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

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 robust Pose Estimation Module leveraging learned 3D priors; and (3) an efficient Octree Anchor Formation mechanism that converts dense point clouds into anchors based on spatial density. 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/

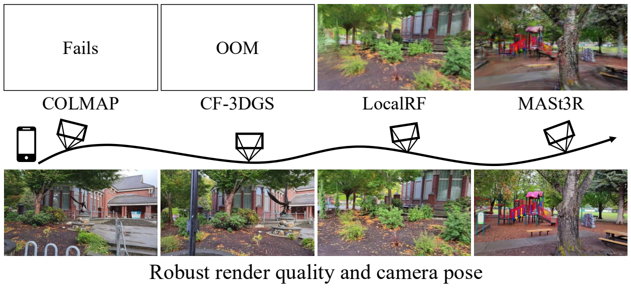


Figure 1. **Novel view synthesis for casual long videos.** Existing methods struggle on casually captured long videos: COLMAP [16] fails in pose estimation, CF-3DGS [5] runs out of memory, LocalRF [11] breaks under complex trajectories, and MASt3R [8]+Scaffold-GS [10] yields inaccurate poses. In contrast, LongSplat delivers robust pose estimation and high-quality novel view synthesis without memory issues.

1. Introduction

High-quality 3D reconstruction and novel view synthesis (NVS) are crucial for applications in VR/AR, digital tourism, and video editing. With the rise of smartphones and action cameras, casually captured videos have become a major source of 3D content, but they are difficult to handle due to irregular trajectories, long sequences, and the lack of reliable camera poses. Existing approaches face two key limitations: reliance on Structure-from-Motion pipelines like COLMAP [16], which often fail in casual settings as shown in Fig. 1, or COLMAP-free methods such as CF-3DGS [5] and LocalRF [11], which struggle with memory constraints or fragmented reconstructions. Even foundation models like MASt3R [8] provide fast initialization but drift significantly on long videos. We present **LongSplat**, a robust unposed 3D Gaussian Splatting (3DGS) [6] framework for casual long videos. LongSplat jointly optimizes camera poses and 3DGS in a unified framework, combining correspondence-guided pose estimation with photometric refinements to maintain accuracy under unstructured motion. In addition, an efficient *Octree Anchor Formation* compresses dense point clouds into anchors, reducing memory usage while preserving fine scene details. These components are integrated in an incremental joint optimization strategy that enforces global consistency across long sequences. Extensive experiments on Tanks and Temples, Free, and Hike datasets show that LongSplat achieves sharper reconstructions and more accurate pose estimates than prior methods, effectively mitigating drift and memory issues. Our main contributions are:

- An incremental joint optimization framework for simultaneous camera pose and 3DGS reconstruction.
- A robust pose estimation module guided by learned 3D priors.
- An adaptive octree-based anchor formation strategy for efficient and scalable reconstruction.

2. Related Work

Novel View Synthesis. Novel view synthesis (NVS) has evolved from early interpolation and geometry-based rendering [3, 4] to neural representations such as NeRF [12]. Extensions improve sampling [1], sparse-view training [14],

