Active3D: Active High-Fidelity 3D Reconstruction via Multi-Level Uncertainty Quantification

Yan Li[1], **Yingzhao Li**[2], **Gim Hee Lee**[1]

[1]National University of Singapore
[2]Harbin Institute of Technology

Abstract

In this paper, we present an active exploration framework for high-fidelity 3D reconstruction that incrementally builds a multi-level uncertainty space and selects next-best-views through an uncertainty-driven motion planner. We introduce a *hybrid implicit–explicit representation* that fuses neural fields with Gaussian primitives to jointly capture global structural priors and locally observed details. Based on this hybrid state, we derive a *hierarchical uncertainty volume* that quantifies both implicit global structure quality and explicit local surface confidence. To focus optimization on the most informative regions, we propose an *uncertainty-driven keyframe selection* strategy that anchors high-entropy viewpoints as sparse attention nodes, coupled with a *viewpoint-space sliding window* for uncertainty-aware local refinement. The planning module formulates next-best-view selection as an *Expected Hybrid Information Gain* problem and incorporates a risk-sensitive path planner to ensure efficient and safe exploration. Extensive experiments on challenging benchmarks demonstrate that our approach consistently achieves state-of-the-art accuracy, completeness, and rendering quality, highlighting its effectiveness for real-world active reconstruction and robotic perception tasks.

Website — https://yanyan-li.github.io/project/vlx/active3d

1 Introduction

Visual-based 3D reconstruction (Newcombe et al. 2011; Whelan et al. 2015; Dai et al. 2017; Li et al. 2020) aims to infer the geometry and appearance of previously unseen scenes from 2D imagery, making it a fundamental problem in both computer vision and robotics. Depending on how the sensor moves, reconstruction methods can be clustered into two categories: passive and active. Passive systems process streams of RGB (Schonberger and Frahm 2016) or RGB-D (Li and Tombari 2022) frames to jointly estimate six-degree-of-freedom (6-DoF) camera motions and fuse the measurements into sparse or dense 3D models, under the assumption of a fixed, user-driven path. In contrast, active reconstruction frameworks (Aloimonos, Weiss, and Bandyopadhyay 1988; Chen, Li, and Kwok 2011) integrate

Figure 1: Performance on the Replica dataset. Left: Comparison of rendering quality (PSNR) versus reconstruction completeness (C.R.) across state-of-the-art methods. Right: Qualitative outputs of our method including reconstructed mesh, Gaussian map, and estimated depth.

next-best-view (NBV) planning (Peralta et al. 2020) to autonomously select subsequent viewpoints to maximize information gain (Isler et al. 2016; Kirsch, Van Amersfoort, and Gal 2019) and ensure comprehensive surface coverage. In addition to accurate geometry, next-generation intelligent robots demand dense 3D models with *high fidelity* and photometric consistency for downstream tasks.

The conventional active reconstruction problem (Isler et al. 2016; Huang et al. 2018) is typically cast as an exploration task: select the sequence of viewpoints that will most effectively reveal detailed scene geometry and appearance. Early approaches leverage occupancy-grid (Elfes 2013) or voxel-based (Wu et al. 2014) maps and frontier-driven exploration to push the boundary between known and unknown space, ensuring that new measurements continually reduce map uncertainty (Lee et al. 2022). However, since these approaches focus solely on geometric uncertainty, the resulting reconstructions are ill-suited for high-quality novel-view rendering and often lack the photometrically consistent details required for downstream tasks. Recent advances in scene representation have revealed two complementary paradigms: *implicit neural fields* (Mildenhall et al. 2021; Barron et al. 2022) and *explicit parameterizations* such as 3D Gaussian Splatting (Kerbl et al. 2023; Li et al. 2024), both achieving impressive performance in novel-view synthesis and surface reconstruction. Implicit models encode continuous neural fields that excel at capturing global structure, while explicit Gaussians faithfully preserve observed geometry and fine details. However, existing active frameworks typically adopt only one of these paradigms. Implicit-based active methods (Yan, Yang, and Zha 2023; Kuang et al. 2024) leverage neural priors for view planning, but their continuous fields tend to hallucinate missing surfaces (e.g., transparent or mirrored areas), leading to persistent high uncertainty and planner oscillation. Conversely, GS-based active approaches (Li et al. 2025; Jin et al. 2025) directly reflect observations into the map, providing reliable local geometry but lacking the ability to reason about occluded or unseen regions, resulting in suboptimal exploration coverage.

These complementary strengths and limitations motivate a hybrid implicit–explicit formulation for active reconstruction, unifying global priors and local textured surface within a single information-theoretic planning framework. First, given a posed RGB-D stream, Active3D constructs a **hybrid implicit–explicit scene state** and derives a **hierarchical uncertainty map** to jointly quantify global structural entropy and local surface uncertainty. Based on this hybrid uncertainty, the planner is further proposed to formulate *next-best-view selection* as an Expected Hybrid Information Gain (**EHIG**) optimization and executes viewpoint-aware trajectory planning. Keyframes are promoted via a dual-uncertainty intersection criterion, selecting viewpoints that observe regions where both implicit and explicit uncertainties are high. This establishes a sparse attention mechanism over the hybrid scene state. A viewpoint-space sliding window then performs uncertainty-aware local refinement of Gaussian primitives with respect to implicit priors, maintaining global–local consistency throughout the reconstruction process. Our contributions are summarized as follows:

- We propose a *hybrid implicit–explicit scene representation* for active 3D reconstruction, unifying neural fields and Gaussian primitives into a joint entropy minimization framework and introducing the *Hybrid Scene State Entropy*.
- We design a *hierarchical uncertainty map* that fuses global implicit variance, local depth residuals, local photometric residuals, and temporal SDF changes via Bayesian fusion, providing a principled multi-scale signal to drive exploration and refinement.
- We formulate next-best-view planning as an *Expected Hybrid Information Gain (EHIG)* problem, combining global structural exploration and local detail preservation with risk-aware path optimization.
- We introduce a viewpoint-aware keyframe selection strategy driven by the intersection of implicit and explicit uncertainties, anchoring high-information regions as sparse attention nodes in the hybrid map. Integrated with a spatial (non-temporal) sliding window, this enables uncertainty-aware local refinement and consistent reconstruction of the hybrid scene state.

