NopeRoomGS: Indoor 3D Gaussian Splatting Optimization without Camera Pose Input

Wenbo Li[1,2*] **Yan Xu**[1,3*] ✉ **Mingde Yao**[1,2] **Fengjie Liang**[4] **Jiankai Sun**[5]
Menglu Wang[6] **Guofeng Zhang**[7] **Linjiang Huang**[8] **Hongsheng Li**[1,2]
[1]MMLab, The Chinese University of Hong Kong [2]CPII [3]University of Michigan
[4]Hong Kong Polytechnic University [5]Stanford [6]USTC [7]Zhejiang University [8]BUAA
wenboli@zju.edu.cn, yxumich@umich.edu,
zhangguofeng@zju.edu.cn, ljhuang@buaa.edu.cn, hsli@ee.cuhk.edu.hk

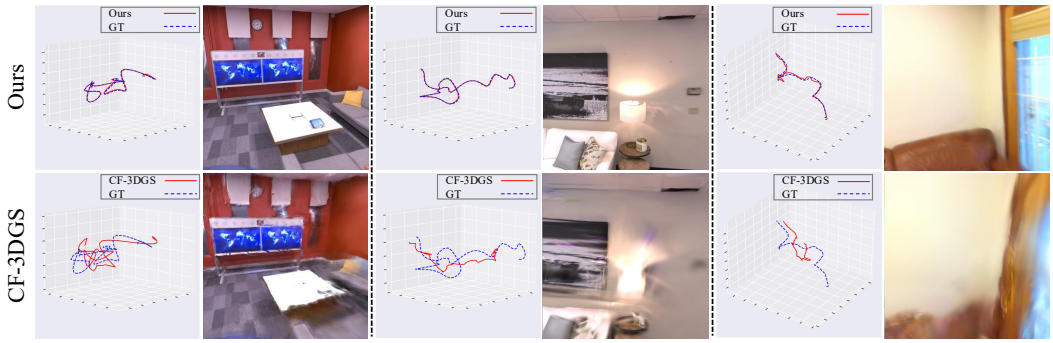

Figure 1: Comparison of camera pose estimation and novel view synthesis with the state-of-the-art. Compared to CF-3DGS [10] (bottom), our method (top) achieves more robust pose estimation and more photorealistic novel view synthesis in challenging indoor scenes with textureless regions or abrupt camera motion. Each example contains the camera trajectory estimation (left) and a sampled synthesized view (right).

Abstract

Recent advances in 3D Gaussian Splatting (3DGS) have enabled real-time, high-fidelity view synthesis, but remain critically dependent on camera poses estimated by Structure-from-Motion (SfM), which is notoriously unreliable in textureless indoor environments. To eliminate this dependency, recent pose-free variants have been proposed, yet they often fail under abrupt camera motion due to unstable initialization and purely photometric objectives. In this work, we introduce **NopeRoomGS**, an optimization framework with **no** need for camera **p**os**e** inputs, which effectively addresses the textureless regions and abrupt camera motion in indoor **room** environments through a local-to-global optimization paradigm for 3D**GS** reconstruction. In the local stage, we propose a lightweight local neural geometric representation to bootstrap a set of reliable local 3D Gaussians for separated short video clips, regularized by multi-frame tracking constraints and foundation model depth priors. This enables reliable initialization even in textureless regions or under abrupt camera motions. In the global stage, we fuse local 3D Gaussians into a unified 3DGS representation through an alternating optimization strategy that jointly refines camera poses and Gaussian parameters, effectively mitigating gradient interference between them. Furthermore, we decompose camera pose optimization based on a piecewise planarity assumption, further enhancing robustness under abrupt camera motion. Extensive experiments on Replica, ScanNet and

* Co-first authorship.
✉ Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Tanks & Temples demonstrate the state-of-the-art performance of our method in both camera pose estimation and novel view synthesis.

1 Introduction

The reconstruction of 3D representations from images for photorealistic novel view synthesis is a longstanding challenge in computer vision and graphics. It has broad applications in augmented reality, virtual reality, and robotics. Recent advances in learning-based methods [35, 2, 49, 51, 25, 31] have significantly improved rendering fidelity and generalization. However, the majority of these approaches still rely on externally estimated camera poses, typically obtained via Structure-from-Motion (SfM) pipelines [35, 49]. This dependency introduces a critical vulnerability: SfM is notoriously unreliable in low-texture regions and under non-smooth camera motion [12, 13, 32], where sparse or ambiguous feature matches lead to pose estimation failures. Such inaccuracies corrupt reconstruction optimization, but also degrade synthesis quality [38, 18, 42, 7], entangling final novel view synthesis quality with the success or failure of an external, heuristic-heavy preprocessing step.

This misalignment has motivated efforts to remove SfM entirely and jointly optimize pose and scene representation [5, 31, 53, 62, 20, 10]. However, this joint optimization problem presents a classic chicken-and-egg dilemma, requiring careful algorithmic design. BARF [31] addresses this by employing a progressive frequency scheduling strategy, gradually increasing the Fourier components in NeRF optimization to stabilize camera pose estimation while preserving accurate scene reconstruction. Nope-NeRF [5] introduces inter-frame geometric constraints and enhances the pose representation to improve stability. More recently, 3D Gaussian Splatting (3DGS) [25] has emerged as a powerful alternative to NeRF, offering both higher rendering efficiency and superior visual quality. Building on this representation, Fu et al. [10] propose the first framework for pose-free novel view synthesis. To facilitate convergence, their method incrementally grows 3D Gaussians frame by frame as the camera moves, optimizing each frame's pose solely through photometric error minimization.

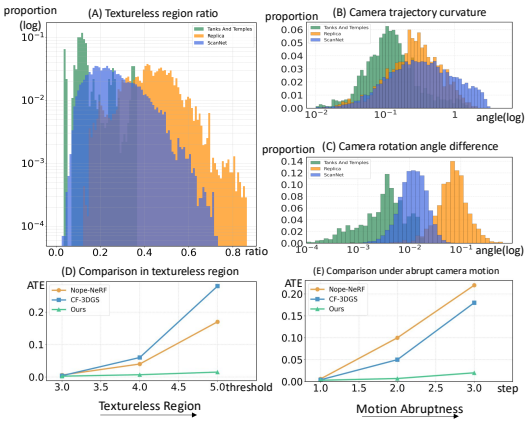

Figure 2: The distributions of low-texture regions (A), camera trajectory curvature (B), and camera rotation difference (C) vary significantly across different scenarios: Tanks & Temples [27], Replica [44], and ScanNet [9]. The indoor datasets, Replica and ScanNet, contain more textureless regions and exhibit more abrupt camera motions than Tanks & Temples, which is generally texture-rich and captured under more controlled conditions. To create more challenging scenarios, we further mask textured regions and sample frames with larger temporal steps on Tanks & Temples. Under these settings, our method demonstrates strong robustness to large textureless areas and sudden camera motions, while Nope-NeRF [5] and CF-3DGS [10] show notable performance drops (D, E).

