ALONG AND ACROSS THE RAY: POST-OPTIMIZATION FOR FEED-FORWARD RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D reconstruction models typically process long videos in windows and align their predictions across windows, yet reconstructions of the same surface can still form offset layers and produce surface ghosting. We show that these errors decompose into transverse and radial components in viewing-ray coordinates, where the transverse component is determined by the cameras and ray directions, while with fixed cameras, projection-preserving corrections can only adjust depth along the rays. Based on this structure, we propose RayCon, a training-free global-to-local post-optimization framework. Globally, RayCon exploits the dependence of reconstruction errors on window boundaries through staggered dual-window inference and fuses two trajectories with complementary errors. Locally, with the cameras fixed, Temporal–Range Consistency Filtering removes inconsistent predictions, while Surface-Consensus Refinement infers surface membership from multi-view consistency and refines depths along the viewing rays. Without retraining the backbone, RayCon reduces mean ATE on KITTI from 31.91 m to 9.59 m, achieves the lowest mean ATE among the compared methods on KITTI and Oxford Spires, and reduces local Chamfer distance.

1 INTRODUCTION

Feed-forward reconstruction models recover dense scene structure, camera poses, and intrinsics from images in a single forward pass. DUSt3R regresses pairwise pointmaps (Wang et al., 2024b), VGGT jointly predicts cameras, depth, and pointmaps from multiple views (Wang et al., 2025a), and recurrent and streaming variants extend this paradigm to video (Wang et al., 2025c; Cheng et al., 2026a; Chen et al., 2026). Unlike classical structure-from-motion and multi-view stereo (Schönberger & Frahm, 2016; Schönberger et al., 2016), these models replace explicit correspondence search and iterative optimization with feed-forward prediction.

Long videos, however, remain difficult. Joint inference requires memory that grows with sequence length, while streaming models propagate a bounded state whose errors can accumulate over time (Cheng et al., 2026a;b). Large scenes are therefore typically reconstructed in overlapping windows and refined globally. Existing methods register adjacent windows with similarity transforms and loop closure (Deng et al., 2026), optimize projective transformations between submaps (Maggio et al., 2025), or jointly refine trajectories and pointmaps in a backend (Fang et al., 2026).

A common remaining failure is surface ghosting. When several windows observe the same physical surface, their predictions often form offset layers, producing duplicated or thickened geometry after fusion (Figure 1). Averaging nearby points may merge distinct surfaces, while removing inconsistent points can discard valid geometry. Resolving these ghosts therefore requires determining which observations belong to the same physical surface.

We observe that this error is strongly anisotropic: predictions of the same surface disagree much more along their viewing rays than across them. A point is unprojected as $x = c + \rho q$, where c is the camera center, q the viewing direction, and ρ the predicted range. Its error naturally decomposes into radial and transverse components. The transverse component depends only on the camera center and ray direction. Once the cameras are fixed, only ρ can change while preserving the original image observation.

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.

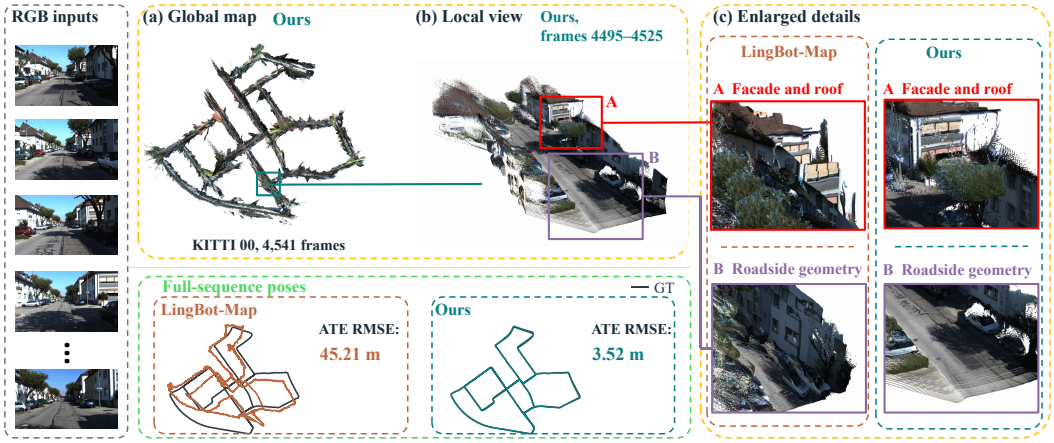

Figure 1: **Global-to-local reconstruction on KITTI 00 (4,541 frames).** From left to right, the figure shows sampled inputs, the complete map, a local oblique view, and enlarged facade and roadside regions comparing LingBot-Map with RayCon (Ours). The bottom row compares full trajectories with ground truth after one Sim(3) alignment per method, where ATE decreases from 45.21 m for LingBot-Map to 3.52 m for RayCon.

Independent windows also retain projective inconsistency. Each window predicts its own cameras and intrinsics and thus forms a local uncalibrated reconstruction, which is determined only up to a 15-degree-of-freedom projective transformation (Hartley & Zisserman, 2004). This ambiguity has also been identified between feed-forward submaps (Maggio et al., 2025). Similarity registration removes only seven degrees of freedom, leaving residual distortion that can preserve within-window reprojections while changing the recovered depths of shared surfaces. This residual appears as cross-window radial disagreement and produces layered geometry after fusion.

Based on this decomposition, we propose **RayCon**, a training-free global-to-local post-optimization framework. Its global stage exploits a simple property of windowed reconstruction: the window partition determines where inter-window alignment errors enter the trajectory. We therefore reconstruct the same video under two staggered window partitions, so that boundary frames in one run become interior frames in the other. The two runs produce complementary camera errors, which we fuse after in-window trajectory smoothing to improve the camera centers and reduce transverse inconsistency.

With the cameras fixed, RayCon then addresses the remaining radial inconsistency. Temporal–Range Consistency Filtering (TRCF) removes depth predictions that contradict nearby observations. Surface-Consensus Refinement (SCR) represents observations with ray-aligned anisotropic ellipsoids, uses multi-view parallax to test one- and two-surface explanations on held-out tracks, and applies the minimum required depth correction along the original rays. Regions without sufficient evidence are left unchanged.

Experiments on KITTI and Oxford Spires show that RayCon improves long-sequence reconstruction without retraining the feed-forward model. On KITTI, it reduces the backbone's mean ATE from 31.91 m to 9.59 m, substantially improving camera trajectory accuracy through post-optimization alone.

Our contributions are:

- We decompose feed-forward pointmap error into transverse and radial components, relate residual projective inconsistency to radial disagreement, and exploit partition-dependent errors through staggered dual inference to improve camera trajectories.
- With the cameras fixed, we introduce TRCF and SCR to remove contradicted points, infer whether observations belong to one or two surfaces from multi-view consensus, and refine their depths along the original viewing rays.
- Experiments on KITTI and Oxford Spires show that RayCon, without any training, achieves the lowest mean ATE among the compared methods and reduces local Chamfer distance.

2 Related Work

Feed-forward 3D reconstruction. Learned models predict dense geometry directly from images, from monocular depth (Yang et al., 2024; Wang et al., 2024a; 2025b; 2026) to pairwise pointmaps (Wang et al., 2024b; Leroy et al., 2024) and joint multi-view prediction of cameras, depth, and tracks (Wang et al., 2025a). Joint inference has been scaled to over a thousand images per pass (Yang et al., 2025) and accelerated without retraining (Shen et al., 2025), and temporal extensions process video incrementally through spatial memory (Wang & Agapito, 2024), a recurrent state (Wang et al., 2025c) whose update can be cast as test-time training (Chen et al., 2025), streaming transformers (Zhuo et al., 2025; Lan et al., 2026; Yuan et al., 2026), keyframe-relative prediction with cache-consistent training (Cheng et al., 2026a), anchor- and memory-organized attention (Chen et al., 2026), or bounded-memory long-horizon attention (Cheng et al., 2026b). All operate within a bounded temporal context, and their windowed predictions are what RayCon post-optimizes.