2 Related Work

Neural Implicit and Explicit Representation. Traditionally, 3D reconstructed models have been represented using various geometric formats, including meshes (Kazhdan, Bolitho, and Hoppe 2006; Li et al. 2021), surfels (Whelan et al. 2015; Stückler and Behnke 2014), and truncated signed distance fields (TSDF) (Osher, Fedkiw, and Piechor 2004; Izadi et al. 2011). With the advent of differentiable radiance fields, these representations have been significantly extended to support high-quality novel view synthesis. In particular, NeRF (Mildenhall et al. 2021) have emerged as a powerful paradigm for photorealistic rendering and scene understanding. Specifically, iMAP (Sucar et al. 2021) utilizes MLP as the only scene representation for both tracking and mapping. To address the over smoothed reconstruction problem of only-MLP representation in large-scale environments, NeuralRecon (Sun et al. 2021) integrates neural TSDF volumes with learned features to enhance 3D reconstruction quality in indoor scenes. Similarly, ConvONet (Peng et al. 2020) predicts occupancy probabilities in 3D space using 3D convolutional architectures (Çiçek et al. 2016; Ronneberger, Fischer, and Brox 2015; Niemeyer et al. 2020), combining the strengths of spatially aware feature encoding and implicit shape modeling.

In contrast to implicit and hybrid approaches, explicit representations directly encode scene geometry and appearance in structured forms such as voxel grids (Müller et al. 2022) or Gaussian primitives (Kerbl et al. 2023), enabling efficient rendering and fast optimization. Plenoxels (Fridovich-Keil et al. 2022) replace MLPs with a sparse voxel grid that stores density and spherical harmonics coefficients. TensoRF (Chen et al. 2022) further improves scalability and memory efficiency by applying low-rank tensor decomposition. More recently, 3D Gaussian Splatting (Kerbl et al. 2023; Li et al. 2024) introduces a point-based explicit method where each Gaussian encodes position, orientation, scale, and radiance attributes, supporting high-fidelity rendering with real-time performance and continuous surfaces.

Active High-quality 3D Modeling. Active reconstruction methods (Yan, Yang, and Zha 2023; Kuang et al. 2024; Pan et al. 2022; Li et al. 2025; Jin et al. 2024; Feng et al. 2024; Chen et al. 2025) autonomously select viewpoints during iterative mapping to maximize coverage and reconstruction quality. NeRF-based NBV strategies (Lee et al. 2022; Pan et al. 2022) use pixel-wise rendering variance as uncertainty cues, while FisherRF (Jiang, Lei, and Daniilidis 2024) introduces Fisher information for view planning. ANM (Yan, Yang, and Zha 2023) maintains weight-space uncertainty in a continually learned neural field, and NARUTO (Feng et al. 2024) extends this paradigm to 6-DoF exploration in large-scale scenes.

Recently, Gaussian primitives have been adopted for active scene modeling. ActiveGAMER (Chen et al. 2025) incorporates rendering quality into the information gain metric. GS-Planner (Jin et al. 2024) detects unobserved regions in the Gaussian map and employs a sampling-based NBV policy. HGS (Xu et al. 2024) proposes an adaptive hierarchical planning strategy balancing global and local refinement. ActiveSplat (Li et al. 2025) extends Gaussian-based SLAM to active mapping with decoupled viewpoint orientation.

Uncertainty estimation plays a central role in NBV selection. NeRF-based methods typically derive voxel or pixel-wise variance from density fields (Pan et al. 2022; Lee et al. 2022), while Gaussian-based methods rely on observation completeness or visibility priors (Jin et al. 2024; Li et al. 2025). In contrast, we fuse *global implicit variance*, *local surface residuals*, and *temporal SDF variation*, constructing a hierarchical uncertainty map that simultaneously guides global exploration and local refinement.

3 Methodology

In the active reconstruction task, the core of the problem is to decide the position and orientation of the i^{th} viewpoint based on the information captured by the previous posed RGB-D stream $\mathcal{S}_{i-1} = \{\mathbf{S}_k\}, \mathbf{S}_k = [I_k, D_k, \mathbf{T}_{c_k,w}, \mathbf{K}], k \in [0, 1, \ldots, i-1]$. Therefore, the problem can be defined as determining how to leverage the previously posed RGB-D stream to guide the selection of the current viewpoint in order to achieve high-quality reconstruction. This process first involves the data organization of the previous RGB-D stream, followed by quantifying the historical information to evaluate the current reconstruction state and predicting potential information gain. By modeling the scene coverage, uncertainty distribution, and geometric consistency from $\mathcal{S}_{i-1}$, the system can actively plan the next viewpoint that maximizes scene completeness and reconstruction fidelity. Fig. 2 depicts the algorithm's workflow.

Hybrid Implicit-explicit Space

To simultaneously capture continuous global priors and high-quality local surface, we construct a **hybrid implicit–explicit space** that integrates implicit neural fields with explicit Gaussian primitives. Given a posed RGB-D observation $\mathbf{S}_k = [I_k, D_k, \mathbf{T}_{c_k,w}, \mathbf{K}]$, this hybrid space provides a unified state representation for incremental active reconstruction.

Definition of Hybrid Scene State.

We introduce a state formulation for the incremental active reconstruction task, where the state $\mathcal{M}_k$ at step k is designed to represent the currently reconstructed portion of the scene:

$$\mathcal{M}_k = \{\mathcal{F}_\theta, \mathcal{G}_k\}, \tag{1}$$

where $\mathcal{F}_\theta : \mathbb{R}^3 \to$ SDF is the implicit neural field, and $\mathcal{G}_k = \{G_i\}_{i=1}^{N_k}$ is the set of 3D Gaussian primitives. Each primitive G_i is parameterized as $G_i = (\mu_i, \Sigma_i, \alpha_i, c_i)$, where $\mu_i \in \mathbb{R}^3$ is the mean position, $\Sigma_i \in \mathbb{R}^{3\times 3}$ the covariance, $\alpha_i \in [0, 1]$ the opacity, and $c_i \in \mathbb{R}^3$ the color vector. And for the implicit neural field, we employ a One-blob encoder (Wang, Wang, and Agapito 2023; Müller et al. 2019) to extract deep features from input point clouds. The implicit representation subsequently maps world coordinates $\mathbf{x} \in \mathbb{R}^3$ to SDF values and color attributes via the MLP:

$$s = f_\tau\big(\gamma(\mathbf{x}), \mathcal{V}_\alpha(\mathbf{x})\big) \tag{2}$$

where $\gamma(\mathbf{x})$ denotes tri-plane decomposition of spatial coordinates, and $\mathcal{V}_\alpha(\mathbf{x})$ represents position feature vectors obtained through volumetric trilinear interpolation. The function $f_\tau(\cdot)$ corresponds to the geometry decoder.

