FAST ACTIVE RECONSTRUCTION BY OMNIDIRECTIONAL VIEW

Anonymous authors
Paper under double-blind review

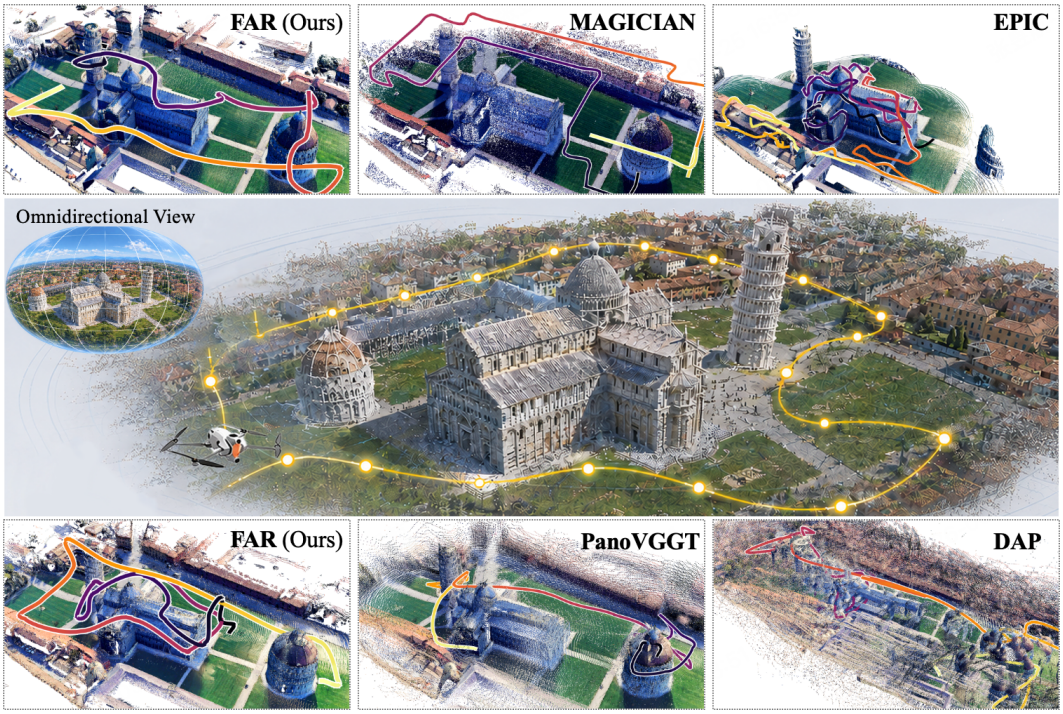


Figure 1: **Fast Active Reconstruction (FAR).** Controlled comparisons of planning and reconstruction on Pisa. Top: FAR, MAGICIAN, and EPIC planning on the same ground-truth depth. Bottom: the FAR planner driven by depth from our reconstruction model, PanoVGGT, and DAP.

ABSTRACT

In this work, we explore panoramic vision for fast active reconstruction, leveraging its omnidirectional observations to enable efficient UAV-based 3D scanning beyond existing perspective-based methods that require frequent camera adjustments due to their limited fields of view. Specifically, we propose FAR, a closed-loop perception–planning–capture framework with two key components. A Spherical Ray Adapter (SRA) incorporates synchronized camera poses into perspective-based VGGT-Ω, enabling streaming metric panoramic depth prediction with cross-window consistency. Adaptive Initialization automatically configures planning parameters from the scene scale for efficient exploration across diverse environments. To support systematic evaluation, we further develop a unified panoramic simulation platform with real-time point-cloud output and standardized planner interfaces, and construct a large-scale benchmark with panoramic observations, accurate poses, and ground-truth geometry. Extensive experiments show that FAR outperforms existing methods in both planning and reconstruction; with RGB panoramas only, it reaches 83.3% coverage with a 1017 m mean path, exceeding MAGICIAN with ground-truth depth (79.4%, 1400 m). The proposed simulator, benchmark, and models will be publicly released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Active reconstruction aims at autonomously acquiring informative observations to efficiently reconstruct unknown 3D environments. Among various applications, UAV-based scanning (Zhou et al., 2021; Geng et al., 2025; Guédon et al., 2023) is a representative active reconstruction task, where the 3D mobility of UAVs provides substantial freedom for active viewpoint selection.

Despite remarkable progress, existing active mapping methods rely mainly on perspective cameras or range sensors with limited fields of view and address active capture through exploration, coverage estimation, or utility prediction (Zhou et al., 2021; Geng et al., 2025; Guédon et al., 2023; Li et al., 2026a). Their limited fields of view require frequent adjustments of camera position and orientation, introducing substantial overhead and hindering fast active reconstruction, i.e., high coverage per meter flown, in large-scale outdoor scenes.

An intuitive solution is to adopt omnidirectional sensing (Wang et al., 2026c) and adapt existing perspective-based active reconstruction modules to panoramic observations. However, such an adaptation is non-trivial on both the perception and planning sides. Specifically, streaming panoramic reconstruction requires stable and metric geometric prediction over time from perspective-trained backbones, while noisy predicted point clouds and the scale variation of outdoor scenes make fixed planning configurations inefficient. Moreover, a comprehensive benchmark remains lacking, as existing panoramic simulators such as AirSim360 (Ge et al., 2026) do not support closed-loop active reconstruction, which requires online interaction among perception, planning, and acquisition.

To address these challenges, we propose FAR, a closed-loop perception–planning–capture framework for fast active reconstruction with panoramic vision, which consists of two core modules: a Spherical Ray Adapter (SRA) for panoramic perception and Adaptive Initialization for view planning. SRA incorporates synchronized camera poses into perspective-based VGGT-Ω, and its depth is aligned to metric scale and refined in a persistent voxel map inspired by AMB3R (Wang & Agapito, 2026), enabling streaming metric panoramic geometry with cross-window consistency. Meanwhile, Adaptive Initialization derives the scene scale from the given task boundary and automatically configures planning parameters, such as voxel resolution and viewpoint-sampling radius.

To support systematic evaluation, we further develop a unified simulation platform and construct FAR26, a large-scale benchmark with panoramic observations, accurate poses, and ground-truth geometry. Compared with AirSim360 (Ge et al., 2026), our platform further supports real-time point-cloud output and a standardized planner interface for closed-loop active reconstruction. Experiments on 11 Macarons++ scenes and FAR26 show that FAR outperforms existing planners and geometry models; with RGB panoramas only, it reaches 83.3% coverage with a $1017\,\mathrm{m}$ mean path, exceeding MAGICIAN with ground-truth depth (79.4%, $1400\,\mathrm{m}$; Figure 1).

Our main contributions are as follows:

- We propose FAR, a closed-loop perception–planning–capture framework for fast active reconstruction with panoramic vision, leveraging omnidirectional observations to enable efficient UAV-based acquisition in large-scale outdoor scenes.
- We introduce two key components: SRA, which enables streaming metric panoramic geometry prediction with cross-window consistency, and Adaptive Initialization, which automatically configures planning parameters according to scene scale.
- We build a unified simulation platform and FAR26, a large-scale benchmark for panoramic active reconstruction, with real-time point-cloud output and standardized planner interaction. Experiments on diverse outdoor scenes demonstrate the effectiveness of FAR.

