, 66, 68], faster training [41, 45, 49], and rendering speed [15, 34, 53, 71]. Recent works have extended NeRF to few-shot scenarios without learned priors [32], domain-specific applications such as autonomous driving environments [51], and dynamic scenes with human pose variations [38]. Point-based methods [22, 37, 67, 74], particularly 3D Gaussian Splatting (3DGS) [22], enable real-time rendering through explicit representations. Recent advances have extended 3DGS capabilities to dynamic specular scenes with physically-based rendering [13], developed compression techniques for efficient storage and transmission [72], and improved robustness for unconstrained image scenarios [19]. However, most approaches still rely on pre-computed camera parameters from SfM [16, 32, 42, 50, 55].

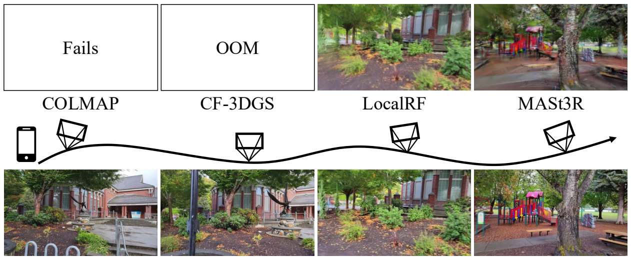


Figure 2. **Motivation for LongSplat.** Existing methods encounter significant challenges when reconstructing scenes from casually captured long videos: COLMAP [50] fails due to incorrect camera pose estimation, CF-3DGS [14] suffers from out-of-memory issues, MASt3R [27]+Scaffold-GS [36] provides inaccurate poses leading to degraded rendering quality, and LocalRF [39] struggles with error drift. In contrast, LongSplat robustly handles these challenges, yielding accurate camera poses and high-quality novel view synthesis without memory constraints.

Unposed Novel View Synthesis. Recent work has aimed to eliminate camera estimation preprocessing. i-NeRF [70] predicts camera poses using pre-trained NeRF. NeRFmm [64] jointly optimizes NeRF and camera poses for forward-facing scenes, with SiNeRF [65] offering improvements. BARF [31] and GARF [10] address gradient inconsistency through coarse-to-fine positional encoding but require good initialization. Advanced approaches [5, 9, 35, 39] leverage pre-trained networks for geometric priors, with NoPe-NeRF [5] incorporating monocular depth priors and CF-3DGS [14] using progressive optimization. Recent methods have improved robustness in joint optimization of camera poses and scene geometry using decomposed low-rank tensorial representations [7] and dynamic radiance fields [35]. These methods typically assume small pose perturbations [10, 31], limited camera motion [64, 65], or additional priors [5, 11, 21, 39], struggling with challenging trajectories, like Free dataset[26, 44].

Large-scale Novel View Synthesis. Extending NVS to large-scale environments introduces memory and computational

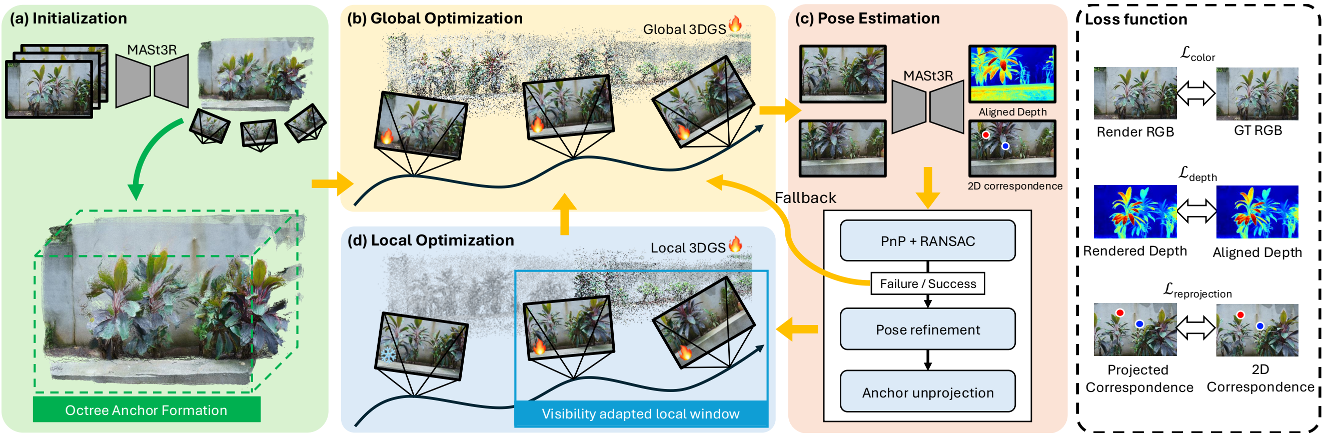


Figure 3. **Overview of the LongSplat framework.** Given a casually captured long video without known poses, LongSplat incrementally reconstructs the scene through tightly coupled pose estimation and 3D Gaussian Splatting. (a) Initialization converts MASt3R [27] depth and correspondences into an octree-anchored 3DGS. (b) Global Optimization jointly refines all camera poses and Gaussians for global consistency. (c) Frame Insertion estimates each new frame pose via correspondence-guided PnP, updates octree anchors using unprojected points, and applies photometric refinement. If PnP fails, a fallback triggers global re-optimization to recover. (d) Incremental Optimization alternates between Local Optimization within a visibility-adapted window and periodic Global Optimization to propagate consistent updates across frames.

challenges that NeRF's implicit global representation struggles with. Recent research employs scene partitioning strategies for managing large scenes [23, 54, 56]. Progressive optimization techniques have been developed for robust view synthesis in large-scale scenes from casually captured videos [39]. At the same time, MVS-based approaches have been enhanced to handle generalizable view synthesis at scale [52]. For indoor environments, methods like GenRC [30] enable room-scale 3D reconstruction from sparse image collections. 3DGS offers explicit representation advantages through Gaussian primitive. VastGaussian [33] divides scenes into separately optimized blocks[22]. Scaffold-GS [36] introduces anchor-based Gaussian representation with fixed-resolution grids, though it requires SfM initialization. Octree-GS [46] implements fixed-level octrees with preset resolutions but similarly depends on SfM. Unlike these approaches, our method dynamically adjusts voxel size based on point cloud density, without dependency on SfM, and addresses unposed, large-scale, casually captured videos through adaptive Octree Anchor Formation.