Hybrid State Quantification. At state k, the key objective is to quantify the **current scene knowledge** and guide the next-best-view selection. This hybrid formulation bridges *global structural exploration* driven by $\mathcal{F}_\theta$ and *local high-fidelity surface* enabled by $\mathcal{G}_k$. Casting NBV planning as an *expected hybrid information gain* optimization, we formalize active reconstruction in a probabilistic information-theoretic context.

We define the voxel-wise hybrid entropy as:

$$H_{\text{hybrid}}(v) = \lambda_{\text{imp}} H[p_{\mathcal{F}_\theta}(v)] + \lambda_{\text{exp}} H[p_{\mathcal{G}_k}(v)], \tag{3}$$

where $H[p]$ denotes Shannon entropy and $\lambda_{\text{imp}}, \lambda_{\text{exp}}$ balance global priors and local observations.

The NBV reward for $\mathbf{c}$ is accumulated over all visible voxels:

$$R(\mathbf{c}) = \sum_{v \in \mathcal{V}_\mathbf{c}} w(v|\mathbf{c})(1 - O(v)), \tag{4}$$

where $\mathcal{V}_\mathbf{c}$ is the set of voxels visible from $\mathbf{c}$, and $O(v)$ is the occupancy probability used to discount free-space ambiguity.

Hierarchical Uncertainty Map Construction

To drive the hybrid NBV objective in Eq. 3, we construct a hierarchical uncertainty volume $\mathcal{V}_u \in \mathbb{R}^{L\times W\times H}$ that fuses **global implicit priors**, **local view-dependent surface**, and **temporal consistency cues**. Each voxel v stores a scalar $u(v) \in \mathbb{R}^+$ representing the hybrid reconstruction confidence. **Global Structure Uncertainty.** The implicit branch $\mathcal{F}_\theta$ encodes a continuous SDF-based representation that provides *global structural entropy*. We approximate per-voxel variance using an uncertainty head $f_\delta(\cdot)$:

$$u_{\text{imp}}(v) = \phi\big(f_\delta(\gamma(\mathbf{x}_v), \mathcal{V}_\alpha(\mathbf{x}_v))\big), \tag{5}$$

where $\mathbf{x}_v$ denotes the voxel center, $\gamma(\cdot)$ is the tri-plane encoder, and $\phi(\cdot)$ applies a softplus normalization. Upon receiving new observations, the structural uncertainty is updated, encouraging coverage-driven exploration and mitigating local greedy behavior during the early stages of mapping.

View-dependent Local Uncertainty. The explicit Gaussian map $\mathcal{G}_k$ provides *local observation entropy* through photometric and geometric residuals. At each step, we select top-K high-uncertainty candidate viewpoints $\mathcal{C}_{\text{high}}$ and compute depth and color errors:

$$E_t^{\text{depth}} = \left\|D_t^{\text{render}} - D_t^{\text{gt}}\right\| \odot M_t, \tag{6}$$

$$E_t^{\text{rgb}} = \sum_{c \in \{R,G,B\}} \left\|I_{t,c}^{\text{render}} - I_{t,c}^{\text{gt}}\right\| \odot M_t, \tag{7}$$

where M_t masks valid pixels. The 2D errors are back-projected into the 3D voxel space to estimate the uncertainty of the local surface using the following formulation:

$$u_{\text{exp}}(v) = \frac{1}{|\mathcal{C}_{\text{high}}|} \sum_{t \in \mathcal{C}_{\text{high}}} \mathcal{P}(E_t^{\text{depth}}, E_t^{\text{rgb}}; v), \tag{8}$$

where $\mathcal{P}(\cdot)$ denotes voxel-wise backprojection with bilinear interpolation.

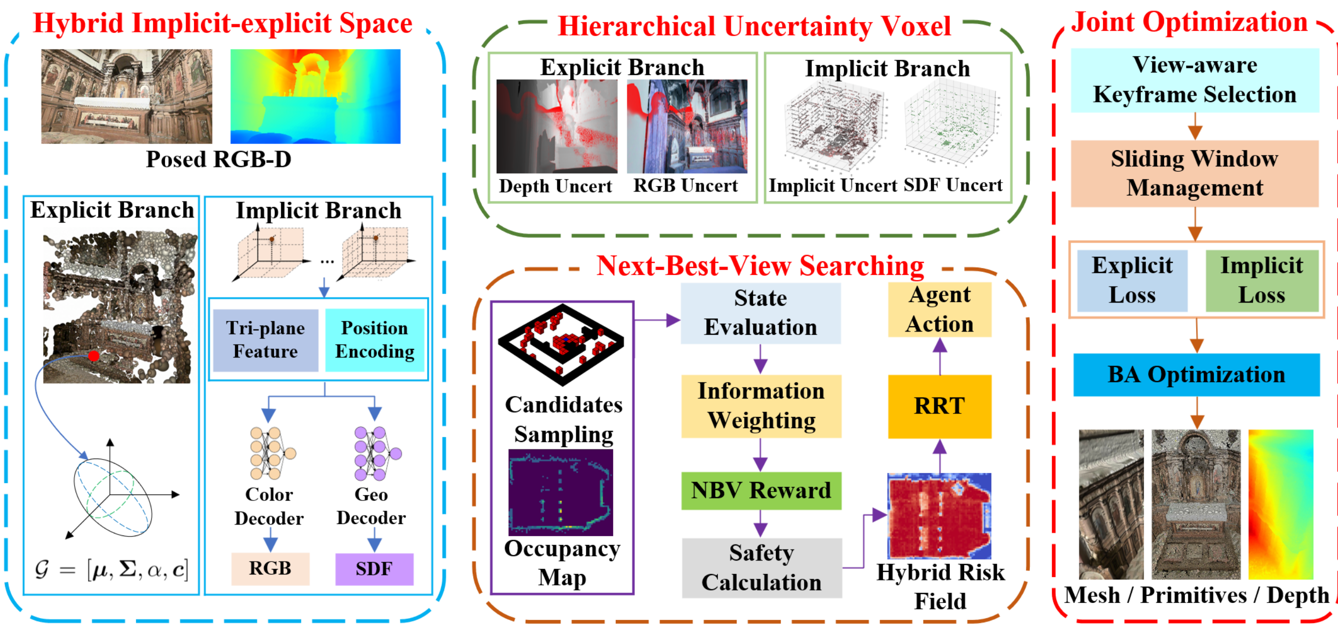

Figure 2: Our method processes the RGB-D stream through dual explicit and implicit reconstruction branches. The explicit branch projects data into a 3D Gaussian model, while the implicit branch employs an encoder-decoder architecture to regress RGB values and SDF. Subsequently, the discrepancy between the rendered RGB-D and the GT RGB-D is computed. Another mlp predicts global uncertainty, while temporal variations on the SDF surface are characterized to derive uncertainty for the hybrid explicit-implicit representation. This representation then drives NBV selection and path planning. Finally, keyframes are selected within a sliding window for joint optimization of the explicit and implicit maps.