Post-optimization at scale. SLAM systems correct drift with loop closure and global optimization (Campos et al., 2021; Teed & Deng, 2021; Lipson et al., 2024; Murai et al., 2025). For feed-forward models, VGGT-Long aligns overlapping chunks with robust estimation and loop constraints (Deng et al., 2026), and VGGT-SLAM aligns submaps from uncalibrated cameras on the SL(4) manifold to absorb their projective ambiguity (Maggio et al., 2025). Other systems jointly optimize poses and patch depths (Wu et al., 2026), refine trajectories and pointmaps in a backend (Fang et al., 2026), or couple depth priors with global structure-from-motion (Pataki et al., 2026). We keep window-level similarity alignment and correct the spatially varying residual along the viewing rays.

Depth fusion and directional uncertainty. Dense mapping fuses depth into signed distance fields (Newcombe et al., 2011; Curless & Levoy, 1996) or point-based maps (Keller et al., 2013), with range uncertainty weighted along rays (Curless & Levoy, 1996); anisotropic point fusion uses directional uncertainty for correspondence and merging (Lefloch et al., 2015), and multi-view stereo checks occlusion and depth consistency (Schönberger et al., 2016). Bounded-error estimation represents measurements by feasible sets (Milanese & Vicino, 1991), and learned systems gate pointmap corrections by predicted uncertainty (Zhu et al., 2026) or predict when refinement helps (Ahmedt-Aristizabal et al., 2026). SCR uses directional uncertainty to decide surface membership on held-out tracks before any point moves.

3 Method

RayCon post-optimizes the cameras and pointmaps predicted by a frozen feed-forward model from monocular video. We decompose point errors into transverse and radial components in viewing-ray coordinates. The global stage reduces transverse error through trajectory smoothing and staggered dual inference; with the cameras fixed, TRCF filters contradicted depths and SCR resolves surface membership before refining radial error along the viewing rays.

3.1 Ray-Decomposed Error of Feed-Forward Pointmaps

Predictions of the same surface from different windows often disagree primarily along their viewing rays. With camera-to-world rotation R_t, center c_t, predicted intrinsics K_t, and axial depth $z_t(u)$, pixel u with homogeneous coordinate $\bar{u}$ is unprojected as

$$x_t(u) = c_t + \rho_t(u) q_t(u), \qquad q_t(u) = \frac{R_t K_t^{-1} \bar{u}}{\|K_t^{-1} \bar{u}\|_2}, \qquad \rho_t(u) = z_t(u) \|K_t^{-1} \bar{u}\|_2. \tag{1}$$

For an observation $y_i = c_i + \rho_i^0 q_i$ of a true surface point $x_i^\star$, let $c_i^\star$ be the true camera center, $r_i = \|x_i^\star - c_i^\star\|_2$, and $\delta\theta_i$ the angular error of q_i. The point error decomposes as

$$y_i - x_i^\star = e_i^{\parallel} q_i + e_i^{\perp}, \qquad e_i^{\perp} = -(I - q_i q_i^\top)(x_i^\star - c_i), \qquad \|e_i^{\perp}\|_2 \le \|c_i - c_i^\star\|_2 + r_i \delta\theta_i. \tag{2}$$

The transverse component contains no predicted depth and is determined by errors in the camera center and ray direction. Once the cameras are fixed, any correction that preserves the original image projection can only change the range along the same ray.

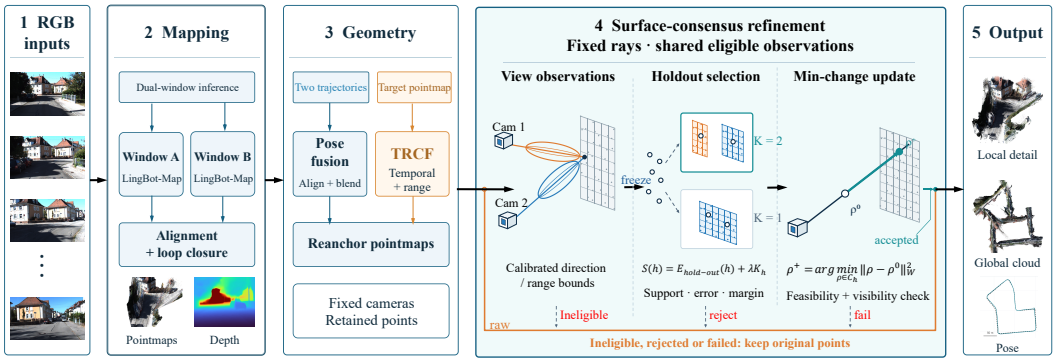

Figure 2: **RayCon overview.** Two staggered window partitions produce complementary camera trajectories, which are aligned and fused after in-window smoothing. The target pointmap is reanchored to the fused cameras, TRCF removes contradicted and redundant points, and SCR selects a surface explanation from multi-view tracks before refining depths along the original rays. Regions without sufficient evidence retain their input.

Independent windows also retain projective inconsistency. Each window predicts its own cameras and intrinsics and therefore forms a local uncalibrated reconstruction, which is determined only up to a 15-degree-of-freedom projective transformation (Hartley & Zisserman, 2004). This ambiguity has also been identified between feed-forward submaps (Maggio et al., 2025). Similarity registration accounts for only seven degrees of freedom, leaving residual distortion that can preserve within-window reprojections while changing the recovered ranges of shared surfaces.

RayCon therefore reduces transverse error through global camera refinement and, with the cameras fixed, resolves the remaining radial inconsistency locally.

3.2 Global Post-Optimization

Our global stage improves the camera centers to reduce transverse pointmap error. Following the divide-and-conquer framework of VGGT-Long (Deng et al., 2026), we split the video into windows of length L with overlap O, predict cameras, intrinsics, and depth in each window with LingBot-Map (Chen et al., 2026), register consecutive windows by similarity transforms on their shared frames, and correct drift with loop closure. Camera-center error then has two sources: variation within each window and alignment residuals at window boundaries. The latter depend on the window partition, since changing the overlap changes where independently reconstructed windows are connected.

We first smooth the camera centers within each window. For stacked centers $C \in \mathbb{R}^{L\times 3}$,

$$C_\lambda = \arg\min_Z \|Z - C\|_F^2 + \lambda \|D_2 Z\|_F^2 = (I + \lambda D_2^\top D_2)^{-1} C, \tag{3}$$

where D_2 is the second-difference operator. We select λ by generalized cross-validation (Golub et al., 1979) and apply a linear correction that restores the first and last centers of each window.

Smoothing reduces within-window variation but does not remove boundary-dependent alignment error. We therefore reconstruct the same video under two staggered window partitions. Run A uses $L = 150$ and $O_A = 90$, while run B uses $L = 150$ and $O_B = 75$. The shifted boundaries place many boundary frames of one run inside a window of the other, producing complementary trajectory errors.

We align the centers of run B to run A with the closed-form Umeyama similarity $(s^\star, Q^\star, b^\star)$ (Umeyama, 1991) and fuse the two trajectories as

$$c_t^F = \tfrac{1}{2}\left(c_t^A + s^\star Q^\star c_t^B + b^\star\right), \qquad R_t^F = R_t^A. \tag{4}$$

In a common frame, the triangle inequality bounds the RMS error of the fused centers by the mean RMS error of the two runs, with equality only when their error sequences are positively proportional.

The output geometry is taken from run A and reanchored to the fused cameras: a registered point x^g from camera (R_t^n, c_t^n) moves to $x^f = R_t^F (R_t^n)^\top (x^g - c_t^n) + c_t^F$, which preserves its camera-local coordinates and keeps it on the same ray.

