2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ActMVS: Active Scene Reconstruction with Monocular Multi-View Stereo

Guo Pu[1*], Yixuan Han[1*], Zhouhui Lian[1†]

***Abstract*— Active scene reconstruction enables robots/UAVs to autonomously plan trajectories and reconstruct environments without costly manual data acquisition. Unlike passive methods, active reconstruction requires real-time construction of high-confidence occupancy maps for collision-free navigation. Existing approaches rely on depth sensors for occupancy map updates, increasing platform cost and weight. To advance spatial intelligence, we aim for a vision-only monocular solution. However, current monocular scene reconstruction methods operate offline and fail to deliver globally consistent dense depth at the frame rates required for robots/UAVs navigation. To bridge this gap, we introduce ActMVS, the first framework for monocular active reconstruction. Our framework integrates a view factor graph construction for informed Multi-View Stereo depth prediction, along with a global depth optimization, to enable the online generation of high-quality, globally consistent dense depth maps. This enables monocular robots/UAVs to maintain reliable occupancy maps for safe trajectory planning during reconstruction. Experiments on Replica datasets demonstrate performance competitive with RGB-D methods. Our code and data are available at `https://github.com/TrickyGo/ActMVS`.**

I. Introduction

Active scene reconstruction is a critical capability for robots/UAVs (Unmanned Aerial Vehicles) to achieve autonomous navigation and reconstruction in unknown environments [1], indispensable for applications like industrial inspection and disaster response, where manual data acquisition is prohibitively costly. While passive reconstruction methods such as SfM [2] and SLAM [3] process manually-collected data, active reconstruction methods dynamically plan sensor trajectories to achieve online scene reconstruction while maximizing scene coverage, thus fundamentally requiring real-time generation of high-confidence occupancy maps to distinguish navigable space from obstacles, ensuring collision-free trajectory planning during exploration [4], [5].

Existing active reconstruction methods [6], [7], [8], [9], [10], [11] predominantly rely on depth sensors (e.g., structured-light or Lidars) to directly acquire accurate depth for efficient construction of volumetric occupancy maps like TSDF [12] and OctoMap [13] for safe spatial reasoning. However, such sensors introduce significant cost, weight, and power constraints that are prohibitive for resource-limited robots/UAVs.

This work was supported by National Natural Science Foundation of China (Grant No.: 62372015), Leading Projects in Key Research Fields of Language Funded by the National Language Commission, and Key Laboratory of Intelligent Press Media Technology.

[1]Wangxuan Institute of Computer Technology, Peking University
*Equal contribution.
†Corresponding author: `lianzhouhui@pku.edu.cn`

To advance spatial intelligence, we aim for a vision-only monocular solution. The fundamental challenge is how to acquire dense depth predictions at frame-rate with metric-scale accuracy for safe spatial reasoning. While monocular depth estimation methods [14], [15] lack metric accuracy, recent Multi-View Stereo (MVS) techniques [16], [17] can estimate high-quality metric depth from well-conditioned image tuples. However, predicting depth with MVS using adjacent frames as a reference is prone to generating sub-optimal results due to insufficient baselines or covisibility. Another issue with MVS is that recent learning-based MVS methods typically process only 8 reference frames, lacking global context, which is a critical limitation since incremental mapping requires globally consistent depth to prevent error accumulation that compromises planning safety and reconstruction quality.

To address these challenges, we introduce ActMVS, the first framework for monocular active scene reconstruction. Inspired by ActiveGS [11], we maintain a voxel map for spatial modeling while leveraging Gaussian splatting for high-fidelity reconstruction. In the absence of depth sensors, we develop two key innovations for reliable metric depth estimation: A view factor graph with voxel-frame visibility modeling for informed MVS depth prediction, and a global depth optimization algorithm enforcing cross-view consistency through depth warping and alignment. This enables ActMVS to generate high-quality, globally consistent dense depth maps online, allowing safe trajectory planning and efficient scene reconstruction.

In summary, our contributions are threefold:

- The first work for monocular active reconstruction, with performance competitive with RGB-D methods.
- A novel view factor graph formulation with voxel-frame visibility modeling for informed MVS depth prediction.
- A novel global depth optimization method leveraging view factor graph to enhance spatial consistency and geometric accuracy.

II. Related Work

A. Scene Reconstruction

Classic 3D scene reconstruction methods [18] such as COLMAP [2] utilize multi-view images to reconstruct scenes through complex pipelines involving keypoint detection [19], image matching [20], camera localization [21], triangulation [22], and bundle adjustment [23], typically representing scenes as sparse point clouds. Classical visual SLAM methods [24], [25], [26] like ORB-SLAM [3]

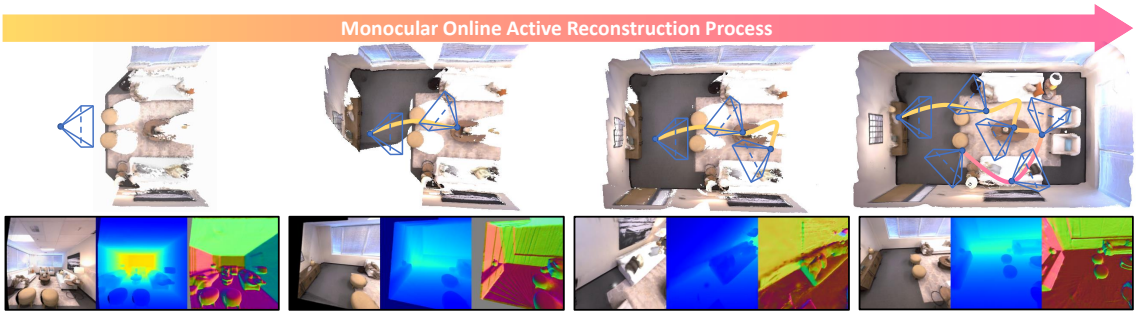

Fig. 1: The ActMVS monocular active reconstruction process which shows the intermediate mesh, camera trajectory, and rendered image with depth and surface normal maps.

adopt similar pipelines with greater emphasis on real-time simultaneous localization and mapping. Recent learning-based approaches have revolutionized scene reconstruction. Neural Radiance Field (NeRF) [27] and 3DGS (3D Gaussian Splatting) [28] utilize differentiable scene representations. Feedforward methods [29], [30], [31], [32], [33] such as VGGT [34] directly predict 3D scenes from images, significantly simplifying reconstruction pipelines. Single-image methods [35], [36], [37], [38] like CAST [39] achieve scene reconstruction from only a single image, greatly enhancing accessibility. These passive reconstruction methods rely on manually captured data.