2 RELATED WORK

2.1 UAV ACTIVE MAPPING AND VIEW PLANNING

UAV mapping has evolved from frontier- and sampling-based exploration (Yamauchi, 1997; Cieslewski et al., 2017; Bircher et al., 2016) to hierarchical systems with incremental frontiers, global route ordering, and local viewpoint refinement (Zhou et al., 2021) on probabilistic volumetric

maps (Hornung et al., 2013). Reconstruction-oriented planners value observations by their contribution to surfaces, via volumetric information gain (Isler et al., 2016), trajectory optimization over a prior model (Roberts et al., 2017; Smith et al., 2018), or surface frontiers (Senarathne & Wang, 2016; Song & Jo, 2018; Hardouin et al., 2020; Zhang et al., 2024); EPIC (Geng et al., 2025) further evaluates surface observation quality in large-scale LiDAR exploration. Learning-based methods predict coverage gain (Guédon et al., 2023) or long-horizon trajectory utility (Li et al., 2026a), and radiance-field methods select views by model uncertainty (Jiang et al., 2024). Omnidirectional sensing (Liu et al., 2024) removes heading search, as in position-only exploration with multi-fisheye cameras (Wang et al., 2026c).

2.2 Feed-Forward 3D Geometry

Feed-forward reconstruction replaces iterative SfM/MVS matching with direct geometric prediction (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026b;a), and MapAnything (Keetha et al., 2026) further accepts optional geometric inputs for metric reconstruction. Geometry-aware attention and relative camera encodings inject camera geometry into multi-view transformers (Miyato et al., 2024; Li et al., 2025b; Zhang et al., 2026a); our SRA extends this idea to per-token spherical rays with known poses. Monocular panoramic methods handle spherical distortion via distortion-aware designs, projection fusion, or panoramic transformers (Zioulis et al., 2018; Wang et al., 2020; Jiang et al., 2021; Shen et al., 2022), and recent feed-forward models accept wide-field-of-view and equirectangular inputs (Jung et al., 2026; Li et al., 2026b; Guo et al., 2026). Streaming and chunk-based models propagate long-sequence context through spatial or pointer memory, persistent states, test-time-trained memory, or chunk alignment (Wang & Agapito, 2025; Wang et al., 2025b; Wu et al., 2025; Zhuo et al., 2026; Deng et al., 2026; Zhang et al., 2026b; Xie et al., 2026), inferring poses jointly with geometry and carrying context mostly in learned memory.

3 Method

3.1 Problem Formulation and System Overview

Problem setup. A UAV with a $360^\circ \times 180^\circ$ panoramic camera reconstructs an unknown outdoor scene within a task boundary $B \subset \mathbb{R}^3$. The i-th capture provides an equirectangular projection (ERP) panorama I_i with a synchronized metric camera-to-world pose (R_i, c_i); a pixel $\mathbf{u}$ with unit ray $r(\mathbf{u}) \in \mathbb{S}^2$ and radial depth $d_i(\mathbf{u})$ maps to the world point $X_i(\mathbf{u}) = R_i\, r(\mathbf{u})\, d_i(\mathbf{u}) + c_i$. As each panorama covers the full sphere, the planner only selects the next capture position $p \in B$, aiming to cover the scene structures in B with minimal flight distance.

System overview. Reconstruction and planning alternate and exchange only metric world points (Figure 2, Algorithm 1): for each window $\mathcal{W}$ of consecutive captures, Section 3.2 predicts metric depth, and Section 3.3 updates a probabilistic voxel map and selects the next capture position p.

3.2 Streaming Metric Panoramic Geometry

3.2.1 Pose-Conditioned Spherical Ray Adapter

Inspired by the relative ray encoding of UCPE (Zhang et al., 2026a), we insert SRA into a pretrained VGGT-Ω backbone (Wang et al., 2026a) to obtain a panoramic feed-forward geometry model. SRA addresses the mismatch between features learned under planar perspective projection and the non-linear ERP projection by conditioning attention on the spherical rays and the relative camera poses; for each window $\mathcal{W}$, the network outputs raw radial depth $d_i^{\mathrm{raw}}(\mathbf{u})$ along each ERP ray, which is aligned to metric scale and refined in Section 3.2.2.

Specifically, for token n in frame i with unit ray $r_n = r(\mathbf{u}_n)$ through its patch center $\mathbf{u}_n$, camera rotation R_i, center c_i, and downward camera axis in the world frame $\mathbf{y}_i = R_i(0,1,0)^\top$, SRA constructs a local ray frame $R_n^{wr} = [\mathbf{x}_n, \mathbf{y}_n, \mathbf{z}_n]$ and a ray-to-world transform T_n^{wr}:

$$\begin{aligned} \mathbf{z}_n &= R_i r_n, \quad \mathbf{x}_n = \frac{\mathbf{y}_i \times \mathbf{z}_n}{\|\mathbf{y}_i \times \mathbf{z}_n\|}, \quad \mathbf{y}_n = \mathbf{z}_n \times \mathbf{x}_n, \\ T_n^{wr} &= \begin{bmatrix} R_n^{wr} & c_i - c_{\mathrm{ref}} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \end{aligned} \tag{1}$$

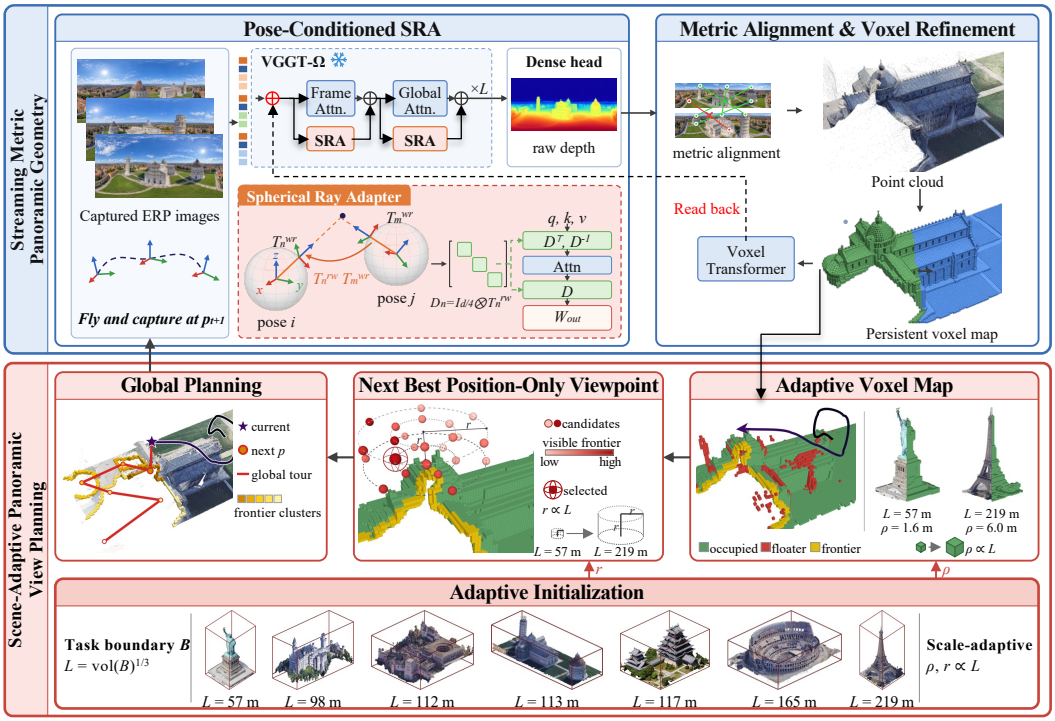


Figure 2: Overview of FAR. Top: SRA adapts VGGT-Ω to posed ERP panoramas; the depth is aligned to metric scale and refined in a persistent voxel map. Bottom: with parameters set by Adaptive Initialization, the refined points update a probabilistic map whose structure-anchored frontiers determine the next capture position.

where c_{ref} is a reference position of the window. Let $T_n^{rw} = (T_n^{wr})^{-1}$. Following UCPE and GTA (Zhang et al., 2026a; Miyato et al., 2024), SRA views the d-dimensional query, key, and value features of each attention head as $d/4$ homogeneous 4-vectors and applies the block-diagonal operator $D_n = I_{d/4} \otimes T_n^{rw}$. Keys and values are lifted from their ray frames to a shared frame, and the output is mapped back into the ray frame of the query:

$$o_n = D_n \sum_m \mathrm{softmax}_m \left(\frac{(D_n^\top q_n)^\top D_m^{-1} k_m}{\sqrt{d}} \right) D_m^{-1} v_m, \tag{2}$$

where q_n, k_n, and v_n are linear projections of the feature h_n of token n, and m ranges over the tokens attended by n. Both the logits and o_n depend on the poses only through the relative transform $T_n^{rw} T_m^{wr}$, and are thus invariant to any rigid change of the world frame. The token update is $h_n \leftarrow h_n + W_{\mathrm{out}} o_n$ with a zero-initialized output projection W_{out}, so SRA initially preserves the behavior of the pretrained backbone and gradually learns a pose-aware panoramic correction.

