R^3: 3D Reconstruction via Relative Regression

Congrong Xu[1,2]* **Huachen Gao**[1] **Xingyu Chen**[1]
Yuliang Xiu[1] **Jun Gao**[2,3] **Anpei Chen**[1]†

[1]Westlake University [2]University of Michigan [3]NVIDIA

Abstract

Recent feed-forward geometry foundation models have demonstrated impressive generalization by recovering depth and poses in a single forward pass. However, these models are typically constrained by a global coordinate frame assumption. This dependency becomes a significant bottleneck for long-context and streaming reconstruction, as it forces the network to maintain an arbitrary temporal origin and handle translation magnitudes that grow unbounded over time. Our solution, which we call R^3, employs relative regression. We employ a lightweight MLP to predict confidence-weighted relative constraints. These confidences serve as a unified anchor: weighting losses during training and guiding pose aggregation during inference. R^3 supports both full-context offline reconstruction and causal, bounded-memory streaming. Our evaluation in both offline and streaming settings validates the effectiveness of our relative mechanism. Project Page: *kevinxu02.github.io/r3-site*

1 Introduction

Feed-forward geometry foundation models have made camera and geometry prediction much more robust. Systems such as DUSt3R, VGGT, π^3, and DA3 [67, 65, 70, 34] are able to recover depth, poses, or pointmaps from images in one forward pass. However, most of these models are built around a global coordinate frame: they either predict pointmaps in a shared frame or regress every camera pose relative to one sequence-level world frame, such as the first camera or a learned canonical frame. This design works well for short offline clips, but it becomes problematic for long videos and streaming reconstruction, where the model must keep a consistent global frame as new views arrive.

Recent work has begun to relax this global-frame assumption. In particular, π^3 [70] replaces the fixed absolute pose loss with relative pose supervision, which reduces the bias toward a hard-coded temporal origin. Yet the output is still a set of global poses in one coordinate system. The model can choose the origin more flexibly, but it still has to represent an entire trajectory in a single frame, and translation magnitudes can grow as the sequence becomes longer. For online reconstruction, this keeps a difficult coordinate-choice problem inside the neural network: future poses are heavily biased toward the frame coordinate initialized from earlier observations.

We suggest that feed-forward 3D reconstruction should learn local relations between views and assemble the global trajectory afterward. Relative pose is a better target for long-context and streaming settings for two simple reasons: its scale depends on the baseline between two frames rather than on the total trajectory length, which keeps the regression target closer to the training distribution as sequences grow; and an N-frame sequence yields many pairwise constraints instead of only N absolute poses. However, raw pairwise prediction is not sufficient: pairs differ in visibility, texture, motion, and baseline, and the model must regress all of them, and global assembly provides

*Research done during an internship at Westlake University.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

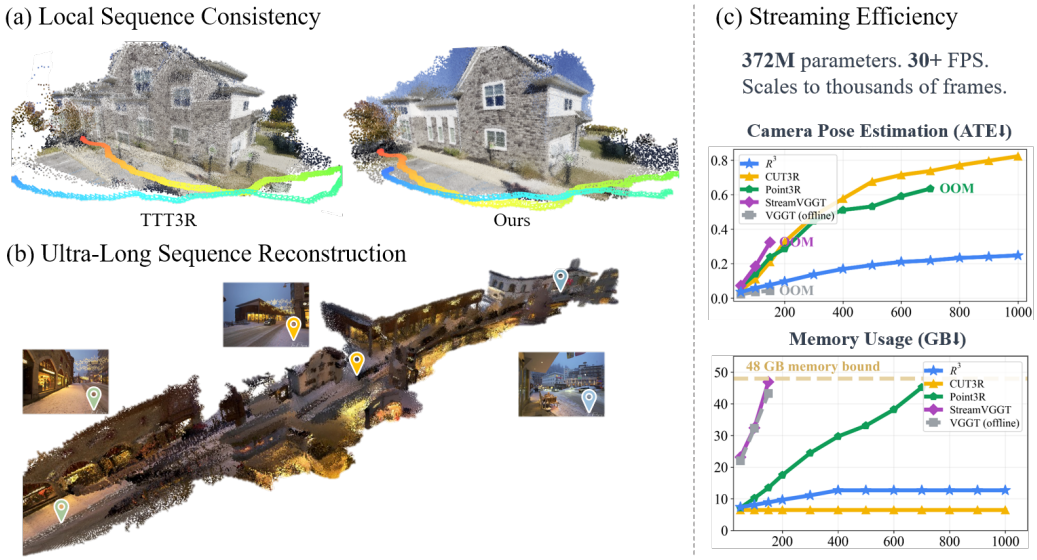


Figure 1: **Consistent, scalable, and efficient streaming geometry via relative pose regression.** R^3 reconstructs camera poses and dense geometry from unbounded video streams via feed-forward relative pose regression. It maintains local consistency, scales to ultra-long sequences with bounded memory, and runs at 20+ FPS with 372M parameters.

no signal for which predictions to trust. To track this issue, we incorporate a confidence modeling for each relative pose, allowing the model to emphasize reliable pairs. This enables any registered frame to serve as an anchor for a new one, with the system selecting the most confident reference from the registered set instead of chaining all poses to a fixed origin.

More specifically, we introduce 3D Reconstruction via Relative Regression (R^3 for short), a feed-forward reconstruction framework based on pairwise relative pose. Fig. 2 contrasts our supervision target with VGGT's first-frame-anchored absolute poses and π^3's model-chosen world frame: R^3 does not ask the network to regress global poses directly, but instead assembles them downstream from pairwise predictions with learned per-pair confidence. Using camera tokens from a DA3 backbone, a lightweight MLP predicts relative rotation, relative translation, and separate confidence scores for rotation and translation. These confidences are used during training as reliability weights and during inference as aggregation weights that fuse pairwise predictions into a consistent trajectory. By focusing on local relations, the model treats absolute poses as a downstream assembly task.

R^3 is a causal, bounded-memory streaming model. Each incoming frame is placed by aggregating confident relative-pose predictions against a confidence-driven keyframe bank that serves as the active context. For offline use, the same checkpoint can run with the causal mask disabled, providing a full-context inference mode without retraining or a second model.

With only 372M parameters and trained on 6 48GB GPUs, R^3 is roughly a third the size of recent 1B-class feed-forward reconstruction models and uses substantially smaller training resources, yet matches or surpasses 1B-class streaming models on pose estimation and dense reconstruction, and remains competitive with offline ones. Compared with state-of-the-art streaming methods [66, 71, 29, 82, 10, 77], it does so while preserving bounded memory usage, even for streams containing thousands of frames.

In summary, we make the following contributions: (i) We reformulate feed-forward 3D reconstruction as a pairwise relative-pose regression task, reducing direct dependence on a fixed global coordinate frame and enabling confidence-weighted global assembly. (ii) We introduce a lightweight MLP that predicts separate rotation and translation confidences, reusing them for loss weighting, streaming pose aggregation, and keyframe-bank management. (iii) We present a single causal architecture that supports both bounded-memory streaming and full-context inference via a simple test-time attention switch, achieving high accuracy across short and long sequences.