B. Active Scene Reconstruction

Active reconstruction optimizes sensor trajectories to maximize scene coverage during mapping. Traditional methods [5] focus on enhancing coverage through path planning on voxel maps or meshes, aiming to efficiently navigate through the environment. NeRF-based methods like NARUTO [6] learn uncertainty grids through hybrid neural representations to guide exploration. 3DGS-based approaches, including GS-Planner [7] and HGS-Planner [8], fuse voxel maps with GS-rendered depth estimation. View planning techniques utilize Voronoi diagrams [9] or Fisher information [10] to strategically select optimal viewpoints. ActiveGS [11] represents the state of the art of RGB-D approaches, combining Gaussian splatting with voxel mapping for high-fidelity reconstruction and spatial reasoning. Critically, all existing active reconstruction methods require depth sensors.

C. Depth Estimation

Monocular depth networks [14], [15] predict scale-ambiguous depth from single images but lack metric accuracy. Multi-View Stereo estimates geometrically consistent depth from posed multi-view images using epipolar geometry [40]. Traditional MVS [41] optimizes per-pixel depth hypotheses via photometric consistency, suffering from $O(n^2)$ view-pair complexity. Learning-based MVS methods [42], [43], [16], [17] enable dense depth prediction and enhance robustness through learnable cost volume aggregation. However, these MVS methods typically process around only 8 reference frames, lacking global context. We adapt the feedforward MVS method MVSA [17] for initial dense depth prediction, and further enhance it with a global depth optimization to achieve improved global depth consistency.

III. Method

A. Overview

Inspired by ActiveGS [11], we leverage a voxel map for spatial representation and a Gaussian-splatting map for high-fidelity surface reconstruction. In the absence of a depth sensor, we introduce a view factor graph incorporating voxel-frame visibility modeling to guide Multi-View Stereo depth prediction during online map updates.

Operating as an online monocular active reconstruction framework, our method sequentially processes image-pose pairs ($\mathbf{I}_t$, $\mathbf{P}_t$) from $t = 0$ to $t = T$, where the camera pose $\mathbf{P}_t$ is obtained via simulator measurements during simulation, and via odometry from robots/UAVs during real-world testing, respectively. The algorithm runs until the preset time budget T is reached, and T can be adjusted based on scene size. At each timestep, we generate a reconstructed 3D scene representation comprising the view factor graph $\mathcal{G}_f$, voxel map $\mathcal{M}_v$, and Gaussian map $\mathcal{M}_g$. The overview of our pipeline is shown in Fig. 2.

B. Hybrid Scene Representations

We employ a View Factor Graph $\mathcal{G}_f = (\mathcal{V}, \mathcal{E})$ to model inter-frame geometric relationships essential for MVS prediction. Nodes $\mathbf{v}_t \in \mathcal{V}$ store view images, optimized depth maps, camera poses and camera intrinsics $\{\mathbf{I}_t, \mathbf{D}_t, \mathbf{P}_t, \mathbf{K}_t\}$, while edges $\mathcal{E}$ encode voxel-frame visibility weights indicating geometric overlap.

Following [4], [5], our Voxel Map $\mathcal{M}_v$ builds upon OctoMap [13]. Each voxel stores an occupancy probability $p \in [0, 1]$ and updates incrementally using depth-derived point clouds. This representation serves as both an occupancy map for collision-free planning and a visibility model for co-visible surfaces.

We adopt Gaussian surfels [44] to construct the Gaussian Splatting Map $\mathcal{M}_g$ for high-fidelity surface modeling. This differentiable representation employs alpha-blending rendering to generate color $\mathbf{I_g}$, depth $\mathbf{D_g}$, normal $\mathbf{N_g}$, and opacity $\mathbf{O_g}$ maps.