3.2.2 Metric Alignment and Persistent Voxel Refinement

SRA produces geometrically structured predictions, but its raw depth still deviates from metric scale. We therefore align it to metric scale before refining it globally. Sparse correspondences obtained by SphereGlue (Gava et al., 2023) that pass a spherical epipolar check with the known poses (Appendix A.2) are triangulated with the known camera baselines to provide metric depth anchors, and we estimate the scale factor $\sigma_{\mathcal{W}}$ as the exponentiated median of the log-ratios between triangulated and raw depths, giving $d_i^{\mathrm{m}} = \sigma_{\mathcal{W}} d_i^{\mathrm{raw}}$ (Eq. (11)). Back-projecting this depth along each spherical ray with the known camera pose yields the metric window point cloud $\mathcal{P}_{\mathcal{W}}$.

The window point cloud $\mathcal{P}_{\mathcal{W}}$ and the 2D features $\{G_i\}_{i \in \mathcal{W}}$ are then fused into a persistent global voxel map. Like the voxel backend of AMB3R (Wang & Agapito, 2026), we discretize metric space with the resolution ρ shared with the planner (Eq. (5)), mapping each point $X_i(\mathbf{u})$ to the voxel $v = \lfloor X_i(\mathbf{u})/\rho \rfloor$; unlike AMB3R, which resets its voxel map for each window, our map persists

and keeps accumulating voxels over the entire trajectory. The map $\mathcal{M}$ stores, for each voxel, the cumulative feature sum S_v^{hist} and point count w_v^{hist}, from which the contributions of frames shared with the previous window are removed before accumulation:

$$H_v = \frac{S_v^{\text{hist}} + \sum_{(i,\mathbf{u})\in\mathcal{P}_v^{\mathcal{W}}} G_i[\mathbf{u}]}{w_v^{\text{hist}} + |\mathcal{P}_v^{\mathcal{W}}|}, \tag{3}$$

where $\mathcal{P}_v^{\mathcal{W}} = \{(i,\mathbf{u}) : i \in \mathcal{W},\ \lfloor X_i(\mathbf{u})/\rho \rfloor = v\}$ indexes the current-window points in voxel v. A voxel transformer f_θ (Wu et al., 2024) maps the aggregated features $\{H_v\}$ of the active voxels near the current observations to refined voxel features, aggregating structural context in Euclidean neighborhoods; these are interpolated back to the observation points as $\tilde{G}$ and injected into the frozen 2D backbone, a step we refer to as read-back:

$$h \leftarrow h + \beta\ \text{Conv}_{\text{zero}}(\tilde{G}), \tag{4}$$

where β is a learnable gate. Decoding the updated features yields refined depth and world points, which update $\mathcal{M}$ and are passed directly to the planner. The model is trained in two stages, first SRA and then the voxel refinement (Appendix A.3).

3.3 Scene-Adaptive Panoramic View Planning

Given the metric world points of Section 3.2, the planner selects the next capture position p among candidates that sufficiently observe structure-anchored frontiers while minimizing flight distance, on a probabilistic voxel map that tolerates noisy predicted geometry and whose metric parameters are initialized from the task scale.

3.3.1 Adaptive Initialization and Probabilistic Voxel Mapping

Fixed parameters either miss details in small scenes or waste computation in large ones. We therefore take the cube root of the volume of the task boundary B as a characteristic scene length and scale all metric parameters with it:

$$L = \text{vol}(B)^{1/3}, \qquad \rho = \alpha L, \quad r = \gamma L, \quad d_{\text{ret}} = \epsilon L, \tag{5}$$

where ρ is the voxel resolution, r the radius of the concentric rings for viewpoint sampling, d_{ret} the distance for identifying unobservable frontier voxels, and α, γ, and ϵ dimensionless coefficients. As all lengths scale with L and the map size $L/\rho = \alpha^{-1}$ is independent of the scene size, we use a single set of coefficients for all scenes, from isolated sculptures to large bridges, without the scene-specific metric settings required by conventional frontier planners (Zhou et al., 2021; Geng et al., 2025).

Within B, the planner maintains a probabilistic three-state voxel map with occupied, free, and unknown sets $\mathcal{V}_{\text{occ}}$, $\mathcal{V}_{\text{free}}$, and $\mathcal{V}_{\text{unk}}$. At planning step t, rays cast from the camera center to the predicted points add $\Delta\ell_t(v) = \Delta_o$ to each end voxel and $-\Delta_f$ to each traversed voxel, accumulated as a clipped log-odds value:

$$\ell_t(v) = \text{clip}(\ell_{t-1}(v) + \Delta\ell_t(v),\ \ell_{\min}, \ell_{\max})\,. \tag{6}$$

A voxel is occupied if $\ell_t(v) > \tau_o$, free if $\ell_t(v) < -\tau_f$, and unknown otherwise, so later observations can correct an erroneous state caused by a single inaccurate prediction.

3.3.2 Panoramic View Selection and Convergence

Classical frontiers (Yamauchi, 1997), the boundaries between free and unknown space, often lie in the sky or distant free space outdoors. Following surface-frontier formulations (Song & Jo, 2018; Zhang et al., 2024), we define the structure-anchored frontier as the free voxels that border both unknown and occupied space:

$$\mathcal{F} = \{v \in \mathcal{V}_{\text{free}} \,|\, \mathcal{N}_6(v) \cap \mathcal{V}_{\text{unk}} \neq \emptyset \wedge \mathcal{N}_6(v) \cap \mathcal{V}_{\text{occ}} \neq \emptyset\}\,, \tag{7}$$

where $\mathcal{N}_6(v)$ is the 6-connected neighborhood; unlike Zhang et al. (2024), the two neighbors need not be adjacent. $\mathcal{F}$ is grouped into connected clusters $\{F_j\}$, and a frontier voxel approached within d_{ret} without being resolved is deemed unobservable and excluded from viewpoint selection.

Table 1: Planning efficiency on the 11 Macarons++ scenes, ordered by scene scale $\mathrm{vol}(B)^{1/3}$. In each metric block, the upper group compares planners with ground-truth depth, and the lower group compares geometry sources under the FAR planner. Values are means over five start positions per scene. Bold: best per group. Path length should be read together with coverage.