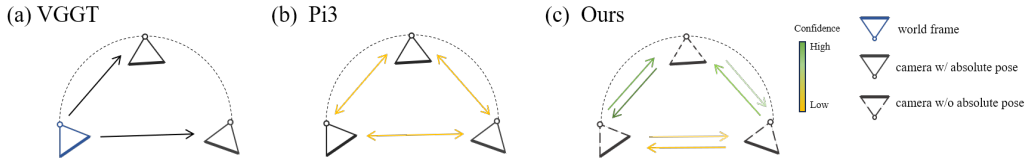


Figure 2: **Three feed-forward pose paradigms, viewed as pose graphs.** Edges denote supervised pairwise pose terms; arrowheads encode directional supervision. (a) VGGT [65] fixes the world frame to the first camera and supervises only edges from this anchor to every other camera. (b) π^3 [70] regresses absolute poses in a model-chosen world frame and supervises every unordered pair with uniform weight. (c) R^3 drops the global-pose head and supervises every directed pair (i, j) with a learned per-edge confidence, yielding a fully-connected directed pose graph.

2 Related Work

Traditional 3D Reconstruction and Learned Back-ends. Classical pipelines split reconstruction into SfM [52, 46], MVS [20, 53, 75], and SLAM [28, 41, 19, 7], with learning gradually folded in via features and matchers [15, 51, 35, 58, 69], learned SfM/VO/SLAM components [64, 17, 60, 61], and learned-prior systems that integrate dense 3D predictions with keyframes, pose graphs, sparse volumes, or trajectory smoothing [43, 39, 37, 62, 59, 31]. We follow the learned-prior direction but keep most of the 3D backbone frozen and use a relative-pose head as a lightweight front-end for aggregation and graph refinement.

Feed-forward 3D Geometry Models. Feed-forward 3D models replace matching, registration, and optimization with direct geometric prediction. DUSt3R/MASt3R [67, 30] made pointmaps a general target, and the paradigm has expanded to multi-view, global, permutation-equivariant, dynamic, and prior- or metric-aware reconstruction [74, 6, 80, 65, 70, 24, 27, 78, 9, 11]. DA3 [34] is the most relevant prior for us: built on DINOv2 [44], it predicts spatially consistent geometry through a unified depth-ray target. Instead of training another foundation model, we ask whether DA3 features can support a control layer for view trust, aggregation, and pair rejection through a small relative-pose head with task-aligned confidence.

Streaming 3D Reconstruction. Feed-forward streaming methods either introduce learned state or explicit memory [63, 66, 71], adapt transformer aggregation through long-context inference, caches, windows, or token-budget management [12, 29, 82, 32, 13, 55, 77, 38], or use test-time training as scene memory [10, 18, 25, 79, 73]. Concurrent with our work, LingBot-Map [8] achieves very strong results with a geometric context transformer combining anchor context, pose-reference windows, and trajectory memory, but is trained on a hundred-GPU cluster with large scale internal datasets, so a fair comparison is unpractical. R^3 is architecturally orthogonal: it introduces no recurrent state, TTT modules, or additional transformers. Instead, it reformulates absolute pose prediction as a system of confidence-weighted relative constraints that aggregate into consistent global trajectories for both streaming and offline inference.

Confidence in 3D Reconstruction. Per-pixel or per-correspondence confidences are standard in pointmap and SLAM models [67, 30, 65, 34, 60, 61], but tied to local uses like depth weighting, correspondence filtering, or optimization residuals. Our contribution is the *role* of confidence: a pairwise pose confidence decoupled into rotation and translation that serves as loss weight, pairwise-aggregation weight, keyframe-bank utility, and outlier gate in streaming, and as per-pair reliability when fusing pairs in full-context refinement. This cross-module reuse lets a thin head on the backbone act as a learned front-end for online reconstruction and offline refinement, without hand-designed information matrices or a new memory architecture.

3 Method

Feed-forward 3D reconstruction from a sequence of N images $\{I_1, \ldots, I_N\}$ aims to recover a consistent global 3D representation of the scene. Our model (R^3) predicts per-frame depth maps $\{D_i\}$ and focal lengths $\{f_i\}$, and derives camera-to-world poses $\{\mathbf{T}_i \in SE(3)\}$ to produce a consistent global point cloud. Unlike prior feed-forward systems

Attributes	**VGGT [65]**	π^3 **[70]**	R^3 **(Ours)**
Relative Pose	✗	✗	✓
Relative Loss	✗	✓	✓
All-pair Edges	✗	✓	✓
Learned Confidence	✗	✗	✓

Table 1: Pose Paradigm Comparison

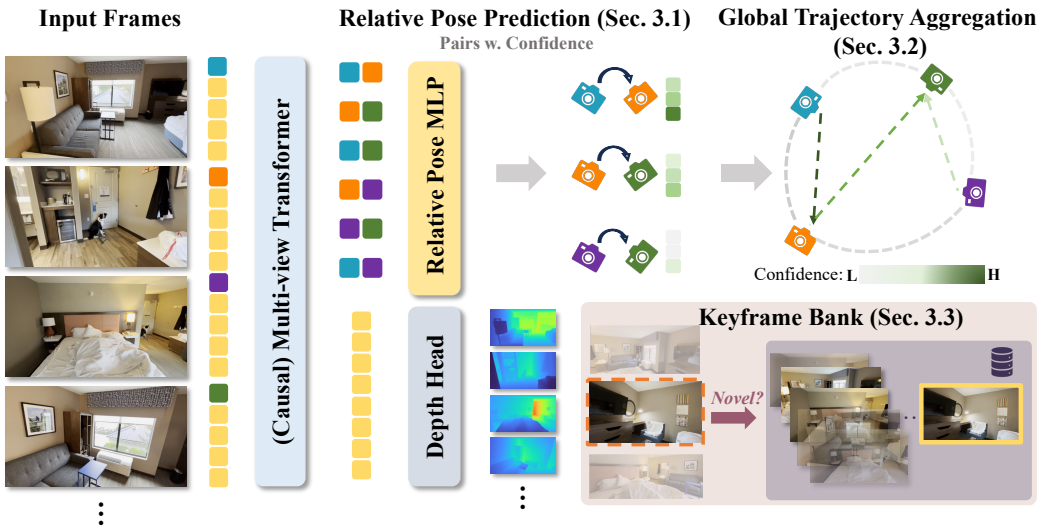


Figure 3: **Overview of R^3.** A causal geometry backbone extracts a single camera token from each frame. A lightweight pairwise pose head then predicts directed relative-pose edges from token pairs, along with separate rotation and translation confidences. These confidence-weighted edges are fused into a coherent trajectory, enabling streaming inference with a bounded active keyframe bank.

that directly regress every $\mathbf{T}_i$ in one global coordinate system [65, 70, 34], R^3 predicts a relative transform and two confidences for every queried frame pair. These predictions form a directed pose graph: frames are nodes, and each queried pair is an edge carrying its relative pose and confidences (Fig. 2c, Tab. 1). We recover camera poses in a common coordinate system by aggregating these predictions. Fig. 3 gives an overview.