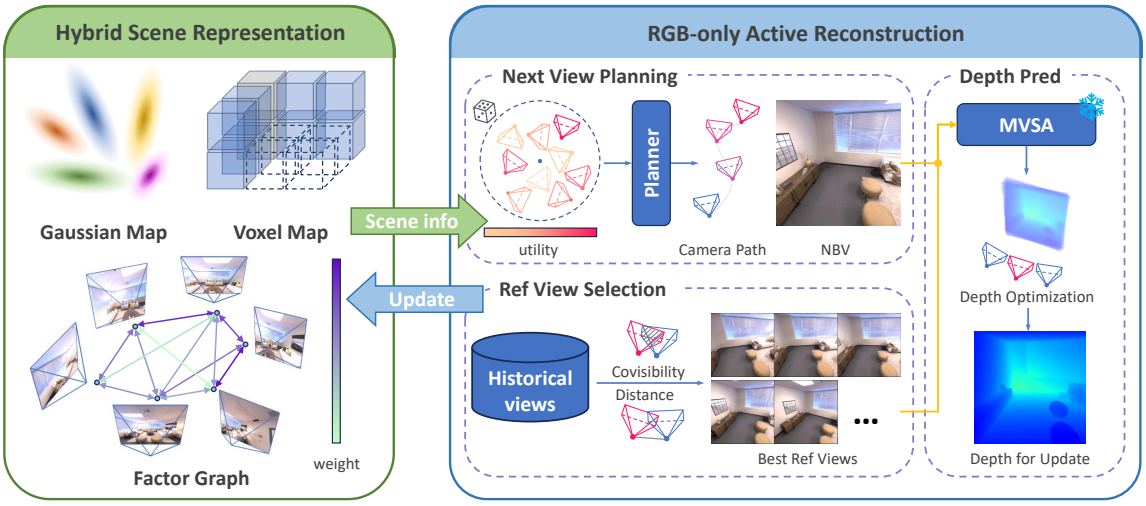

Fig. 2: ActMVS actively reconstructs environments through iterative view planning and incremental mapping. At each timestep t, the system accepts posed RGB images ($\mathbf{I}_t$, $\mathbf{P}_t$) as input and outputs a dense depth map $\mathbf{D}_t$ while concurrently updating the View Factor Graph $\mathcal{G}_f$, Voxel Map $\mathcal{M}_v$, and Gaussian Splatting Map $\mathcal{M}_g$.

C. Online Active Reconstruction Process

During initialization, k viewpoints featuring uniform baselines are sampled near the origin to initialize the scene representation. Subsequently, the pipeline iterates between the planning and mapping stages.

1) Planning Phase: We determine the next camera extrinsics $\mathbf{P}$ by selecting the Next-Best-View (NBV) based on a weighted score combining view completeness and travel distance.

First, candidate viewpoint positions $\mathbf{p}$ are sampled outside the safety radius ($\|\mathbf{p}\| > R_{\min}$), the exploration boundary ($\|\mathbf{p}\| < R_{\max}$), and traversable regions ($\mathcal{M}_v^{(\text{free})}$).

For each candidate, we calculate the view completeness using the number of visible unexplored voxels, identified by comparing Gaussian-rendered depth maps against the projections of voxel centroids. The travel distance d_{travel} to the candidate is computed via an A* path search [45] over the voxel map $\mathcal{M}_v$.

Finally, the candidate viewpoint with the highest weighted score is chosen as the NBV. Instead of navigating directly to the NBV, we interpolate collision-free trajectories along the computed A* path. This ensures sufficient covisibility for subsequent MVS depth prediction.

2) Mapping Phase: First, visibility modeling computes pairwise overlaps Ω_{tj} via voxel covisibility. The graph $\mathcal{G}_f$ updates by inserting $\{\mathbf{I}_t, \mathbf{P}_t\}$ and connecting to the top-k neighbors with edge weights Ω_{tj}.

Subsequently, we adapt MVSA [17] to estimate dense depth $\widehat{\mathbf{D}}_t$ with selected neighbors as reference views, followed by N_{iter} steps of global depth optimization optimizes over $\mathbf{D}_t$ through pairwise depth warping in $\mathcal{G}_f$.

Finally, we update $\mathcal{M}_v$ via probabilistic occupancy integration. For $\mathcal{M}_g$, we perform Gaussian primitive initialization via depth point-cloud projection in low-opacity areas where rendered opacity satisfies $\mathbf{O}_\mathbf{g} < 0.1$, following $\mathcal{M}_g$ optimization by minimizing photometric and depth $\mathcal{L}_1$ losses against all $\{\mathbf{I}_t, \mathbf{D}_t\}$ in $\mathcal{G}_f$.

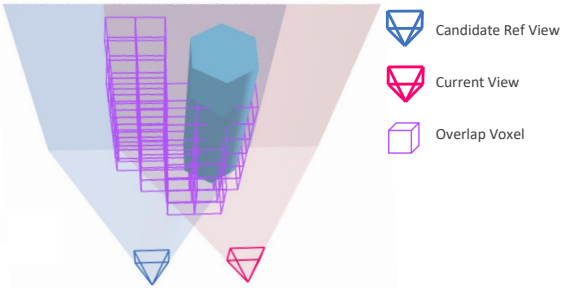

Fig. 3: Visualization of voxels in the overlapped region of the current view and candidate reference view. Obstructed voxels are excluded to prevent interference caused by the area invisible to either of the two viewpoints.

D. Voxel-Frame Visibility Modeling

MVSA predicts 3D-consistent depth by effectively incorporating multi-view geometry constraint and monocular feature from multiple posed images. Therefore, evaluating the covisibility between viewpoints during active reconstruction process is crucial for selecting reference frames that provide effective information, ultimately achieving more accurate depth prediction. A naive solution is to rely purely on similarity in camera position and orientation. However, this method does not account for occlusions. Even viewpoints with similar camera positions and orientations may fail to provide sufficient RGB overlap due to obstructions nearby. Another approach involves feature extraction and matching to find the suitable viewpoints directly in the feature space. Nevertheless, this method lacks robustness in regions with similar textures and introduces significant computational overhead. Based on the geometric priors contained in the gaussian map and the efficient scene representation provided by our voxel map, we propose a frame visibility assessment method that considers occlusions while ensuring high computational efficiency, as shown in Fig. 3.

To be specific, we compute per-frame frustum masks M_t

that identify voxels visible from each viewpoint in 3 steps. (1) Render gaussian depth using current camera extrinsic and intrinsic matrix (P_t, K_t) to obtain depth map D_t. (2) Project voxels onto the camera plane, obtaining UV coordinates (u_k, v_k) and depth values d_k of each voxel center ϑ_k. (3) Select voxels on the camera plane with depth values that lie within corresponding rendered gaussian depth. The whole process can be formulated as:

$$(u_k, v_k, d_k) = \psi_{P_t,K_t}(\vartheta_k), \vartheta_k \in \mathcal{M}_v, \tag{1}$$

$$M_t = \{k | d_k \in (0, D_t[u_k, v_k]) \wedge u_k \in [0, w) \wedge v_k \in [0, h)\}, \tag{2}$$

where ψ_{K_t,P_t} denotes the rendering process using camera extrinsic and intrinsic (P_t, K_t).