Temporal Variation Uncertainty. To detect emerging surfaces and inconsistencies, we evaluate SDF changes between consecutive keyframes: $\Delta S_t = S_t - S_{t-1}$. According to the varying states of surfaces, define masks for new surfaces, geometry changes, and novel free space:

$$\begin{cases} M_{\text{new}} = \mathbb{I}(0 \le S_t \le \tau_s) \odot \mathbb{I}(\Delta S_t > \tau_n), \\ M_{\text{change}} = \mathbb{I}(|\Delta S_t| > \tau_c), \\ M_{\text{free}} = \mathbb{I}(S_t > \tau_f) \odot \mathbb{I}(S_{t-1} < -\tau_f). \end{cases} \tag{9}$$

The temporal uncertainty term is:

$$u_{\text{time}}(v) = \beta_1 |\Delta S_t(v)| + \beta_2 \cdot \mathbb{I}(v \in M_{\text{focus}}), \tag{10}$$

where $M_{\text{focus}} = M_{\text{new}} \cup M_{\text{change}} \cup M_{\text{free}}$.

Then, the final hierarchical uncertainty is fused as:

$$u_{\text{final}}(v) = \alpha_1 u_{\text{imp}}(v) + \alpha_2 u_{\text{exp}}(v) + \alpha_3 u_{\text{time}}(v), \tag{11}$$

where α_i are weights estimated via evidence maximization, interpreted as a fusion of global priors, local observations, and temporal consistency. This **hierarchical map** directly links to the NBV reward in Eq. 4, providing a multi-scale uncertainty signal that balances exploration coverage and model fidelity.

Next-Best-View Searching

With the hybrid scene state $\mathcal{M}_k = \{\mathcal{F}_\theta, \mathcal{G}_k\}$ and hierarchical uncertainty map $u_{\text{final}}(v)$ defined in Eq. 11, the goal of active planning is to select the next viewpoint $\mathbf{c}_i$ that maximizes the expected *hybrid information gain*.

EHIG Objective. Based on the final hierarchical uncertainty, we cast NBV selection as:

$$\mathbf{c}_i^* = \arg\max_{\mathbf{c} \in \mathcal{C}} \mathbb{E}\left[\Delta \mathcal{I}_{\text{hybrid}}(\mathbf{c})\right], \tag{12}$$

where $\mathcal{C}$ is the candidate viewpoint set, and $\Delta \mathcal{I}_{\text{hybrid}}$ measures the reduction of hybrid entropy:

$$\Delta \mathcal{I}_{\text{hybrid}}(\mathbf{c}) = \Delta H[\mathcal{F}_\theta] + \Delta H[\mathcal{G}_k], \tag{13}$$

corresponding to global implicit and local explicit uncertainty reduction, respectively.

Voxel-wise Information Weighting. For a voxel v visible from candidate $\mathbf{c}$, we define its contribution as:

$$w(v|\mathbf{c}) = \alpha U(v) + \beta H_{\text{hybrid}}(v), \tag{14}$$

where $U(v)$ is the hierarchical uncertainty estimate from Eq. 11 and $H_{\text{hybrid}}(v)$ is the hybrid entropy in Eq. 3. α and β are weights. This formulation unifies multi-scale uncertainty into a single information-theoretic weight.

NBV Reward. Given the information weight of the voxel, the expected reward of candidate $\mathbf{c}$ is obtained via Eq. 4.

Risk-Aware Path Planning. After obtaining the next goal, we employs an enhanced RRT* algorithm (LaValle and Kuffner 2001) for active path planning. To generate physically feasible trajectories, we integrate the NBV reward into a risk-aware cost function:

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \int_{\mathbf{p}} \Big(C_{\text{travel}}(x) - \eta R(\mathbf{c}_x) + \lambda C_{\text{risk}}(x) \Big) dx, \tag{15}$$

where $\mathbf{p}$ is the planned path, C_{travel} the navigation cost, C_{risk} the collision probability, and $R(\mathbf{c}_x)$ the NBV reward at pose x.

This proposed NBV searching bridges hybrid scene representation, multi-scale uncertainty, and active trajectory optimization into a single expected information gain framework. By combining global implicit entropy reduction and local explicit observation gain, the planner achieves coverage-aware and detail-preserving exploration.

Uncertainty-driven Keyframe Selection

Unlike conventional keyframe strategies that are tightly coupled with temporal ordering, we propose a **Uncertainty-driven** selection criterion that anchors high-information observations in the hybrid scene state $\mathcal{M}_k$. Rather than merely ensuring temporal coverage, the proposed keyframes act as a *sparse attention mechanism*, focusing optimization on regions where the hybrid uncertainty is maximized.

Viewpoint-Based Keyframe Selection. By decoupling keyframe selection from temporal sampling and binding it to viewpoint-space information gain, our method avoids redundant observations and focuses optimization capacity on spatially complementary views, which is crucial for active reconstruction. For a newly acquired RGB-D frame $\mathbf{S}_c$ with camera pose $\mathbf{T}_{c,w}$, we compute its viewpoint divergence relative to the active keyframe set $\mathcal{S}_{\text{KF}}$:

$$\delta_c = \min_{\mathbf{S}_j \in \mathcal{S}_{\text{KF}}} d_{\text{view}}\big(\mathbf{T}_{c,w}, \mathbf{T}_{j,w}\big), \tag{16}$$

where d_{view} measures the viewpoint baseline in SE(3) space, combining angular separation and projected frustum overlap.