Despite significant advancements, existing pose-free methods often assume that the photometric error from visual textures alone can provide strong and reliable gradients for accurate pose optimization. However, this assumption frequently breaks down in the presence of textureless regions or abrupt camera, e.g., in indoor environments. We evaluated several state-of-the-art methods on indoor datasets and observed notable performance degradation (see Tab. 1 for details). Furthermore, when we artificially mask texture-rich regions or increase the camera motion speed within the datasets originally used by these methods, we observe similarly pronounced performance drops (Fig. 2). These findings empirically validate our hypothesis regarding the limitations of the current approaches in indoor environments.

To overcome these limitations, we propose NopeRoomGS, a robust pose-free 3DGS framework designed for scenarios where textureless regions and abrupt camera motion prevail, especially indoor environments. Our method employs a local-to-global scheme to progressively build up the 3DGS.

In the local stage, we introduce a lightweight local neural geometric representation that jointly optimizes depth and camera pose on short, overlapping video clips extracted from the input sequence. To recover the consistent geometry of each video clip, the optimization is supervised by multi-frame tracking constraints derived from CoTracker [23], which remain effective even under abrupt camera

motion. Furthermore, to enhance robustness in textureless regions, we incorporate regularization from a pretrained monocular depth foundation model [24]. This design enables reliable recovery of local 3D geometry and camera poses under challenging conditions with textureless regions and abrupt camera motion, thereby providing a stable initialization of local 3D Gaussians for subsequent global fusion.

In the global stage, we progressively fuse these local 3D Gaussians into a unified global 3DGS representation over entire video sequence. This fusion process is formulated as an optimization problem, where the global 3DGS is supervised by a combination of photometric loss, depth alignment loss, and pose constraints derived from piecewise planarity assumption, ensuring global geometric consistency. Furthermore, we adopt an alternating optimization strategy that updates Gaussian parameters and camera poses in turn, rather than jointly, to reduce gradient interference and enhance convergence stability. This strategy improves the robustness of global reconstruction in challenging scenarios with textureless regions and abrupt camera motion.

Our contributions are summarized as follows:

- We propose a pose-free 3DGS framework with a local-to-global scheme, which exhibits strong robustness in textureless regions and under abrupt camera motion.
- In the local stage, we introduce a lightweight neural geometric representation that jointly optimizes depth and camera pose over short video clips. It yields locally consistent geometry and accurate poses, even in textureless regions and under abrupt camera motion, effectively bootstrapping a set of local 3D Gaussians.
- In the global stage, we progressively fuse local 3D Gaussians into a unified global 3DGS representation with alternative optimization strategy. To ensure stable convergence, we supervise the model with the combination of photometric loss, depth alignment loss, and pose constraints derived from piecewise planarity assumption. These components collectively enforce structural consistency in the global reconstruction.
- Experiments on public datasets, including Replica [44] and ScanNet [9] and Tanks & Temples [27], demonstrate that our method achieves state-of-the-art performance, with particularly strong results in indoor scenes and competitive results in general scenarios.

2 Related Work

Novel view synthesis. Generating photorealistic images from novel viewpoints is a key challenge in computer vision, which has been approached using diverse 3D representations. Among them, Neural Radiance Fields (NeRFs) [35] have emerged as a leading method, with extensions tackling challenges such as aliasing [2, 3, 4], surface reflectance [48, 1], sparse views [26, 37, 55, 60, 19]. In parallel, explicit point-based or mesh-based representations have gained increasing attention for their efficiency and interpretability [56, 65, 25, 34, 28, 64, 33]. In particular, 3DGS [25] demonstrates that real-time, high-fidelity view synthesis can be achieved by representing scenes as sets of anisotropic Gaussian primitives optimized through differentiable rendering. However, many of these methods still rely on pre-computed camera poses. Although the camera pose estimation has been studied for decades [36, 43, 58, 57, 47, 16, 17, 50], it is still challenging to robustly estimate the camera poses with low-cost sensors. Most recent novel view synthesis methods rely on Structure-from-Motion tools like COLMAP [11, 43, 36, 46, 33], which limit their applicability in scenarios where obtaining accurate camera poses is challenging, such as in low-texture environments or when only sparse or unstructured image collections are available.

Radiance field without camera pose prior. Recent studies have integrated pose estimation into NeRF training to remove dependence on pre-computed camera poses. Early methods, such as i-NeRF [63], refined camera parameters by aligning keypoints to a pre-trained NeRF, while NeRFmm [54] jointly optimized both the NeRF model and camera poses, albeit with limitations to forward-facing scenes. BARF [31] addressed gradient inconsistency in positional encoding with coarse-to-fine optimization, but still required initial pose estimates within 15°of the ground truth.

Nope-NeRF [5] leveraged monocular depth priors for both scene reconstruction and pose estimation. Despite these advances, NeRF-based methods remain computationally intensive and face challenges in achieving real-time rendering performance [40].

3DGS without camera pose prior. Recent work such as CF-3DGS [10] first removes COLMAP dependency by jointly optimizing poses and Gaussians under inter-frame photometric supervision, making them susceptible to failure under large camera motion or textureless regions. In parallel, a growing line of feed-forward, pose-free 3DGS methods [61, 22, 15, 8] propose generalizable

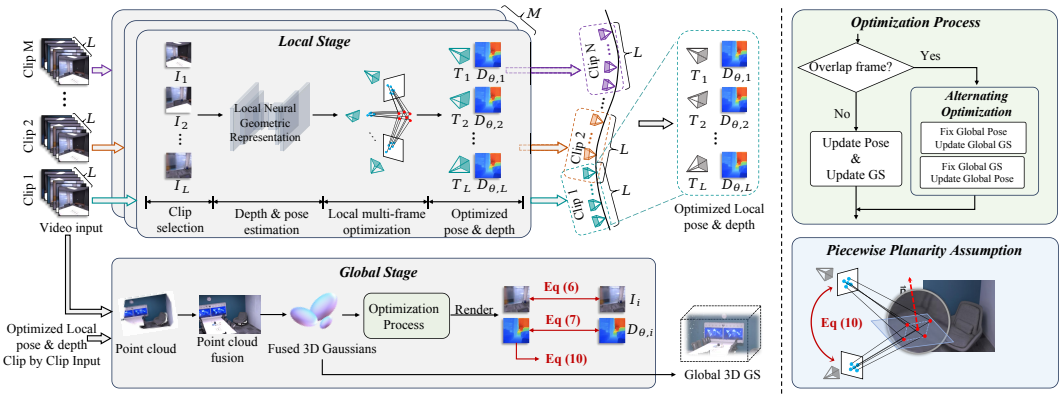

Figure 3: The pipeline of NopeRoomGS. Our NopeRoomGS framework adopts a Local-to-Global optimization scheme to address textureless regions and abrupt camera motion. In the local stage (Sec. 3.1), a lightweight Local Neural Geometric Representation jointly optimizes depth and camera pose for each clip, yielding consistent geometry and local 3D Gaussians. In the global stage (Sec. 3.2), these are fused into a unified 3DGS via differentiable optimization with photometric, depth, and pose constraints under the Piecewise Planarity Assumption. An Alternating Optimization strategy mitigates gradient interference, ensuring stable and accurate reconstruction.

networks to predict Gaussians (and often poses) without per-scene optimization. NoPoSplat [61] predicts 3D Gaussians directly in a canonical space, whereas SelfSplat [22] jointly learns camera poses and scene geometry via self-supervised depth and photometric consistency. Although these approaches remove the need for external SfM, they still face limitations when processing very large image collections, where computational and memory costs grow sharply and maintaining stable global pose/geometry estimates becomes challenging.