3.3 TEMPORAL–RANGE CONSISTENCY FILTERING

TRCF removes depth predictions that are directly contradicted by neighboring frames and reduces redundant distant points before surface reasoning. For a point x from frame t and a neighboring frame $t' \in \{t-1, t+1\}$ in the same window, let $z_{t'}$ be the axial depth of x in camera t' and $\hat{z}_{t'}$ the predicted depth at the nearest pixel to $\pi_{t'}(x)$. The comparison is testable, denoted $A_{t'}$, when the projection is valid, both depths are positive and finite, and the derivative of $z_{t'}$ with respect to the source depth exceeds 0.5.

With $\epsilon_{t'} = |\hat{z}_{t'} - z_{t'}| / \max(\hat{z}_{t'}, z_{t'})$, we define support, occlusion, and contradiction as

$$\begin{gathered} \mathsf{S}_{t'} = A_{t'} \wedge [\epsilon_{t'} \le 0.1], \qquad \mathsf{O}_{t'} = A_{t'} \wedge [\hat{z}_{t'} < 0.9 z_{t'}], \qquad \mathsf{C}_{t'} = A_{t'} \wedge \neg \mathsf{S}_{t'} \wedge \neg \mathsf{O}_{t'}, \\ M^{\text{time}} = (\mathsf{S}_{t-1} \vee \mathsf{S}_{t+1}) \wedge \neg(\mathsf{C}_{t-1} \vee \mathsf{C}_{t+1}). \end{gathered} \tag{5}$$

One supporting neighbor suffices, and occluded or untestable views are neutral. The first and last frames of each window skip this test.

We then thin distant points after normalizing camera-frame ranges by their per-frame median. All points up to the 95th percentile are retained. Between the 95th and 98.5th percentiles, we keep the nearest point in each normalized voxel of size 0.05, and use size 0.125 beyond the 98.5th percentile. These operations are camera-local and median-normalized, so the selection is independent of the global similarity frame.

Offset layers that remain consistent over time pass the filter, and SCR decides whether they form one surface or two.

3.4 SURFACE-CONSENSUS REFINEMENT

SCR refines the radial error of the points retained by TRCF within bounded local regions, each observation belonging to one region. RGB tracks $k(i)$ connect observations of the same physical point across views. Whole tracks are deterministically split into a fitting set $\mathcal{K}_{\text{fit}}$ and a held-out validation set $\mathcal{K}_{\text{val}}$; observations that fail validity or track-reliability checks keep their input.

Ray-aligned observation sets. With the cameras fixed, an observation constrains its point more tightly across the viewing ray than along it. We therefore associate observation i with

$$\mathcal{E}_i = \{x : \|L_i(y_i - x)\|_2 \le \kappa\}, \qquad L_i^\top L_i = \Sigma_i^{-1}, \qquad \Sigma_i = \sigma_{r,i}^2 q_i q_i^\top + \sigma_{t,i}^2 (I - q_i q_i^\top), \tag{6}$$

with $\sigma_{r,i} = a_r \rho_i^0$, $\sigma_{t,i} = a_t \rho_i^0$, and $a_r > a_t$.

A single ellipsoid leaves the depth nearly free. Across a track, however, observations are seen from different directions, and the intersection of their elongated sets constrains a shared landmark X_k through parallax. With projection matrix H_i, tracked pixel u_i, and $b_i(X) = H_i[X^\top, 1]^\top$, the landmark must also agree with the tracked pixels:

$$\|b_i(X_k)_{xy} - u_i b_i(X_k)_z\|_2 \le \tau \, b_i(X_k)_z. \tag{7}$$

Surface hypothesis selection. The offset layers that survive TRCF admit two explanations: radial disagreement on one surface, or two surfaces such as a foreground structure in front of a wall. SCR considers hypotheses h with $K_h \in \{1, 2\}$ planes, namely one plane, two parallel planes, or two planes with different orientations.

Fitting tracks generate the candidate hypotheses. Each plane has a fixed unit normal n_j and an optimizable offset d_j, while all observations of a track share a surface label s_k. For a fixed normal, the offsets compatible with a fitting track form an interval. Labels are assigned to the largest group of tracks with overlapping intervals, with at least four fitting tracks per plane.

For observations $\mathcal{A}$ with labels $\zeta_i = s_{k(i)}$, we define the feasible set $\mathcal{F}(h; \mathcal{A}, \zeta)$ over depths ρ, plane offsets d, and landmarks X:

$$\begin{aligned} n_{\zeta_i}^\top (c_i + \rho_i q_i) = d_{\zeta_i}, \quad n_{s_k}^\top X_k = d_{s_k}, \quad X_{k(i)} \in \mathcal{E}_i, \\ \|c_i + \rho_i q_i - X_{k(i)}\|_2 \le \eta_i, \quad c_i + \rho_i q_i \in \mathcal{E}_i, \quad |\rho_i - \rho_i^0| \le \Delta \rho_i^0. \end{aligned} \tag{8}$$

Equation 7, a common support box, a cross-view projection bound, and a grazing limit $|n_{\zeta_i}^\top q_i| \geq \gamma$ complete the set (Appendix A). Here κ sets the ellipsoid size, η_i bounds the distance between an output point and its landmark, and Δ caps the relative depth change. With normals and labels fixed, the constraints are affine or second-order cones, so $\mathcal{F}$ is convex.

Feasibility on the fitting tracks cannot distinguish one from two surfaces, since the more flexible model can fit ghosted layers in-sample. We therefore fit each hypothesis on $\mathcal{K}_{\text{fit}}$, freeze its plane parameters, and evaluate all hypotheses on the same held-out tracks $\mathcal{K}_{\text{val}}$. Plane j is admissible for a validation track k with observations $\mathcal{I}_k$ when $\mathcal{F}(h; \mathcal{I}_k, j\mathbf{1})$ remains feasible with the fitted plane fixed. For the admissible set $\mathcal{J}_k(h)$, we define

$$
\begin{aligned}
r_i(h,j) &= \frac{|n_j^\top y_i - d_j^{\text{fit}}|}{\sqrt{n_j^\top \Sigma_i n_j}}, \qquad \ell_k(h) = \min_{j \in \mathcal{J}_k(h)} \sum_{i \in \mathcal{I}_k} \alpha_{ki}\, \psi(r_i(h,j))\,, \\
S(h) &= \frac{1}{|\mathcal{K}_{\text{val}}|} \sum_{k \in \mathcal{K}_{\text{val}}} \ell_k(h) + \lambda_K K_h,
\end{aligned}
\tag{9}
$$

where $\psi(r) = \min(r^2, M_{\text{fail}})$, $\ell_k(h) = M_{\text{fail}}$ when no plane is admissible, α_{ki} normalizes repeated observations from the same frame, and $n_j^\top \Sigma_i n_j = \sigma_{r,i}^2 (n_j^\top q_i)^2 + \sigma_{t,i}^2 (1 - (n_j^\top q_i)^2)$ accounts for the directional uncertainty of each observation. Tracks that a hypothesis fails to explain remain in the average. We accept the best hypothesis only if its score falls below an absolute threshold, every plane is supported by at least two validation tracks, and it improves over the runner-up by a fixed margin. Otherwise, the region retains its input.

Minimum-change radial correction. An accepted explanation changes the input only as much as needed to satisfy it. Fitting tracks retain their consensus labels, validation tracks use their admissible planes, and unlabeled tracks keep their input. With normals and labels fixed, we jointly refit the plane offsets, landmarks, and depths by solving

$$
\rho^+ = \arg\min_{\rho \in \mathcal{C}} (\rho - \rho^0)^\top W (\rho - \rho^0), \quad \mathcal{C} = \{\rho : \exists\, d, X,\ (\rho, d, X) \in \mathcal{F}(h; \mathcal{A}, \zeta)\}, \tag{10}
$$

where $W = \operatorname{diag}(v_i / \sigma_{r,i}^2)$ and v_i accounts for duplicate observations from the same frame. The refined point is $p_i^+ = y_i + (\rho_i^+ - \rho_i^0) q_i$.