3.1 Relative Pose Prediction with Learned Confidence

We realize this pairwise framing with a *pairwise pose head*: a single lightweight MLP, shared across all frame pairs, that sits on top of the DA3 [34] geometry backbone. The backbone produces one latent camera token $\mathbf{z}_i$ per frame. Given two camera tokens $(\mathbf{z}_i, \mathbf{z}_j)$, the pairwise pose head predicts the relative pose from frame i to frame j, together with two scalar confidences:

$$\left(\hat{\mathbf{q}}_{i\to j},\ \hat{\mathbf{t}}_{i\to j},\ c^{\mathrm{R}}_{i\to j},\ c^{\mathrm{T}}_{i\to j}\right) = \mathrm{MLP}_{\mathrm{rel}}([\mathbf{z}_i;\mathbf{z}_j])\,.$$

Here $\hat{\mathbf{q}}_{i\to j}$ is the relative rotation (unit quaternion), $\hat{\mathbf{t}}_{i\to j}$ is the translation of camera j expressed in i's coordinate frame, and $c^{\mathrm{R}}_{i\to j}, c^{\mathrm{T}}_{i\to j} > 0$ are the rotation and translation confidences. Each prediction defines one directed edge (i, j) of the pose graph, weighted by its two confidences. A separate per-frame head predicts the focal length $\hat{f}_i$ from each token $\mathbf{z}_i$.

Relative Versus Absolute Pose Prediction. An absolute pose specifies a camera's position and orientation in a chosen global coordinate system. Relative translation instead describes the displacement between two cameras, so its magnitude depends on their separation rather than on their distance from a fixed origin. Widely separated cameras can still form difficult pairs; the confidence-weighted loss lets the model downweight unreliable predictions. The head also exposes $O(N^2)$ supervised pose constraints from an N-frame training clip while reusing one shared MLP across all queried pairs.

Confidence Design. Rotation and translation exhibit different failure modes: translation is often less reliable when geometric cues are weak, whereas rotation can remain well constrained. A single shared confidence cannot express this asymmetry, so the head predicts separate reliability estimates for rotation and translation, which enter both the training loss (Sec. 3.4) and the trajectory aggregation (Sec. 3.2). App. D compares pose-supervision choices and analyzes the learned confidences as pair-reliability estimates.

3.2 Global Trajectory Aggregation

The pairwise pose head predicts relative poses between cameras. Depth fusion and point-cloud reconstruction require camera poses in a common coordinate system. We obtain these poses by aggregating the pairwise predictions.

In the causal streaming setting, frame 1 defines the origin and orientation of this coordinate system, with $\mathbf{T}_1 = I$. Let $\mathcal{R}_j$ be the set of reference frames available for estimating the pose of a later frame j. Each reference frame i has an already-estimated camera pose. Composing it with the predicted relative pose from i to j gives a candidate pose for j:

$$\mathbf{q}_j^{(i)} = \mathbf{q}_i \otimes \hat{\mathbf{q}}_{i\to j}, \quad \mathbf{t}_j^{(i)} = \mathbf{t}_i + \mathbf{q}_i(\hat{\mathbf{t}}_{i\to j}),$$

where $\mathbf{q}_i(\cdot)$ rotates a vector by $\mathbf{q}_i$. In streaming, the reference set contains the earlier frames in the active context $\mathcal{C}_t$ described in Sec. 3.3.1, so $\mathcal{R}_j = \mathcal{C}_t$.

To reduce the cost of fusion, we can select up to K reference frames from $\mathcal{R}_j$ with the highest average confidence, $\bar{c}_{i\to j} = \frac{1}{2}(c^{\mathrm{R}}_{i\to j} + c^{\mathrm{T}}_{i\to j})$. We denote this selected set by $\mathcal{N}_K(j)$. When $|\mathcal{R}_j| \le K$, all available reference frames are used and $\mathcal{N}_K(j) = \mathcal{R}_j$. This selection only determines which candidate poses contribute to the current fusion; it does not remove frames from the active context. We fuse the selected candidates using separate confidence weights for rotation and translation:

$$\mathbf{t}_j = \sum_{i\in\mathcal{N}_K(j)} \tilde{c}^{\mathrm{T}}_{i\to j}\mathbf{t}_j^{(i)}, \quad \mathbf{q}_j = \mathrm{Norm}\left(\sum_{i\in\mathcal{N}_K(j)} \tilde{c}^{\mathrm{R}}_{i\to j}\mathbf{q}_j^{(i)}\right),$$

where $\tilde{c}^{\mathrm{R}}_{i\to j}, \tilde{c}^{\mathrm{T}}_{i\to j}$ are the softmax normalizations of $c^{\mathrm{R}}_{i\to j}, c^{\mathrm{T}}_{i\to j}$ over $\mathcal{N}_K(j)$, and $\mathrm{Norm}(\cdot)$ renormalizes its argument to a unit quaternion. Candidate quaternions are sign-aligned before averaging. App. E.1 evaluates how the number of selected reference frames affects reconstruction accuracy.

3.3 Causal Streaming and Full-Context Inference

3.3.1 Streaming Inference with an Active Context

In streaming mode, frames arrive sequentially. For each incoming frame j, we maintain a small set of earlier frames as its active context. This set contains frame 1 and a dynamic keyframe bank $\mathcal{B}_t$, giving $\mathcal{C}_t = \{1\} \cup \mathcal{B}_t$. The camera poses of all frames in the active context have already been estimated. We predict the relative pose from each context frame i to frame j, then estimate j's pose using the aggregation described above, with $\mathcal{R}_j = \mathcal{C}_t$.

After its pose has been estimated, frame j is considered for admission to the bank. The following rules determine which frames are added and removed.

Adding Keyframes. Frame j is admitted to $\mathcal{B}_t$ only if its pre-attention backbone token is sufficiently different from the current bank:

$$\max_{i\in\mathcal{B}_t} \cos(\mathbf{tok}_i, \mathbf{tok}_j) < \tau,$$

where $\mathbf{tok}_j$ is the average of j's backbone encoder tokens before any cross-frame interaction, and τ is the novelty threshold. Because the backbone is pretrained for 3D reconstruction, this token provides a cue for geometric and appearance redundancy.

Culling. When $\mathcal{B}_t$ reaches its capacity $M_{\max}$, the entry with the lowest utility $u_j = d_j\, c_j$ is evicted. Here $d_j = \min_{i\in\mathcal{B}_t\setminus\{j\}}\big(1 - \cos(\mathbf{tok}_i, \mathbf{tok}_j)\big)$ is the token-level distinctiveness of frame j from the closest other bank entry, and c_j is the strongest pair confidence between j and the rest of the bank (App. C). The first frame is excluded from eviction.

Like classical keyframe sparsification [28, 42, 41, 19, 7], the bank keeps a compact set of frames, but uses token novelty and pose-head confidence in place of hand-designed motion and covisibility rules.