In contrast, we fully exploit 3D structural-consistency constraints and foundation-model priors with a lightweight local neural geometric representation, and fuse the resulting local 3D Gaussians using alternating optimization strategy of camera poses and 3D Gaussian parameters, which stabilizes initialization, reduces gradient interference, and yields globally consistent reconstructions in challenging scenarios with textureless regions and abrupt camera motion.

3 Method

Given a sequence of N unposed images $\mathcal{I} = \{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{N}$ with known camera intrinsic, our goal is to recover the camera poses $\mathcal{T} = \{\mathbf{T}_i \in \mathbb{SE}(3)\}_{i=1}^{N}$ and reconstruct a unified global 3DGS representation $\mathcal{G}$ of the scene.

This task is fundamentally challenging because both accurate camera poses and reliable 3DGS representation must be recovered simultaneously from unposed inputs. The explicit nature of the 3DGS representation, though beneficial for efficient rendering, makes this joint optimization highly non-convex and sensitive to initialization [21, 6]. In particular, the problem is exacerbated in indoor environments characterized by textureless surfaces and abrupt camera motion, where standard feature-based methods fail and gradient descent is prone to local minima [14]. To improve robustness, we adopt a local-to-global optimization scheme, as illustrated in Fig. 3. In the local stage (Sec. 3.1), we introduce a learnable local neural geometric representation that jointly estimates depth and camera pose within each video clip. This yields reliable scene structure and camera pose to bootstrap a set of local 3D Gaussians as an initialization for global optimization, which is critical for stable convergence [59]. In the global stage (Sec. 3.2), these local 3D Gaussians are progressively fused into a unified global 3DGS representation. This fusion is formulated as a differentiable optimization problem, supervised by a combination of photometric loss, depth alignment loss, and pose constraints based on a piecewise planarity assumption. In addition, we adopt an alternating optimization strategy to updates camera poses and Gaussian parameters in turn, rather than jointly [30], to mitigate gradient interference and enhance convergence.

3.1 Local Stage

In the local stage, to address the challenging scenario with textureless regions and abrupt camera motion, we propose a local neural geometric representation that jointly optimizes camera poses

and per-frame depth via gradient descent. The optimization is supervised by multi-frame tracking constraints [23], which enhance geometric coherence under abrupt camera motion. Additionally, we incorporate strong depth priors from a pretrained foundation model [24] to regularize the solution in textureless regions. Unlike prior pairwise methods [10], our representation is shared across all frames within a clip, enabling consistent geometric reasoning across different viewpoints.

Local neural geometric representation. To ensure consistent geometry under challenging conditions such as textureless regions and abrupt camera motion, we parameterize the local scene structure using a lightweight local neural geometric representation $\mathcal{F}_\theta$, where θ denotes learnable parameters. To make this representation both robust and rapidly deployable, $\mathcal{F}_\theta$ is initialized from a fast monocular depth estimator [41] and shared across all frames within a short video clip, enabling coherent geometry over the video clip.

Given a short clip $\mathcal{I} = \{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{L}$ of length L, the local neural geometric representation $\mathcal{F}_\theta$ takes as input a color image I_i to predict the corresponding depth map $D_{\theta,i}$:

$$D_{\theta,i} \;=\; \mathcal{F}_\theta(I_i). \tag{1}$$

We denote the depth maps of the clip as $\mathcal{D}_\theta = \{D_{\theta,i} \in \mathbb{R}^{H\times W}\}_{i=1}^{L}$. By sharing the network $\mathcal{F}_\theta$ across the local video clip, consistent geometric cues can be learned during optimization. The intuition of this optimization process is similar to the classic bundle adjustment, but our approach is based on neural representations.

Joint optimization of local camera pose and depth. After obtaining the initial depth maps above, we jointly optimize the local camera poses $\mathcal{T} = \{\mathbf{T}_i \in \mathrm{SE}(3)\}_{i=1}^{L}$ and the depth maps $\mathcal{D}_\theta$ by enforcing geometric consistency across frames.

Specifically, for any pixel $\mathbf{u}_i$ in the frame i, we can get its correspondences $\{\mathbf{u}_j | j \in \mathcal{N}(\mathbf{u}_i)\}$ in the neighboring frames $\mathcal{N}(\mathbf{u}_i)$ by using the quasi-dense off-the-shelf multi-frame tracking constraints derived from CoTracker [23]. To enforce the geometric constraints for pose optimization, we first unproject the pixel $\mathbf{u}_i$ to get its 3D position $\mathbf{p}_i$ in the current frame:

$$\mathbf{p}_i = \pi^{-1}(\mathbf{u}_i) = D_{\theta,i}(\mathbf{u}_i)\mathbf{K}^{-1}\begin{pmatrix}\mathbf{u}_i\\ 1\end{pmatrix}. \tag{2}$$

where $\mathbf{K}$ denotes the intrinsic matrix and $D_{\theta,i}(\mathbf{u}_i)$ is its estimated depth value in Eq. 1. Then, we map it to the nearby frames and enforce the geometric consistency by minimizing the projection error:

$$\mathcal{L}^i_{\text{proj}} = \sum_{j\in\mathcal{N}(\mathbf{u}_i)} \Big\| \pi\big(\mathbf{T}_{j\leftarrow i}\pi^{-1}(\mathbf{u}_i)\big) - \mathbf{u}_j \Big\|^2, \tag{3}$$

where $\pi(\cdot)$ denotes the pinhole camera projection function given the intrinsic matrix $\mathbf{K}$, and $\mathbf{T}_{j\leftarrow i} \in \mathrm{SE}(3)$ denotes the relative pose transformation from frame i to frame j.