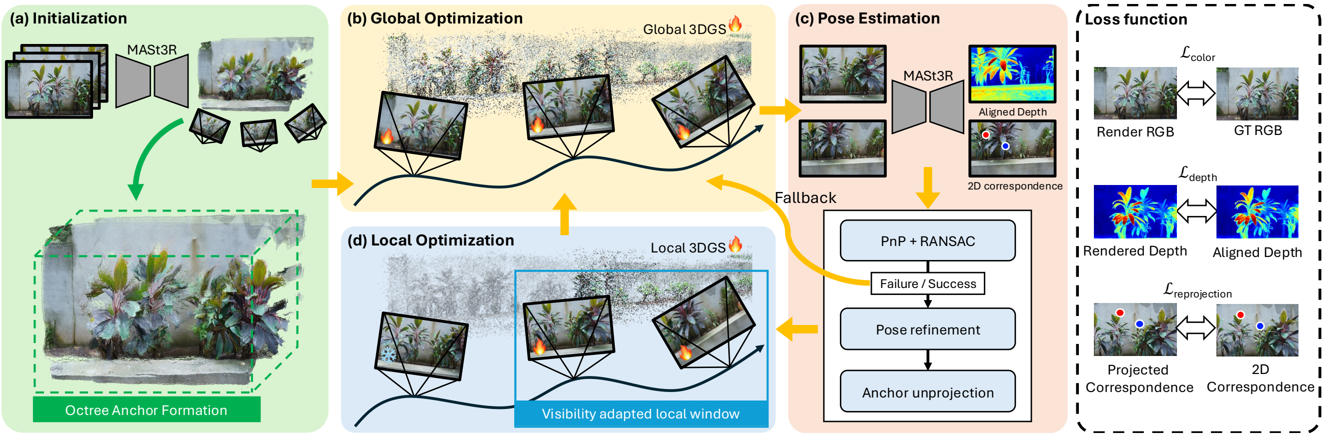


Figure 2. **Overview of the LongSplat framework.** Given a casually captured long video without known poses, LongSplat incrementally reconstructs the scene through tightly coupled pose estimation and 3D Gaussian Splatting. (a) Initialization converts MASt3R [8] global aligned point cloud into an octree-anchored 3DGS. (b) Global Optimization jointly refines all camera poses and 3D Gaussians for global consistency. (c) Pose estimation estimates each new frame pose via correspondence-guided PnP, applies photometric refinement, and updates octree anchors using unprojected points. 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. (e) All optimization stages leverage a unified objective composed of photometric loss, depth loss, and reprojection loss to ensure accurate geometry and appearance reconstruction.

and efficiency [13]. Recently, point-based methods, particularly 3D Gaussian Splatting (3DGS) [6], enable real-time rendering, but typically require pre-computed camera poses.

Unposed NVS. To remove dependence on SfM, works such as BARF [9], NeRFmm [21], and iNeRF [22] jointly optimize poses and radiance fields, though they often assume limited motion or good initialization. Recent approaches incorporate depth or learned priors [2, 5], but robustness degrades on challenging trajectories.

Large-scale NVS. Scaling NVS introduces memory and consistency issues. Block-based NeRFs [17] and hierarchical or octree-based 3DGS [10, 15] improve scalability but still rely on SfM initialization. Our work instead adapts voxel resolution dynamically from input point clouds, enabling scalable reconstruction without pose supervision.

Casual Long Videos. Long, unconstrained videos pose difficulties due to drift, irregular motion, and scene growth. LocalRF [11] mitigates drift with progressive optimization but produces fragmented reconstructions. 3D foundation models (e.g., DUSt3R [19], MASt3R [8]) provide useful priors but accumulate errors over long sequences. LongSplat leverages such priors as initialization and progressively refines both poses and 3DGS through joint optimization.

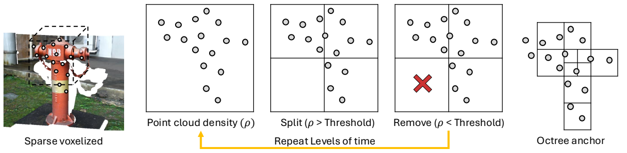
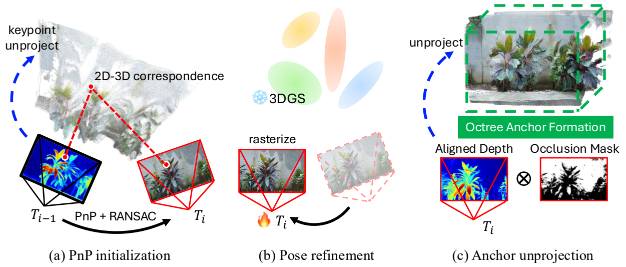


Figure 3. **Octree Anchor Formation.** Starting from a voxelized point cloud, voxels are adaptively split when density exceeds a threshold and pruned otherwise. Iterating across octree levels yields a compact anchor structure that reduces memory and enables efficient large-scale reconstruction.



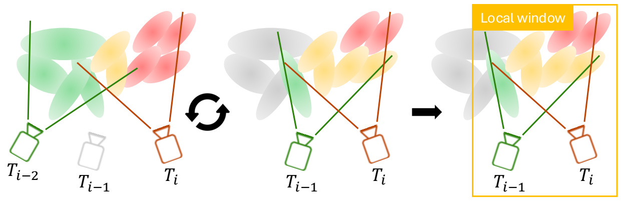
(a) PnP initialization (b) Pose refinement (c) Anchor unprojection

Figure 4. **Camera pose estimation.** (a) PnP initialization from 3D–2D correspondences with RANSAC. (b) Pose refinement by minimizing reprojection error in the 3DGS scene. (c) Anchor unprojection: newly visible regions are detected via occlusion masks and converted into anchors with Octree Anchor Formation.