Method	Red.	Lib.	Pan.	Neu.	Alh.	Pisa	Fus.	Dun.	Col.	Eif.	Bri.	Mean
$\mathrm{vol}(B)^{1/3}$ (m)	*45*	*57*	*73*	*98*	*113*	*113*	*117*	*163*	*165*	*219*	*233*	–
Coverage at $\varepsilon = 1.0\,m$ (%)↑												
EPIC (w/ GT depth)	**94.1**	84.4	**84.4**	68.9	62.2	63.0	61.6	84.3	40.2	52.9	25.5	65.6
MAGICIAN (w/ GT depth)	93.5	81.7	70.8	78.2	75.3	69.4	94.2	94.0	67.2	67.7	81.2	79.4
FAR (w/ GT depth)	93.2	**85.2**	82.0	**99.0**	**77.6**	**71.3**	**96.4**	**98.9**	**71.9**	**75.6**	**90.3**	**85.6**
DAP	17.7	43.1	30.6	21.2	12.5	13.1	9.9	0.2	1.0	0.8	1.4	13.8
PanoVGGT	58.0	94.0	69.1	73.5	79.7	73.2	72.4	77.2	65.1	32.5	66.7	69.2
Ours	**100.0**	**99.8**	**74.2**	**89.6**	**82.9**	**79.4**	**86.3**	**86.7**	**66.7**	**72.8**	**77.7**	**83.3**
Coverage–path AUC↑												
EPIC (w/ GT depth)	0.615	0.475	0.591	0.471	0.271	0.427	0.396	0.310	0.194	0.288	0.111	0.377
MAGICIAN (w/ GT depth)	0.486	0.574	0.500	0.494	0.486	0.467	0.589	0.539	0.422	0.393	0.554	0.500
FAR (w/ GT depth)	**0.836**	**0.708**	**0.654**	**0.785**	**0.596**	**0.565**	**0.789**	**0.738**	**0.530**	**0.451**	**0.638**	**0.663**
DAP	0.056	0.272	0.159	0.138	0.081	0.076	0.047	0.002	0.008	0.007	0.008	0.078
PanoVGGT	0.205	0.603	0.506	0.429	0.497	0.485	0.565	0.472	0.436	0.036	0.398	0.421
Ours	**0.802**	**0.899**	**0.608**	**0.683**	**0.574**	**0.614**	**0.651**	**0.664**	**0.450**	**0.521**	**0.512**	**0.634**
Path length (m)↓												
EPIC (w/ GT depth)	408	725	1507	1046	1291	2519	1444	3775	2214	2398	2547	1807
MAGICIAN (w/ GT depth)	547	592	**729**	806	1058	1254	1465	1821	1985	2706	**2440**	1400
FAR (w/ GT depth)	**180**	**368**	758	**685**	**832**	**1006**	**903**	**1278**	**1493**	**1887**	2836	**1111**
DAP	2957	2392	1719	2164	1399	1576	1412	2384	2405	**400**	6864	2334
PanoVGGT	**124**	**411**	605	**579**	837	**743**	**528**	1538	**1137**	913	**1632**	**822**
Ours	433	434	**456**	698	**829**	1086	621	**1413**	1622	1632	1961	1017

For each cluster F_j, candidate positions are sampled on concentric rings of radius r at multiple heights (Appendix A.4). As the panorama observes all directions, each candidate is scored by position alone, without yaw search, through its visibility score $S(p, F_j)$, the number of voxels of F_j with an unobstructed line of sight from p. Among candidates reaching a prescribed score threshold, a global–local route planner adapted from FUEL (Zhou et al., 2021) selects the next capture position while minimizing flight distance.

Since occlusions and unreachable surfaces can keep $\mathcal{F}$ nonempty, acquisition stops when the number $N_t = |\mathcal{F}^{\mathrm{a}}|$ of active frontier voxels $\mathcal{F}^{\mathrm{a}} \subseteq \mathcal{F}$, i.e., those not deemed unobservable, falls below a fraction η of its peak N_{peak} so far:

$$N_t < \eta N_{\mathrm{peak}}, \tag{8}$$

a relative criterion that is invariant to the scene scale.

4 Experiments

We evaluate planning efficiency (Section 4.2) with two controlled comparisons, one of planners under the same ground-truth geometry and one of geometry sources under the same planner, and then assess the geometry itself (Section 4.3), its downstream use in 3D Gaussian Splatting (Kerbl et al., 2023) (3DGS, Section 4.4), and the reconstruction components (Section 4.5).

4.1 Experimental Setup

Simulation platform and benchmark. Our platform renders panoramic UAV observations, as in AirSim360 (Ge et al., 2026), and provides real-time interfaces for external planners and perception modules, supporting both closed-loop and open-loop evaluation. On this platform, we build FAR26, 20 high-fidelity Unreal Engine outdoor scenes with panoramic observations, accurate poses, and ground-truth geometry (Appendix A.5).

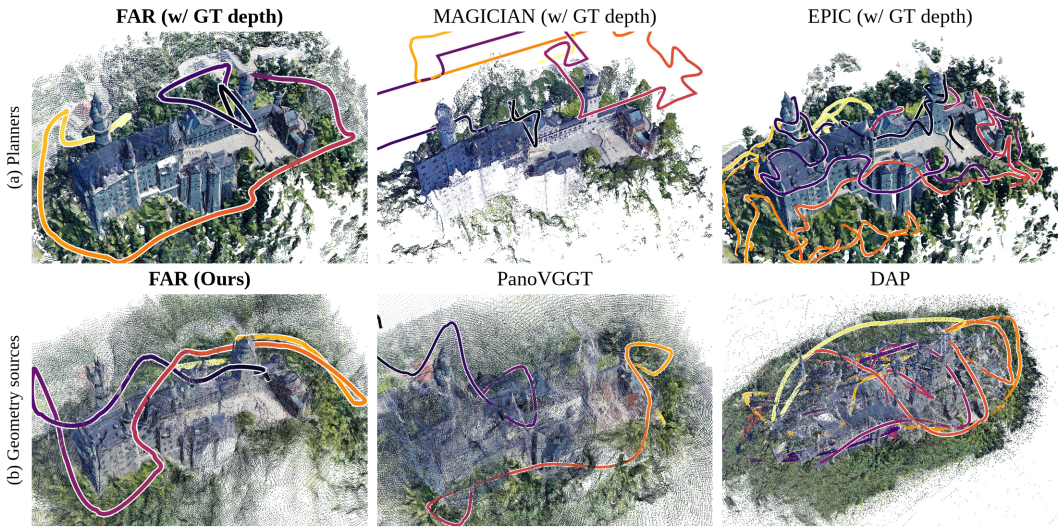
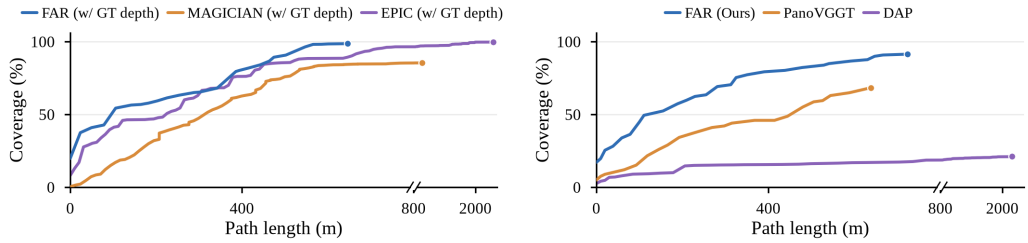


Figure 3: Trajectories on Neuschwanstein, showing the highest-coverage run of each method. (a) Planners with ground-truth depth. (b) The FAR planner driven by each geometry source, rendered with the point clouds from predicted depth. Color encodes flight progress.



(a) Planners with ground-truth depth

(b) Geometry sources under the FAR planner

Figure 4: Coverage ($\varepsilon = 1.0\,\mathrm{m}$) versus path length on Neuschwanstein for the runs in Figure 3. Path-length axes are compressed beyond the break marks.