3.3.2 Full-context Inference

When the complete clip is available, we run the same checkpoint in full-context mode by removing the causal attention mask at test time. The backbone can then see all frames at once, and the pairwise

pose head can be queried on every pair (i, j) with $i \neq j$. We initialize the trajectory with a causal streaming pass, then refine the camera poses using confidence-weighted relative-pose constraints from all pairs. This is a pose-only optimization: it does not perform bundle adjustment, point reprojection, or depth optimization. Details of the keyframe bank and refinement solver are given in App. C.

3.4 Training Objective

The pairwise pose head and depth head are supervised with confidence-weighted residual losses. The total objective is $\mathcal{L} = \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}}$.

Confidence-aware Camera Loss. For each pair (i, j) with $i \neq j$, rotation and translation are supervised by confidence-weighted L_1 residuals:

$$\mathcal{L}_{\text{rot}}(i,j) = c^{\text{R}}_{i\to j}\,\ell^{L_1}_{\text{rot}}(i,j) - \alpha \log c^{\text{R}}_{i\to j}, \quad \mathcal{L}_{\text{trans}}(i,j) = c^{\text{T}}_{i\to j}\,\ell^{L_1}_{\text{trans}}(i,j) - \alpha \log c^{\text{T}}_{i\to j},$$

where $\ell^{L_1}_{\text{rot}}, \ell^{L_1}_{\text{trans}}$ are L_1 residuals against the ground-truth rotation and translation (App. D). In the causal setting, $\mathcal{L}_{\text{cam}}$ averages these terms over past-to-current ordered pairs and adds a plain per-frame L_1 term for the focal-length head.

A large residual makes the product $c\,\ell$ costly, so the optimizer reduces c for inaccurate predictions. The $-\log c$ regularizer prevents confidence from collapsing to zero. Together, these terms encourage higher confidence for accurate predictions and lower confidence for inaccurate ones. Unlike a fixed-weight loss $\lambda_{\text{R}}\ell_{\text{rot}} + \lambda_{\text{T}}\ell_{\text{trans}}$, this lets the rotation and translation weights vary across pairs.

Depth Loss. The depth head is supervised in median-normalized space: each prediction D and target D^* is divided by its own per-frame median to remove scene-scale ambiguity, yielding $\tilde{D}$ and $\tilde{D}^*$. The supervision target D^* depends on the data source: ground-truth depth on synthetic scenes, and the depth output of a frozen pretrained DA3 [34] model on real-world sequences. The loss is

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \Big(\Sigma_p \left| \tilde{D}(p) - \tilde{D}^*(p) \right| - \alpha \log \Sigma_p \Big),$$

where Σ_p is the learned per-pixel depth confidence and the supervision mask $\mathcal{M}$ uses ground-truth validity for synthetic data and DA3-teacher confidence for real-world data.

Architecture updates, optimization schedules, datasets, and other training details are given in App. F.

4 Experiments

We evaluate R^3 across four dimensions: (i) camera pose accuracy, (ii) dense point-map reconstruction quality, (iii) scalability for long sequences, and (iv) robustness to distractor frames.

Unless otherwise specified, all evaluations follow the standard protocols of the respective benchmarks. Our single causal checkpoint runs in the default streaming mode; when the complete clip is available, we also report a lightweight full-context switch obtained by removing the causal mask at test time and running one confidence-weighted pose-graph refinement. To assess scalability, all methods are tested under a fixed 48 GiB GPU memory budget, where "OOM" indicates failure to meet this constraint. Streaming pose and point-map tests retain the full KV cache; long-sequence tests use keyframe selection, and robustness tests use outlier rejection (Tab. 9). Baselines use their released default configurations.

4.1 Camera Pose Estimation

We evaluate camera pose accuracy on Sintel [4], TUM-dynamics [57], and ScanNet [14], comparing streaming R^3 against existing streaming baselines and also reporting full-sequence context results.

As shown in Tab. 2, streaming R^3 provides the strongest overall pose accuracy among streaming methods while using a compact 372M-parameter model. This is substantially smaller than most recent streaming reconstruction baselines with reported model sizes, yet it improves global trajectory accuracy across all three benchmarks, with only one short-term translation metric slightly trailing the best baseline. When the complete sequence is available, the same checkpoint can be switched to

Table 2: **Camera pose estimation on offline (Top) and online (Bottom) settings.** ATE / RPE-T / RPE-R, all lower is better; RPE-R is in degrees. VGGT, π^3, DA3-Large, and R^3 (full context) are full-sequence references. **Bold**/underlined marks best/second-best within each block. † marks offline methods. * indicates values taken from the paper.

Method	#Params	Sintel (50 frames)			TUM-dynamics (90 frames)			ScanNet (90 frames)		
		ATE↓	RPE-T↓	RPE-R↓	ATE↓	RPE-T↓	RPE-R↓	ATE↓	RPE-T↓	RPE-R↓
VGGT† [65]	1.26B	0.172	0.061	0.471	**0.012**	0.010	0.309	0.035	0.015	0.376
π^3† [70]	959M	**0.073**	0.061	**0.288**	0.014	**0.009**	**0.307**	**0.030**	**0.013**	**0.345**
DA3-Large† [34]	385M	0.140	0.059	0.450	0.013	0.011	0.310	0.039	0.017	0.590
$\mathbf{R}^3$ †	**372M**	0.130	**0.047**	0.523	**0.012**	**0.009**	0.318	0.037	0.015	0.472
Spann3R [63]	—	0.329	0.110	4.471	0.056	0.021	0.591	0.096	0.023	0.661
CUT3R [66]	793M	0.213	0.066	0.621	0.046	0.015	0.473	0.099	0.022	0.600
Point3R [71]	—	0.351	0.128	1.822	0.075	0.029	0.642	0.106	0.035	1.946
StreamVGGT [82]	1.26B	0.251	0.149	1.894	0.061	0.033	3.209	0.161	0.057	3.647
STream3R [29]	1.26B	0.213	0.076	0.868	0.026	0.013	0.330	0.052	0.021	0.850
TTT3R [10]	793M	0.201	**0.063**	0.617	0.028	0.012	0.379	0.064	0.021	0.592
ZipMap-stream* [25]	1.40B	0.159	0.065	0.750	—	—	—	—	—	—
$\mathbf{R}^3$	**372M**	**0.115**	0.068	**0.548**	**0.018**	**0.010**	**0.320**	**0.038**	**0.016**	**0.480**

full-context inference and remains competitive with DA3/VGGT references without retraining; π^3, a larger model trained at a substantially larger scale, remains stronger in this setting. The slightly worse ATE on Sintel is due to the outlier sequence `sintel-temple-3`.

4.2 Point-Map Reconstruction

Following CUT3R [66], we evaluate dense reconstruction on 7-Scenes [56] and NRGBD [2] with uniform keyframe sampling (strides 200 and 100). In Tab. 3, streaming R^3 leads on 7-Scenes, showing that the pose estimates translate into accurate geometry under sparse sampling, and remains competitive with the strongest streaming baseline on NRGBD. The full-context switch improves the primary mean Acc/Comp values on both datasets from the same checkpoint.