Casual Long Videos. Casual long videos present unique challenges: free-moving trajectories, lack of pose information, and continuously expanding scenes. LocalRF [39] addresses these through progressive localized field construction but suffers from slow training and fragmentation under irregular camera movements. 3D Foundation Models [59], including DUSt3R [62], MASt3R [27], Fast3r [69], and CUT3R [61], estimate poses and geometry directly but accumulate errors in long sequences. LongSplat treats foundation model outputs as soft priors, jointly optimizing them with 3D Gaussian Splatting while progressively correcting poses and geometry through combined PnP and optimization strategies.

3. Method

LongSplat reconstructs long video sequences with unknown camera poses and unconstrained trajectories through a fully incremental pipeline based on octree-anchored 3D Gaussian Splatting. The process begins with octree anchor formation, where per-frame dense point clouds are structured into an adaptive representation. Next, camera poses are estimated and refined using correspondence-guided initialization and photometric alignment. Finally, the reconstruction alternates between local optimization, which updates Gaussians within a visibility-adapted window, and global refinement, which ensures long-term consistency. This design allows LongSplat to robustly handle long, unconstrained trajectories while adapting to scene complexity and minimizing drift.

3.1. Preliminaries

Gaussian Splatting. 3D Gaussian Splatting (3DGS) [22] represents the scene as a set of 3D Gaussians, each defined by a center $\mu \in \mathbb{R}^3$, a covariance matrix Σ, color, scale, rotation, and opacity. The covariance is factorized into a rotation $R \in SO(3)$ and a diagonal scale matrix S, giving:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \quad \Sigma = RSS^\top R^\top. \tag{1}$$

This parameterization allows each Gaussian to adaptively capture local scene geometry.

To render the scene, each Gaussian is projected onto the image plane using the camera pose W, resulting in a 2D Gaussian with covariance $\Sigma_{\text{2D}} = JW\Sigma W^\top J^\top$, where J is the Jacobian of the projective transformation. The final rendered color and depth are computed via alpha blending:

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

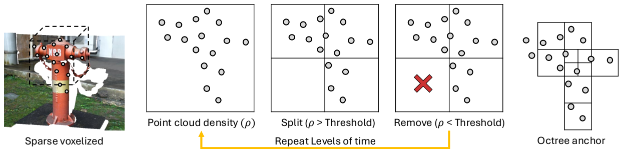

Sparse voxelized
Point cloud density (ρ)
Split (ρ > Threshold)
Remove (ρ < Threshold)
Repeat Levels of time
Octree anchor

Figure 4. **Visualization of our proposed Octree Anchor Formation strategy.** Given an initial sparse voxelized point cloud, we iteratively perform density-guided adaptive voxel splitting and pruning. Voxels with point cloud density (ρ) exceeding a threshold are split, while those with density below the threshold are pruned. Repeated across multiple octree levels, this adaptive octree anchor design significantly reduces memory usage, allowing efficient representation and rendering of large-scale scenes.

where c_i and α_i denote the color and opacity of the i-th Gaussian, respectively. d_i denotes the depth value along the ray at the Gaussian's center.

Anchor-based 3D Gaussian Splatting. To enhance memory efficiency and robustness in large scenes, Scaffold-GS [36] introduces the anchor-based 3DGS representation. Instead of directly maintaining individual Gaussians, the scene is first divided into sparse voxels, each acting as an anchor. From each anchor, k Gaussians are initialized with positions relative to the anchor center:

$$\{\mu_0, \mu_1, \dots, \mu_{k-1}\} = x_v + \{O_0, O_1, \dots, O_{k-1}\} \cdot l_v, \tag{3}$$

where x_v denotes the anchor position, $\{O_i\}$ are offset vectors, and l_v is a scaling factor. Each Gaussian's opacity, rotation, scale, and color are decoded from an anchor feature through lightweight MLPs. For opacity, the formulation is:

$$\{\alpha_0, \alpha_1, \dots, \alpha_{k-1}\} = F_\alpha(\hat{f}_v, \Delta v_c, \hat{d}_v), \tag{4}$$

where F_α is an MLP taking the anchor feature $\hat{f}_v$, the relative view distance Δv_c, and the view direction $\hat{d}_v$ as inputs.

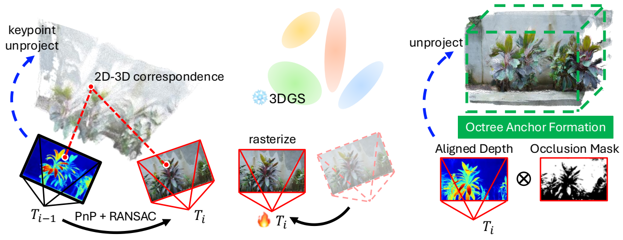
Anchor Initialization. In traditional Scaffold-GS, initial anchors are derived from sparse SfM point clouds. Points are voxelized to form anchor centers:

$$V = \{v \mid v = \lfloor \frac{p}{\epsilon} \rfloor \cdot \epsilon, \forall p \in P\}, \tag{5}$$

where P is the SfM point cloud and ϵ is the voxel size. Each anchor holds a local feature, managing its associated Gaussians. This design ensures structured densification and pruning, adapting Gaussian density to scene complexity and improving both memory and rendering efficiency.