This establishes the fundamental voxel-to-frame relationship that underpins our spatial reasoning and provides efficient overlap calculation by a single Logical AND operation.

Inter-Frame Relation Modeling: Without ground truth depth, ensuring geometrically consistent depth predictions across multiple views is particularly crucial for accurate geometry reconstruction. MVSA incorporates the camera pose information of the reference frame into the cost volume, providing the model with implicit cues about scale derived from camera translation and rotation. To enhance the 3D consistency of depth predictions, we calculate covisibility between source and reference frames as:

$$\Omega_{ij} = |M_i \cap M_j| = \sum_{k=1}^{|\mathcal{M}_v|} \mathbb{I}[M_i^k = 1 \wedge M_j^k = 1], \tag{3}$$

where $\mathbb{I}$ is the indicator function.

Then, we quantify the correlation score between frame i and frame j using the following formula balancing both covisibility and scale information contained in camera translation:

$$\begin{cases} \lambda_1 \Omega_{ij} + \lambda_2 \|\mathbf{t}_i - \mathbf{t}_j\|_2, & \Omega_{ij} \geq \epsilon \\ 0, & else \end{cases}, \tag{4}$$

where ϵ is a threshold set to 50, discarding viewpoints with no frustum overlap.

Although MVSA performs camera metadata normalization to maintain scene scale-agnostic properties, our experiments reveal that selecting reference frames with poses too close to the input frame can magnify discrepancies in depth predictions. Conversely, reference frames with excessively distant camera poses introduce features with poor similarity and degrade depth accuracy. To address this, we impose a spatial-angular constraint on camera translation $\delta_{\min} \leq \|\mathbf{t}_i - \mathbf{t}_j\|_2 \leq \delta_{\max}$ and orientation $\theta_{\min} \leq \arccos(\mathbf{n}_i \cdot \mathbf{n}_j) \leq \theta_{\max}$, where $\mathbf{t}$ denotes the camera position and $\mathbf{n}$ is the optical axis direction. This ensures that the frames have meaningful baselines for triangulation.

E. Globally Optimized Depth

Even with our searched well-conditioned reference frames, MVSA typically processes only 8 reference frames for prediction, lacking global context, which is prone to error accumulation that compromises planning safety and reconstruction quality. To address this issue, we propose a depth optimization algorithm based on our view factor graph $\mathcal{G}_f$ to achieve globally consistent depth estimates. We optimize initial depth predictions by applying pixel-level multi-view constraints to adjacent frames within $\mathcal{G}_f$. This approach enforces 3D consistency in co-visible regions via depth warping alignment, as illustrated in Fig. 4.

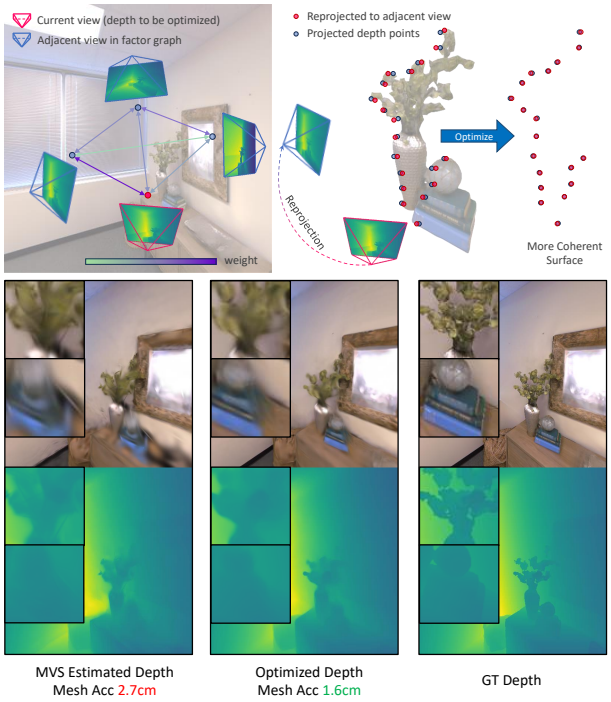

Fig. 4: Our global depth optimization over view factor graph enforces 3D consistency in co-visible regions through depth warping alignment, resulting in enhanced mesh quality.

1) Depth Consistency Factor: The core constraint $\mathcal{F}_{\text{depth}}$ enforces geometric consistency between frame pairs $(i, j) \in \mathcal{E}$ of the view factor graph $\mathcal{G}_f$. For a pair (i, j) with corresponding nodes $\mathbf{v}_i = \{\mathbf{I}_i, \mathbf{D}_i, \mathbf{P}_i, \mathbf{K}_i\}$ and $\mathbf{v}_j = \{\mathbf{I}_j, \mathbf{D}_j, \mathbf{P}_j, \mathbf{K}_j\}$, we aim to align their co-visible 3D structures. This is achieved by backprojecting depth estimates to 3D, warping points between views, and imposing alignment constraints on valid correspondences. Specifically, the depth consistency factor is computed as follows. For a pixel $\mathbf{p}_i = (u_i, v_i)$ in $\mathbf{v}_i$, its 3D position in the camera coordinate system is:

$$\mathbf{X}_i^{\text{c}}(\mathbf{p}_i) = \mathbf{K}_i^{-1} \begin{bmatrix} u_i \cdot z_i \\ v_i \cdot z_i \\ z_i \end{bmatrix}, \quad \text{where} \quad z_i = \mathbf{D}_i(\mathbf{p}_i) \tag{5}$$

with $\mathbf{K}_i$ denoting the camera intrinsic matrix. This point transforms to the frame j's camera coordinate system by:

$$\mathbf{X}_j^{\text{c}} = \mathbf{P}_j \mathbf{P}_i^{-1} \mathbf{X}_i^{\text{c}} \tag{6}$$

where $\mathbf{P}$ denotes world-to-camera transformation matrices.