Aggressive active motion planning may cause an agent to overskip salient textural structures, we introduce a dual-uncertainty intersection criterion. Define the *high-uncertainty intersection set* as:

$$\mathcal{V}_{\text{high}} = \{v \in \mathcal{V}_u \mid u_{\text{exp}}(v) > \tau_h \wedge u_{\text{imp}}(v) > \tau_h\}. \tag{17}$$

For frame $\mathbf{S}_c$, we compute its *uncertainty coverage ratio* ρ_c as the fraction of $\mathcal{V}_{\text{high}}$ visible in its frustum:

$$\rho_c = |\mathcal{V}_{\text{high}} \cap \mathcal{V}_{\text{vis}}(\mathbf{S}_c)| / |\mathcal{V}_{\text{vis}}(\mathbf{S}_c)|,$$

with $\mathcal{V}_{\text{vis}}$ being the visible voxel set. A frame is promoted to keyframe if:

$$(\delta_c > \tau_{\text{view}}) \wedge (\Delta\mathcal{I}_{\text{hybrid}}(\mathbf{S}_c) > \tau_{\text{info}}) \wedge (\rho_c > \tau_\rho), \tag{18}$$

where τ_{view}, τ_{info}, τ_ρ are viewpoint, information-gain and coverage threshold, respectively. The uncertainty-driven keyframe selection scheme actually establishes a sparse attention mechanism toward scene structures. This ensures selection of frames observing regions where both geometric and neural uncertainties are high.

Viewpoint-Space Sliding Window. Employing all keyframes for joint optimization still incurs excessive computational burden, prior approaches maintained a sliding window over continuous time. However, this strategy exhibits significant viewpoint redundancy as agent approaches the target, while failing to establish sufficient covisibility constraints upon revisiting similar locations. We maintain a local optimization window $\mathcal{W}_k = \{\mathbf{S}_{c_1}, \dots, \mathbf{S}_{c_m}\}$ indexed by spatially selected keyframes, not constrained by temporal adjacency. The hybrid state $\mathcal{M}_k$ is jointly refined via:

$$E_{\text{total}} = E_{\text{photo}} + E_{\text{geo}} + \lambda E_{\text{reg}}, \tag{19}$$

where E_{photo} enforces multi-view photometric consistency on $\mathcal{G}_k$, E_{geo} aligns Gaussian primitives with the implicit SDF $\mathcal{F}_\theta$, and E_{reg} prevents overfitting across non-overlapping viewpoints.

Method	Metric	Off0	Off1	Off2	Off3	Off4	R0	R1	R2
ANM-S	Acc ↓	1.44	1.03	1.60	1.80	1.50	1.47	1.29	1.28
	Com. ↓	1.98	1.55	6.65	1.13	1.08	0.91	1.02	0.85
	C.R. ↑	95.43	92.66	79.20	94.98	95.35	96.71	95.66	96.79
NARUTO	Acc ↓	1.26	1.04	×	34.84	1.67	1.75	×	1.50
	Com. ↓	1.41	1.30	×	2.96	2.01	1.56	×	1.49
	C.R. ↑	97.63	96.88	×	91.27	95.14	94.58	×	97.56
	PSNR ↑	31.01	31.43	×	26.63	28.57	26.55	×	25.56
	SSIM ↑	0.892	0.897	×	0.831	0.882	0.782	×	0.818
	LPIPS ↓	0.299	0.283	×	0.283	0.284	0.354	×	0.367
ActiveSplat	Acc ↓	1.16	1.11	1.47	1.70	1.50	1.67	1.43	1.36
	Com. ↓	0.63	0.94	5.59	1.83	1.06	0.84	0.74	1.04
	C.R. ↑	97.54	94.54	80.69	91.49	95.34	97.04	96.84	95.65
	PSNR ↑	24.48	26.95	22.72	20.96	27.88	26.16	29.00	28.86
	SSIM ↑	0.857	0.871	0.888	0.804	0.878	0.823	0.877	0.894
	LPIPS ↓	0.145	0.130	0.113	0.232	0.147	0.199	0.136	0.113
OURS	Acc ↓	1.12	1.02	1.34	1.56	1.38	1.59	1.13	1.26
	Com. ↓	1.34	1.17	1.66	1.97	1.87	1.75	1.32	1.52
	C.R. ↑	97.76	98.21	96.86	94.70	96.80	97.28	98.09	98.18
	PSNR ↑	40.51	40.54	33.72	34.14	37.37	33.80	34.63	36.00
	SSIM ↑	0.980	0.979	0.951	0.949	0.964	0.948	0.954	0.962
	LPIPS ↓	0.030	0.034	0.067	0.075	0.054	0.072	0.056	0.053

Table 1: Quantitative comparison of 3D reconstruction and view synthesis quality between the proposed method and state-of-the-art approaches on the Replica dataset. The symbol × indicates that the method fails to complete exploration within five trials.

4 Experiments

Implementation and Simulator

We implement the proposed method within the Habitat simulator (Savva et al. 2019) as an active exploration system. The agent captures posed RGB-D observations along planned viewpoints. The camera field-of-view is set to 60° vertically and 90° horizontally, and the system processes sequences online with on-policy planning and incremental reconstruction. Further implementation details are presented in the supplementary material.

Datasets, Metrics, and Baselines

Following prior active mapping benchmarks (Yan, Yang, and Zha 2023), we evaluate on two widely used datasets:(i) **Replica** (Straub et al. 2019) with 8 indoor scenes, and (ii) **Matterport3D (MP3D)** (Chang et al. 2017) with 5 large-scale scenes exhibiting significant occlusion and spatial complexity. All methods are run for 2000 exploration steps on Replica and MP3D.