3.2. Octree Anchor Formation

In large-scale casual video settings, memory efficiency and scene adaptability are essential. Our Octree Anchor Formation dynamically adjusts spatial resolution based on observed geometry, enabling scalable and redundant-free anchor management. LongSplat constructs structured anchors from MASt3R's per-frame dense point clouds using an adaptive octree (Fig. 3 (a)). Unlike Scaffold-GS, which relies on a fixed-resolution grid, we progressively subdivide space based on local point density. Each point cloud $\mathbf{P} = \{p_i\}$ is voxelized into a sparse grid at resolution ϵ_0. Voxels exceeding a density threshold τ_{split} split into 8 smaller voxels:



(a) PnP initialization (b) Pose refinement (c) Anchor unprojection

Figure 5. **Detailed illustration of our camera pose estimation.** (a) PnP initialization: Given correspondences between the predicted 3D anchor points from frame T_{i-1} and the 2D keypoints detected in frame T_i we employ PnP with RANSAC to robustly estimate an initial camera pose. (b) Pose refinement: The estimated pose is further refined by rasterizing the 3DGS scene and iteratively minimizing reprojection error to enhance pose accuracy. (c) Anchor unprojection: Newly observed regions are detected via an occlusion mask, computed by forward-warping the previous frame's rendered depth. These regions are unprojected into 3D and converted into anchors via Octree Anchor Formation.

$$\epsilon_{l+1} = \frac{1}{2}\epsilon_l. \tag{6}$$

This process repeats up to a maximum level L. Low-density voxels (density $\rho_v < \tau_{\text{prune}}$) are removed to reduce redundancy (Fig. 4).

Each anchor inherits a spatial scale proportional to its voxel size, ensuring coarse anchors for sparsely observed areas and finer anchors for detailed regions:

$$s_v \propto \epsilon_v. \tag{7}$$

To further prevent unnecessary duplication, newly generated anchors are compared to existing ones. If significant spatial overlap exists, the new anchor is discarded. This density-adaptive, duplication-free octree formation ensures compact memory usage while preserving adaptive resolution across scenes.

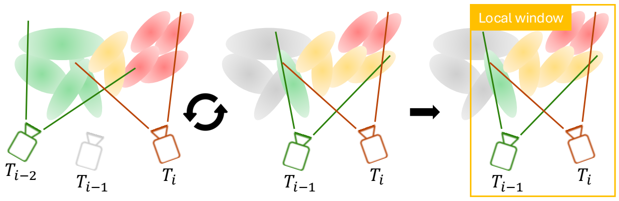
3.3. Pose Estimation module

Accurate and robust camera pose estimation is essential for consistent reconstruction in unposed long video settings. We estimate each pose using 2D-3D correspondences derived from MASt3R, followed by photometric refinement against the current 3D Gaussian scene to maintain coherence across evolving 3D structures (Fig. 3 (c)).

For each new frame t, MASt3R provides 2D correspondences $\{(x_i, x_i')\}$ between frame t and $t-1$, allowing back-projection of matched points x_i to 3D via:

$$X_i = D_{t-1}(x_i) \cdot K^{-1}\tilde{x}_i. \tag{8}$$

These 2D-3D correspondences $\{(x_i', X_i)\}$ are used to solve the initial pose T_t via PnP (Fig. 5 (a)), followed by photometric refinement that minimizes (Fig. 5 (b)):



(a) IOU < Threshold (b) IOU > Threshold (c) Local optimize

Figure 6. **Illustration of our Visibility-Adapted Local Window strategy for local optimization.** To ensure balanced training of the 3D Gaussians, we dynamically define the optimization window based on anchor visibility overlap. Specifically, we compute the Intersection-over-Union (IoU) of visible anchors between consecutive views. Suppose the visibility IoU is below a certain threshold (a). In that case, the local optimization window is adjusted by removing the earliest frame, iteratively repeating until a suitable window with IoU above the threshold is found (b). This approach ensures balanced training coverage and enhances local reconstruction details during optimization (c).

$$\mathcal{L}_{\text{photo}} = \sum_{p \in \Omega} \|I_t(p) - \hat{I}_t(p)\|^2, \tag{9}$$

where I_t is the observed frame and $\hat{I}_t$ is the rendering using the current 3DGS. This ensures the pose aligns with the evolving scene.

To correct MASt3R's depth scale drift, we compute a scale factor $\hat{s}_t$ by comparing the rendered depth D_{t-1} and MASt3R's aligned depth D_t^{align}:

$$\hat{s}_t = \frac{\langle D_{t-1}, D_t^{\text{align}} \rangle}{\langle D_t^{\text{align}}, D_t^{\text{align}} \rangle}. \tag{10}$$

This rescaled depth ensures consistent scale across frames.

As the camera moves, newly visible regions are detected via an occlusion mask M_{occ}, derived by forward-warping D_{t-1} to frame t and comparing it to the rescaled depth D_t^{MASt3R} (Fig. 5 (c)). Newly visible pixels are unprojected into 3D using:

$$p_i = D_{t,\mathbf{u}_i}^{\text{MASt3R}} \cdot \mathbf{K}^{-1} \mathbf{u}_i. \tag{11}$$

These new points are converted into hierarchical octree anchors using the *Octree Anchor Formation* described in Sec. 3.2, with overlapping anchors removed to avoid redundancy (Fig. 5 (c)). This process incrementally expands the scene while maintaining structural regularity.

3.4. Incremental Joint Optimization

To handle casually captured long videos, LongSplat adopts a progressive incremental optimization framework that alternates between per-frame local reconstruction and cross-frame global consistency refinement.

Initialization. We begin with a small set of initial frames. Camera poses and dense point clouds for these frames are estimated using MASt3R [27], followed by converting the point cloud into an initial octree-anchored 3DGS using the proposed Octree Anchor Formation (Fig. 3 (a). When camera intrinsics are unavailable, we directly adopt MASt3R's predicted focal length.

Global Optimization. After initialization, we jointly optimize all 3D Gaussian parameters and camera poses across all processed frames (Fig. 3 (b)). This global optimization ensures geometric consistency across the entire sequence, reducing accumulated pose drift and local misalignments.