The projective correspondence $\mathbf{p}_j = (u_j, v_j)$ in frame j

is given by perspective projection:

$$\mathbf{p}_j = \pi\left(\mathbf{K}_j \mathbf{X}_j^{\mathrm{c}}\right) = \left(\frac{\mathbf{K}_{j(1,:)}\mathbf{X}_j^{\mathrm{c}}}{\mathbf{K}_{j(3,:)}\mathbf{X}_j^{\mathrm{c}}}, \frac{\mathbf{K}_{j(2,:)}\mathbf{X}_j^{\mathrm{c}}}{\mathbf{K}_{j(3,:)}\mathbf{X}_j^{\mathrm{c}}}\right), \quad (7)$$

where $\mathbf{K}_{j(k,:)}$ denotes the k-th row of $\mathbf{K}_j$.

For robustness against outliers, we formulate the depth consistency factor using a log-space Huber loss:

$$\mathcal{F}_{\text{depth}}^{i,j} = \sum_{\mathbf{p}_i \in \mathbf{X}_i^{\mathrm{c}}} \rho_\delta\left(\log \mathbf{D}_j(\mathbf{p}_j) - \log z_j'\right) \cdot \mathbf{M}_{i,j}(\mathbf{p}_i) \quad (8)$$

where $z_j' = \mathbf{X}_j^{\mathrm{c}}(z)$ is the transformed depth value. ρ_δ denotes the Huber norm (δ specifies the transition point). $\mathbf{M}_{i,j}$ is a binary validity mask excluding occlusions $z_j' > \mathbf{D}_j(\mathbf{p}_j) + \tau_o$, boundary violations $\mathbf{p}_j \notin [1, W-1] \times [1, H-1]$ and invalid depths $\mathbf{D}_j(\mathbf{p}_j) \le \tau_d \vee z_j' \le \tau_d$, where τ_o and τ_d are predefined tolerance thresholds.

2) Regularization Factor: As depth warping yields sparse pixel-level depth map supervision, we incorporate total variation regularization to ensure smoothness and geometric plausibility:

$$\mathcal{F}_{\text{reg-TV}}^{i} = \lambda_{\text{TV}} \sum_{\mathbf{p}_i} \left(\|\nabla_u \mathbf{D}_i\|_\gamma + \|\nabla_v \mathbf{D}_i\|_\gamma\right), \quad (9)$$

where $\|\cdot\|_\gamma$ denotes the Charbonnier penalty $\|\cdot\|_\gamma = \sqrt{(x)^2 + \gamma^2}$, ∇_u and ∇_v represent horizontal/vertical gradient operators. γ controls edge sensitivity.

3) Optimization: The complete optimization minimizes:

$$\min_{\{\mathbf{D}_i\}_{i\in\mathcal{K}}} \left[\sum_{(i,j)\in\mathcal{E}} \mathcal{F}_{\text{depth}}^{i,j} + \sum_{i\in\mathcal{K}} \mathcal{F}_{\text{reg-TV}}^{i}\right], \quad (10)$$

where $\mathcal{K}$ denotes keyframes selected from $\mathcal{V}$ based on spatial distribution and visual covisibility.

We solve this non-linear optimization via iterative back-propagation using the Adam optimizer. Sequential per-frame optimization across $\mathcal{G}_f$ ensures global consistency through incremental geometric constraint propagation while maintaining computational efficiency.

IV. Experiments

A. Evaluation

1) Implementation Details: The occupancy grid $\mathcal{M}_v$ employs 20 cm^3 voxels. Exploration radius bounds are $R_{\min} = 0.3$ m and $R_{\max} = 0.5$ m. The NBV score weights view completeness ($w_{\text{comp}} = 1.0$) against travel distance ($w_{\text{dist}} = 0.5$). Depth estimation employs MVSA [17] at 512×512 resolution. Spatial-angular constraints enforce baseline thresholds $\delta_{\min} = 0.1$ m, $\delta_{\max} = 0.5$ m, and angular bounds $\theta_{\min} = 0°$, $\theta_{\max} = 25°$. Correlation parameters $\lambda_1 = 1.0$ and $\lambda_2 = 0.1$ prioritize overlap percentage over translation, with voxel overlap threshold $\epsilon = 50$ preventing degenerate pairs. Each input frame selects the top-16 reference frames by correlation scores. Each keyframe undergoes $N_{\text{iter}} = 20$ global depth optimization steps with Huber norm $\delta = 0.1$, Charbonnier penalty $\gamma = 0.001$, TV regularization $\lambda_{\text{TV}} = 0.1$, occlusion detection threshold $\tau_o = 0.1$ m, and depth validity threshold $\tau_d = 0.01$ m. Gaussian map $\mathcal{M}_g$ undergoes 10 optimization iterations per step with batch size 8 over randomly selected views. We conduct evaluation experiments of all methods on an NVIDIA L40 48GB GPU with Intel Xeon Platinum 8370C 2.80GHz CPU.

Fig. 5: Reconstruction results of UAV deployment in Airsim simulation.

2) Baselines and Protocols: Due to the absence of prior monocular methods for active reconstruction, we evaluate against two state-of-the-art RGB-D baselines ActiveGS and NARUTO. Experiments utilize the Habitat simulator [46] and Replica dataset [47] with 90° field of view and 512×512 resolution. Rendering quality is evaluated using standard metrics PSNR ↑, SSIM ↑, LPIPS ↓. Mesh reconstruction performance is assessed via Accuracy ↓ (cm), Completion ↓ (cm), and Chamfer distance ↓ (cm).