This convex quadratic program with second-order-cone constraints (Boyd & Vandenberghe, 2004) moves points only along their original rays. We retain the update only when the solver returns an optimal solution with bounded constraint violation, all source-view reprojections remain within the calibrated tolerance, and no new free-space conflict with the fixed predicted depths is introduced. Otherwise, the entire region keeps its input. Point identities, counts, and cameras are preserved.

For nonempty compact convex $\mathcal{F}$, the metric projection satisfies (Bauschke & Borwein, 1996)

$$
\|\rho^+ - \rho^\star\|_W^2 \leq \|\rho^0 - \rho^\star\|_W^2 - \|\rho^+ - \rho^0\|_W^2, \qquad \forall\, \rho^\star \in \mathcal{C}. \tag{11}
$$

Thus, whenever the true depths lie in $\mathcal{C}$, the refined depths are no farther from them than the input, while an input that already satisfies the selected explanation remains unchanged. SCR leaves the cameras fixed and therefore does not affect trajectory metrics.

4 Experiments

4.1 Experimental Setup

Implementation. RayCon uses the LingBot-Map long checkpoint (Chen et al., 2026) as a frozen frontend with windows of $L = 150$ frames and overlaps of 90 and 75, and requires no training. The frontend runs on an NVIDIA GeForce RTX 5090 GPU with 32 GB of memory, while TRCF and SCR run on the CPU with CVXPY and Clarabel. Appendix A lists all parameters and stage runtimes.

Datasets. We evaluate trajectories on KITTI Odometry 00–10 (Geiger et al., 2012) and ten Oxford Spires sequences (Tao et al., 2026), and surface quality on local regions of 7-Scenes Heads (Shotton et al., 2013). Sampling and sequence details are given in Appendix B.

Table 1: ATE RMSE (m, lower is better) on KITTI 00–10, with loop correction (LC) marked for each method. Green and blue mark the best and second-best values in each column, and TL, OOM, and – denote tracking loss, out of memory, and reported failure.

	Methods	LC	Avg.	00	01	02	03	04	05	06	07	08	09	10
Classic	ORB-SLAM2	✓	54.82	6.03	508.34	14.76	1.02	1.57	4.04	11.16	2.19	38.85	8.39	6.63
	LDSO	✓	22.43	9.32	11.68	31.98	2.85	1.22	5.10	13.55	2.96	129.02	21.64	17.36
	ORB-SLAM2	×	69.73	40.65	502.20	47.82	0.94	1.30	29.95	40.82	16.04	43.09	38.77	5.42
	ORB-SLAM3	✓	16.30	8.94	103.76	24.90	0.76	0.79	5.33	12.87	2.55	4.45	8.39	6.57
Learning-based	DROID-VO	×	54.19	98.43	84.20	108.80	2.58	0.93	59.27	64.40	24.20	64.55	71.80	16.91
	DPVO	×	53.61	113.21	12.69	123.40	2.09	0.68	58.96	54.78	19.26	115.90	75.10	13.63
	DROID-SLAM	✓	100.28	92.10	344.60	107.61	2.38	1.00	118.50	62.47	21.78	161.60	72.32	118.70
	DPV-SLAM	✓	53.03	112.80	11.50	123.53	2.50	0.81	57.80	54.86	18.77	110.49	76.66	13.65
	DPV-SLAM++	✓	25.75	8.30	11.86	39.64	2.50	0.78	5.74	11.60	1.52	110.90	76.70	13.70
	MASt3R-SLAM	✓	–	TL	TL	TL	TL	TL	TL	TL	TL	TL	TL	TL
	CUT3R	×	–	OOM	OOM	OOM	148.07	22.31	OOM	OOM	OOM	OOM	OOM	OOM
	VGGT-Long	✓	27.64	8.67	121.17	32.08	6.12	4.23	8.31	5.34	4.63	53.10	41.99	18.37
	Fast3R	×	–	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	VGGT	×	–	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	SLAMFormer-∞	✓	23.01	15.39	96.58	28.81	4.97	4.16	11.36	13.54	8.53	37.00	18.85	13.93
	FastVGGT	×	189.29	–	705.39	–	62.38	10.27	157.74	124.43	69.27	–	190.10	194.75
	VGGT-SLAM	✓	263.37	–	607.16	–	169.83	13.12	–	–	–	–	–	–
	TTT3R	×	177.73	190.93	546.84	218.77	105.28	11.62	153.12	132.94	70.95	180.57	211.01	133.00
	STream3R	×	227.77	190.98	681.95	301.40	158.25	102.73	159.85	135.03	90.37	261.15	216.31	207.49
	StreamVGGT	×	226.15	191.93	653.06	303.35	157.50	108.24	160.46	133.71	89.00	263.95	216.69	209.80
	LongStream	×	51.90	92.55	46.01	134.70	3.81	1.95	84.69	23.12	14.93	62.07	85.61	21.48
	DROID+M3D	✓	21.26	12.84	150.44	20.03	2.08	0.83	7.78	3.19	18.32	9.73	3.76	4.83
	ViSTA-SLAM	✓	143.41	160.90	414.40	257.30	39.32	25.85	110.63	58.68	51.99	215.60	154.60	88.24
	LoGeR*	×	18.64	30.47	47.91	36.32	5.30	1.95	26.34	6.60	5.55	24.41	10.12	10.11
	LingBot-Map	×	31.91	24.24	83.63	65.75	2.50	1.02	74.77	7.51	9.87	19.45	42.55	19.69
	RayCon	✓	**9.59**	**3.52**	**26.92**	**31.68**	**1.90**	**0.94**	**3.41**	**2.34**	**2.47**	**16.54**	**10.89**	**4.84**

Metrics. Trajectory accuracy is the ATE RMSE of camera centers after one full-sequence Sim(3) alignment per trajectory, and dataset means weight all sequences equally. Surface quality is measured by Chamfer distance (CD), the mean of the two directed mean nearest-neighbor distances, and by F1 and Recall at a 10 cm threshold. Appendix B.1 gives the complete definitions.

Baselines. On KITTI, we compare with classical SLAM and with learning-based SLAM, feed-forward, and streaming reconstruction methods, using results reported by SCE-SLAM (Wu et al., 2026), SLAMFormer-∞ (Fang et al., 2026), LongStream (Cheng et al., 2026a), and Scalix (Laina et al., 2026). On Oxford Spires, we compare with optimization-based and streaming methods reported by ABot-Recon (Han et al., 2026). The source protocols differ in detail, so the tables summarize published comparisons; Appendix B documents the source of every row.

4.2 Long-Sequence Trajectory Reconstruction

KITTI. RayCon reaches a mean ATE of 9.59 m, the lowest among the displayed methods and less than a third of the 31.91 m of its LingBot-Map backbone (Table 1). The reduction is not confined to a few sequences, since RayCon has a lower error than LingBot-Map on every sequence. Figure 3 shows that its trajectories follow the ground truth more closely and that its local reconstructions are more coherent, although sparse and difficult regions remain. Appendix C shows the full map of KITTI 00.

Oxford Spires. RayCon reaches a mean ATE of 6.04 m, compared with 7.32 m for LingBot-Map, and again has the lowest mean among the displayed methods (Table 2). The improvement therefore extends from driving scenes to handheld sequences around complex architecture. In Figure 4, the recovered routes stay close to the reference, while facade and boundary details show the remaining local artifacts.