Frame Insertion and Pose Estimation. As new frames arrive, we estimate their poses using the correspondence-guided PnP initialization and refinement strategy described in Sec. 3.3. If PnP fails due to insufficient feature correspondences or poor initialization, we trigger a fallback mechanism that re-optimizes all past frames globally before retrying pose estimation. This iterative fallback enhances robustness under challenging motion or weak texture (Fig. 3 (c)).

Local Optimization with Visibility-Adaptive Window. Once the pose is estimated, we optimize only the Gaussians visible in the new frame's frustum, while constraining them with observations from nearby frames in a dynamically selected *visibility-adapted local window* (Fig. 6). Covisibility between frames is measured by:

$$\text{IoU}(t, t') = \frac{|\mathcal{V}(t) \cap \mathcal{V}(t')|}{|\mathcal{V}(t) \cup \mathcal{V}(t')|}, \tag{12}$$

where $\mathcal{V}(t)$ denotes the set of Gaussians visible in frame t. Frames with covisibility below a threshold τ are excluded from the window. This adaptive mechanism ensures local Gaussians are consistently supervised by reliable multi-view constraints, balancing efficiency and accuracy.

Final Global Refinement. In the final step, a final global refinement jointly optimizes all Gaussians and camera poses over the sequence. This final pass further improves both rendering quality and long-range pose consistency.

Monocular Depth and Keypoint Losses. To provide additional supervision in newly revealed regions, where multi-view observations are insufficient, we introduce two regularization terms. A monocular depth loss encourages rendered depth to match MASt3R's scale-aligned depth prior:

$$\mathcal{L}_{\text{depth}} = \|D^{\text{rendered}} - D^{\text{MASt3R}}\|^2. \tag{13}$$

Additionally, a keypoint reprojection loss enforces alignment between projected 3D keypoints and their 2D observations:

$$\mathcal{L}_{\text{reprojection}} = \sum_k \|\pi(\mathbf{X}_k) - \mathbf{u}_k\|^2, \tag{14}$$

where $\pi(\cdot)$ denotes projection using the current pose.

Total Loss. Throughout the entire incremental reconstruction pipeline, each processed frame is optimized using the following objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}} + \lambda_{\text{reprojection}} \mathcal{L}_{\text{reprojection}}, \tag{15}$$

This combined loss applies to both local and global optimization stages, ensuring coherent multi-view, robust pose refinement, and stable geometry reconstruction across the evolving scene.

Table 1. **Quantitative comparison on the Free dataset [60] across various baseline methods.** Methods such as CF-3DGS [14] frequently encounter out-of-memory issues, denoted by “-”. Our method consistently outperforms all baselines across diverse scenes, delivering superior rendering quality and robustness, especially in challenging environments characterized by complex camera trajectories and varied geometric structures. “*”: Initialized with MASt3R poses, then jointly optimized.

Scenes	COLMAP [50] + F2-NeRF [60]			COLMAP [50] + Scaffold-GS [36]			MASt3R [27] + Scaffold-GS [36]			MASt3R [27] + Scaffold-GS [36]*			CF-3DGS [14]			NoPe-NeRF [5]			LocalRF [39]			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Grass	23.44	0.58	0.45	26.75	0.82	0.20	22.65	0.61	0.34	25.06	0.79	0.21	-	-	-	16.39	0.27	0.81	18.84	0.35	0.60	**26.16**	**0.80**	**0.22**
Hydrant	23.75	0.74	0.28	26.66	0.86	0.12	23.22	0.71	0.21	25.68	0.83	0.12	-	-	-	17.94	0.43	0.66	19.19	0.48	0.48	**24.69**	**0.79**	**0.18**
Lab	24.34	0.83	0.26	28.27	0.92	0.10	20.66	0.74	0.25	22.42	0.80	0.18	-	-	-	17.42	0.52	0.63	17.22	0.55	0.47	**27.11**	**0.87**	**0.15**
Pillar	28.05	0.79	0.23	31.75	0.90	0.12	23.95	0.70	0.28	22.88	0.67	0.24	14.55	0.40	0.66	18.88	0.44	0.75	22.98	0.59	0.49	**30.44**	**0.88**	**0.16**
Road	26.03	0.80	0.27	30.45	0.92	0.10	24.23	0.73	0.25	25.05	0.78	0.27	-	-	-	17.48	0.44	0.79	20.68	0.54	0.56	**27.73**	**0.84**	**0.20**
Sky	25.10	0.86	0.24	28.34	0.92	0.12	23.26	0.80	0.22	25.37	0.88	0.14	-	-	-	16.18	0.51	0.65	18.76	0.60	0.46	**28.07**	**0.91**	**0.13**
Stair	28.14	0.84	0.22	32.13	0.93	0.10	23.35	0.71	0.30	24.46	0.79	0.28	13.41	0.41	0.63	19.14	0.47	0.69	23.55	0.66	0.38	**31.00**	**0.89**	**0.16**
Avg.	25.55	0.78	0.28	29.19	0.90	0.12	23.05	0.72	0.27	24.42	0.79	0.21	13.98	0.41	0.65	17.63	0.44	0.71	20.17	0.54	0.49	**27.88**	**0.85**	**0.17**