Datasets. Planning is evaluated on the 11 Macarons++ outdoor scenes (Guédon et al., 2023; Li et al., 2026a), real-scan landmark meshes with $\mathrm{vol}(B)^{1/3}$ from 45 m to 233 m, and reconstruction also on the FAR26 test scenes, using posed panoramas from FAR flights with ground-truth depth.

Baselines. The first planning comparison feeds the same ground-truth depth to EPIC (Geng et al., 2025) (LiDAR), MAGICIAN (Li et al., 2026a) (perspective camera), and FAR (ERP panorama) through their native sensors; the second drives the FAR planner with depth from DAP (Lin et al., 2026), PanoVGGT (Guo et al., 2026), or our model. Reconstruction baselines are VGGT-Ω (Wang et al., 2026a), Argus (Li et al., 2026b), Wid3R (Jung et al., 2026), and PanoVGGT, all aligned to metric scale with our pose-based triangulation (Section 3.2.2), while monocular DAP is used as is. Planning results are averaged over five start positions per scene (details in Appendix A.6).

Metrics. For planning, following MACARONS (Guédon et al., 2023) and MAGICIAN (Li et al., 2026a), coverage is the fraction of 10^5 ground-truth surface points within $\varepsilon = 1.0\,\mathrm{m}$ of the point cloud from the depth consumed by the planner, and path length is the total flight distance, both at the end of a run; AUC is the normalized area under the coverage–path curve up to a per-scene budget, the shortest mean path of the planners with ground-truth depth, so a higher AUC means faster coverage per meter. For reconstruction, depth is evaluated with AbsRel, RMSE, and $\delta_{1.25}$, fused point clouds with Rel, Acc, Comp, and $F@1\mathrm{m}$, and 3DGS renders with PSNR, SSIM, and LPIPS.

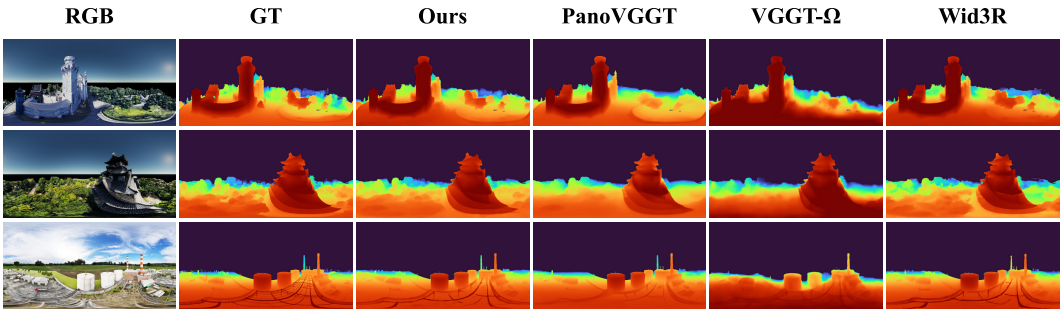


Figure 5: Qualitative multi-view metric depth on Macarons++ and FAR26. Baseline predictions are scale-aligned as in Table 2.

Table 2: Multi-view metric depth (per 8-frame window) and 3D reconstruction (fused over the trajectory) on Macarons++ and the FAR26 test scenes. Baselines are aligned to metric scale as ours. RMSE, Acc, and Comp are in meters; Rel is the relative error and F@1m the F-score at 1 m. Best results are in **bold**.

	Depth estimation			3D reconstruction			
Method	AbsRel↓	RMSE↓	$\delta_{1.25}$ ↑	Rel%↓	Acc↓	Comp↓	F@1m↑
Macarons++							
VGGT-Ω (Wang et al., 2026a)	0.3047	38.29	0.4343	31.14	0.928	0.622	0.292
Argus (Li et al., 2026b)	0.5328	69.61	0.2481	60.35	0.974	0.731	0.227
Wid3R (Jung et al., 2026)	0.1553	40.95	0.7829	14.95	0.900	0.522	0.458
PanoVGGT (Guo et al., 2026)	0.1667	27.18	0.7684	17.10	0.850	0.493	0.499
Ours	**0.1196**	**23.29**	**0.8647**	**11.25**	**0.780**	**0.462**	**0.586**
FAR26							
VGGT-Ω (Wang et al., 2026a)	0.2908	36.16	0.5226	30.37	0.875	0.474	0.349
Argus (Li et al., 2026b)	0.2864	38.17	0.4022	28.75	0.894	0.485	0.373
Wid3R (Jung et al., 2026)	0.1069	19.92	0.8967	9.80	0.812	0.344	0.583
PanoVGGT (Guo et al., 2026)	0.1294	15.18	0.8475	13.14	0.779	0.358	0.587
Ours	**0.0696**	**12.09**	**0.9450**	**6.55**	**0.667**	**0.301**	**0.740**

4.2 Planning Efficiency

Planners on ground-truth geometry. Under identical ground-truth depth (Table 1, top), FAR achieves the highest mean coverage (85.6% vs. 79.4% for MAGICIAN and 65.6% for EPIC), the highest AUC on all 11 scenes (mean 0.663 vs. 0.500 and 0.377), and a 20.6% and 38.5% shorter mean path. Circling the structures in a few loops (Figure 3(a)), it gains coverage fastest (Figure 4(a)).

Geometry sources under the FAR planner. With our model, FAR retains 97% and 96% of its coverage and AUC with ground-truth depth (83.3%, 0.634) and outperforms PanoVGGT and DAP in both metrics on every scene. On Red., Lib., Alh., and Pisa, its coverage even exceeds that with ground-truth depth, which we attribute to the noise and extra points of predicted depth promoting further exploration. PanoVGGT flies shorter but ends lower (Figure 4(b)), and DAP covers little (Figure 3(b)).

4.3 3D Reconstruction

We evaluate this geometry per 8-frame window with depth metrics and fused over the full trajectory with surface metrics. Our model achieves the best result in every metric on both datasets (Table 2); relative to the best baseline in each metric, it lowers AbsRel by 23.0% and 34.9% and Rel by 24.7% and 33.2%, and raises F@1m by 17.4% and 26.1% on Macarons++ and FAR26, respectively. Its depth also better preserves the boundaries of the towers, the tiered roof, and the chimneys (Figure 5).

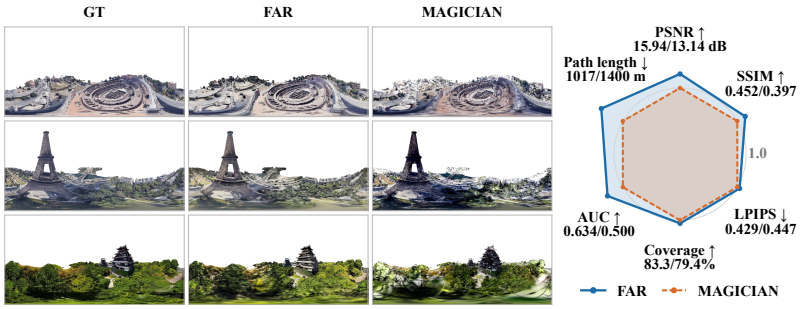


Figure 6: Novel-view synthesis. Left: ground-truth, FAR, and MAGICIAN renders. Right: FAR (predicted depth) vs. MAGICIAN (ground-truth depth) on rendering and planning metrics (Table 1), averaged over the 11 Macarons++ scenes and normalized by MAGICIAN.

Table 3: Ablation on Macarons++ with the metrics of Table 2. Best results are in **bold**.