Table 3: **Sparse-view point-map reconstruction.** Acc and Comp are lower-is-better; NC is higher-is-better. VGGT, DA3-Large, and R^3 (full context) are full-sequence references. **Bold**/underlined marks best/second-best within each block. † marks offline methods.

	7-Scenes [56]						NRGBD [2]					
	Acc↓		Comp↓		NC↑		Acc↓		Comp↓		NC↑	
Method	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
VGGT† [65]	0.087	0.039	0.091	0.039	**0.787**	**0.890**	0.073	**0.018**	0.077	0.021	**0.910**	**0.990**
DA3-Large† [34]	0.086	**0.034**	0.088	0.033	0.756	0.873	0.045	0.020	0.061	0.025	0.896	0.988
$\mathbf{R}^3$ †(full context)	**0.081**	0.043	**0.070**	**0.027**	0.770	0.881	**0.038**	0.020	**0.044**	**0.019**	0.887	**0.990**
Spann3R [63]	0.298	0.226	0.205	0.112	0.650	0.730	0.416	0.323	0.417	0.285	0.684	0.789
CUT3R [66]	0.126	0.047	0.154	**0.031**	0.727	0.834	0.099	0.031	0.076	0.026	0.837	0.971
STream3R [29]	0.122	0.044	0.110	0.038	0.746	0.856	0.057	**0.014**	**0.028**	**0.013**	**0.910**	**0.993**
StreamVGGT [82]	0.129	0.056	0.115	0.041	0.751	0.865	0.084	0.044	0.074	0.041	0.861	0.986
$\mathbf{R}^3$	**0.092**	**0.038**	**0.093**	0.040	**0.755**	**0.868**	**0.047**	0.022	0.050	0.020	0.883	0.991

4.3 Long-Sequence Scalability

We evaluate long-sequence behavior on dense reconstruction of 7-Scenes at sequence lengths up to 1000 frames (Tab. 4) and on ScanNet / TUM-dynamics for pose accuracy (Fig. 4). R^3 processes each stream in a *single pass without reset*. On 7-Scenes, streaming baselines lose Acc/Comp at 1000 frames, while R^3 remains roughly flat. On ScanNet / TUM-dynamics, its ATE grows gradually as views are added but avoids the stronger drift observed in the baselines. We attribute this slower drift to confidence-weighted aggregation over the keyframe bank. Fig. 5 shows the same trend qualitatively on in-the-wild clips. FPS and GPU-memory curves under the same protocol are deferred to App. B.

We further test pose-only trajectory accuracy on a subset of DL3DV-Benchmark [36] (304–439 frames), which contains wider camera baselines and outdoor scenes beyond the mostly indoor TUM-dynamics/ScanNet setting. TTT3R and LongStream [13] are run with the resets of their released default configurations, which outperform no-reset inference; the full protocol is provided in App. H. As shown in Tab. 5, R^3 lowers mean ATE-norm by 59% relative to the strongest streaming baseline, InfiniteVGGT ($2.85\% \rightarrow 1.16\%$).

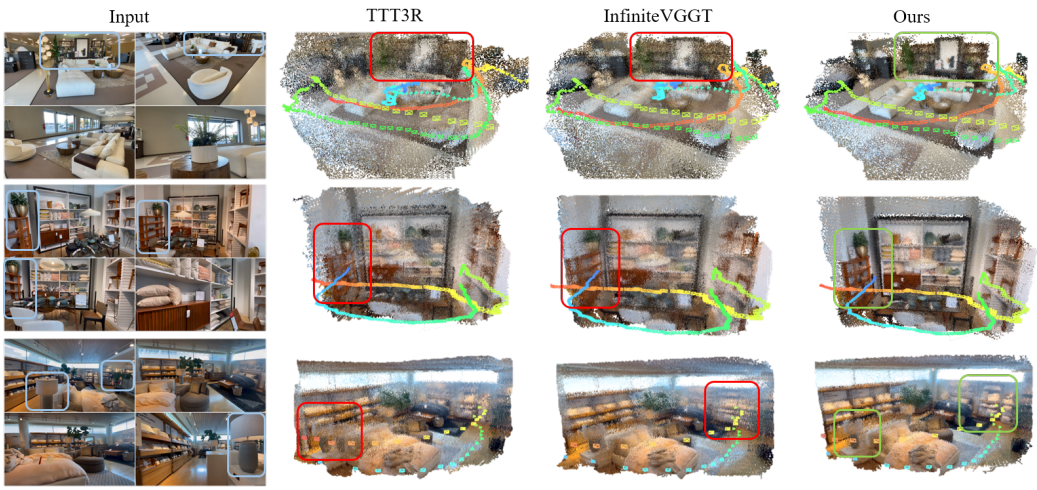


Figure 5: **Long streaming comparison.** Qualitative in-the-wild results show that R^3 maintains more consistent trajectories and point-map alignments over hundreds of frames than baselines.

Table 4: **Long-sequence reconstruction on 7-Scenes.** Mean values at sequence lengths 200/500/1000. Acc / Comp are lower-is-better; NC is higher-is-better. "OOM" marks methods that exceed the 48 GiB GPU memory budget. **Bold**/underlined marks best/second-best per metric column.

	Length 200			Length 500			Length 1000		
Method	Acc↓	Comp↓	NC↑	Acc↓	Comp↓	NC↑	Acc↓	Comp↓	NC↑
Spann3R [63]	0.215	0.122	0.535	0.343	0.154	0.515	0.340	0.154	0.508
CUT3R [66]	0.087	0.045	0.566	0.194	0.092	0.527	0.240	0.102	0.513
Point3R [71]	0.041	0.023	0.579	0.056	0.031	0.555	0.068	0.025	0.533
TTT3R [10]	0.027	0.023	0.582	0.065	0.030	0.552	0.126	0.050	0.525
StreamVGGT [82]	0.038	0.029	0.583	OOM	OOM	OOM	OOM	OOM	OOM
InfiniteVGGT [77]	0.046	0.031	0.582	0.040	0.024	0.561	0.061	0.035	0.537
R^3	**0.021**	**0.018**	**0.600**	**0.022**	**0.017**	**0.562**	**0.022**	**0.017**	**0.538**

Table 5: **Long-sequence camera trajectory accuracy on DL3DV-Benchmark.** ATE-norm / ATE-RMSE / Rot RMSE, all lower is better; ATE-norm is in %, Rot RMSE is in degrees. Win rate is the fraction of 25 scenes with the lowest ATE-norm. †Evaluated with the reset mechanism of its released default configuration; R^3 runs in a single pass without reset. **Bold** marks the best per column.