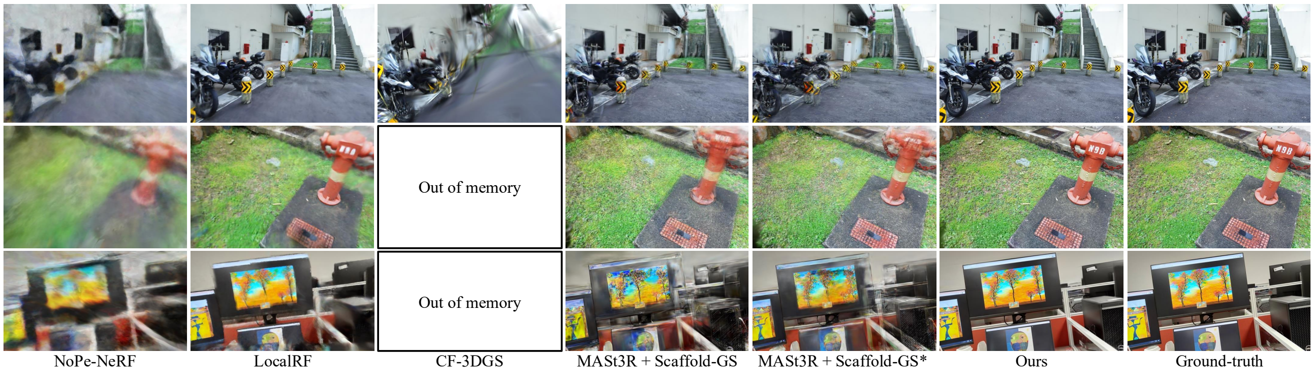


Figure 7. **Qualitative comparison on the Free dataset [60].** We compare our method with state-of-the-art approaches including NoPe-NeRF [5], LocalRF [39], CF-3DGS [14], and MASt3R [27] combined with Scaffold-GS [36]. CF-3DGS fails due to memory constraints (OOM), and other baseline methods exhibit artifacts or blurry reconstructions. In contrast, our method produces results closest to the ground truth, demonstrating clearer details, accurate geometry, and visually consistent rendering, particularly under challenging scene structures and complex camera trajectories. “*”: Initialized with MASt3R poses, then jointly optimized.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate LongSplat on three challenging real-world datasets with varying difficulty levels:

- **Tanks and Temples [25] (Standard)**: Eight scenes with smooth, forward-facing camera trajectories, evaluated at full resolution. Every 8th frame is used for testing.
- **Free Dataset [60] (Moderate)**: Seven handheld videos featuring complex, unconstrained trajectories with multiple foreground objects, evaluated at 1/2 resolution. Frequent scene changes make memory-efficient 3D representation essential. Every 8th frame is tested.
- **Hike Dataset [39] (Hard)**: Long videos with hundreds to thousands of frames, complex trajectories, and detailed geometry, evaluated at 1/4 resolution. The scale and duration demand adaptive memory management. Every 10th frame is used for testing.

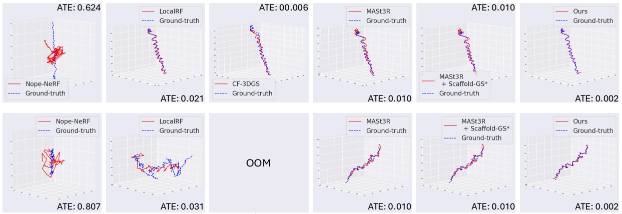
Evaluation Metrics. We evaluate novel view synthesis quality using PSNR, SSIM [63], and LPIPS [75]. Pose accuracy is measured with Absolute Trajectory Error (ATE) and Relative Pose Error (RPE), using COLMAP poses as ground truth. We also report model size, training time, and FPS to assess computational efficiency.

Table 2. **Quantitative evaluation of camera pose estimation accuracy on the Free dataset [60].** Our method achieves superior performance across most scenes, significantly reducing pose errors compared to state-of-the-art approaches. “*”: Initialized with MASt3R poses, then jointly optimized.

Method	RPE$_t$↓	RPE$_r$↓	ATE↓
MASt3R [27] + Scaffold-GS [36]	0.162	0.265	0.013
MASt3R [27] + Scaffold-GS [36]*	0.083	0.176	0.008
CF-3DGS [14]	0.234	3.442	0.022
NoPe-NeRF [5]	6.231	4.822	0.576
LocalRF [39]	0.754	7.086	0.035
Ours	**0.028**	**0.103**	**0.004**

Baselines. We compare LongSplat with COLMAP-based methods (COLMAP [50]+F2-NeRF [60] / 3DGS [22] / Scaffold-GS [36]) and unposed methods (NoPe-NeRF [5], LocalRF [39], CF-3DGS [14]). Additionally, we evaluate a naïve baseline combining MASt3R’s [27] predicted point cloud and poses with Scaffold-GS. During training, camera poses are either fixed (MASt3R + Scaffold-GS) or jointly optimized (MASt3R + Scaffold-GS*).

Implementation Details. We implement LongSplat based on Scaffold-GS [36], using its learning rate schedule and growing/pruning rules. Each anchor emits k Gaussians predicted by a lightweight 2-layer MLP. The initial sparse voxel

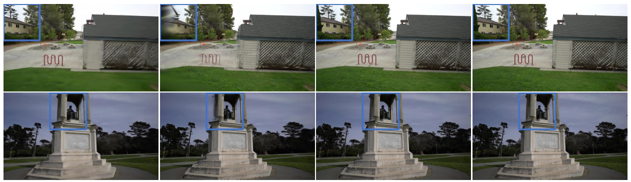


Nope-NeRF LocalRF CF-3DGS MASt3R + Scaffold-GS MASt3R + Scaffold-GS* Ours

Figure 8. **Visualization of camera trajectories on Free dataset [60].** CF-3DGS [14] encounters OOM and fails for long sequences, whereas our method reliably estimates accurate, stable trajectories, demonstrating superior robustness.

Table 3. **Quantitative evaluation of novel view synthesis quality on the Tanks and Temples dataset [25].** Our proposed LongSplat consistently surpasses existing methods across multiple scenes.