To avoid the depth optimization deviate too far away from the realistic solution manifold, we regularize its prediction with the output from a stronger but much heavier monocular depth foundation model [24], striking a balance between performance and efficiency. This regularization is implemented by a scale-and-shift-invariant regularization loss [41]:

$$\mathcal{L}^i_{\text{ssi}} = \rho\Big(D_{\theta,i} - \alpha\,\tilde{D}_i - \beta\Big), \tag{4}$$

where $D_{\theta,i}$ is the output from local neural geometric representation $\mathcal{F}_\theta$, and $\tilde{D}_i$ is the pseudo ground-truth produced by the foundation model, and $\rho(\cdot)$ is a distance function (i.e., Huber loss). The parameters α and β are obtained by solving a least-squares problem [41] to resolve the scale ambiguity of monocular depth estimation.

The overall objective function in the local stage is thus:

$$\mathcal{T}, \mathcal{D}_\theta = \arg\min_{\mathcal{T},\mathcal{D}_\theta} \mathcal{L}_{proj} + w\mathcal{L}_{ssi}, \tag{5}$$

where $\mathcal{L}_{proj}$ and $\mathcal{L}_{ssi}$ are the summation of all the $\mathcal{L}^i_{\text{proj}}$ and $\mathcal{L}^i_{\text{ssi}}$ respectively within the same clip, and the weight w controls the regularization strength. Here, we slightly abuse the notation $\mathcal{D}_\theta$, as we optimize the parameter θ practically.

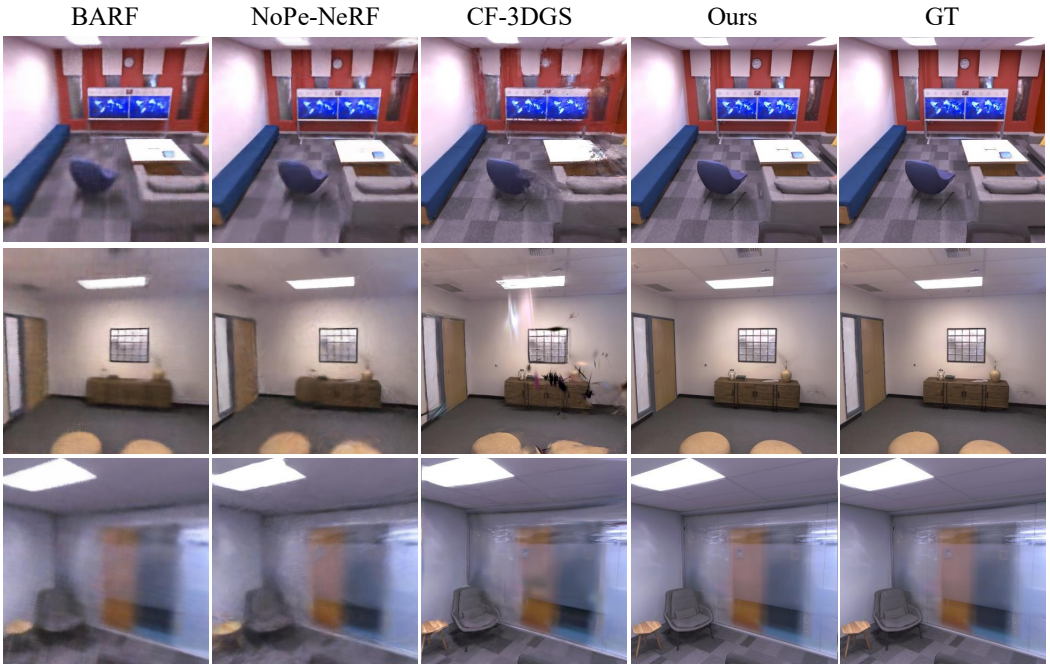

Figure 4: Qualitative comparison of novel view synthesis on Replica [44] dataset. Compared with other cutting-edge counterparts, our method synthesizes more detailed textures under the pose-free setting.

3.2 Global Stage

After obtaining the local camera poses $\mathcal{T}$ and depth maps $\mathcal{D}_\theta$ within each clip via Eq. 5, we initialize local 3D Gaussians accordingly and progressively merge them into a unified global 3DGS representation $\mathcal{G}$.

Specifically, for each frame i within a clip, we unproject its depth map $D_{\theta,i}$ to the 3D space using camera intrinsic $\mathbf{K}$ (Eq. 2) to get the point cloud $\mathbf{P}_i$ and transform it using the estimated pose $\mathbf{T}_i$ to the local frame of this clip. After finishing this process, we get a local point cloud for each clip. To decrease the memory burden, we downsample the point cloud and initialize the local 3D Gaussians with the downsampled point cloud. Then, the local 3D Gaussians are merged into a unified global 3DGS representation $\mathcal{G}$ according to the relative pose between the local frame and the global frame. Thereafter, we optimize the global 3D Gaussian parameters as well as the camera pose based on the input image $\mathcal{I}$ and estimated depth map $\mathcal{D}_\theta$ from the local stage. In the ensuing part, we will elaborate on the objective function we used for optimization.

3.2.1 Objective Function

Photometric loss. Following the original 3DGS [25], we include the photometric loss terms in the objective function:

$$\mathcal{L}^i_{\text{rgb}} = \gamma \left\| I_{\mathcal{G}}(\mathbf{T}_i) - I_i \right\|_1 + (1-\gamma)\mathcal{L}_{D-SSIM}(I_{\mathcal{G}}(\mathbf{T}_i), I_i)\,, \tag{6}$$

which is constituted by a L_1 term and a D-SSIM term [25]. $I_{\mathcal{G}}(\mathbf{T}_i)$ represents the rendered image from the Gaussian Splatting $\mathcal{G}$ with the camera pose $\mathbf{T}_i$, I_i is the i-th input image, and γ is the hyperparameter to balance the two terms.

Depth alignment loss. With the photometric loss, the camera poses can be effectively corrected by the regions with rich textures, as these regions can produce strong gradients if the poses are erroneous. However, the indoor scenes are full of textureless areas, e.g., blank walls, which pose great challenges to pose-free 3DGS. To enhance the robustness in textureless regions, we also add a depth alignment term similar to Eq. 4:

$$\mathcal{L}^i_{\text{depth}} = \rho\Big(D_{\mathcal{G}}(\mathbf{T}_i) - \alpha\, D_{\theta,i} - \beta\Big). \tag{7}$$

The difference is that we here use the optimized depth maps $D_{\theta,i}$ from the local stage as the ground truth to constrain the depth maps $D_{\mathcal{G}}(\mathbf{T}_i)$ rendered with camera pose $\mathbf{T}_i$.

Table 1: Camera pose estimation performance metrics comparison on Replica [44] dataset. All baseline methods are trained with their official implementations and original configurations, and evaluated following the same protocol to ensure fair and consistent comparison. The best results are highlighted in bold.