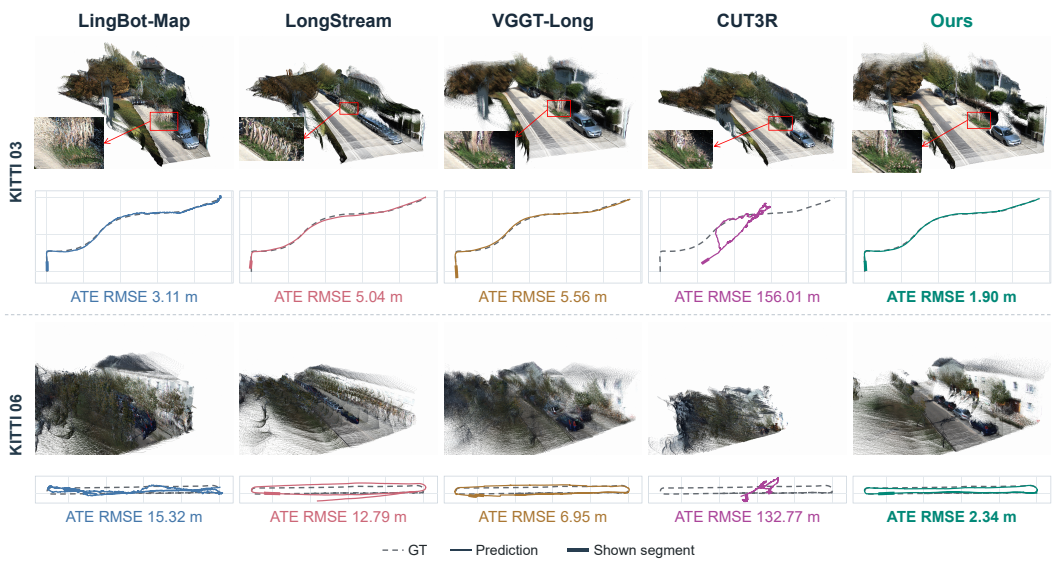

Figure 3: Qualitative comparison on KITTI sequences 03 (top) and 06 (bottom), showing local point clouds and complete trajectories for LingBot-Map, LongStream, VGGT-Long, CUT3R, and RayCon (Ours). Dashed and solid curves denote ground-truth and predicted trajectories, respectively. The inset highlights a roadside region in sequence 03 for closer comparison.

Table 2: ATE RMSE (m, lower is better) on ten Oxford Spires sequences from Keble College (K), Observatory Quarter (O), Christ Church (C), and Bodleian Library (B). Green and blue mark the best and second-best values in each column, and Opt. denotes optimization-based methods.

	Methods	LC	Avg.	K02	K03	K04	K05	O01	O02	C02	C03	C05	B02
Opt.	VGGT-SLAM	✓	6.15	2.64	3.78	3.89	1.86	5.74	3.39	7.41	1.02	12.94	18.83
	MASt3R-SLAM	✓	12.12	7.06	18.03	11.27	14.64	10.26	10.30	2.90	6.95	19.85	19.99
	VGGT-Long	✓	7.41	9.07	9.37	11.77	3.96	6.03	4.05	2.41	0.48	14.38	12.59
Streaming	InfiniteVGGT	×	31.76	25.02	20.39	25.09	24.15	29.10	28.14	33.02	12.95	40.39	79.32
	OVGGT	×	33.19	32.97	23.01	28.53	23.23	28.51	27.75	30.01	13.09	42.35	82.41
	Stream3R (window=5)	×	34.11	31.28	22.39	31.07	24.63	28.38	27.18	38.12	12.99	42.39	82.66
	CUT3R	×	32.79	31.68	21.67	22.76	25.07	26.77	29.41	33.93	13.06	43.66	79.93
	TTT3R	×	25.96	24.66	22.56	27.33	18.67	26.94	28.18	19.83	9.68	26.76	55.01
	LongStream	×	15.92	11.57	9.44	16.88	10.67	12.43	14.34	17.93	6.42	35.49	24.02
	LingBot-Map	×	7.32	2.71	3.04	2.04	3.23	5.38	4.43	2.21	0.74	37.99	11.43
	HorizonStream	×	8.81	5.79	6.47	7.36	6.27	6.33	5.30	10.69	3.11	25.35	11.41
	HorizonStream	✓	7.93	5.79	6.47	1.25	6.24	6.33	5.30	10.69	0.47	25.35	11.41
	RayCon	✓	**6.04**	**1.98**	**1.87**	**2.51**	**1.78**	**1.54**	**1.22**	**5.85**	**1.32**	**34.33**	**8.04**

4.3 Surface Optimization

We test whether surface optimization improves local geometry with the cameras fixed, evaluating each region within its input bounding box padded by 20% of the median viewing range, without deleting or revoxelizing points.

On the region with the largest reduction in CD (Table 3), surface optimization lowers CD and raises F1 and Recall. Across all eight regions, mean CD decreases from 0.0973 to 0.0938 m and mean F1 increases from 43.40% to 45.61%. The gain comes from refining track-supported geometry along the viewing rays; Appendix C reports component controls.

4.4 Ablation Study

Staggered dual-window fusion. Removing the second partition raises the mean ATE on KITTI from 9.59 m to 11.07 m, and fusion improves 8 of the 11 sequences (Table 4). On Oxford Spires, the fused trajectory outperforms both single-partition runs on five of ten sequences and never exceeds

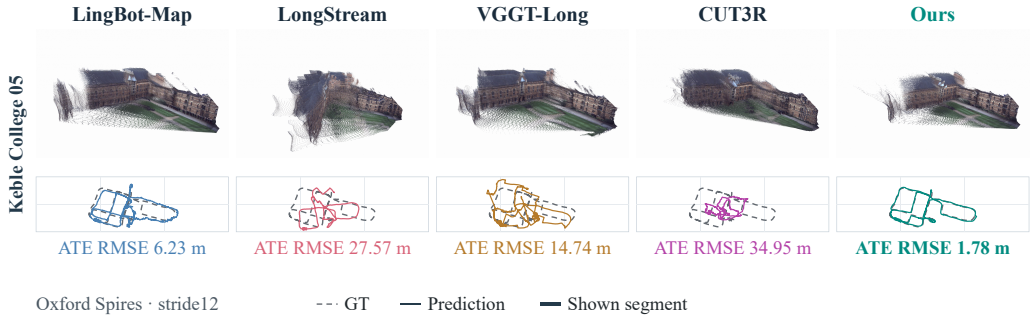

Figure 4: Local point clouds and full trajectories on Oxford Spires Keble College 05, with methods and curves as in Figure 3.

Table 3: Local geometry with fixed cameras for surface optimization on a 7-Scenes Heads region (top) and TRCF on KITTI 01, 02, 03, and 08 (bottom). CD is in meters; F1 and Recall are in percent at 10 cm.

Data	Configuration	CD ↓	F1 ↑	Recall ↑
7-Scenes Heads	Input	0.1531	9.49	8.29
	RayCon, in-sample selection	0.1339	20.32	13.56
	RayCon	**0.1319**	**22.32**	**15.65**
KITTI (4 seq.)	No filtering	4.112	4.03	**16.43**
	Temporal support only	3.914	4.30	14.74
	TRCF	**3.861**	**4.40**	14.20

Table 4: Staggered dual-window pose fusion (ATE, m). Case ATE is reported on the sequence with the largest reduction, and Mean ATE averages each dataset.

Case	Configuration	Overlap	Fusion	Case ATE ↓	Mean ATE ↓
KITTI 03	w/o dual-window fusion	90	×	2.910	11.074
	RayCon	75 + 90	✓	**1.903**	**9.588**
Oxford K05	Single-window source	75	×	1.826	6.844
	w/o dual-window fusion	90	×	2.451	6.060
	RayCon	75 + 90	✓	**1.783**	**6.045**

their mean error, as Section 3.2 predicts. The gain carries over to the dense KITTI map, whose mean CD drops from 2.568 m to 2.497 m (Appendix C), and it comes from the staggered boundaries because the frontend is deterministic.