	Depth estimation			3D reconstruction			
Setting	AbsRel↓	RMSE↓	$\delta_{1.25}$ ↑	Rel%↓	Acc↓	Comp↓	F@1m↑
w/o SRA	0.1914	31.46	0.6821	19.37	0.872	0.561	0.433
w/o alignment	0.2073	37.72	0.5872	18.06	0.930	0.693	0.370
w/o refinement	0.1337	26.54	0.8240	12.16	0.811	0.498	0.545
Ours	**0.1196**	**23.29**	**0.8647**	**11.25**	**0.780**	**0.462**	**0.586**

4.4 3D Gaussian Splatting

The fused point cloud of the shorter, RGB-only FAR flight (Section 4.2) initializes a 3DGS scene, compared with MAGICIAN using ground-truth depth. As perspective RGB-D and RGB panoramas cannot be trained identically, each flight uses a matched splatter without access to test cameras, FastGS (Ren et al., 2026) for MAGICIAN and SPaGS (Li et al., 2025a) for FAR; for a fair comparison, both are scored at the same ten panoramas (Appendix A.6). FAR reaches 15.94 dB PSNR, 0.452 SSIM, and 0.429 LPIPS, against 13.14 dB, 0.397, and 0.447 for MAGICIAN (Figure 6), indicating that its more efficient panoramic acquisition yields better 3DGS despite using RGB only.

4.5 Ablation Study

We ablate the three components of Section 3.2 on Macarons++ (Table 3): **w/o SRA** fine-tunes only the depth head of the frozen VGGT-Ω, **w/o alignment** fuses the raw depth d^{raw} without Eq. (11), and **w/o refinement** skips the voxel refinement and read-back of Eq. (4). Removing any component degrades all seven metrics. Without SRA, Rel is the highest (19.37%), suggesting that the depth head alone cannot adapt perspective features to the ERP projection; without alignment, windows are fused at inconsistent scales, giving the largest errors in six of the seven metrics (AbsRel +73.3%, Comp +50.0%); without refinement, the drops are smaller but consistent (F@1m 0.586 → 0.545).

5 Conclusion

We presented FAR, a closed-loop perception–planning–capture framework that exploits panoramic vision for fast active reconstruction. As each panorama observes the full sphere, FAR reduces view planning to selecting capture positions. SRA adapts a perspective feed-forward backbone to posed ERP panoramas, whose depth is aligned to metric scale and refined in a persistent voxel map; Adaptive Initialization derives all metric planning parameters from the task scale, so that a single set of coefficients serves scenes from 45 m to 233 m. FAR achieves higher coverage per meter flown than existing planners and more accurate metric geometry than existing feed-forward models.

Limitations. Like most active reconstruction systems, FAR assumes accurate synchronized poses, whereas GNSS and visual–inertial estimates drift in large outdoor scenes; refining poses jointly with geometry online remains open. Sky, water, glass, and thin structures remain challenging for feed-forward depth prediction, and dynamic objects violate the static-scene assumption shared by current methods. Real-flight deployment under onboard compute limits is left for future work.

Reproducibility Statement

Section 3 and Algorithm 1 (Appendix A.1) describe the complete closed-loop pipeline. Implementation details of the reconstruction model are given in Appendix A.2, its training data, scene split, objectives, and loss weights in Appendix A.3, and the planning details in Appendix A.4. Appendix A.5 describes the simulation platform and the FAR26 benchmark, and Appendix A.6 lists the sensor settings and metric definitions. The simulator, benchmark, and models will be publicly released.

AI Use Statement

We used generative AI tools only to aid and polish the writing and to check and proofread the paper formatting. Generative AI tools were not used for any other part of this research. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work.

References

Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon "next-best-view" planner for 3D exploration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1462–1468, 2016. 2

Titus Cieslewski, Elia Kaufmann, and Davide Scaramuzza. Rapid exploration with multi-rotors: A frontier selection method for high speed flight. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2135–2142, 2017. 2

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: Chunk it, loop it, align it – pushing VGGT's limits on kilometer-scale long RGB sequences. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. 3

Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and Andreas Dengel. SphereGlue: Learning keypoint matching on high resolution spherical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6134–6144, 2023. 4, 15

Xian Ge, Yuling Pan, Yuhang Zhang, Xiang Li, Weijun Zhang, Dizhe Zhang, Zhaoliang Wan, Xin Lin, Xiangkai Zhang, Juntao Liang, Xiangtai Li, Wenjie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. AirSim360: A panoramic simulation platform within drone view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26931–26940, 2026. 2, 6

Shuang Geng, Zelin Ning, Fu Zhang, and Boyu Zhou. EPIC: A lightweight LiDAR-based AAV exploration framework for large-scale scenarios. *IEEE Robotics and Automation Letters (RA-L)*, 10(5):5090–5097, 2025. 2, 3, 5, 7

Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. MACARONS: Mapping and coverage anticipation with RGB online self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 940–951, 2023. 2, 3, 7

Yijing Guo, Mengjun Chao, Luo Wang, Tianyang Zhao, Haizhao Dai, Yingliang Zhang, Jingyi Yu, and Yujiao Shi. PanoVGGT: Feed-forward 3D reconstruction from panoramic imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36444–36453, 2026. 3, 7, 8

Guillaume Hardouin, Julien Moras, Fabio Morbidi, Julien Marzat, and El Mustapha Mouaddib. Next-best-view planning for surface reconstruction of large-scale 3D environments with multiple UAVs. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1567–1574, 2020. 3

Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 34(3):189–206, 2013. 3

Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3D reconstruction. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3477–3484, 2016. 3

Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. UniFuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters (RA-L)*, 6(2):1519–1526, 2021. 3

Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and mapping with radiance fields using Fisher information. In *European Conference on Computer Vision (ECCV)*, pp. 422–440, 2024. 3

Dongki Jung, Jaehoon Choi, Adil Qureshi, Somi Jeong, Dinesh Manocha, and Suyong Yeon. Wid3R: Wide field-of-view 3D reconstruction via camera model conditioning. In *European Conference on Computer Vision (ECCV)*, pp. 665–684, 2026. 3, 7, 8

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, pp. 499–509, 2026. 3

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 6

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024. 3

Junbo Li, Florian Hahlbohm, Timon Scholz, Martin Eisemann, Jan-Philipp Tauscher, and Marcus Magnor. SPaGS: Fast and accurate 3D Gaussian splatting for spherical panoramas. *Computer Graphics Forum*, 44(4), 2025a. doi: 10.1111/cgf.70171. 9

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 18020–18045, 2025b. 3

Shiyao Li, Antoine Guédon, Shizhe Chen, and Vincent Lepetit. MAGICIAN: Efficient long-term planning with imagined gaussians for active mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21606–21615, 2026a. 2, 3, 7

Xi Li, Linyuan Li, Yan Wu, Tong Rao, Kai Zhang, Xinchen Hui, and Cihui Pan. Argus: Metric panoramic 3D reconstruction for indoor scenes. In *European Conference on Computer Vision (ECCV)*, pp. 628–647, 2026b. 3, 7, 8

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26835–26844, 2026. 7

Peize Liu, Chen Feng, Yang Xu, Yan Ning, Hao Xu, and Shaojie Shen. OmniNxt: A fully open-source and compact aerial robot with omnidirectional visual perception. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10605–10612, 2024. 3

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. GTA: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations (ICLR)*, 2024. 3, 4

Jing Ou, Zidong Cao, Yinrui Ren, Zhuoxiao Li, Jinjing Zhu, Tongyan Hua, Shuai Zhang, Hui Xiong, and Wufan Zhao. Holo360D: A large-scale real-world dataset with continuous trajectories for advancing panoramic 3D reconstruction and beyond. In *European Conference on Computer Vision (ECCV)*, 2026. 15

Shiwei Ren, Tianci Wen, Yongchun Fang, and Biao Lu. FastGS: Training 3D Gaussian splatting in 100 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26094–26103, 2026. 9