3. Method

3.1. Octree Anchor Formation

To efficiently represent large-scale casual videos, LongSplat builds anchors from MASt3R's per-frame point clouds using an adaptive octree (Fig. 3). Voxels exceeding a density threshold τ_{split} are subdivided ($\epsilon_{l+1} = \frac{1}{2}\epsilon_l$) while those below τ_{prune} are pruned, repeating up to a maximum depth L. Each anchor inherits a scale proportional to its voxel size ($s_v \propto \epsilon_v$), yielding coarse anchors for sparse regions and finer ones for detailed areas. Redundant anchors with significant spatial overlap are discarded, producing a compact, density-adaptive representation that contrasts with the fixed-resolution grids of Scaffold-GS.



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

Figure 5. **Visibility-Adapted Local Window.** Local optimization windows are dynamically adjusted using the IoU of visible anchors between consecutive views. When IoU falls below a threshold, earlier frames are removed until sufficient overlap is achieved, ensuring balanced training and improved local reconstruction.

3.2. Pose Estimation Module

To achieve consistent reconstruction in unposed long videos, LongSplat estimates each camera pose using 2D–3D correspondences from MASt3R, refined against the evolving 3DGS scene (Fig. 2 (c)). For a new frame t, correspondences $\{(x_i, x_i')\}$ allow back-projection via

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

which are solved by PnP to obtain the initial pose T_t (Fig. 4 (a)). Photometric refinement then minimizes

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

ensuring alignment with the current 3DGS (Fig. 4 (b)). To correct depth scale drift, a factor

$$\hat{s}_t = \frac{\langle D_{t-1}, D_t^{\text{align}}\rangle}{\langle D_t^{\text{align}}, D_t^{\text{align}}\rangle} \tag{3}$$

rescales MASt3R predictions. Newly visible regions are detected via an occlusion mask M_{occ} and unprojected as

$$p_i = D_{t,\mathbf{u}_i}^{\text{MASt3R}} \cdot \mathbf{K}^{-1}\mathbf{u}_i, \tag{4}$$

then converted into octree anchors (Sec. 3.1), with overlapping anchors removed (Fig. 4 (c)). This incremental strategy maintains global consistency while expanding the scene.

3.3. 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 [8], followed by converting the point cloud into an initial octree-anchored 3DGS using the proposed Octree Anchor Formation (Fig. 2 (a). When camera intrinsics are unavailable, we directly adopt MASt3R's estimated focal length.

Global Optimization. After initialization, we jointly optimize all 3D Gaussian parameters and camera poses across all processed frames (Fig. 2 (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. 3.2. 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. 2 (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. 5). Covisibility between frames is measured by:

$$\text{IoU}(t, t') = \frac{|\mathcal{V}(t) \cap \mathcal{V}(t')|}{|\mathcal{V}(t) \cup \mathcal{V}(t')|}, \tag{5}$$

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.

Depth and Reprojection 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:

$$\mathcal{L}_{\text{depth}} = \|D^{\text{rendered}} - D^{\text{MASt3R}}\|^2. \tag{6}$$

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

$$\mathcal{L}_{\text{reprojection}} = \sum_k \|\pi(\mathbf{X}_k) - \mathbf{u}_k\|^2, \tag{7}$$

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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{reprojection}}\mathcal{L}_{\text{reprojection}}, \tag{8}$$

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.

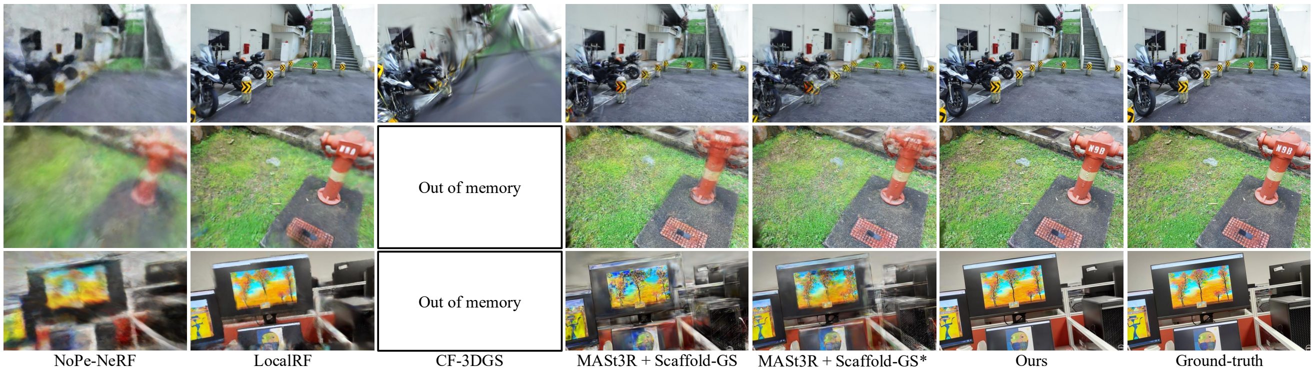
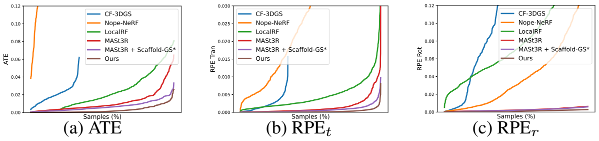