Method	PSNR↑	SSIM↑	LPIPS↓	RPE_t↓	RPE_r↓	ATE↓
COLMAP+3DGS [22]	30.21	0.92	0.10	–	–	–
MASt3R [27] + Scaffold-GS [36]	28.67	0.79	0.21	0.166	0.168	0.006
MASt3R [27] + Scaffold-GS [36]*	30.92	0.90	0.13	0.047	0.103	0.005
NoPe-NeRF [5]	26.34	0.74	0.39	0.080	**0.038**	0.006
CF-3DGS [14]	31.28	0.93	0.09	0.041	0.069	0.004
Ours	**32.83**	**0.94**	**0.08**	0**.032**	0.068	**0.003**



NoPe-NeRF CF-3DGS Ours Ground-truth

Figure 9. **Qualitative comparison on the Tanks and Temples dataset [25].** NoPe-NeRF [5] produces visibly blurred results with inaccurate geometries, while CF-3DGS [14], despite better sharpness, fails to reconstruct fine details accurately. In contrast, our LongSplat method achieves superior rendering quality, closely matching the ground truth with sharper textures, more accurate geometry, and consistent lighting.

grid size is 0.1. Camera poses are optimized via a differentiable CUDA-accelerated rasterizer, parameterized with quaternions and translation vectors. We use 200 local, 500 global, and 10,000 refinement iterations, starting with three initial frames. The octree density thresholds for splitting and removal start at 10 and 5, progressively increasing with depth. Visibility IoU threshold is set to 0.2. All experiments are conducted on a single NVIDIA RTX 4090.

4.2. Comparisons

Tanks and Temples. We evaluate LongSplat on the Tanks and Temples dataset [25], a standard benchmark for novel view synthesis. As shown in Tab. 3, LongSplat achieves state-of-the-art rendering quality (avg. PSNR: 32.83 dB) and superior camera pose estimation accuracy (lowest ATE and RPE). Qualitative results in Fig. 9 confirm sharper textures, accurate geometry, and better visual consistency compared to baselines. Please refer to the supplementary material for the full quantitative evaluation table for each scene.

Table 4. **Quantitative evaluation on the Hike dataset [39].** Our method consistently outperforms baselines across diverse scenes with complex trajectories and extended sequences, highlighting LongSplat's robustness and superior scene representation capability. CF-3DGS [14] encounters OOM in all scenes and is thus omitted.

Method	PSNR↑	SSIM↑	LPIPS↓
MASt3R [27] + Scaffold-GS [36]	17.30	0.42	0.52
MASt3R [27] + Scaffold-GS [36]*	17.90	0.44	0.50
LocalRF [39]	23.56	0.68	0.29
Ours	**25.39**	**0.81**	**0.19**

Table 5. **Ablation on training components.** Removing pose estimation, global optimization, or local optimization significantly degrades performance, highlighting each module's importance. Our full method achieves the best rendering quality and pose accuracy.

Method	PSNR↑	SSIM↑	LPIPS↓	RPE_t↓	RPE_r↓	ATE↓
w/o Pose estimation	20.19	0.56	0.51	0.42	2.71	0.71
w/o Global optimization	20.50	0.58	0.41	0.12	0.50	0.01
w/o Local optimization	25.94	0.77	0.28	0.06	0.31	0.01
w/o Refinement	26.08	0.80	0.25	0.04	0.22	0.01
Ours	**27.88**	**0.85**	**0.17**	**0.03**	**0.11**	**0.00**

Table 6. **Ablation on local window sizes.** Fixed small windows (e.g., 1-frame or 5-frame) or global optimization degrades reconstruction quality and pose accuracy. Our visibility-adaptive window dynamically selects optimal context, achieving the best balance of local detail and global consistency.

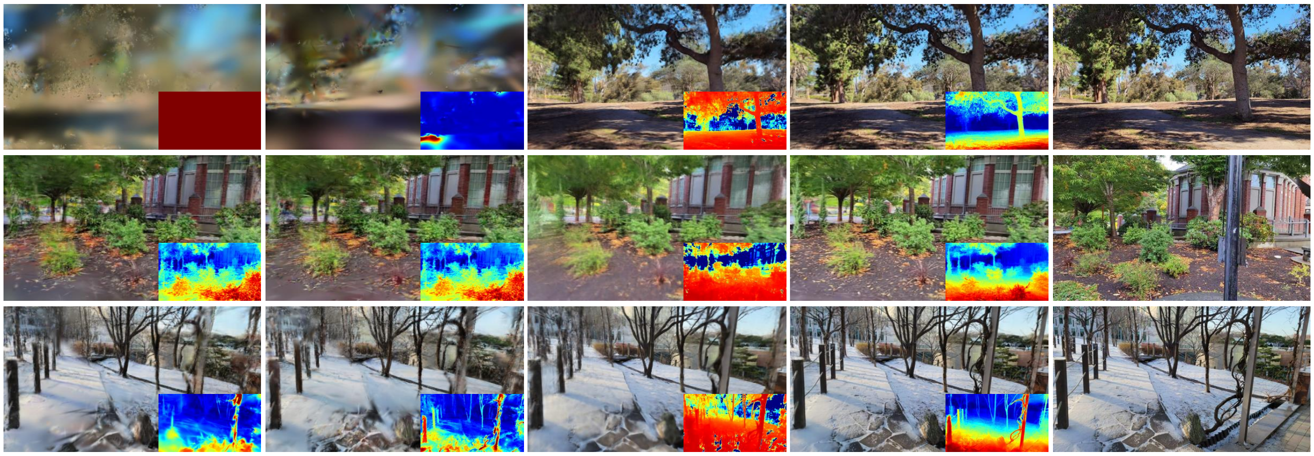
Window size	PSNR ↑	SSIM ↑	LPIPS ↓	RPE_t↓	RPE_r↓	ATE↓
1-frame (Minimum Window)	26.58	0.80	0.23	0.05	0.21	0.01
5-frame (Fixed Window)	26.90	0.82	0.22	0.04	0.18	0.01
All Frames (Global Optimize)	26.15	0.78	0.26	0.06	0.28	0.08
Ours (Visibility-Adaptive)	**27.88**	**0.85**	**0.17**	**0.03**	**0.11**	**0.00**

Free Dataset. We evaluate LongSplat on the challenging Free dataset, achieving superior reconstruction quality (avg. PSNR: 27.88 dB, SSIM: 0.85, LPIPS: 0.17) as shown in Tab. 1 and Fig. 7. Competing methods like CF-3DGS often face OOM issues, while LocalRF produces fragmented geometry and pose drift. Our method also achieves consistently lower pose errors (ATE, RPE) than baselines, as shown quantitatively in Tab. 2 and visually in Fig. 8.