Method	ATE-norm (%)↓	ATE-RMSE↓	Rot RMSE (°)↓	Win rate↑
TTT3R† [10]	4.91	0.601	5.65	1/25
LongStream† [13]	5.40	0.660	6.03	0/25
InfiniteVGGT [77]	2.85	0.367	3.06	0/25
R^3	**1.16**	**0.155**	**1.79**	**24/25**

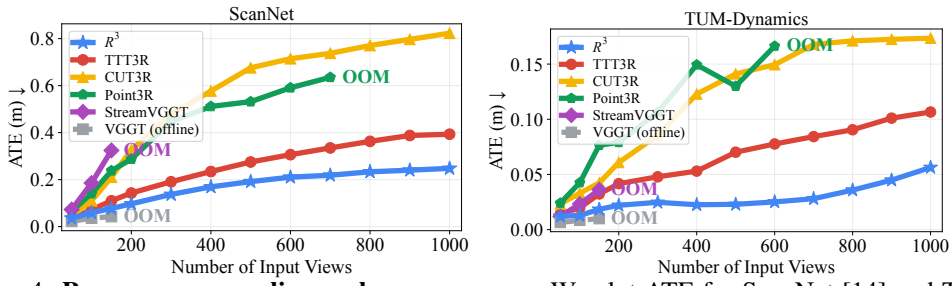


Figure 4: **Pose accuracy scaling on long sequences.** We plot ATE for ScanNet [14] and TUM-dynamics [57] as the number of input frames increases. While several streaming baselines exhibit cumulative drift or trigger out-of-memory (OOM) failures, the error of R^3 grows more slowly.

4.4 Robustness via Confidence Gating

The learned confidences are able to further function as an effective outlier gate. Specifically, when the mean confidence of a new frame j against the active context $\mathcal{C}_t$ falls below a calibrated baseline, we invalidate its KV-cache entries, skip bank admission, and suppress its pose estimation. This mechanism enables R^3 to handle transient failures such as motion blur, occlusions, or sudden scene cuts without polluting the keyframe bank. We provide details on calibration, thresholds, and our segment-reset strategy in App. C.

Table 6: **Robustness with distractors.** Mean over Small/Medium/Large settings (10 seeds each). SR: rejection success; †: offline.

Method	ETH3D [54]		RobustNeRF [50]	
	ATE↓	SR↑	ATE↓	SR↑
RobustVGGT-A†	0.733	0.914	0.138	0.641
RobustVGGT-F†	0.763	0.985	**0.138**	0.586
R^3	0.998	–	0.199	–
$\mathbf{R}^3$ + reject	**0.244**	**1.000**	0.152	**0.986**

Following the Robust-VGGT [22] protocol, we interleave distractor frames into the input stream while maintaining temporal order. As shown in Tab. 6, distractors substantially degrade the base R^3 trajectory (ETH3D ATE 0.998), and the confidence gate reduces the error to 0.244; the improvement therefore comes from rejecting outlier frames from the inference context. The gate uses only the learned pairwise pose confidence, without a separate outlier detector or an outlier-specific training objective. R^3 makes this decision in a single online pass, whereas Robust-VGGT requires the full sequence for a multi-pass offline detection stage.

5 Ablation Study

5.1 Effectiveness of Relative-Pose Formulation

To isolate the intrinsic benefits of the relative-pose formulation, we conduct a controlled comparison against a direct global-pose regression baseline. Within each setting, the two variants share the same initialization, trainable modules, training data, and optimization budget, varying only the prediction target and its associated loss. The streaming block starts from an absolute-pose-pretrained backbone, so its initialization, if anything, favors the direct baseline. The full-context block is initialized from VGGT to match the ablation setting of π^3 [70].

Table 7: **Relative-pose prediction objective.** ATE / RPE-T / RPE-R, all lower is better. Within each block, the two variants share the same initialization, trainable modules, training data, and optimization budget, and differ only in the pose prediction target and loss. **Bold** marks the best result within each block. † marks full-sequence methods.

	Sintel [4]			ScanNet [14]		
Variant	ATE↓	RPE-T↓	RPE-R↓	ATE↓	RPE-T↓	RPE-R↓
Full-context (VGGT init, 64k iters)						
Backbone direct (Pi3 loss)†	0.1583	0.0841	0.9492	0.0683	0.0195	0.5143
Backbone + our rel. target/loss†	**0.1370**	**0.0757**	**0.7282**	**0.0604**	**0.0189**	**0.5039**
Streaming (5k iters, abs-pose-pretrained backbone, pose heads reset)						
Backbone direct (VGGT abs loss)	**0.141**	0.079	1.56	0.063	0.026	1.33
Backbone + our rel. target/loss	0.146	**0.066**	**0.67**	**0.052**	**0.021**	**0.93**

As shown in Tab. 7, the relative formulation improves most metrics in both full-context and streaming scenarios, especially RPE and ScanNet ATE; the direct baseline remains slightly better on Sintel streaming ATE. Overall, the results indicate that supervised relative-pose regression is a strong geometric prior for long-range sequences, even before any complex aggregation is applied. Additional ablations on full-attention fine-tuning, aggregation strategy, and keyframe admission are in App. E.

5.2 Relative-Pose Prediction on Long Trajectories

We freeze the DA3 backbone and train only a relative-pose head to test how accurately relative poses can be read from features learned for absolute-pose prediction (*DA3 frozen + rel. head*). In R^3, we also adapt the backbone under relative-pose supervision. To study how motion errors grow on sequences longer than those seen during training, we compare both models with *DA3 absolute*, a causal model fine-tuned to predict poses in the first-frame coordinate system. All three models are trained on the same data until convergence. We stream 300 frames from each of 5 KITTI-360 [33] scenes, unseen during training. At each prefix length $k \geq 10$, we evaluate the relative pose of the latest adjacent pair, $(k-1, k)$, after Sim(3) alignment using frames $1..k$. This isolates the local motion errors that can contribute to trajectory drift as the camera moves farther from the first frame.

Table 8: **Adjacent-pair translation error on KITTI-360.** Means over 5 scenes and prefix lengths in each interval. Distance is the mean GT distance to the first frame; distances and errors are in meters.

Frames	Distance to first frame	DA3 absolute	DA3 frozen + rel. head	R^3
10–50	14	0.13	0.16	**0.12**
51–100	40	0.59	0.69	**0.38**
101–150	67	1.88	3.10	**0.83**
151–200	89	2.55	2.77	**1.58**
201–250	111	2.80	1.73	**1.56**
251–300	146	3.92	3.88	**1.86**

R^3 achieves the lowest translation error across all six sequence-length ranges (Tab. 8). Its advantage over the frozen readout shows the benefit of adapting DA3 features for relative-pose prediction. Compared with DA3 absolute, R^3 halves the fitted rate of adjacent-pair translation error growth (0.76 vs. 1.52 m per 100 frames). The 99th-percentile rotation error drops from 72° to 12°, and the maximum from 171° to 29°.

6 Conclusion

We presented R^3, a feed-forward 3D reconstruction framework that reformulates global-frame regression as pairwise relative-pose regression. Our approach employs a lightweight MLP to predict relative motion alongside decoupled confidences for rotation and translation. This confidence signal acts as a unified primitive: it weights training losses, guides pose aggregation at inference, and drives keyframe-bank management. Experimental results demonstrate that R^3 is competitive with the baselines, even on sequences exceeding thousands of frames. By offloading the coordinate-frame choice to a relative pose aggregation step, R^3 suggests that local geometric relations are a natural and efficient learning target for feed-forward 3D reconstruction.