Figure 6. **Qualitative comparison on the Free dataset [18].** We compare our method with state-of-the-art approaches including NoPe-NeRF [2], LocalRF [11], CF-3DGS [5], and MASt3R [8] combined with Scaffold-GS [10]. CF-3DGS fails due to memory constraints (OOM), and other baseline methods exhibit artifacts 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.

Table 1. **Quantitative evaluation on the Free dataset [18].** We report average rendering quality and pose accuracy across all scenes. “*”: Initialized with MASt3R poses, then jointly optimized.

Method	PSNR↑	SSIM↑	LPIPS↓	RPE_t↓	RPE_r↓	ATE↓
COLMAP+F2-NeRF [18]	25.55	0.78	0.28	–	–	–
COLMAP+Scaffold-GS [10]	29.19	0.90	0.12	–	–	–
MASt3R [8]+Scaffold-GS	23.05	0.72	0.27	0.162	0.265	0.013
MASt3R [8]+Scaffold-GS*	24.42	0.79	0.21	0.083	0.176	0.008
CF-3DGS [5]	13.98	0.41	0.65	0.234	3.442	0.022
NoPe-NeRF [2]	17.63	0.44	0.71	6.231	4.822	0.576
LocalRF [11]	20.17	0.54	0.49	0.754	7.086	0.035
Ours	**27.88**	**0.85**	**0.17**	**0.028**	**0.103**	**0.004**



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

Figure 7. **Robustness analysis on camera pose estimation (Free dataset [18]).** 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.

4. Experiments

4.1. Experimental Setup

We evaluate LongSplat on three challenging real-world datasets: **Tanks and Temples** [7] with smooth forward-facing trajectories (results in supplementary), the **Free dataset** [18] containing seven handheld videos with unconstrained motion and frequent scene changes, and the long-sequence **Hike dataset** [11] (results in supplementary). Novel view synthesis quality is assessed using PSNR, SSIM [20], and LPIPS [23], while pose accuracy is measured with ATE and RPE against COLMAP ground truth. We compare against COLMAP-based pipelines (COLMAP+F2-NeRF, COLMAP+3DGS, COLMAP+Scaffold-GS), unposed methods (NoPe-NeRF, LocalRF, CF-3DGS), and MASt3R [8]+Scaffold-GS variants, where poses are either fixed or jointly optimized. Computational efficiency is further reported in terms of model size, training time, and FPS.

4.2. Comparisons

Free Dataset. We evaluate LongSplat on the challenging Free dataset, achieving superior reconstruction quality as shown in Tab. 1 and Fig. 6. Competing methods like CF-3DGS often face OOM issues, while LocalRF produces fragmented geometry and pose drift. Although MASt3R + Scaffold-GS avoids OOM errors, its inaccurate global pose estimates from MASt3R result in blurred renderings and structural distortions. Our method also achieves consistently lower pose errors than baselines.

Robustness Analysis of Camera Pose Estimation. We further analyze robustness by comparing cumulative error distributions for ATE and RPE (translation and rotation) in Fig. 7. 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 that integrates incremental optimization, robust tracking, and adaptive octree anchors. It achieves superior pose accuracy, reconstruction quality, and memory efficiency compared to prior methods.

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.

References

[1] 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 *ICCV*, 2021. 1

[2] 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, 4

[3] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *SIGGRAPH*, 1993. 1

[4] 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. 1

[5] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *CVPR*, 2024. 1, 2, 4

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 1, 2

[7] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017. 4

[8] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 1, 2, 3, 4

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

[10] 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. 1, 2, 4

[11] 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. 1, 2, 4

[12] 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. 1

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

[14] 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. 1

[15] 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. 2

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

[17] 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 *CVPR*, 2022. 2

[18] 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. 4

[19] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 4

[21] 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

[22] 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

[23] 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. 4