Scenes	Ours			CF-3DGS [10]			NoPoSplat [61]			SelfSplat [22]			Nope-NeRF [5]			BARF [2]			NeRFmm [54]		
	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE $\downarrow$	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	RPE_t	RPE_r	ATE
office0	**0.231**	**0.153**	**0.011**	1.168	7.282	0.043	1.456	7.456	0.053	1.073	7.654	0.074	1.231	5.462	0.042	1.561	6.273	0.052	1.467	6.834	0.060
office1	**0.020**	**0.045**	**0.001**	1.171	6.288	0.078	1.325	5.946	0.067	0.946	8.678	0.064	1.416	7.922	0.056	1.462	5.234	0.048	1.432	7.236	0.054
office2	**0.404**	**0.260**	**0.027**	0.816	1.668	0.044	1.874	7.764	0.053	1.554	6.832	0.074	1.273	5.234	0.047	1.073	6.235	0.055	1.346	6.892	0.023
office3	**0.056**	**0.062**	**0.003**	1.489	5.994	0.093	2.113	7.235	0.067	2.038	8.245	0.056	1.156	6.002	0.052	1.119	7.345	0.056	1.489	7.231	0.054
office4	**0.083**	**0.126**	**0.003**	1.086	6.901	0.048	1.932	7.567	0.072	1.807	7.113	0.054	0.923	6.231	0.037	1.223	5.987	0.041	1.946	6.923	0.045
room0	**0.051**	**0.055**	**0.003**	1.201	6.545	0.044	1.834	6.593	0.046	1.504	7.832	0.074	1.023	5.432	0.013	1.327	6.432	0.052	1.835	7.235	0.033
room1	**0.230**	**0.175**	**0.012**	1.088	6.961	0.056	2.325	5.745	0.043	2.164	6.883	0.095	0.875	6.256	0.025	1.045	5.436	0.043	1.322	6.923	0.072
room2	**0.111**	**0.128**	**0.008**	1.177	5.584	0.064	1.325	8.385	0.084	1.164	7.224	0.083	1.231	5.467	0.036	1.032	5.467	0.023	1.347	6.342	0.043
mean	**0.148**	**0.126**	**0.009**	1.150	5.903	0.059	1.773	7.086	0.061	1.531	7.56	0.072	1.141	6.001	0.039	1.230	6.051	0.046	1.523	6.962	0.048

Pose constraint based on piecewise planarity assumption. The camera motion could be abrupt in indoor scenes. To further improve the robustness against abrupt camera motion, we propose a cross-frame geometric constraint based on a piecewise planarity assumption.

We assume each point $\mathbf{x}_p$ in the scene lies on a infinitesimal piecewise plane defined by $\mathbf{n}_i^\top \mathbf{x}_p + \delta_i = 0$, where $\mathbf{n}_i$ is the surface normal and δ_i is the displacement coefficient. The normal and the displacement coefficient are calculated from the rendered depth map of 3DGS $\mathcal{G}$, using 4 surrounding pixels (left, right, upper, lower).

Given two consecutive frames, I_i and I_{i+1} with a relative pose estimation $\mathbf{T}_{i\to i+1} = [\mathbf{R}_{i\to i+1} \mid \mathbf{t}_{i\to i+1}]$ between, the plane parameters in frame i, denoted as $(\mathbf{n}_i, \delta_i)$, can be transformed to frame $i+1$ by

$$\hat{\mathbf{n}}_{i+1} = \mathbf{R}_{i\to i+1}\mathbf{n}_i, \tag{8}$$

$$\hat{\delta}_{i+1} = \delta_t - \mathbf{n}_i^\top \mathbf{t}_{i\to i+1}, \tag{9}$$

to get the transformed plane parameters $(\hat{\mathbf{n}}_{i+1}, \hat{\delta}_{i+1})$.

We then enforce consistency between the transformed plane parameters $(\hat{\mathbf{n}}_{i+1}, \hat{\delta}_{i+1})$ and the directly estimated values $(\mathbf{n}_{i+1}, \delta_{i+1})$ in the frame $i+1$, using the following loss function:

$$\mathcal{L}_{\text{plane}}^{i+1} = \lambda_n \left\| 1 - \hat{\mathbf{n}}_{i+1}^\top \mathbf{n}_{i+1} \right\|_2^2 + \lambda_\delta \left\| \hat{\delta}_{i+1} - \delta_{i+1} \right\|_2^2, \tag{10}$$

where λ_n and λ_δ are the regularization parameters that control the relative importance of normal and offset consistency, respectively. This loss function ensures that the geometries in consecutive frames are coherently aligned and decomposes the supervision to rotational and translational components of the camera poses. We found that, by adjusting λ_n and λ_δ, we can achieve more robust pose estimation in indoor environments. To further enhance robustness, we also extract the edge map based on input frame, which is used as a mask to restrict this constraint only to planar regions, while avoiding its effect on the plane edges. Further detail can be found in the supplementary materials.

Overall global optimization. By combining the objective functions mentioned above, the overall optimization process is formulated by summing over all the frames:

$$\mathcal{T}, \mathcal{G} = \arg\min_{\mathcal{T},\mathcal{G}} \sum_{i=1}^{L} \lambda_{\text{plane}}\mathcal{L}_{\text{plane}}^i + \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}}^i + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}^i, \tag{11}$$

where the weights λ_{plane}, λ_{rgb}, and λ_{depth} balance different terms.

3.2.2 Alternating Optimization Strategy

Although Eq. 11 defines a unified objective over both the global camera poses $\mathcal{T}$ and the 3D Gaussian parameters $\mathcal{G}$, jointly optimizing them often results in unstable convergence due to gradient interference between the two parameter spaces [21, 6, 14]. To mitigate this issue, we adopt an alternating optimization strategy, updating camera poses and 3D Gaussian parameters in turn. This decoupled scheme improves convergence stability and preserves global consistency, as validated by our ablation studies in Sec. 4.4. Further implementation details, including the gradient update formulation for camera pose parameters, are provided in the supplementary material.

Table 2: Novel view synthesis performance metrics comparison on Replica [44] dataset. For fair comparison, each baseline is trained using its publicly released code and original hyperparameter settings, and evaluated under the same protocol. The best results are highlighted in bold.