Mike Roberts, Shital Shah, Debadeepta Dey, Anh Truong, Sudipta Sinha, Ashish Kapoor, Pat Hanrahan, and Neel Joshi. Submodular trajectory optimization for aerial 3D scanning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5334–5343, 2017. 3

P. G. C. N. Senarathne and Danwei Wang. Towards autonomous 3D exploration using surface frontiers. In *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 34–41, 2016. 3

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *European Conference on Computer Vision (ECCV)*, pp. 195–211, 2022. 3

Neil Smith, Nils Moehrle, Michael Goesele, and Wolfgang Heidrich. Aerial path planning for urban scene reconstruction: A continuous optimization method and benchmark. *ACM Transactions on Graphics (TOG)*, 37(6), 2018. 3

Soohwan Song and Sungho Jo. Surface-based exploration for autonomous 3D modeling. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4319–4326, 2018. 3, 5

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 459–468, 2020. 3

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, pp. 78–89, 2025. 3

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14612–14625, 2026. 2, 4

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. 3

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schoenberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, 2026a. 3, 7, 8

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b. 3

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 (CVPR)*, pp. 20697–20709, 2024. 3

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026b. 3

Zirui Wang, Xinjia Luo, Haotian Sun, Jun Ma, Jian Guo, and Boyu Zhou. Palm-sized omnidirectional vision-based UAV exploration with sparse topological map guidance. *arXiv preprint arXiv:2605.07275*, 2026c. 2, 3

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851, 2024. 5, 15

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 77686–77710, 2025. 3

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable test-time training for large-scale 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21760–21771, 2026. 3

Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings of the 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, pp. 146–151, 1997. 2, 5

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38027–38037, 2026a. 3, 4

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. LoGeR: Long-context geometric reconstruction with hybrid memory. In *European Conference on Computer Vision (ECCV)*, 2026b. 3

Mingjie Zhang, Chen Feng, Zengzhi Li, Guiyong Zheng, Yiming Luo, Zhu Wang, Jinni Zhou, Shaojie Shen, and Boyu Zhou. SOAR: Simultaneous exploration and photographing with heterogeneous UAVs for fast autonomous reconstruction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10975–10982, 2024. 3, 5

Boyu Zhou, Yichen Zhang, Xinyi Chen, and Shaojie Shen. FUEL: Fast UAV exploration using incremental frontier structure and hierarchical planning. *IEEE Robotics and Automation Letters (RA-L)*, 6(2):779–786, 2021. 2, 5, 6, 15

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations (ICLR)*, 2026. 3

Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. OmniDepth: Dense depth estimation for indoors spherical panoramas. In *European Conference on Computer Vision (ECCV)*, pp. 453–471, 2018. 3

A APPENDIX

This appendix complements Sections 3 and 4 with the following details:

- the complete closed-loop procedure of FAR in Appendix A.1;
- reconstruction implementation details, including the SRA update, metric scale estimation, and persistent voxel refinement, in Appendix A.2;
- the training data and two-stage training objectives in Appendix A.3;
- planning implementation details, including scale adaptation, probabilistic mapping, frontier clustering, routing, and execution, in Appendix A.4;
- the simulation platform and the scenes of the FAR26 benchmark in Appendix A.5;
- experimental details, including sensor settings, metrics, and additional planning analysis, in Appendix A.6.

A.1 CLOSED-LOOP ALGORITHM

Algorithm 1 summarizes the perception–planning–capture loop of Section 3.1. Reconstruction and planning exchange only metric world points; the scale-dependent planning parameters are computed once from the task boundary before the flight.

Algorithm 1 Fast Active Reconstruction (FAR)

Require: Task boundary B, initial capture position $p_0 \in B$, termination ratio η
Ensure: Captured panoramas with poses $\mathcal{C}$ and fused metric world geometry

Compute L, ρ, r, and d_{ret} from B ▷ Eq. (5)
Initialize voxel map $\mathcal{M} \leftarrow \emptyset$, log-odds map $\ell_0 \equiv 0$, $\mathcal{C} \leftarrow \emptyset$, $N_{\mathrm{peak}} \leftarrow 0$, $t \leftarrow 0$
loop
  Capture I_t with pose (R_t, c_t) at p_t and append it to $\mathcal{C}$
  Form the window $\mathcal{W}_t$ from the latest captures
  Predict raw radial depth d^{raw} with SRA ▷ Section 3.2.1
  Align d^{raw} to metric depth d^{m} and back-project to $\mathcal{P}_{\mathcal{W}_t}$ ▷ Eq. (11)
  Fuse $\mathcal{P}_{\mathcal{W}_t}$ into $\mathcal{M}$ and decode refined world points ▷ Eqs. (3), (4)
  Update ℓ_t by spherical ray casting and extract $\mathcal{F}$ and $\{F_j\}$ ▷ Eqs. (6), (7)
  Exclude frontier voxels approached within d_{ret} but not resolved
  $N_t \leftarrow |\mathcal{F}^{\mathrm{a}}|$, $N_{\mathrm{peak}} \leftarrow \max(N_{\mathrm{peak}}, N_t)$
  if $N_t < \eta N_{\mathrm{peak}}$ **then** ▷ Eq. (8)
    break
  end if
  Generate ring candidates of radius r and filter them by $S(p, F_j)$ ▷ Section 3.3.2
  if no frontier cluster has a valid reachable candidate **then**
    break
  end if
  Select p_{t+1} by global–local routing and fly along a collision-free path
  $t \leftarrow t + 1$
end loop
return $\mathcal{C}$ and the fused metric world geometry

A.2 RECONSTRUCTION IMPLEMENTATION DETAILS

Spherical Ray Adapter. SRA is inserted into each transformer block of VGGT-Ω. Queries, keys, and values are projected to a compressed attention space with one eighth of the backbone width and two heads; camera and register tokens, which lie on no ray, bypass SRA. Subtracting c_{ref} in Eq. (1) only keeps the translations small and does not affect the relativity, since $T_n^{rw} T_m^{wr}$ is unchanged when all camera centers are shifted. Eq. (2) writes only the SRA branch, which operates as a residual alongside the frozen attention of the backbone. The complete token update in each block is therefore

$$h_n \leftarrow h_n + \mathrm{Attn}_{\mathrm{bb}}(h_n) + W_{\mathrm{out}}\, o_n, \tag{9}$$

where $\mathrm{Attn}_{\mathrm{bb}}$ denotes the pretrained backbone attention and W_{out} is initialized to zero.

Metric scale estimation. Each sparse correspondence $(\mathbf{u}_i, \mathbf{u}_j)$ between frames i and j obtained by SphereGlue (Gava et al., 2023) is verified against the spherical epipolar constraint

$$|(R_j\, r(\mathbf{u}_j))^\top \hat{\mathbf{n}}_{ij}| < \sin\theta_{\mathrm{epi}}, \tag{10}$$

where $\hat{\mathbf{n}}_{ij}$ is the unit normal of $R_i\, r(\mathbf{u}_i) \times (c_j - c_i)$ and θ_{epi} is an angular tolerance. For each verified correspondence k, triangulating the world rays $R_i\, r(\mathbf{u}_i)$ and $R_j\, r(\mathbf{u}_j)$ from the camera centers c_i and c_j yields a metric depth z_k^{tri} along the ray of frame i. Over the set $\mathcal{A}$ of verified correspondences in window $\mathcal{W}$, the scale factor is

$$\sigma_{\mathcal{W}} = \exp\Big(\,\mathrm{median}_{k\in\mathcal{A}}\,\big(\log z_k^{\mathrm{tri}} - \log d_i^{\mathrm{raw}}(\mathbf{u}_i)\big)\Big), \qquad d_i^{\mathrm{m}} = \sigma_{\mathcal{W}}\, d_i^{\mathrm{raw}}, \tag{11}$$

where $d_i^{\mathrm{raw}}(\mathbf{u}_i)$ is the raw depth at the matched pixel of correspondence k. The median in log space is robust to residual outlier matches, and the alignment introduces no learnable parameters.