Hike Dataset. We evaluate LongSplat on the challenging Hike dataset, achieving state-of-the-art reconstruction quality (avg. PSNR: 25.30 dB, SSIM: 0.81, LPIPS: 0.19) (Tab. 4). Competing methods like CF-3DGS often fail (OOM), while LocalRF produces lower-quality results (PSNR: 23.56 dB). Qualitative comparisons (Fig. 10) further highlight LongSplat's superior visual fidelity and robustness.

4.3. Ablation Studies

Training Components. To analyze the contribution of each training component, we individually disable them and evaluate performance. As shown in Tab. 5, removing pose estimation severely harms reconstruction quality and increases pose errors (ATE: 0.71). Omitting global or local optimization also reduces performance. Our full method achieves the highest quality (PSNR: 27.88 dB, SSIM: 0.85, LPIPS: 0.17) and minimal pose errors (ATE: 0.003).



MASt3R + Scaffold-GS | MASt3R + Scaffold-GS* | LocalRF | Ours | Ground-truth

Figure 10. **Qualitative results on the Hike dataset [39].** Compared to existing methods such as LocalRF [39] and MASt3R [27]+Scaffold-GS [36], our approach significantly improves visual clarity and reconstruction fidelity, accurately capturing complex details and textures in challenging scenes captured during long, casual outdoor trajectories. Notably, our method better preserves structural details and reduces artifacts, demonstrating enhanced robustness and visual quality. "*": Initialized with MASt3R poses, then jointly optimized.

Table 7. **Ablation on anchor unprojection strategies.** Our Adaptive Octree method achieves the best rendering quality and lowest perceptual errors, significantly reducing memory usage (7.92× compression) compared to baselines.

Method	PSNR↑	SSIM↑	LPIPS↓	Size (MB)↓	Compress↑
Per-pixel Unprojection (Dense)	22.47	0.69	0.35	799	1.00x
Fixed-size Voxel Unprojection	26.99	0.81	0.18	591	1.35x
Naive Densification	25.73	0.75	0.31	63	12.66x
Ours (Adaptive Octree)	27.88	0.85	0.17	101	7.92x

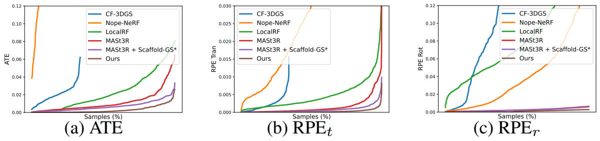
Table 8. **Comparison of training efficiency on the Free dataset.** Our method significantly reduces training time and achieves dramatically higher throughput (FPS) while simultaneously maintaining a compact model size compared to state-of-the-art approaches.

Method	FPS ↑	Training time ↓	Size (MB) ↓
NoPe-NeRF [5]	0.29	36 hr	7
LocalRF [39]	1.17	14 hr	1080
CF-3DGS [14]	9.81	2 hr	1966
Ours	281.71	1 hr	101

Local Window Sizes. We analyze the effect of local window size on reconstruction and pose accuracy (Tab. 6). Small fixed-size windows (e.g., 1 frame) lack sufficient constraints, causing fragmentation and higher errors. Our visibility-adapted window achieves the best balance, yielding the highest reconstruction quality (PSNR: 27.88 dB) and lowest pose drift (ATE: 0.003).

Anchor Unprojection Strategies. We compare our adaptive octree anchor formation to (1) per-pixel initialization, (2) fixed-resolution voxels, and (3) naïve densification (Tab. 7). Our method achieves superior reconstruction quality with significantly reduced memory usage (7.92× compression).

Training Efficiency. We evaluate the computational efficiency of LongSplat (Tab. 8), which achieves 281.71 FPS and trains in just 1 hour on an NVIDIA RTX 4090, nearly 30× faster than LocalRF. Our method also significantly reduces the model size to approximately 101 MB.



(a) ATE (b) RPE_t (c) RPE_r

Figure 11. **Robustness analysis on camera pose estimation (Free dataset [60]).** We plot cumulative error distributions for ATE, RPE translation, and rotation. Our method consistently achieves lower errors compared to existing methods, demonstrating superior robustness and reduced pose drift.

Robustness Analysis of Camera Pose Estimation. We further analyze robustness by comparing cumulative error distributions for ATE and RPE (translation and rotation) in Fig. 11. LongSplat achieves consistently lower errors than baselines, effectively minimizing drift and maintaining stable trajectories, highlighting the advantage of our incremental optimization and robust tracking.

5. Conclusion

We present LongSplat, a robust unposed 3D Gaussian Splatting framework for casual long videos. It integrates incremental joint optimization, a robust tracking module, and adaptive octree anchors, significantly improving pose accuracy, reconstruction quality, and memory efficiency. Extensive experiments confirm that LongSplat consistently outperforms state-of-the-art approaches. Future work includes handling dynamic scenes and enhancing pose estimation robustness.

Limitations. LongSplat shares common limitations of unposed reconstruction methods, assuming static scenes and fixed intrinsics, making it unsuitable for dynamic objects or varying focal lengths.

Acknowledgements. This work was supported by NVIDIA Taiwan AI Research Development Center (TRDC). This research was funded by the National Science and Technology Council, Taiwan, under Grants NSTC 112-2222-E-A49-004-MY2 and 113-2628-E-A49-023-. Yu-Lun Liu acknowledges the Yushan Young Fellow Program by the MOE in Taiwan.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhöfer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16610–16620, 2023. 2

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[3] 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.