Local stage. Temporal support removes contradicted depths and lowers CD, and range thinning lowers it further and raises F1, at a small cost in Recall from the removed points (Table 3). Scoring the frozen candidates on fitting tracks selects two parallel layers, the pattern of ghosting, while held-out scoring selects two planes of different orientation with lower CD and higher F1 and Recall. At equal empirical coverage, isotropic observation sets need 56 times the volume of directional ones (Appendix C).

5 Limitations and Conclusion

Limitations. Dual-window prediction requires a second frontend pass and therefore increases inference cost. Local refinement relies on reliable feature tracks and approximately planar structure, so weakly textured, heavily occluded, and curved regions provide fewer usable constraints. The local stage also keeps the cameras fixed, leaving residual pose error to the trajectory fusion of the global stage; jointly optimizing cameras and surfaces is a natural extension.

Conclusion. We presented RayCon, which improves long-sequence monocular reconstruction from feed-forward models without retraining. It fuses the trajectories of two staggered window partitions and, using the difference between radial and transverse errors, decides which observations a single surface can explain before correcting them with a minimum depth adjustment along the rays. Experiments show that trajectory fusion lowers the mean ATE on KITTI and Oxford Spires, and that surface-consensus refinement improves the geometry of the evaluated local regions.

LLM USAGE STATEMENT

Large language models were used to aid and polish writing, and for retrieval and discovery, such as finding related work.

REFERENCES

David Ahmedt-Aristizabal, Mohammad Ali Armin, Russell Tsuchida, and Lars Petersson. GeoCond: A conditioning-aware reliability adapter for feed-forward 3D reconstruction, 2026. URL https://arxiv.org/abs/2609.18465v1.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. doi: 10.1137/S0036144593251710. URL https://cmps-people.ok.ubc.ca/bauschke/Research/05.pdf.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. doi: 10.1109/TRO.2021.3075644.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training, 2025. URL https://arxiv.org/abs/2509.26645.

Chong Cheng, Xianda Chen, Tao Xie, Wei Yin, Weiqiang Ren, Qian Zhang, Xiaoyang Guo, and Hao Wang. LongStream: Long-Sequence Streaming Autoregressive Visual Geometry, 2026a. URL https://arxiv.org/abs/2602.13172v2.

Chong Cheng, Peilin Tao, Nanjie Yao, Guanzhi Ding, Xianda Chen, Yuansen Du, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, Zhengqing Chen, and Hao Wang. HorizonStream: Long-horizon attention for streaming 3D reconstruction, 2026b. URL https://arxiv.org/abs/2605.23889.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of SIGGRAPH*, 1996. URL https://graphics.stanford.edu/papers/volrange/.

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*, 2026. URL https://arxiv.org/abs/2507.16443v2.

Zhijian Fang, Weicheng Zheng, Yijun Yuan, Weibang Wang, Zhuoguang Chen, Chang Sun, Junhao Huang, Kenan Li, Minghui Qin, and Hang Zhao. SLAMFormer-∞: Infinite SLAM transformer for unbounded frontend and backend processing, 2026. URL https://arxiv.org/abs/2608.03429.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.cvlibs.net/publications/Geiger2012CVPR.pdf.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. doi: 10.1080/00401706.1979.10489751.

Jiarong Han, Jincheng Xiong, Yuzhou Liu, Linzhe Shi, Changjie Wu, Ning Guo, Mu Xu, Hang Zhang, and Ming Qian. Revisiting Local Context for Long-Horizon Streaming 3D Reconstruction, 2026. URL https://arxiv.org/html/2608.27529v1#S4.T4.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. doi: 10.1017/CBO9780511811685.

Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *International Conference on 3D Vision*, pp. 1–8, 2013. doi: 10.1109/3DV.2013.9.

Sebastian Barbas Laina, Tianyi Zhang, Panagiotis Petropoulakis, Simon Schaefer, Simon Boche, Jaehyung Jung, Cedric Le Gentil, and Stefan Leutenegger. Scalix: Uncertainty-Aware Scale-Consistent Monocular SLAM, 2026. URL https://arxiv.org/abs/2608.17553v1.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *International Conference on Learning Representations*, 2026.

Damien Lefloch, Tim Weyrich, and Andreas Kolb. Anisotropic point-based fusion. In *18th International Conference on Information Fusion (FUSION)*, pp. 2121–2128, 2015. URL https://reality.tf.fau.de/projects/kinect/lefloch15anisotropic.html.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R, 2024. URL https://arxiv.org/abs/2406.09756.

Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam, 2024. URL https://arxiv.org/abs/2408.01654.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold, 2025. URL https://arxiv.org/abs/2505.12549.

Mario Milanese and Antonio Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Automatica*, 27(6):997–1009, 1991. doi: 10.1016/0005-1098(91)90134-N. URL https://www.sciencedirect.com/science/article/pii/000510989190134N.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.12392.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–136, 2011. doi: 10.1109/ISMAR.2011.6092378.

Zador Pataki, Paul-Edouard Sarlin, and Marc Pollefeys. VidMap: Exploiting temporal structure for video-based structure-from-motion, 2026. URL https://arxiv.org/abs/2607.27194.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html.

Johannes L. Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. URL https://demuc.de/papers/schoenberger2016mvs.pdf.

You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013. doi: 10.1109/CVPR.2013.377.

Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The Oxford Spires Dataset: Benchmarking Large-Scale LiDAR-Visual Localisation, Reconstruction and Radiance Field Methods. *The International Journal of Robotics Research*, 45(6), 2026. doi: 10.1177/02783649251369905. URL https://journals.sagepub.com/doi/10.1177/02783649251369905.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.nips.cc/paper/2021/file/89fcd07f20b6785b92134bd6c1d0fa42-Paper.pdf.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory, 2024. URL https://arxiv.org/abs/2408.16061.

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*, 2025a. URL https://arxiv.org/abs/2503.11651.

Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. WeatherDepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024a. URL https://arxiv.org/abs/2310.05556.

Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://papers.neurips.cc/paper_files/paper/2025/hash/72bcad2c3fe7625fc233fabe84a6d7ef-Abstract-Conference.html.

Jiyuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, and Xiangxiang Chu. FE2E: From editor to dense geometry estimator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19844–19853, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_FE2E_From_Editor_to_Dense_Geometry_Estimator_CVPR_2026_paper.html.

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*, 2025c. URL https://arxiv.org/abs/2501.12387.

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*, 2024b. URL https://arxiv.org/abs/2312.14132.

Yuchen Wu, Jiahe Li, Xiaohan Yu, Lina Yu, Jin Zheng, and Xiao Bai. SCE-SLAM: Scale-consistent monocular SLAM via scene coordinate embeddings, 2026. URL https://arxiv.org/abs/2601.09665.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass, 2025. URL https://arxiv.org/abs/2501.13928.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. URL https://arxiv.org/abs/2401.10891.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual geometry grounded transformer for endless streams, 2026. URL https://arxiv.org/abs/2601.02281.

Ganlin Zhang, Shenhan Qian, Xi Wang, and Daniel Cremers. Vista-slam: Visual slam with symmetric two-view association, 2025. URL https://arxiv.org/abs/2509.01584.

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, 2026. URL https://arxiv.org/abs/2603.03269.

Zihao Zhu, Wenyuan Zhao, Nuo Chen, Chao Tian, and Zhiwen Fan. Trust it or not: Evidential uncertainty for feed-forward 3D reconstruction with Trust3R, 2026. URL https://arxiv.org/abs/2605.19539v1.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer, 2025. URL https://arxiv.org/abs/2507.11539.