Acknowledgments

We gratefully thank Siyuan Bian and Weiwei Xu for their insightful discussions, and the members of Jun Gao Lab and Inception3D Lab for their support throughout this project. This work was supported by NSFC program 62606405.

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision (ECCV)*, 2022.

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, June 2022.

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[4] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012.

[5] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020.

[6] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[7] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, J. 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.

[8] 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.

[9] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2503.24391.

[10] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.

[11] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.06219.

[12] Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. LONG3R: Long sequence streaming 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2507.18255.

[13] Chong Cheng, Xianda Chen, Tao Xie, Wei Yin, Weiqiang Ren, Qian Zhang, Xiaoyang Guo, and Hao Wang. LongStream: Long-sequence streaming autoregressive visual geometry. *arXiv preprint arXiv:2602.13172*, 2026.

[14] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[16] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2024.

[17] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. MASt3R-SfM: A fully-integrated solution for unconstrained structure-from-motion. In *International Conference on 3D Vision (3DV)*, 2025.

[18] Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3D reconstruction at scale. *arXiv preprint arXiv:2602.23361*, 2026.

[19] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(3):611–625, 2018.

[20] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(8):1362–1376, 2010.

[21] 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 (CVPR)*, 2012.

[22] Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. *arXiv preprint arXiv:2512.04012*, 2025.

[23] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jérôme Revaud. Pow3R: Empowering unconstrained 3D reconstruction with camera and scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1071–1081, 2025.

[25] Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T. Barron, Noah Snavely, and Aleksander Hołyński. ZipMap: Linear-time stateful 3D reconstruction via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.04385.

[26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[27] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

[28] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.

[29] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

[30] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024.

[31] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Hołyński, and Noah Snavely. MegaSaM: Accurate, fast, and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10486–10496, 2025.

[32] Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. WinT3R: Window-based streaming reconstruction with camera token pool. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.05296.

[33] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3292–3310, 2023.

[34] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[35] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17627–17638, 2023.

[36] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[37] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.09401.

[38] Si-Yu Lu, Po-Ting Chen, Hui-Che Hsu, Sin-Ye Jhong, Wen-Huang Cheng, and Yung-Yao Chen. OVGGT: O(1) constant-cost streaming visual geometry transformer. *arXiv preprint arXiv:2603.05959*, 2026.

[39] Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

[40] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[41] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[42] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[43] 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 (CVPR)*, 2025. arXiv:2412.12392.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[45] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[46] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

[47] Project Aria. Aria synthetic environments dataset. `https://www.projectaria.com/datasets/ase/`, 2024. Meta Reality Labs Research.

[48] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[49] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[50] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, and Andrea Tagliasacchi. RobustNeRF: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20626–20636, June 2023.

[51] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[52] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[53] Johannes Lutz Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[54] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[55] You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Training-free acceleration of visual geometry transformer. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.02560.

[56] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[57] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[58] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[59] Marwan Taher, Ignacio Alzugaray, Kirill Mazur, Xin Kong, and Andrew J. Davison. KV-Tracker: Real-time pose tracking with transformers. *arXiv preprint arXiv:2512.22581*, 2025.

[60] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[61] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[62] Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. *arXiv preprint arXiv:2511.20343*, 2025.

[63] Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, 2025.

[64] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

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

[67] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[68] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[69] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[70] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.13347.

[71] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[72] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. RGBD objects in the wild: Scaling real-world 3D object learning from RGB-D videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[73] Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable test-time training for large-scale 3D reconstruction. *arXiv preprint arXiv:2604.08542*, 2026.

[74] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[75] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2018.

[76] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[77] Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026.

[78] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.03825.

[79] 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. *arXiv preprint arXiv:2603.03269*, 2026.

[80] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. FLARE: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21936–21947, 2025.

[81] Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. OmniWorld: A multi-domain and multi-modal dataset for 4D world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

[82] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.11539.

A Reconstruction Gallery



Figure 6: **Reconstruction gallery.** Qualitative reconstruction results from R^3 in streaming mode across diverse indoor and outdoor scenes. The reconstructed point clouds remain geometrically coherent and visually consistent across varied scene layouts, object scales, and camera trajectories, demonstrating R^3's ability to maintain stable scene structure during online reconstruction of long sequences.

B Long-Sequence Compute Cost

Fig. 7 reports inference FPS and GPU memory usage under the same 7-Scenes protocol used in Sec. 4.3. Global-regression baselines (e.g., StreamVGGT) either hit OOM or slow sharply as N grows; R^3 replaces this $O(N^2)$ growth with a bounded memory increase and a gentler FPS decline, consistent with the bounded keyframe bank.

C Streaming Details and Optional Pose-Graph Refinement

This appendix provides additional details on the streaming machinery of R^3: its connection to classical keyframe selection, the train/inference length-gap argument, outlier rejection, pose-graph

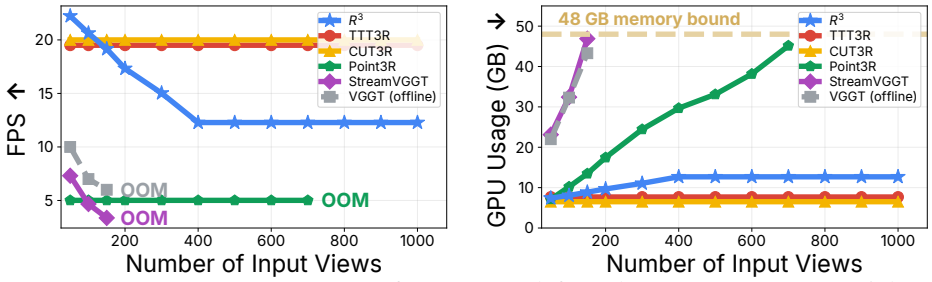


Figure 7: **Long-sequence compute cost.** Inference FPS (left) and GPU memory usage (right) as the number of input views increases. "OOM" denotes out-of-memory.

refinement for full-context inference and optional streaming cleanup, and segment resets for long streams. The core learned mechanism—the confidence-driven keyframe bank—is described in the main text (Sec. 3.3.1); the components below are auxiliary and follow common practice. Tab. 9 summarizes which of them are enabled in each reported experiment.

Table 9: **Inference mechanisms used in each experiment.** All rows use relative-pose prediction with confidence-weighted aggregation; each experiment enables at most one additional mechanism to isolate its effect. Without keyframe selection, the full KV cache is kept. Within each benchmark, all sequences share one fixed configuration with no per-sequence tuning.

Experiment	Keyframe τ	Outlier reject	Segment reset	Pose-graph ref.
Pose & point map (Tabs. 2, 3)	✗	✗	✗	✗
Long sequence (Tabs. 4, 5; Fig. 4)	✓	✗	✗	✗
Robustness (Tab. 6)	✗	✓	✗	✗
Full-context (R^3-Offline rows)	n/a	✗	✗	✓

Comparison to classical keyframe selection. $\mathcal{B}_t$ plays the role of a *keyframe database* in classical SLAM systems such as PTAM [28], ORB-SLAM [42, 41, 7], and DSO [19]: it is a compact subset of frames carried through the pipeline instead of the full stream. Classical systems trigger keyframe insertion using hand-tuned translation, rotation, or covisibility thresholds. Our admission rule replaces these cues with a learned-feature similarity test plus a maximum-staleness cap, leveraging the fact that the DA3 [34] frame-only encoder is pretrained on large-scale multi-view geometry, so its averaged frame token already reflects both visual content and scene-level geometry.