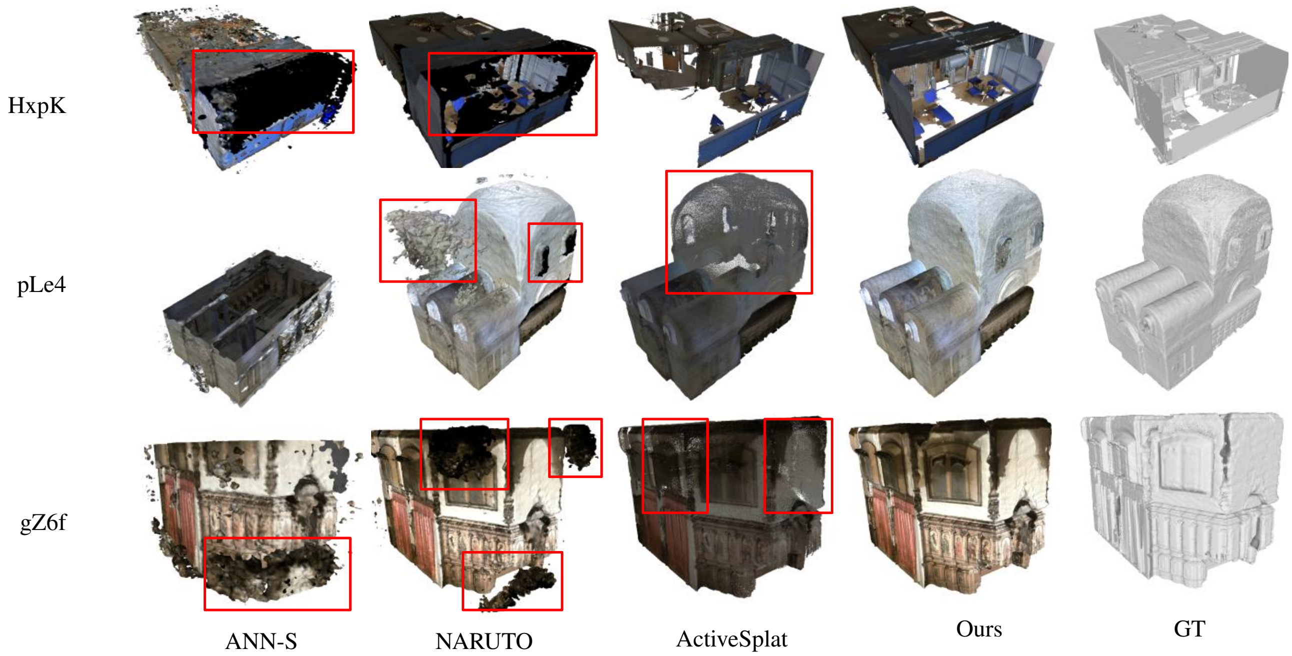

Figure 3: Qualitative comparison of 3D reconstruction results on representative MP3D sequences. Additional results and detailed comparisons for all Replica and MP3D sequences are provided in the supplementary material.

We report metrics targeting the critical objectives of active reconstruction: *accuracy* (Acc, cm), *completion* (Com, cm), and *completion ratio* (C.R., %), where Acc/Com are computed with a $5cm$ threshold. To evaluate rendering quality, we report PSNR, SSIM, and LPIPS on held-out viewpoints. For additional geometric consistency analysis, we compute the Mean Absolute Distance (MAD) between the reconstructed SDF and ground-truth surfaces.

We compare our method against state-of-the-art active reconstruction frameworks: ActiveNR (Yan, Yang, and Zha 2023), ANM-S (Kuang et al. 2024), NARUTO (Feng et al. 2024), and ActiveSplat (Li et al. 2025). We further compare passive baselines in the supplemental material. All baselines are re-trained and evaluated locally for fair comparison.

Evaluation on Replica

Table 1 reports 3D reconstruction and view synthesis metrics on the Replica dataset. Our method consistently achieves the best or second-best performance across all metrics. For reconstruction, it yields the highest completion ratio (C.R.) and lowest Acc/Com error, reaching 98.09% C.R. on R1 and 98.18% on R2. For view synthesis, it achieves the highest PSNR (up to 40.51) and SSIM (0.980) while maintaining the lowest LPIPS, demonstrating sharp textures and photometric consistency.

Evaluation on MP3D

Table 2 evaluates our method on the MP3D dataset. Compared to ActiveSplat, our approach significantly improves both geometry and rendering fidelity. We achieve the highest combined reconstruction score in nearly all scenes, exceeding 98% on three out of five sequences. For photometric metrics, our method delivers the best PSNR and SSIM in four out of five cases, while maintaining the lowest LPIPS, reflecting perceptually consistent rendering.

Figure 3 visualizes reconstructions on MP3D. Compared to NARUTO and ActiveSplat, our method produces sharper edges, fewer ghosting artifacts, and consistent textures under dynamic occlusion.

Ablation Study

As summarized in Table 3, ablation studies are conducted on the challenging MP3D YmJk scene—characterized by significant occlusion and complex geometry.

Uncertainty Setting. Removing multi-resolution triplane encoding causes system failure due to complete loss of spatial perception. Eliminating the MLP-predicted uncertainty volume severely degrades reconstruction completeness (Com: 4.37 cm vs. 2.81 cm) by impeding global scene understanding. Exclusion of depth uncertainty induces erratic reconstruction (Acc: 4.75 cm vs. 2.66 cm) due to compromised surface fidelity estimation, which destabilizes optimization. Omission of RGB uncertainty substantially deteriorates rendering metrics (PSNR: 28.35 dB vs. 30.93 dB), attributable to degraded color/texture perception. Disabling the time-varying SDF representation markedly decreases reconstruction completeness.

Searching and Planning. Replacing the risk-aware path planner with naive uncertainty-volume aggregation degrades reconstruction coverage (C.R.: 89.23% vs. 91.73%), as this suboptimal strategy prompts excessive surface proximity, reducing global observability while increasing collision risk. Finally, disabling keyframe management guided by spatial