A Implementation Details

Frontend and global post-optimization. Frames are resized to width 518 with Lanczos interpolation, and the height is rounded down to a multiple of 14. Each window clears the model cache, processes its first eight frames jointly, and streams the remaining frames. Consecutive windows are registered on the pointmaps of their shared frames with confidence weights $\sqrt{\mathrm{conf}^x \mathrm{conf}^y}$ and Huber reweighting ($\delta = 0.1$, five iterations). Loop candidates are retrieved with SALAD descriptors (similarity threshold 0.85, top five, non-maximum suppression over 25 frames), and each candidate pair is reconstructed from two 20-frame neighborhoods to form a similarity constraint. The similarity graph is optimized for at most 30 Levenberg–Marquardt iterations. The smoothing weight in Equation 3 is selected from 97 log-spaced values in $[10^{-4}, 10^{8}]$; windows with fewer than five frames are not smoothed.

Complete SCR feasible set. Coordinates are normalized by the region median point and the median ray distance. Beyond Equation 8, every output point and landmark lies in a common box Q that pads the bounding box of the region inputs by $0.16 \max_i \rho_i^0$, offsets satisfy $|d_j| \le 2B_Q + 1$ for the box radius B_Q, and depths satisfy $\max(10^{-8}, 0.85\rho_i^0) \le \rho_i \le 1.15\rho_i^0$. For each output point P_i and each view t of its track, the projection satisfies the cone of Equation 7 with tolerance $\max(\tau, e_{it}^0 + 0.25) - 10^{-5}$ pixels, where e_{it}^0 is the reprojection error of the raw point. Fitting tracks with a maximum ray angle below 0.25° or with $|n^\top q_i| < 0.15$ are treated as unobservable.

Solver and checks. All problems are solved with CVXPY and Clarabel (at most 150 iterations, 10 s, tolerances 10^{-9}). Only optimal solutions with maximum constraint violation below 2×10^{-6} are accepted; any other status in candidate fitting, scoring, or the final refit returns the region to its input. The final check requires finite outputs with positive depth inside every source image, a reprojection error at most $\max(\tau, e_{it}^0 + 0.25)$, and no new free-space conflict, defined as an output depth below 0.9 times the fixed predicted depth at the projected pixel where the raw point had none.

Table 5: Parameters of RayCon, all fixed before evaluation.

Stage	Parameter	Value
Windows	length L; overlaps of runs A / B	150; 90 / 75
Fusion	center weights; rotation	0.5 / 0.5; target run
TRCF	relative depth tolerance; occlusion ratio; depth derivative	0.10; 0.9; 0.5
TRCF	range percentiles; normalized voxel sizes	95 / 98.5; 0.05 / 0.125
SCR sets	a_r; a_t; κ; η_i	0.08; 0.01; 2; $0.025\rho_i^0$
SCR sets	pixel tolerance τ	3.954 px
SCR bounds	Δ; grazing limit γ; minimum parallax	0.15; 0.15; 0.25°
Selection	minimum fitting tracks per surface; minimum validation tracks	4; 2
Selection	M_{fail}; λ_K; score threshold; margin	9; 0.15; 2; 0.03
Selection	parallel-layer separation; orientation split cosine	0.03; 0.94

The pixel tolerance τ is the 90% empirical quantile of held-out reprojection errors on independent calibration tracks, each triangulated from the remaining views.

Runtime. Stage times are taken from the archived run logs. On KITTI 00–10, TRCF together with cache loading, coordinate synchronization, voxelization, and map export takes 2092.43 s in total, or 190.22 s per sequence. On Oxford Keble College 05 (962 frames), the same stage takes 190.79 s, including mask writing and export of the display window. On this sequence, the pose-only frontend takes 1895.1 s for the single overlap-90 run and 3189.3 s summed over the two runs of staggered dual chunking, excluding fusion, map export, and evaluation. SCR takes 38.40 s on the Heads region of Table 9, which contains 180 observations, and 225.31 s over all eight development regions (28.16 s per region), including solving, checks, output writing, and regions whose update was rejected.

B Evaluation Protocol and Baseline Sources

B.1 Evaluation protocol

On KITTI, we process the complete monocular left-color stream (`image_2`) of each sequence at stride 1. On Oxford Spires, we use Keble College 02–05, Observatory Quarter 01–02, Christ Church 02, 03, and 05, and Bodleian Library 02, sampling each complete sequence at stride 12; the published results use clips of 3,821–3,840 frames at stride 1 (Han et al., 2026). Surface quality is evaluated on eight track-supported local regions from the first 150 frames of 7-Scenes Heads.

For every local run, ATE RMSE is the root-mean-square camera-position error after one full-sequence Sim(3) alignment of the frozen predictions. With valid timestamp associations $\mathcal{I}_s$, estimated positions $\hat{\mathbf{t}}_{s,i}$, reference positions $\mathbf{t}_{s,i}$, and fitted similarity $(a_s, \mathbf{R}_s, \mathbf{b}_s)$,

$$e_s = \left(\frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \left\| a_s \mathbf{R}_s \hat{\mathbf{t}}_{s,i} + \mathbf{b}_s - \mathbf{t}_{s,i} \right\|_2^2 \right)^{1/2}, \qquad \bar{e} = \frac{1}{S} \sum_{s=1}^{S} e_s, \tag{12}$$

where $S = 11$ for KITTI and $S = 10$ for Oxford Spires, and unrounded sequence errors are averaged. The alignment measures trajectory shape and drift. External averages retain the published aggregates, and we report sequence-level ATE instead of the segment-based KITTI odometry errors.

On Oxford Spires, 6,187 of the 6,240 predicted poses have a reference pose within 0.1 s. The 53 unmatched poses come from Observatory Quarter 02 (one pose) and Christ Church 02 (52 poses) and are excluded from ATE.

Surface quality is evaluated per region within the bounding box of the input region padded by 20% of its median viewing range. All configurations share this support and the archived full-trajectory Sim(3), use no local alignment, point deletion, or revoxelization, and are frozen before reference evaluation. CD is the mean of the two directed mean nearest-neighbor distances without squaring, and F1 and Recall use a 10 cm threshold.

B.2 Baseline sources

Each external row is transcribed from a single source. On KITTI, Table 1 of SCE-SLAM (Wu et al., 2026) supplies ORB-SLAM2, LDSO, DROID-VO, DPVO, DROID-SLAM, DPV-SLAM, DPV-SLAM++, MASt3R-SLAM, CUT3R, and VGGT-Long; Table 1 of SLAMFormer-∞ (Fang et al., 2026) supplies ORB-SLAM2 without LC, Fast3R, VGGT, and SLAMFormer-∞; Table 1 of LongStream (Cheng et al., 2026a) supplies FastVGGT, VGGT-SLAM, TTT3R, STream3R, StreamVGGT, and LongStream; and the Sim(3) rows of Table I of Scalix (Laina et al., 2026) supply ORB-SLAM3, DROID+M3D, ViSTA-SLAM, LoGeR*, and LingBot-Map. On Oxford Spires, Table 4 of ABot-Recon (Han et al., 2026) supplies all external rows with Umeyama Sim(3) alignment. Stream3R (window=5) is the official five-frame inference mode, DPV-SLAM++ adds image retrieval and pose-graph correction (Lipson et al., 2024), and the asterisk in LoGeR* is part of the method name (Zhang et al., 2026).