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *arXiv preprint arXiv:2304.06706*, 2023. 2

[5] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *CVPR*, 2023. 2, 6, 7, 8

[6] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, 2001. 2

[7] 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*, pages 990–1000, 2024. 2

[8] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *SIGGRAPH*, 1993. 2

[9] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, and Ameesh Makadia. Lu-nerf: Scene and pose estimation by synchronizing local unposed nerfs. *arXiv preprint arXiv:2306.05410*, 2023. 2

[10] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Garf: Gaussian activated radiance fields for high fidelity reconstruction and pose estimation. *arXiv e-prints*, 2022. 2

[11] Wenyan Cong, Kevin Wang, Jiahui Lei, Colton Stearns, Yuanhao Cai, Dilin Wang, Rakesh Ranjan, Matt Feiszli, Leonidas Guibas, Zhangyang Wang, Weiyao Wang, and Zhiwen Fan. Videolifter: Lifting videos to 3d with fast hierarchical stereo alignment, 2025. 2

[12] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *SIGGRAPH*, 1996. 2

[13] Cheng-De Fan, Chen-Wei Chang, Yi-Ruei Liu, Jie-Ying Lee, Jiun-Long Huang, Yu-Chee Tseng, and Yu-Lun Liu. Spectromotion: Dynamic 3d reconstruction of specular scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21328–21338, 2025. 2

[14] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *CVPR*, 2024. 2, 6, 7, 8

[15] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, 2021. 2

[16] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. 2003. 2

[17] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, pages 577–584, 2005. 2

[18] Youichi Horry, Ken-Ichi Anjyo, and Kiyoshi Arai. Tour into the picture: using a spidery mesh interface to make animation from a single image. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 225–232, 1997. 2

[19] Hao-Yu Hou, Chia-Chi Hsu, Yu-Chen Huang, Mu-Yi Shen, Wei-Fang Sun, Cheng Sun, Chia-Che Chang, Yu-Lun Liu, and Chun-Yi Lee. 3d gaussian splatting with grouped uncertainty for unconstrained images. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[20] Ronghang Hu, Nikhila Ravi, Alex Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *ICCV*, 2020. 2

[21] Bo Ji and Angela Yao. Sfm-free 3d gaussian splatting via hierarchical training. *arXiv preprint arXiv:2412.01553*, 2024. 2

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 2, 3, 6, 7

[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] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022. 2

[25] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017. 6, 7

[26] Johannes Kopf, Michael F. Cohen, and Richard Szeliski. First-person hyper-lapse videos. *TOG*, 2014. 2

[27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2, 3, 5, 6, 7, 8

[28] Du-Hsiu Li, Hsueh-Ming Hang, and Yu-Lun Liu. Virtual view synthesis using backward depth warping algorithm. In *2013 Picture Coding Symposium (PCS)*, pages 205–208. IEEE, 2013. 2

[29] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 2

[30] Ming-Feng Li, Yueh-Feng Ku, Hong-Xuan Yen, Chi Liu, Yu-Lun Liu, Albert YC Chen, Cheng-Hao Kuo, and Min Sun. Genrc: Generative 3d room completion from sparse image collections. In *European Conference on Computer Vision*, pages 146–163. Springer, 2024. 3

[31] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 2

[32] Chin-Yang Lin, Chung-Ho Wu, Chang-Han Yeh, Shih-Han Yen, Cheng Sun, and Yu-Lun Liu. Frugalnerf: Fast convergence for few-shot novel view synthesis without learned priors. *CVPR*, 2025. 2

[33] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024. 3

[34] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[35] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *CVPR*, 2023. 2

[36] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *CVPR*, 2024. 2, 3, 4, 6, 7, 8

[37] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2

[38] Caoyuan Ma, Yu-Lun Liu, Zhixiang Wang, Wu Liu, Xinchen Liu, and Zheng Wang. Humannerf-se: A simple yet effective approach to animate humannerf with diverse poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1460–1470, 2024. 2

[39] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023. 2, 3, 6, 7, 8

[40] 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*, 2021. 2

[41] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 2022. 2

[42] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 2015. 2

[43] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 2

[44] Yunlong Ran, Yanxu Li, Qi Ye, Yuchi Huo, Zechun Bai, Jiahao Sun, and Jiming Chen. Ct-nerf: Incremental optimizing neural radiance field and poses with complex trajectory. *arXiv preprint arXiv:2404.13896*, 2024. 2

[45] 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*, pages 14335–14345, 2021. 2

[46] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning 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

[47] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *ECCV*, 2020. 2

[48] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *CVPR*, 2021. 2

[49] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2, 6

[51] Mu-Yi Shen, Chia-Chi Hsu, Hao-Yu Hou, Yu-Chen Huang, Wei-Fang Sun, Chia-Che Chang, Yu-Lun Liu, and Chun-Yi Lee. Driveenv-nerf: Exploration of a nerf-based autonomous driving environment for real-world performance validation. *arXiv preprint arXiv:2403.15791*, 2024. 2

[52] Chih-Hai Su, Chih-Yao Hu, Shr-Ruei Tsai, Jie-Ying Lee, Chin-Yang Lin, and Yu-Lun Liu. Boostmvsnerfs: Boosting mvs-based nerfs to generalizable view synthesis in large-scale scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[53] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 2

[54] Teppei Suzuki. Fed3dgs: Scalable 3d gaussian splatting with federated learning. *arXiv preprint arXiv:2403.11460*, 2024. 3

[55] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9(1):1–11, 2017. 2

[56] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 3

[57] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 2

[58] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[59] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3

[60] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *CVPR*, 2023. 6, 7, 8

[61] 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. 3

[62] 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. 3

[63] 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*, 2004. 6

[64] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[65] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. 2022. 2

[66] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[67] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2

[68] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 2

[69] 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. 3

[70] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *IROS*, 2021. 2

[71] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[72] Yu-Ting Zhan, Cheng-Yuan Ho, Hebi Yang, Yi-Hsin Chen, Jui Chiu Chiang, Yu-Lun Liu, and Wen-Hsiao Peng. Cat-3dgs: A context-adaptive triplane approach to rate-distortion-optimized 3dgs compression. *arXiv preprint arXiv:2503.00357*, 2025. 2

[73] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 2

[74] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–12, 2022. 2

[75] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[76] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. 2018. 2