3) Quantitative Evaluation: Quantitative results across eight Replica scenes in Table I categorize methods by depth-sensor dependency (RGB vs. RGB-D). The RGB group consists of our method and relevant ablation studies. Rendering results are computed from 1000 novel viewpoints randomly sampled per scene. On average, our approach achieves the best performance within the RGB group across both rendering and mesh quality metrics while closely approaching RGB-D baselines. These results validate that our multi-view constrained depth optimization effectively compensates for the absence of depth sensors.

4) Qualitative Evaluation: Visual comparisons in Fig. 6 showcase the camera path and the reconstructed mesh of ours, along with ground truth RGB-D and rendered RGB-D of ActiveGS, NARUTO, our full method, and ablation variants from randomly sampled novel viewpoints. While ActiveGS exhibits the closest visual resemblance to ground truth, our method produces closely high-quality outputs. NARUTO's NeRF-based renderings suffer significant degradation in novel views, due to the sparse viewpoints inherent in fast UAV motion trajectories.

5) Discussions: ActiveGS achieves the fastest convergence with an average runtime of 300 seconds. Due to NeRF's slower performance, NARUTO requires 1800 sec-

TABLE I: Quantitative evaluation on Replica scenes. Methods are grouped by input modality (RGBD vs. RGB).

Method	Scene											
	Room0						Room1					
	Render			Mesh			Render			Mesh		
RGBD												
ActiveGS	**28.14**	**0.868**	**0.199**	**1.081**	**1.278**	**1.180**	**29.55**	**0.881**	**0.178**	**0.902**	**1.063**	**0.983**
Naruto	25.51	0.769	0.380	1.907	1.619	1.763	27.23	0.792	0.372	1.642	1.296	1.469
RGB												
ActMVS	26.32	0.839	**0.263**	**2.211**	**2.583**	**2.397**	**28.05**	**0.857**	**0.242**	**1.654**	2.207	**1.930**
w/o OPT	**27.17**	**0.843**	0.266	5.066	2.608	3.837	26.24	0.830	0.258	2.705	**2.081**	2.393
w/o REF	23.35	0.792	0.314	4.387	4.758	4.573	24.06	0.792	0.329	32.12	4.421	18.271
w/o MVS	11.47	0.521	0.723	43.66	48.79	46.225	12.64	0.561	0.702	38.04	41.21	39.625
	Room2						Office0					
	Render			Mesh			Render			Mesh		
RGBD												
ActiveGS	**30.01**	**0.906**	**0.154**	**0.940**	**0.970**	**0.955**	**34.62**	**0.950**	**0.089**	**0.896**	**1.004**	**0.950**
Naruto	27.50	0.827	0.322	1.553	1.543	1.548	30.55	0.877	0.304	1.489	1.350	1.420
RGB												
ActMVS	27.33	0.875	0.217	**1.854**	**1.901**	**1.878**	**31.63**	**0.927**	**0.148**	**2.111**	2.343	**2.227**
w/o OPT	26.90	0.864	0.230	2.359	2.267	2.313	31.62	0.925	0.168	8.411	**2.000**	5.206
w/o REF	**28.69**	**0.891**	**0.204**	2.086	2.116	2.101	28.01	0.890	0.224	5.036	3.914	4.475
w/o MVS	10.60	0.465	0.703	45.98	52.31	49.145	15.15	0.511	0.631	40.17	43.48	41.825
	Office1						Office2					
	Render			Mesh			Render			Mesh		
RGBD												
ActiveGS	**34.15**	**0.949**	**0.104**	**0.752**	**0.876**	**0.814**	**30.10**	**0.915**	**0.135**	**1.059**	**1.143**	**1.101**
Naruto	30.21	0.885	0.260	1.188	1.082	1.135	25.17	0.807	0.334	2.618	2.409	2.514
RGB												
ActMVS	**31.96**	**0.921**	**0.147**	**2.227**	4.590	**3.409**	**26.82**	**0.884**	**0.197**	**3.246**	**3.347**	**3.297**
w/o OPT	27.67	0.873	0.216	35.97	**4.352**	20.161	22.54	0.798	0.308	6.787	4.213	5.500
w/o REF	28.93	0.894	0.190	5.106	5.948	5.527	18.94	0.777	0.336	7.909	8.745	8.327
w/o MVS	16.46	0.423	0.498	38.38	54.56	46.47	10.10	0.526	0.675	41.61	39.77	40.69
	Office3						Office4					
	Render			Mesh			Render			Mesh		
RGBD												
ActiveGS	**29.91**	**0.910**	**0.146**	**1.201**	**1.578**	**1.390**	**31.75**	**0.918**	**0.145**	**1.166**	**1.337**	**1.252**
Naruto	24.93	0.793	0.369	2.706	2.255	2.481	27.60	0.844	0.330	2.148	1.747	1.948
RGB												
ActMVS	**26.18**	**0.869**	**0.222**	**2.880**	3.751	**3.316**	29.75	0.896	0.201	2.382	3.189	2.786
w/o OPT	24.06	0.833	0.290	38.31	**3.258**	20.784	**29.99**	**0.903**	**0.199**	**2.247**	**2.336**	**2.292**
w/o REF	22.69	0.824	0.310	10.66	5.612	8.136	28.87	0.889	0.208	3.293	4.105	3.699
w/o MVS	11.00	0.529	0.675	38.90	45.51	42.205	13.06	0.589	0.612	42.94	41.21	42.075

onds on average. Compared to ActiveGS, our method necessitates denser camera trajectories and additional computation for view factor graph construction, depth prediction and optimization, resulting in an average runtime of 1200 seconds. We recommend checking our supplementary video for detailed camera paths, as well as depth prediction results and Gaussian update process during the reconstruction.