Table notation. LC marks geometric loop correction; temporal memory alone does not count. On KITTI, the LC entries of ORB-SLAM3, DROID-SLAM, MASt3R-SLAM, CUT3R, FastVGGT, TTT3R, STream3R, StreamVGGT, LongStream, DROID+M3D, ViSTA-SLAM, LoGeR*, and LingBot-Map follow method defaults instead of a documented run setting, as do all external Oxford Spires rows except HorizonStream, as detailed below. ORB-SLAM3, DROID+M3D, ViSTA-SLAM, and LingBot-Map on KITTI are re-evaluations by the Scalix authors, and the LongStream source does not specify the ATE alignment of FastVGGT, VGGT-SLAM, TTT3R, STream3R, StreamVGGT, and LongStream. A dash in a numeric cell marks a source-reported failure, and the averages of FastVGGT and VGGT-SLAM cover only their available sequences. Green and blue cells mark the best and second-best distinct values per column; ties share a color and missing entries are excluded.

Loop-correction entries. Explicit labels in the source tables take precedence: SCE-SLAM and SLAMFormer-∞ state the LC setting of their rows, and ABot-Recon distinguishes the base and LC

variants of HorizonStream. For the remaining rows, the entry follows the published method. ORB-SLAM3 enables loop closure by default, DROID-SLAM adds long-range edges on revisits and performs global bundle adjustment (Teed & Deng, 2021), MASt3R-SLAM retrieves and verifies loop candidates (Murai et al., 2025), ViSTA-SLAM corrects verified loop edges on a Sim(3) pose graph (Zhang et al., 2025), and VGGT-SLAM and VGGT-Long enable loop retrieval in their default configurations. InfiniteVGGT, OVGGT, STream3R, CUT3R, FastVGGT, TTT3R, StreamVGGT, LongStream, LingBot-Map, and LoGeR* predict through attention or recurrent state without a geometric loop-correction backend. Both frontend runs of RayCon enable loop closure.

C Additional Results

Full-sequence reconstruction. Figure 5 shows the complete KITTI 00 map. RayCon has an ATE of 3.52 m, compared with 45.21, 90.01, 10.82, and 191.12 m for LingBot-Map, LongStream, VGGT-Long, and CUT3R, whose compact map reflects its collapsed trajectory.

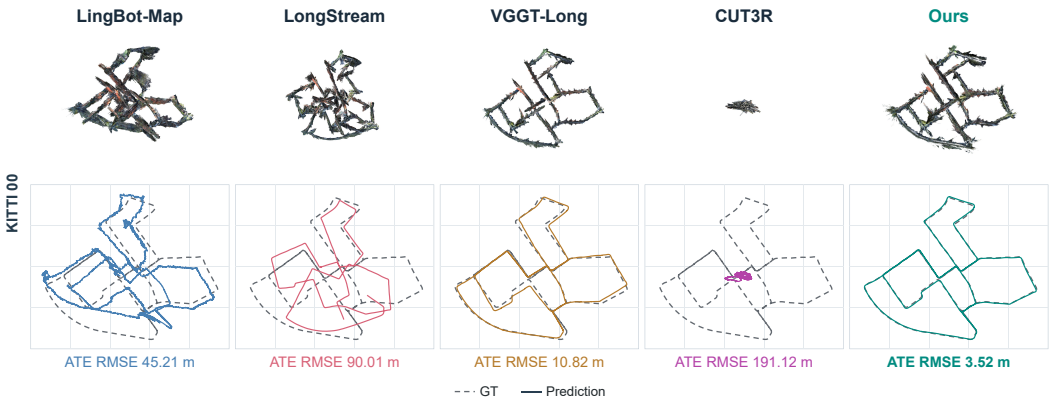

Figure 5: Global point clouds and trajectories over all 4,541 frames of KITTI 00. All methods share the top-view camera, display scale, sky mask, 100 m range limit, and 0.20 m display voxels.

Per-sequence pose results. Tables 6 and 7 list every sequence of the dual-window ablation. Fusion improves 8 of 11 KITTI sequences and 6 of 10 Oxford Spires sequences, and on Oxford Spires its error never exceeds the mean of the two single-partition runs.

Table 6: Per-sequence ATE RMSE (m) on KITTI with and without staggered dual-window fusion.

Configuration	Avg.	00	01	02	03	04	05	06	07	08	09	10
w/o fusion	11.074	3.76	38.44	32.37	2.91	**0.83**	**3.16**	2.65	**2.36**	19.25	11.19	4.89
RayCon	**9.588**	**3.52**	**26.92**	**31.68**	**1.90**	0.94	3.41	**2.34**	2.47	**16.54**	**10.89**	**4.84**

Table 7: Per-sequence ATE RMSE (m) on Oxford Spires for the two single-partition runs with overlaps 75 and 90 and for their fusion.

Configuration	Avg.	K02	K03	K04	K05	O01	O02	C02	C03	C05	B02
Single / 75	6.844	2.17	2.48	3.31	1.83	1.80	1.44	7.86	1.33	38.21	**8.01**
Single / 90	6.060	1.99	**1.59**	**2.45**	2.45	**1.53**	1.44	7.51	1.67	**31.55**	8.42
RayCon	**6.045**	**1.98**	1.87	2.51	**1.78**	1.54	**1.22**	**5.85**	**1.32**	34.33	8.04

Dense map geometry. Table 8 evaluates the exported KITTI maps against the sparse LiDAR reference. Each map is aligned with the full-trajectory Sim(3), voxelized at 0.10 m, and restricted to points within 1–80 m of the reference camera path, without ICP; $P \to G$ and $G \to P$ are the two directed mean nearest-neighbor distances. Fusion lowers CD and both directed distances and raises F1 at 20 cm, while F1 at 5 and 10 cm changes by less than 0.02 points.

Table 8: Dense map geometry on KITTI 00–10 with and without staggered dual-window fusion, averaged over sequences. Distances are in meters, and F1@5, F1@10, and F1@20 are percentages at 5, 10, and 20 cm.

Configuration	CD	$P \to G$	$G \to P$	F1@5	F1@10	F1@20
w/o fusion	2.568	4.100	1.037	**1.839**	**6.532**	13.152
RayCon	**2.497**	**4.024**	**0.970**	1.829	6.525	**13.266**

TRCF ablation. The KITTI rows of Table 3 use fixed single-window cameras on sequences 01, 02, 03, and 08, where all three configurations are available. The temporal-support and TRCF runs share code, configuration, and cameras except for the range switch, and the unfiltered row comes from an earlier run of the same pipeline.

SCR component controls. Table 9 reports controls that replace one component at a time on the region of Table 3, selected among the eight regions for the largest reduction in CD with nondecreasing F1 and Recall. All configurations share fixed cameras, source point identities, candidate normals, and evaluation support. Directional and isotropic bounds are calibrated to the same 73.66% coverage on 448 separate tracks using leave-one-view-out RGB ray intersections, and the isotropic set is 56.43 times larger in volume. The pairwise control replaces each shared track landmark with pair-specific witnesses at the same marginal tolerances, the scoring control scores the same frozen candidates on fitting tracks, and the robust control applies a Huber loss to normalized ray-depth changes under the same accepted hypothesis and feasible domain. On this region, in-sample scoring selects two parallel planes and gives higher CD and lower F1 and Recall than held-out scoring, which selects two planes of different orientation; the other three controls give the same scores as the full method.

Table 9: Component controls on the 7-Scenes Heads region of Table 3. D, J, H, and M denote directional bounds, joint track compatibility, held-out-track scoring, and minimum ray-depth change; CD is in meters and F1 and Recall are percentages at 10 cm.

Configuration	D	J	H	M	CD ↓	F1 ↑	Recall ↑
Unmodified input	–	–	–	–	0.153130	9.494	8.288
Isotropic bounds	×	✓	✓	✓	**0.131895**	**22.315**	**15.647**
Pairwise compatibility	✓	×	✓	✓	**0.131895**	**22.315**	**15.647**
In-sample candidate scoring	✓	✓	×	✓	0.133906	20.319	13.555
Robust-fit objective	✓	✓	✓	×	**0.131895**	**22.315**	**15.647**
Full surface refinement	✓	✓	✓	✓	**0.131895**	**22.315**	**15.647**