Scenes	Ours			CF-3DGS [10]			NoPoSplat [61]			SelfSplat [22]			Nope-NeRF [5]			BARF [2]			NeRFmm [54]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
office0	**34.32**	**0.92**	**0.09**	28.50	0.87	0.28	17.34	0.57	0.47	16.50	0.56	0.35	28.26	0.85	0.51	25.23	0.78	0.48	23.52	0.58	0.49
office1	**31.78**	**0.92**	**0.10**	29.11	0.85	0.18	18.52	0.41	0.41	18.11	0.59	0.40	27.35	0.69	0.40	26.38	0.63	0.53	25.33	0.61	0.53
office2	**30.77**	**0.91**	**0.11**	24.97	0.78	0.29	18.54	0.56	0.34	15.97	0.45	0.43	26.77	0.76	0.35	26.52	0.73	0.43	23.28	0.61	0.54
office3	**32.37**	**0.90**	**0.12**	23.70	0.75	0.25	15.84	0.54	0.43	21.70	0.57	0.45	26.01	0.74	0.41	26.37	0.61	0.41	25.25	0.78	0.45
office4	**32.12**	**0.90**	**0.09**	27.49	0.78	0.28	18.93	0.59	0.35	15.49	0.56	0.38	27.64	0.84	0.26	26.48	0.72	0.36	24.08	0.70	0.43
room0	**33.21**	**0.89**	**0.06**	22.60	0.68	0.38	19.96	0.47	0.45	18.63	0.54	0.47	25.33	0.72	0.38	25.53	0.64	0.32	22.93	0.48	0.51
room1	**32.13**	**0.87**	**0.08**	24.50	0.74	0.35	23.92	0.53	0.46	22.15	0.48	0.43	29.42	0.80	0.38	25.54	0.56	0.54	25.40	0.69	0.43
room2	**30.40**	**0.88**	**0.10**	25.34	0.74	0.34	22.08	0.55	0.36	23.42	0.55	0.48	28.96	0.61	0.47	24.83	0.53	0.60	24.16	0.43	0.40
mean	**32.14**	**0.90**	**0.09**	25.40	0.77	0.29	19.39	0.52	0.41	18.90	0.54	0.42	27.46	0.75	0.40	25.86	0.65	0.46	24.24	0.61	0.47

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate our method on three public datasets: Replica [44], ScanNet [9], and Tanks & Temples [27], covering both synthetic and real-world scenes. The Replica [44] dataset offers high-fidelity synthetic indoor scenes with precise ground-truth camera poses. Its large textureless regions and complex camera trajectories make it well-suited for evaluating camera pose estimation and novel view synthesis. The ScanNet [9] dataset consists of real-world RGB-D indoor scenes captured in unconstrained environments, presenting challenges such as sensor noise and motion blur. Tanks & Temples [27] dataset contains texture-rich scenes with relatively controlled camera motions. Most previous methods are evaluated on this dataset. We include it to test the generalizability of our method.

Metrics. Following prior works [10, 5], we evaluate our method on two key tasks: camera pose estimation and novel view synthesis. For camera pose estimation, we use standard visual odometry metrics [29, 45], including Absolute Trajectory Error (ATE) and rotational Relative Pose Error (RPE_r) and translational Relative Pose Error (RPE_t). For novel view synthesis, we use standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [52], and Learned Perceptual Image Patch Similarity (LPIPS) [66].

Implementation Details. Our method is implemented using PyTorch [39], building on the optimization settings from 3DGS [25] with necessary adjustments. A key feature is synchronizing new frame additions with point densification intervals to ensure steady scene expansion. For detailed information, please refer to the supplementary materials.

4.2 Replica

In this subsection, we conduct a comparative analysis of our method against several established baselines, including NoPoSplat [61], SelfSplat [22], CF-3DGS [10], Nope-NeRF [5], BARF [2] and NeRFmm [54], all of which are widely recognized for their contributions to camera pose estimation and novel view synthesis.

For camera pose estimation, the optimized camera poses are aligned with Procrustes Analysis, as described in prior works [5, 54], and compared against the ground-truth poses from training views. Quantitative results are summarized in Tab. 1, where our method achieves performance superior to current state-of-the-art approaches.

For novel view synthesis, we adopt the evaluation protocol of CF-3DGS [10] and NeRFmm [54]. The optimized 3DGS model, trained exclusively on the training views, is kept fixed, while the camera poses of the test views are refined by minimizing the photometric reconstruction error between synthesized and ground-truth images. As reported in Tab. 2, our method consistently outperforms all baselines. Moreover, the qualitative comparisons in Fig. 4 highlight that our synthesized images preserve finer details and exhibit sharper textures than those produced by existing methods.

Table 3: Camera pose estimation and novel view synthesis performance metrics comparison on ScanNet [9] dataset. For fair comparison, each baseline is trained using its publicly released code and original hyperparameter settings, and evaluated under the same evaluation protocol. The best results are highlighted in bold.

Methods	RPE_t ↓	RPE_r ↓	ATE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRFmm [54]	1.153	0.963	0.123	15.50	0.59	0.53
BARF [2]	1.145	0.894	0.134	21.42	0.54	0.45
Nope-NeRF [5]	0.763	0.688	0.040	22.11	0.64	0.39
SelfSplat [22]	0.984	0.932	0.231	18.28	0.52	0.45
NoPoSplat [61]	1.167	0.875	0.124	17.28	0.54	0.67
CF-3DGS [10]	0.653	0.684	0.040	23.26	0.68	0.24
Ours	**0.579**	**0.524**	**0.020**	**25.38**	**0.72**	**0.20**

4.3 ScanNet

To further validate the effectiveness of our method in realistic and unconstrained environments, we evaluate it on ScanNet [9] dataset, which consists of real-world indoor scenes. Compared to Replica [44] dataset, ScanNet [9] dataset introduces additional challenges such as sensor noise, cluttered layouts, and motion blur, making it a more demanding benchmark for pose and reconstruction quality. We follow the evaluation protocol of [10, 5] to assess both camera pose estimation accuracy and novel view synthesis quality. As shown in Tab. 3, our method consistently achieves superior results across all evaluation metrics.

4.4 Tanks & Temples

While Replica [44] and ScanNet [9] datasets provide indoor scenes, we further evaluate the generalization ability and robustness of our method on Tanks & Temples [27] dataset, which consists of photogrammetric reconstructions of both indoor and outdoor scenes with texture variations. Following standard evaluation protocols established in [10, 5], we assess both camera pose estimation accuracy and novel view synthesis quality. As reported in Tab. 4, our method performs competitively with state-of-the-art approaches across nearly all metrics. Qualitative results are presented in Fig. 5.

Table 4: Camera pose estimation and novel view synthesis performance metrics comparison on Tanks & Temples [27] dataset. All baseline methods are trained with their official implementations and original configurations, and evaluated following the same protocol to ensure fair and consistent comparison. The best results are highlighted in bold.

Methods	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRFmm [54]	1.735	0.477	0.123	22.50	0.59	0.54
BARF [2]	1.046	0.441	0.078	23.42	0.61	0.54
Nope-NeRF [5]	0.080	**0.038**	0.006	26.34	0.74	0.39
SelfSplat [22]	1.046	0.489	0.094	22.42	0.58	0.56
NoPoSplat [61]	1.832	0.488	0.117	20.15	0.53	0.47
CF-3DGS [10]	0.041	0.069	0.004	31.28	0.93	0.09
Ours	**0.034**	0.043	**0.003**	**31.68**	**0.94**	**0.07**

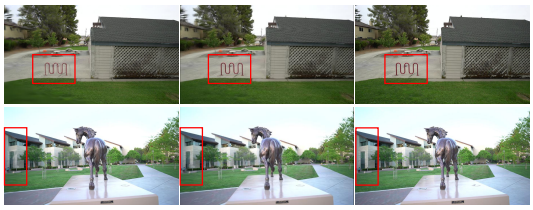

Figure 5: Qualitative results of novel view synthesis on Tanks & Temples [27] dataset. Our NopeRoomGS produces more realistic rendering results than other baselines.