B. Ablation Studies

Ablation studies are quantified in Table I and visualized in Fig. 6, validating critical component contributions.

1) w/o OPT: Removing global depth optimization causes depth instability and surface degradation. This significantly reduces mesh quality, demonstrating that global optimization enforces consistency across viewpoints and stabilizes reconstructed geometry.

2) w/o REF: Replacing our voxel-based reference frame search with simple adjacent-frame selection reduces the quality and stability of MVS depth predictions. The suboptimal baseline lengths between frames often cause MVS failures due to insufficient parallax for reliable scale recovery.

3) w/o MVS: We replace the MVS prediction in our method with the state-of-the-art monocular metric depth estimator DepthAnythingV2 [15]. Sole reliance on DepthAnythingV2 metric depth estimation model causes catastrophic reconstruction failures due to significant error accumulation. This validates that integrating Multi-View Stereo is essential for achieving plausible geometry under strong perspective changes.

C. Extended Simulations

Real-world applicability is demonstrated through AirSim simulations featuring quadrotor UAV navigation in more realistic and complex environments compared to Habitat. However, the large amount of video memory required by 3DGS-based reconstruction exceeds the capacity of current onboard devices. To approximate real-world deployment, we run ActMVS on a server equipped with a single NVIDIA L40 GPU and use Flask for the network transmission of waypoint and camera image data. Instead of relying on controller APIs provided by Airsim, we opted for the widely used open-source flight control system PX4 with Robot Operating System (ROS) and MAVLink for sensor data processing and transmission, simulating a more authentic operating environment for real-world drones. In Airsim simulation, we utilize Ego-planner [48] to execute the camera path generated by ActMVS, ensuring compliance with the dynamical constraints of a real drone. Reconstruction results in Fig. 5 demonstrate the effectiveness of ActMVS in various scenarios, where the drone did not encounter any collisions during the experiment, further validating the feasibility of the ActMVS path planning.

V. Conclusion

We present the first monocular scene reconstruction framework that enables robust autonomous aerial reconstruction without depth sensors, by integrating a view factor graph with voxel-frame visibility modeling and a global depth optimization mechanism that leverages the connectivity within this graph to enforce cross-view consistency. Experiment results demonstrate competitive performance against RGB-D baselines, closely approaching them in both rendering fidelity and geometric accuracy. Future work will focus on real-time deployment optimization for fully autonomous drone operations under physical constraints.

References

[1] S. Chen, Y. Li, and N. M. Kwok, "Active vision in robotic systems: A survey of recent developments," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1343–1377, 2011.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[4] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, "An information gain formulation for active volumetric 3d reconstruction," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3477–3484.

[5] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon" next-best-view" planner for 3d exploration," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1462–1468.

[6] Z. Feng, H. Zhan, Z. Chen, Q. Yan, X. Xu, C. Cai, B. Li, Q. Zhu, and Y. Xu, "Naruto: Neural active reconstruction from uncertain target observations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 572–21 583.

[7] R. Jin, Y. Gao, Y. Wang, Y. Wu, H. Lu, C. Xu, and F. Gao, "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)*. IEEE, 2024, pp. 11 202–11 209.

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

[9] Y. Li, Z. Kuang, T. Li, Q. Hao, Z. Yan, G. Zhou, and S. Zhang, "Activesplat: High-fidelity scene reconstruction through active gaussian splatting," *IEEE Robotics and Automation Letters*, 2025.

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

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

[12] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 303–312.

[13] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous robots*, vol. 34, no. 3, pp. 189–206, 2013.

[14] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[15] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.

[16] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman, and C. Godard, "Simplerecon: 3d reconstruction without 3d convolutions," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–19.

[17] S. Izquierdo, M. Sayed, M. Firman, G. Garcia-Hernando, D. Turmukhambetov, J. Civera, O. Mac Aodha, G. Brostow, and J. Watson, "Mvsanywhere: Zero-shot multi-view stereo," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 493–11 504.

[18] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "Openmvg: Open multiple view geometry," in *International Workshop on Reproducible Research in Pattern Recognition*. Springer, 2016, pp. 60–74.

[19] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[20] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[21] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6684–6692.

[22] R. I. Hartley and P. Sturm, "Triangulation," *Computer vision and image understanding*, vol. 68, no. 2, pp. 146–157, 1997.

[23] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, "Bundle adjustment in the large," in *European conference on computer vision*. Springer, 2010, pp. 29–42.

[24] H. Zhou, D. Zou, L. Pei, R. Ying, P. Liu, and W. Yu, "Structslam: Visual slam with building structure lines," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1364–1375, 2015.

[25] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[26] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE transactions on robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[28] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[29] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[30] Y. Cabon, L. Stoffl, L. Antsfeld, G. Csurka, B. Chidlovskii, J. Revaud, and V. Leroy, "Must3r: Multi-view network for stereo 3d reconstruction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1050–1060.

[31] J. Yang, A. Sax, K. J. Liang, M. Henaff, H. Tang, A. Cao, J. Chai, F. Meier, and M. Feiszli, "Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 924–21 935.

[32] B. Duisterhof, L. Zust, P. Weinzaepfel, V. Leroy, Y. Cabon, and J. Revaud, "Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion," *arXiv preprint arXiv:2409.19152*, 2024.

[33] J. Wang, N. Karaev, C. Rupprecht, and D. Novotny, "Vggsfm: Visual geometry grounded deep structure from motion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 686–21 697.

[34] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[35] G. Pu, P.-S. Wang, and Z. Lian, "Sinmpi: Novel view synthesis from a single image with expanded multiplane images," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10.

[36] H. Liang, J. Cao, V. Goel, G. Qian, S. Korolev, D. Terzopoulos, K. N. Plataniotis, S. Tulyakov, and J. Ren, "Wonderland: Navigating 3d scenes from a single image," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 798–810.