Persistent voxel refinement. The voxel transformer f_θ is implemented with Point Transformer V3 (Wu et al., 2024), which serializes the active voxels along a Hilbert curve so that each attention window remains local in metric space. Refined voxel features are interpolated to the observation points from their $K = 16$ nearest voxels to form $\tilde{G}$. $\mathrm{Conv}_{\mathrm{zero}}$ in Eq. (4) consists of two 1×1 convolutions separated by a GELU, with the second convolution initialized to zero, and the gate β of Eq. (4) is initialized to one. The read-back residual is added at the patch embedding and before each intra-frame and cross-frame attention block. The voxel map $\mathcal{M}$ resides in host memory, and only the active neighborhood of the current window is transferred to the GPU.

A.3 Training Data and Objectives

Training data. The 20 FAR26 scenes are split into training and test scenes at a ratio of 4:1, giving 16 training and 4 test scenes; all FAR26 results are reported on the test scenes. The reconstruction model is trained on the FAR26 training scenes together with the open-source Holo360D dataset (Ou et al., 2026), which provides real-world 360° images with aligned camera poses and depth maps.

Objectives. Both training stages supervise metric radial depth. Stage 1 freezes the VGGT-Ω backbone and trains SRA and the dense depth head with

$$\mathcal{L}_1 = \mathcal{L}_{\log} + \lambda_\nabla \mathcal{L}_\nabla + \lambda_m \mathcal{L}_m, \tag{12}$$

where $\mathcal{L}_{\log}$ is the log-depth residual, $\mathcal{L}_\nabla$ is the edge-aware gradient penalty, and $\mathcal{L}_m$ is the metric residual. Stage 2 freezes all Stage 1 parameters, trains f_θ and $\mathrm{Conv}_{\mathrm{zero}}$ on the post-read-back predictions, and adds a surface-normal term $\mathcal{L}_n$ that penalizes angular normal deviations:

$$\mathcal{L}_2 = \mathcal{L}_1 + \lambda_n \mathcal{L}_n. \tag{13}$$

Owing to zero initialization, Stage 2 initially reproduces the Stage 1 predictions. We use $\lambda_\nabla = 1.0$, $\lambda_m = 0.05$, and $\lambda_n = 0.1$. Each training sequence starts from an empty voxel map, whereas at inference the map accumulates over the entire flight.

A.4 Planning Implementation Details

Scale adaptation and probabilistic mapping. The parameterization of Eq. (5) is scale-equivariant: when a scene is uniformly enlarged, ρ, r, and d_{ret} grow by the same factor. In the probabilistic map, voxels not traversed by any ray receive no update, and never-observed voxels, with $\ell_t(v) = 0$, are unknown. The ratio η of Eq. (8) controls the trade-off between flight distance and residual coverage.

Frontier clustering. Frontier voxels of Eq. (7) are grouped by 26-connectivity, and oversized clusters are recursively split along their principal direction, following FUEL (Zhou et al., 2021).

Candidate generation and routing. For each cluster F_j, candidates are sampled on concentric rings of radius r (Eq. (5)) around the cluster center at multiple heights and retained only when they lie in known free space, satisfy the task-boundary and safety-distance constraints, and reach the visibility threshold of Section 3.3.2. The global layer orders the frontier clusters, and the local layer selects one retained candidate per cluster to minimize the resulting flight distance.

Table 4: FAR26 scenes, ordered by scene scale $\mathrm{vol}(B)^{1/3}$ in meters.

Scene	Type	$\mathrm{vol}(B)^{1/3}$	Scene	Type	$\mathrm{vol}(B)^{1/3}$
Madrid Street	Urban	113	Factory	Industrial	272
Factory Exterior	Industrial	119	Modern Hospital	Urban	310
Industrial Harbour	Industrial	153	Residential Houses	Residential	311
Industrial City	Industrial	159	Airport	Airport	316
Old Factory	Industrial	173	New York City	Urban	343
Modular European Town	Urban	175	Miami Beach	Coastal	394
Park Island	Park	187	Modular Neighborhood	Residential	404
Suburbs City Pack	Residential	200	Seaside Town	Coastal	444
Rome	Urban	211	Train Station	Station	486
Modern City	Urban	237	Japanese City	Urban	562



Figure 7: Equirectangular panoramas rendered in FAR26 scenes, covering city districts, parks, industrial areas, and residential neighborhoods.

Execution and termination. The UAV follows a direct segment when it remains in known free space and respects the maximum step length; otherwise, a 26-connected A* search is performed on the voxel map, where unknown voxels are non-traversable and occupied voxels are inflated by the safety distance. Intermediate nodes are issued as waypoints, and each waypoint triggers a panoramic capture that is integrated in the same way as the capture at p_{t+1}. Acquisition terminates when Eq. (8) is satisfied or when no frontier cluster has a valid reachable candidate.

A.5 Simulation Platform and FAR26 Benchmark

Platform. A standardized planner interface lets external planners, such as EPIC, drive the simulated UAV along their planned paths in real time without modification, and a perception interface streams point clouds to reconstruction modules in real time. The platform thus supports closed-loop evaluation, in which perception, planning, and capture interact online (Section 4.2), and open-loop evaluation of perception modules on recorded flights (Section 4.3). Ground-truth geometry is obtained for any scene by densely sampling the engine-exported mesh.

FAR26. The 20 scenes range in scale $\mathrm{vol}(B)^{1/3}$ from 113 m to 562 m and include city districts, parks, industrial areas, and residential neighborhoods; each is provided with a calibrated task boundary and start frame. The scenes are split into 16 training and 4 test scenes (Appendix A.3).

Table 4 lists the scenes ordered by scene scale, and Figure 7 shows panoramas rendered in several of them.

A.6 Experimental Details

Datasets and sensors. The Macarons++ scenes range from isolated statues to a river-spanning bridge and are deployed in our platform at metric scale. In the planner comparison, EPIC is a LiDAR-based exploration planner for large-scale environments ($50\,\mathrm{m}$ range, $\pm 45^\circ$ vertical band, 40° pitch); MAGICIAN, the state of the art on Macarons++, uses a 456×256 perspective camera with a 60° field of view in its Macarons++ configuration; and FAR uses a 1036×518 ERP panorama with a $130\,\mathrm{m}$ depth range.

Metrics. The 10^5 ground-truth points are sampled uniformly by area on the ground-truth mesh within the task boundary B of each scene, and coverage is normalized by the visible surface ratio of each scene. Since flight distance dominates UAV acquisition cost, AUC is computed over path length rather than over time steps as in MAGICIAN, and a run that ends before the budget keeps its final coverage. For the fused point cloud, Rel measures the relative shape error, Acc and Comp the metric accuracy and completeness, and $F@1\mathrm{m}$ their balance at $1\,\mathrm{m}$. For 3DGS, the ten test panoramas are selected by greedy coverage of the ground-truth surface, and all renders are scored inside the surface mask.

Additional planning analysis. With ground-truth depth, FAR flies a mean path of $1111\,\mathrm{m}$; as Figure 3(a) shows, MAGICIAN repeatedly flies away from the structures, whereas EPIC winds densely over them. Under the FAR planner, PanoVGGT reaches 69.2% coverage and 0.421 AUC and flies a shorter mean path than our model ($822\,\mathrm{m}$ against $1017\,\mathrm{m}$), while DAP reaches 13.8% and 0.078 despite the longest mean path ($2334\,\mathrm{m}$).