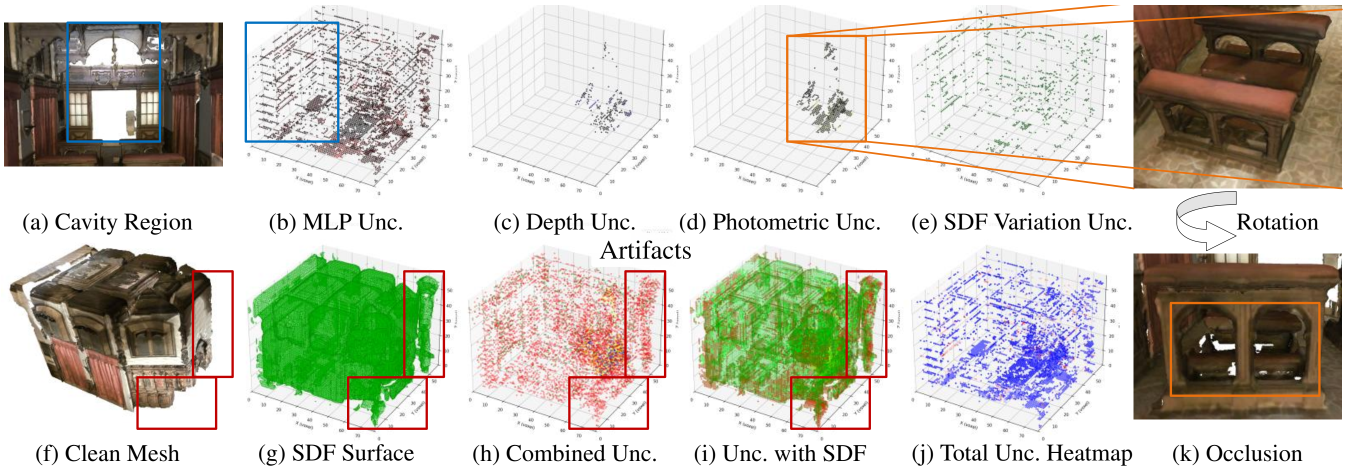

Figure 4: Visualization of uncertainties and their spatial relationship to real scene. Our proposed hybrid strategy not only endows the agent with global optimization capabilities, but also enables it to perceive intricate structures and textures while handling occlusions.

Method	Metric	Gdvg	gZ6f	HxpK	pLe4	YmJk	Avg.
ActiveINR	Acc ↓	5.09	4.15	15.60	5.56	8.61	7.80
	Com. ↓	5.69	7.43	15.96	8.03	8.46	9.11
	C.R. ↑	80.99	80.68	48.34	76.41	79.35	73.15
ANM-S	Acc ↓	5.52	1.62	2.13	4.54	4.50	3.66
	Com. ↓	3.95	2.01	12.49	2.51	3.53	4.90
	C.R. ↑	91.00	94.58	60.39	95.02	88.65	85.93
NARUTO	Acc ↓	2.34	3.57	7.29	4.46	9.52	5.44
	Com. ↓	4.93	2.47	2.84	3.14	5.68	3.81
	C.R. ↑	84.88	93.26	92.15	82.67	78.99	86.39
	PSNR ↑	23.42	23.84	23.32	27.15	23.64	24.27
	SSIM ↑	0.742	0.719	0.734	0.767	0.735	0.739
	LPIPS ↓	0.416	0.523	0.492	0.554	0.517	0.500
ActiveSplat	Acc ↓	2.39	1.74	2.53	4.09	9.52	4.05
	Com. ↓	3.76	1.34	24.28	1.07	2.84	6.66
	C.R. ↑	92.11	97.61	44.45	99.10	90.78	84.81
	PSNR ↑	22.77	16.40	18.33	23.49	24.57	21.12
	SSIM ↑	0.700	0.601	0.776	0.667	0.852	0.719
	LPIPS ↓	0.264	0.342	0.236	0.345	0.156	0.269
OURS	Acc ↓	1.68	1.90	1.61	2.68	2.66	2.11
	Com. ↓	1.59	1.96	2.09	2.38	2.81	2.27
	C.R. ↑	98.23	97.94	98.12	94.55	91.73	96.11
	PSNR ↑	31.12	32.43	29.53	33.14	30.93	31.43
	SSIM ↑	0.912	0.939	0.905	0.920	0.923	0.920
	LPIPS ↓	0.160	0.168	0.176	0.222	0.179	0.181

Table 2: Quantitative comparison on the MP3D dataset for 3D reconstruction and novel view synthesis.

co-visibility and uncertainty underutilizes historical observations upon revisit, leading to rendering degradation.

Advantages of Hierarchical Uncertainties. Fig. 4 visualizes the Hierarchical Uncertainty Map. The fully implicit uncertainty (b, e) provides the agent with global optimization capability. However, as the MLP-predicted SDF tends to generate redundant structures (f, g), it induces excessively high uncertainty in void regions (a) and redundant structure areas (g). This results in the agent allocating excessive attention to non-existent uncertainties (h). Conversely, the fully explicit uncertainty (c, d) aids the agent in identifying complex structures and textures. Nevertheless, due to its inability to perceive occluded regions (k) via α-blending, it leads the agent to prematurely conclude optimization completeness and initiate subsequent planning. Our hybrid approach synergistically combines the strengths of both explicit and implicit representations. By adaptively weighting the explicit and implicit uncertainties, it enhances the agent's perceptual awareness across all local and global regions (g).

Method	PSNR↑	SSIM↑	LPIPS↓	MAD↓	Acc↓	Com↓	C.R.↑
final	30.93	0.923	0.179	1.53	2.66	2.81	91.73
w.o. Tri-plane Encoder	×	×	×	×	×	×	×
w.o. MLP Uncert	30.89	0.917	0.191	1.80	2.65	4.37	84.84
w.o. Depth Uncert	29.28	0.907	0.218	1.88	4.75	5.12	83.97
w.o. RGB Uncert	28.35	0.901	0.201	1.78	2.69	4.02	86.18
w.o. SDF Temp	31.23	0.921	0.187	1.58	2.71	3.11	90.83
w.o. Risk Planning	30.78	0.916	0.179	1.61	2.67	3.40	89.23
w.o. Uncert Keyframe	29.43	0.917	0.182	1.54	2.77	2.88	88.91
w. Temporal Sliding Window	28.69	0.910	0.186	1.62	2.69	3.72	87.02

Table 3: Ablation study on MP3D dataset. The best results are highlighted in the table.

5 Conclusion

We have introduced Active3D, an active 3D reconstruction framework that unifies implicit neural fields and explicit Gaussian primitives into a hybrid information-theoretic formulation. By deriving a hierarchical uncertainty volume from this hybrid scene state, our method simultaneously captures global structural priors and local observation confidence, enabling principled next-best-view selection. An uncertainty-driven keyframe selection strategy anchors high-entropy viewpoints as sparse attention nodes, while a viewpoint-space sliding window performs uncertainty-aware local refinement to maintain global–local consistency.

Acknowledgments

This research was supported by the Tier 2 Grant (MOE-T2EP20124-0015) from the Singapore Ministry of Education.

References

Aloimonos, J.; Weiss, I.; and Bandyopadhyay, A. 1988. Active vision. *International journal of computer vision*, 1: 333–356.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, 333–350. Springer.

Chen, L.; Zhan, H.; Chen, K.; Xu, X.; Yan, Q.; Cai, C.; and Xu, Y. 2025. ActiveGAMER: Active GAussian Mapping through Efficient Rendering. *arXiv preprint arXiv:2501.06897*.