4.5 Ablation Study

In this section, we analyze the effectiveness of different key components proposed in our pipeline through systematic ablation studies.

Effectiveness of local neural geometric representation. To evaluate the contribution of the local neural geometric representation (LNGR), we conduct an ablation study where LNGR is removed and replaced with a naive initialization strategy: the camera pose of each frame is initialized using that of the previous frame, and depth is directly estimated from a monocular foundation model [24] without LNGR refinement.

As shown in Tab. 5 (row "w/o LNGR") and Fig. 6, this configuration causes a noticeable drop in both pose accuracy and rendering quality, which underscores the essential role of LNGR $\mathcal{F}_\theta$ in handling challenging indoor scenarios with textureless regions and abrupt camera motion. As described in Sec. 3.1, by jointly optimizing depth and pose across short video clips, $\mathcal{F}_\theta$ produces consistent and robust local scene geometry and camera pose. These outputs provide reliable initializations for local 3D Gaussians, which in turn supply high-quality geometry and pose priors for the subsequent global optimization stage. Such strong initializations are crucial for ensuring stable convergence and achieving high reconstruction fidelity.

Effectiveness of piecewise planarity assumption. To assess the impact of the piecewise planarity assumption (PPA), we ablate the PPA constraint from the global optimization stage and replace it with pairwise depth consistency loss computed via reprojection between adjacent frames.

Table 5: Ablation study on Replica [44] dataset. A comparison of our full pipeline and variants without Local Neural Geometric Representation (LNGR), Piecewise Planarity Assumption (PPA), Alternating Optimization Strategy (AOS), and Depth Alignment Loss (DAL) (described in Eq. 7), respectively.

Methods	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ours	0.148	0.126	0.009	32.14	0.90	0.09
w/o LNGR	1.532	7.422	0.063	14.82	0.65	0.42
w/o PPA	0.192	0.179	0.023	30.34	0.89	0.14
w/o AOS	0.205	0.189	0.019	28.28	0.89	0.15
w/o DAL	0.156	0.158	0.078	31.42	0.90	0.14

As shown in Tab. 5 (row "w/o PPA") and Fig. 6, this substitution results in a decline in both camera pose accuracy and novel view synthesis quality, thereby substantiating the efficacy of PPA in guiding global optimization. As detailed in Sec. 3.2.1, PPA improves camera pose estimation in textureless regions by introducing geometry-

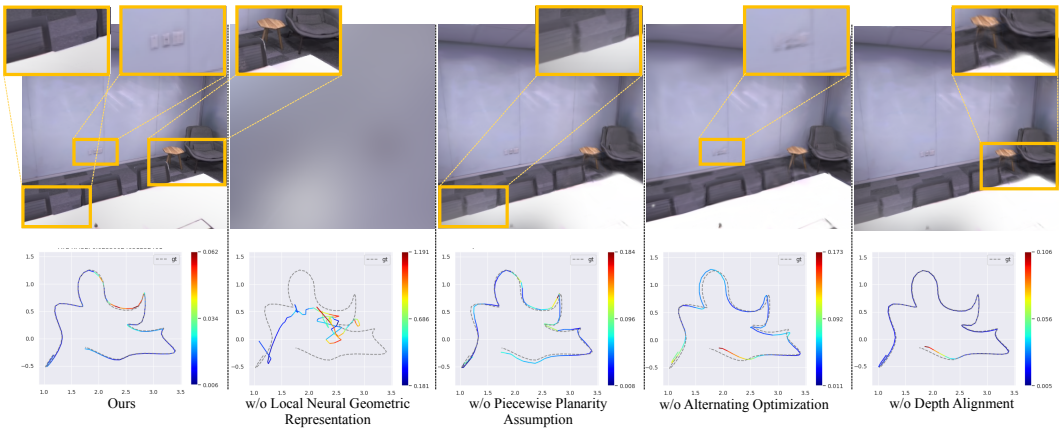

Figure 6: Ablation study for camera pose estimation and novel view synthesis on Replica [44] dataset. We compare our full pipeline with the variants without local neural geometric representations, piecewise planarity assumption, alternating optimization strategy, and depth alignment loss as described in Eq. 7, respectively. The 3D trajectories are projected onto the XY plane of the coordinate system.

aware supervision, and stabilizes optimization under abrupt camera motion by decomposing the supervision signal into rotational (normal alignment) and translational (offset consistency) components, which contributes to the robustness of the overall reconstruction process.

Effectiveness of alternating optimization strategy. To evaluate the impact of our incremental and Alternating Optimization Strategy (AOS), we replace it with a joint global optimization method. As shown in Tab. 5 (row "w/o AOS") and Fig. 6, this substitution leads to a drop in both camera pose accuracy and novel view synthesis quality, highlighting the effectiveness of our proposed design. Given the same reliable initialization, alternating optimization plays a critical role in achieving stable convergence and higher reconstruction quality by mitigating interference between camera pose and 3D Gaussian parameters updates.

Effectiveness of depth alignment loss. To assess the effectiveness of the Depth Alignment Loss (DAL) in Eq. 7, we replace the optimized depth from the local stage with monocular predictions from a pretrained foundation model [24] to supervise the depth rendered from the global 3DGS stage. As shown in Tab. 5 (row "w/o DAL") and Fig. 6, this leads to degraded camera pose accuracy and novel view synthesis quality. These results highlight the importance of depth alignment supervision and demonstrate the effectiveness of our local neural geometric representation in recovering consistent scene geometry.

5 Conclusion and Limitation

In this work, we propose NopeRoomGS, a fully pose-free (i.e., no pose priors) 3D Gaussian Splatting framework that progressively recovers both camera poses and 3DGS representation via a local-to-global optimization paradigm. In the local stage, we introduce a lightweight local neural geometric representation that is jointly optimized on short video clips under supervision from multi-frame tracking and foundation-model depth priors, thereby enabling accurate reconstruction in textureless regions and under abrupt camera motion. In the global stage, we fuse the resulting local 3D Gaussians into a unified 3DGS representation through alternating optimization, ensuring geometric consistency under challenging conditions. Extensive experiments on public datasets demonstrate that our method achieves state-of-the-art performance in both camera pose estimation and novel view synthesis, extending the applicability of 3DGS to real-world, unconstrained environments.

Despite the improved robustness in indoor scenarios and better handling of complex camera motion, several limitations remain. First, the local stage introduces additional computational overhead compared with vanilla 3DGS. Second, even with local-to-global optimization and multiple loss constraints, the method can fail under extremely rapid camera motion or severely sparse inputs. We aim to address these issues in future work.