[37] G. Pu, Y. Zhao, and Z. Lian, "Pano2room: Novel view synthesis from a single indoor panorama," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[38] Z. Huang, Y.-C. Guo, X. An, Y. Yang, Y. Li, Z.-X. Zou, D. Liang, X. Liu, Y.-P. Cao, and L. Sheng, "Midi: Multi-instance diffusion for single image to 3d scene generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 646–23 657.

[39] K. Yao, L. Zhang, X. Yan, Y. Zeng, Q. Zhang, W. Yang, L. Xu, J. Gu, and J. Yu, "Cast: Component-aligned 3d scene reconstruction from an rgb image," *arXiv preprint arXiv:2502.12894*, 2025.

[40] F. Wang, Q. Zhu, D. Chang, Q. Gao, J. Han, T. Zhang, R. Hartley, and M. Pollefeys, "Learning-based multi-view stereo: A survey," *arXiv preprint arXiv:2408.15235*, 2024.

[41] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 519–528.

[42] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[43] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "Patchmatchnet: Learned multi-view patchmatch stereo," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 194–14 203.

[44] P. Dai, J. Xu, W. Xie, X. Liu, H. Wang, and W. Xu, "High-quality surface reconstruction using gaussian surfels," in *ACM SIGGRAPH 2024 conference papers*, 2024, pp. 1–11.

[45] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

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

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

[48] X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao, "Ego-planner: An esdf-free gradient-based local planner for quadrotors," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 478–485, 2020.

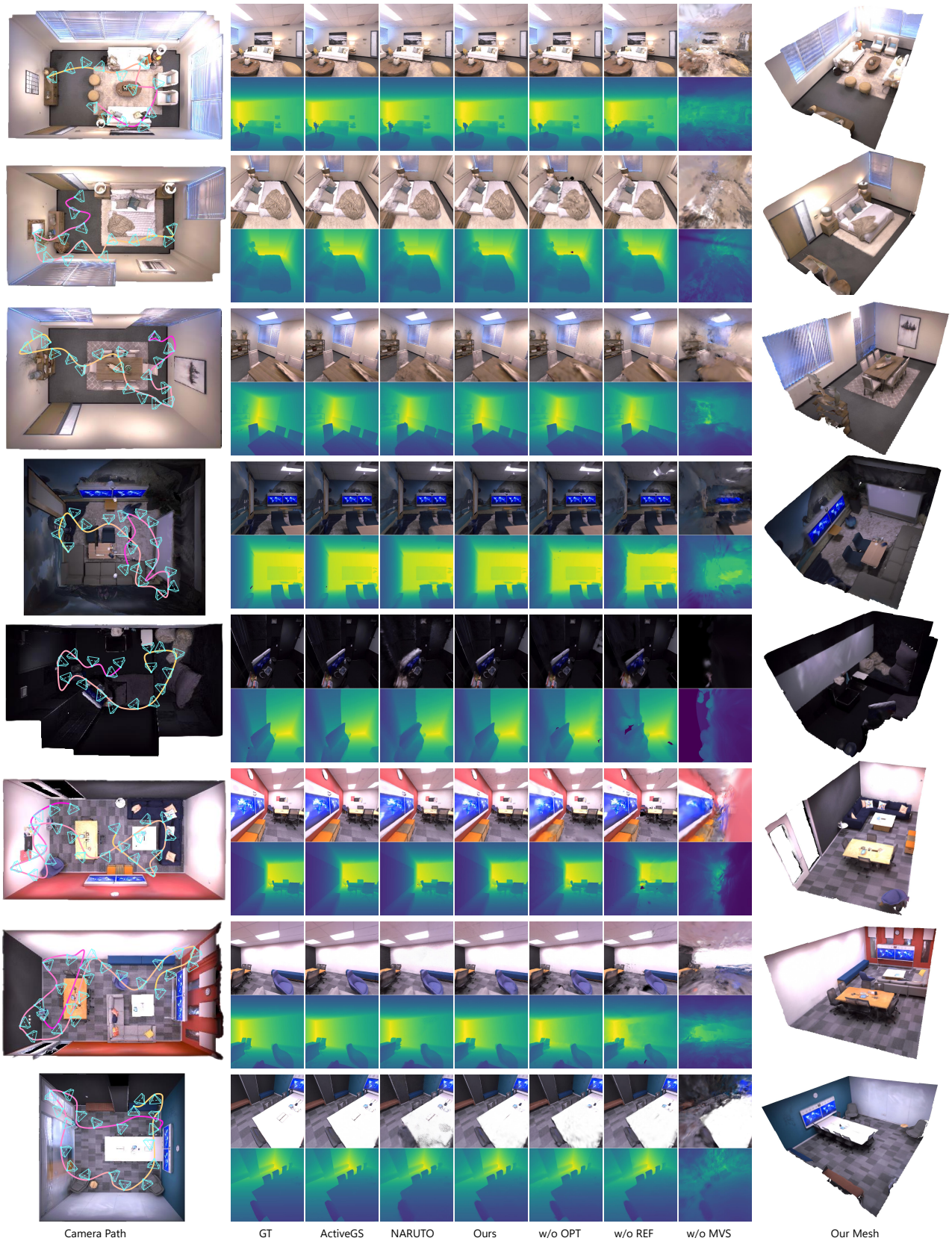

Fig. 6: Qualitative comparison on Replica scenes showcase the camera path and the reconstructed mesh of ours, along with ground truth RGB-D and rendered RGB-D of ActiveGS, NARUTO, our full method, and ablation variants from randomly sampled novel viewpoints. While ActiveGS exhibits the closest visual resemblance to ground truth, our method produces closely high-quality outputs. For better visualization of the 3D scene in all figures in this paper, mesh faces blocking the room interior are deleted. Please zoom in for detailed inspections.