Chen, S.; Li, Y.; and Kwok, N. M. 2011. Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, 30(11): 1343–1377.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Dai, A.; Nießner, M.; Zollhöfer, M.; Izadi, S.; and Theobalt, C. 2017. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4): 1.

Elfes, A. 2013. Occupancy grids: A stochastic spatial representation for active robot perception. *arXiv preprint arXiv:1304.1098*.

Feng, Z.; Zhan, H.; Chen, Z.; Yan, Q.; Xu, X.; Cai, C.; Li, B.; Zhu, Q.; and Xu, Y. 2024. Naruto: Neural active reconstruction from uncertain target observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21572–21583.

Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5501–5510.

Huang, R.; Zou, D.; Vaughan, R.; and Tan, P. 2018. Active image-based modeling with a toy drone. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 6124–6131. IEEE.

Isler, S.; Sabzevari, R.; Delmerico, J.; and Scaramuzza, D. 2016. An information gain formulation for active volumetric 3D reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 3477–3484. IEEE.

Izadi, S.; Kim, D.; Hilliges, O.; Molyneaux, D.; Newcombe, R.; Kohli, P.; Shotton, J.; Hodges, S.; Freeman, D.; Davison, A.; et al. 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 559–568.

Jiang, W.; Lei, B.; and Daniilidis, K. 2024. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *European Conference on Computer Vision*, 422–440. Springer.

Jin, L.; Zhong, X.; Pan, Y.; Behley, J.; Stachniss, C.; and Popović, M. 2025. Activegs: Active scene reconstruction using gaussian splatting. *IEEE Robotics and Automation Letters*.

Jin, R.; Gao, Y.; Wang, Y.; Wu, Y.; Lu, H.; Xu, C.; and Gao, F. 2024. Gs-planner: A gaussian-splatting-based planning framework for active high-fidelity reconstruction. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11202–11209. IEEE.

Kazhdan, M.; Bolitho, M.; and Hoppe, H. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Kuang, Z.; Yan, Z.; Zhao, H.; Zhou, G.; and Zha, H. 2024. Active neural mapping at scale. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7152–7159. IEEE.

LaValle, S. M.; and Kuffner, J. J. 2001. Rapidly-exploring random trees: Progress and prospects: Steven m. lavalle, iowa state university, a james j. kuffner, jr., university of tokyo, tokyo, japan. *Algorithmic and computational robotics*, 303–307.

Lee, S.; Chen, L.; Wang, J.; Liniger, A.; Kumar, S.; and Yu, F. 2022. Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. *IEEE Robotics and Automation Letters*, 7(4): 12070–12077.

Li, Y.; Brasch, N.; Wang, Y.; Navab, N.; and Tombari, F. 2020. Structure-slam: Low-drift monocular slam in indoor environments. *IEEE Robotics and Automation Letters*, 5(4): 6583–6590.

Li, Y.; Kuang, Z.; Li, T.; Hao, Q.; Yan, Z.; Zhou, G.; and Zhang, S. 2025. ActiveSplat: High-Fidelity Scene Reconstruction Through Active Gaussian Splatting. *IEEE Robotics and Automation Letters*, 10(8): 8099–8106.

Li, Y.; Lyu, C.; Di, Y.; Zhai, G.; Lee, G. H.; and Tombari, F. 2024. Geogaussian: Geometry-aware gaussian splatting for scene rendering. In *European Conference on Computer Vision*, 441–457. Springer.

Li, Y.; and Tombari, F. 2022. E-graph: Minimal solution for rigid rotation with extensibility graphs. In *European Conference on Computer Vision*, 306–322. Springer.

Li, Y.; Yunus, R.; Brasch, N.; Navab, N.; and Tombari, F. 2021. RGB-D SLAM with structural regularities. In *2021 IEEE international conference on Robotics and automation (ICRA)*, 11581–11587. IEEE.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Müller, T.; McWilliams, B.; Rousselle, F.; Gross, M.; and Novák, J. 2019. Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5): 1–19.

Newcombe, R. A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A. J.; Kohi, P.; Shotton, J.; Hodges, S.; and Fitzgibbon, A. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, 127–136. Ieee.

Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3504–3515.

Osher, S.; Fedkiw, R.; and Piechor, K. 2004. Level set methods and dynamic implicit surfaces. *Appl. Mech. Rev.*, 57(3): B15–B15.

Pan, X.; Lai, Z.; Song, S.; and Huang, G. 2022. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, 230–246. Springer.

Peng, S.; Niemeyer, M.; Mescheder, L.; Pollefeys, M.; and Geiger, A. 2020. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 523–540. Springer.

Peralta, D.; Casimiro, J.; Nilles, A. M.; Aguilar, J. A.; Atienza, R.; and Cajote, R. 2020. Next-best view policy for 3d reconstruction. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 558–573. Springer.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9339–9347.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Stückler, J.; and Behnke, S. 2014. Multi-resolution surfel maps for efficient dense 3D modeling and tracking. *Journal of Visual Communication and Image Representation*, 25(1): 137–147.

Sucar, E.; Liu, S.; Ortiz, J.; and Davison, A. J. 2021. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6229–6238.

Sun, J.; Xie, Y.; Chen, L.; Zhou, X.; and Bao, H. 2021. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15598–15607.

Wang, H.; Wang, J.; and Agapito, L. 2023. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13293–13302.

Whelan, T.; Leutenegger, S.; Salas-Moreno, R. F.; Glocker, B.; and Davison, A. J. 2015. ElasticFusion: Dense SLAM without a pose graph. In *Robotics: science and systems*, volume 11, 3. Rome, Italy.

Wu, S.; Sun, W.; Long, P.; Huang, H.; Cohen-Or, D.; Gong, M.; Deussen, O.; and Chen, B. 2014. Quality-driven poisson-guided autoscanning. *ACM Trans. Graph.*, 33(6): 203–1.

Xu, Z.; Jin, R.; Wu, K.; Zhao, Y.; Zhang, Z.; Zhao, J.; Gao, F.; Gan, Z.; and Ding, W. 2024. Hgs-planner: Hierarchical planning framework for active scene reconstruction using 3d gaussian splatting. *arXiv preprint arXiv:2409.17624*.

Yan, Z.; Yang, H.; and Zha, H. 2023. Active neural mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10981–10992.