Societal Impact. This technology can benefit AR/VR, robotics, digital content creation, telepresence, and cultural heritage preservation. However, its computational demands may contribute to a higher carbon footprint.

Acknowledgment. This study was supported in part by National Key R&D Program of China Project 2022ZD0161100, in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, in part by NSFC-RGC Project N_CUHK498/24, and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515130008, XW).

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, June 2023.

[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.

[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.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023.

[6] Keng-Wei Chang, Zi-Ming Wang, and Shang-Hong Lai. Keygs: A keyframe-centric gaussian splatting method for monocular image sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1989–1997, 2025.

[7] Shu Chen, Yang Zhang, Yaxin Xu, and Beiji Zou. Structure-aware nerf without posed camera via epipolar constraint. *arXiv preprint arXiv:2210.00183*, 2022.

[8] Zequn Chen, Jiezhi Yang, and Heng Yang. Pref3r: Pose-free feed-forward 3d gaussian splatting from variable-length image sequence. *arXiv preprint arXiv:2411.16877*, 2024.

[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.

[10] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, June 2024.

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

[12] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. *CVPR*, 2024.

[13] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21594–21603, 2024.

[14] Lukas Höllein, Aljaž Božič, Michael Zollhöfer, and Matthias Nießner. 3dgs-lm: Faster gaussian-splatting optimization with levenberg-marquardt. *arXiv preprint arXiv:2409.12892*, 2024.

[15] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024.

[16] Zhaoyang Huang, Yan Xu, Jianping Shi, Xiaowei Zhou, Hujun Bao, and Guofeng Zhang. Prior guided dropout for robust visual localization in dynamic environments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2791–2800, 2019.

[17] Zhaoyang Huang, Han Zhou, Yijin Li, Bangbang Yang, Yan Xu, Xiaowei Zhou, Hujun Bao, Guofeng Zhang, and Hongsheng Li. Vs-net: Voting with segmentation for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6101–6111, 2021.

[18] Zhisheng Huang, Peng Wang, Jingdong Zhang, Yuan Liu, Xin Li, and Wenping Wang. 3r-gs: Best practice in optimizing camera poses along with 3dgs. *arXiv preprint arXiv:2504.04294*, 2025.

[19] Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9187–9198, 2023.

[20] 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.

[21] Jaewoo Jung, Jisang Han, Honggyu An, Jiwon Kang, Seonghoon Park, and Seungryong Kim. Relaxing accurate initialization constraint for 3d gaussian splatting. *arXiv preprint arXiv:2403.09413*, 2024.

[22] Gyeongjin Kang, Jisang Yoo, Jihyeon Park, Seungtae Nam, Hyeonsoo Im, Sangheon Shin, Sangpil Kim, and Eunbyung Park. Selfsplat: Pose-free and 3d prior-free generalizable 3d gaussian splatting. *arXiv preprint arXiv:2411.17190*, 2024.

[23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024.

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[25] 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.

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

[27] 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.

[28] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022.

[29] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[30] Linfei Li, Lin Zhang, Zhong Wang, and Ying Shen. Gs3lam: Gaussian semantic splatting slam. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3019–3027, 2024.

[31] 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.

[32] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021.

[33] Fan Lu, Yan Xu, Guang Chen, Hongsheng Li, Kwan-Yee Lin, and Changjun Jiang. Urban radiance field representation with deformable neural mesh primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 465–476, 2023.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[40] AKM Rabby and Chengcui Zhang. Beyondpixels: A comprehensive review of the evolution of neural radiance fields. *arXiv preprint arXiv:2306.03000*, 2023.

[41] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[42] Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Look gauss, no pose: Novel view synthesis using gaussian splatting without accurate pose initialization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8732–8739. IEEE, 2024.

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

[44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[45] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.

[46] 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.

[47] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[48] 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.

[49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[50] 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.

[51] Zhengren Wang. 3d representation methods: A survey. *arXiv preprint arXiv:2410.06475*, 2024.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Yan Xu, Zhaoyang Huang, Kwan-Yee Lin, Xinge Zhu, Jianping Shi, Hujun Bao, Guofeng Zhang, and Hongsheng Li. Selfvoxelo: Self-supervised lidar odometry with voxel-based deep neural networks. In *Conference on robot learning*, pages 115–125. PMLR, 2021.

[58] Yan Xu, Junyi Lin, Jianping Shi, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Robust self-supervised lidar odometry via representative structure discovery and 3d inherent error modeling. *IEEE Robotics and Automation Letters*, 7(2):1651–1658, 2022.

[59] Yueming Xu, Haochen Jiang, Zhongyang Xiao, Jianfeng Feng, and Li Zhang. Dg-slam: Robust dynamic gaussian splatting slam with hybrid pose optimization. *arXiv preprint arXiv:2411.08373*, 2024.

[60] 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.

[61] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

[62] 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 2021 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330.

[63] 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.

[64] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019.

[65] 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.

[66] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe our two main contributions: (1) A local-to-global optimization framework for pose-free 3DGS, and (2) A neural geometric representation that enhances robustness under textureless regions and abrupt motion. These are consistently supported by our experiments and analysis.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper did talk about the limitations of the work in the last section

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This work is primarily empirical, focus on the application, and does not present any formal theorems or theoretical analysis requiring assumptions or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The architecture details, training settings, dataset splits, and evaluation metrics are all clearly described in the main text and appendix. These provide sufficient information for reproducing the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Although no code is released, it includes detailed instructions, data preparation, training, and evaluation. These are sufficient to faithfully reproduce the main experimental results and we will open source upon the paper acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All training details, including optimizer choice, learning rate, batch size, and data splits, are described in paper and Appendix to ensure reproducibility and clarity.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The experimental results in the paper are not accompanied by any statistical significance tests, as prior work in this area typically only reports aggregate results without such analysis. To ensure comparability and clarity, we follow the reporting conventions in the related literature.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides detailed information on the compute infrastructure used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and ensured that the work complies with all guidelines, including data usage, model deployment considerations, and societal impact awareness.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work may enable more robust 3D reconstruction in low-texture or dynamic environments, which could benefit AR/VR applications and autonomous systems. We discuss this at the end of this paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve models or data with high risk of misuse. It focuses on standard scene reconstruction tasks without releasing generative or language models, or using scraped or sensitive data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use existing datasets and code under their respective licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new framework.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research does not involve crowdsourcing or any form of human subject participation.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve any human subjects, personal data, or crowdsourced data. All experiments are conducted on publicly available datasets and simulated environments.

    Guidelines: The paper does not involve any human subjects or crowdsourced data, and thus does not require IRB or equivalent approval.

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The research does not involve the use of large language models (LLMs) as part of the core methodology, nor are they used as important, original, or non-standard components in the work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.