Train/inference length gap. The novelty gate keeps the active context informative without increasing memory unboundedly and slows long-term saturation, helping bridge the train/inference length gap: we train on sequences of at most 32 frames but deploy on streams of hundreds. A strict novelty threshold keeps the effective context length close to the training regime even as wall-clock sequence length grows. To prevent staleness when scene content evolves slowly and no incoming frame is novel enough to pass the threshold, we force-admit a frame whenever no keyframe has been added in the last $\Delta_{\max} = 20$ incoming frames. The culling cap is $M_{\max} = 100$. When this cap is exceeded, we evict the keyframe with the lowest utility

$$u_j = d_j c_j, \qquad e_{ij} = \min(i,j) \to \max(i,j),$$
$$d_j = \min_{i \in \mathcal{B}_t \setminus \{j\}} \big(1 - \cos(\mathbf{tok}_i, \mathbf{tok}_j)\big),$$
$$c_j = \max_{i \in \mathcal{B}_t \setminus \{j\}} \tfrac{1}{2}\big(c^{\mathrm{R}}_{e_{ij}} + c^{\mathrm{T}}_{e_{ij}}\big).$$

Here e_{ij} is the edge evaluated when the later of the two frames arrived. Since evicted frames never re-enter the bank, the earlier frame was in the active context at that time, so this edge is always available and culling evaluates no additional edges. d_j is small for frames that are redundant with another bank entry, and c_j is small when the frame has no strong pose edge to the bank. The first frame is excluded from eviction.

Outlier rejection and segment reset. Let $\bar{c}_j = \frac{1}{|\mathcal{C}_t|} \sum_{i \in \mathcal{C}_t} \bar{c}_{i \to j}$ denote the mean averaged-pair confidence of frame j against its current context. We calibrate a baseline $\bar{c}^{\mathrm{ref}} = \frac{1}{N_{\mathrm{cal}}} \sum_{j=1}^{N_{\mathrm{cal}}} \bar{c}_j$

from the first N_{cal} frames of the stream and reject any subsequent frame whose mean falls below $\tau_{\text{out}}\,\bar{c}^{\text{ref}}$. This check is post-hoc: rejected frames have already passed through the backbone because their tokens are needed to score the pair head. Upon rejection, however, their entries are evicted from the backbone KV cache so they do not pollute later causal attention; they also do not enter $\mathcal{B}_t$, contribute no edges to $\mathcal{E}_t$, or receive an estimated absolute pose. The threshold is loose enough to retain difficult but recoverable frames while filtering catastrophic cases such as severe motion blur, near-total occlusion, or scene cuts mid-frame. The same test also serves as a tracking-failure detector: after N_{rej} consecutive rejected frames, we declare loss of track and trigger a segment reset (below) rather than continue with a degraded context. The keyframe-bank cap $M_{\max}$ is a secondary bound that rarely fires before this confidence trigger in practice. We use $N_{\text{cal}} = 3$, $\tau_{\text{out}} = 0.15$, and $N_{\text{rej}} = 3$ in our experiments.

Optional pose-graph refinement for streaming runs. The full-context mode (Sec. 3.3) closes with a confidence-weighted pose-graph optimization over the full pair edge set. The same step can be invoked as an end-of-stream refinement on a streaming run — using the streaming trajectory as initialization and the edge set $\mathcal{E}$ restricted to pairs that were evaluated against the keyframe bank during the stream — or at every segment boundary for jobs that tolerate occasional latency spikes in exchange for tighter mid-stream geometry. We do not apply this refinement to any reported streaming main-body number; it appears only as an ablation. Given an edge set $\mathcal{E}$ and an aggregated initialization $\{\mathbf{T}_i\}$, let e_{ij}^{R} and e_{ij}^{T} denote the rotation and translation residuals between the relative pose induced by $(\mathbf{T}_i, \mathbf{T}_j)$ and the predicted edge $(\hat{\mathbf{q}}_{i\to j}, \hat{\mathbf{t}}_{i\to j})$. We solve

$$\min_{\{\mathbf{T}_i\}} \sum_{(i,j)\in\mathcal{E}} \Big(c_{i\to j}^{\text{R}}\, H_{\delta_{\text{R}}}(e_{ij}^{\text{R}}) + c_{i\to j}^{\text{T}}\, H_{\delta_{\text{T}}}(e_{ij}^{\text{T}}) \Big),$$

with Huber losses $H_{\delta_{\text{R}}}, H_{\delta_{\text{T}}}$, $\mathbf{T}_1$ fixed, and L-BFGS as the solver. This is a lightweight pose-only refinement: edges carry only relative-pose residuals and scalar confidences, with no bundle adjustment, point reprojection, or depth optimization.

Confidence for loop-closure retrieval. R^3 has no loop closure, but its pair confidence can rank candidate frames for one. On DL3DV, we rank all past frames of a query by pair confidence and measure the point-visibility IoU of the top-ranked frame relative to the best achievable one, computed from COLMAP [52] reconstructions. Over 754 queries, the top-confidence frame reaches a median of 98.8% of the best IoU and exceeds 90% in 90.5% of cases. This covisibility proxy ignores occlusion, but indicates that the confidence is a useful retrieval signal for a downstream loop-closure back-end.

Segment reset and bridges. A segment ends either when sustained low confidence triggers the reset above or when the sequence length cap $L_{\max}$ is reached on a very long stream. In both cases, we link consecutive segments with a short *bridge* of 3–10 shared frames. When advancing to a new segment, we clear the DA3 cache and keyframe bank, then rerun the backbone on the bridge frames together with the incoming stream.

Global pose continuity follows by composition: each bridge frame b carries an absolute pose $\mathbf{T}_b^{\text{abs}}$ from the previous segment, and for each new frame j we recover $\mathbf{T}_j^{\text{abs}} = \mathbf{T}_b^{\text{abs}} \cdot \mathbf{T}_{b\to j}$, without requiring Sim(3) or $\mathrm{SE}(3)$ registration across segments.

Scale can drift between independent forward passes. We anchor it by aligning the DA3 depth prediction on each segment's first frame to a pretrained metric-depth model with a single median-based scalar, which is applied uniformly to all pose translations and depths in that segment. Per-boundary error is therefore reduced to one metric-anchored scalar rather than the joint rotation, translation, and scale drift of Sim(3)-style window fusion.

D Pose Supervision and Learned Pair Reliability

D.1 Pose Loss Formulations

Fig. 2 compares three feed-forward pose paradigms; here we spell out their corresponding supervision targets. Let $\mathbf{T}_i = (\mathbf{q}_i, \mathbf{t}_i)$ denote the camera-to-world pose of frame i, and let $\mathbf{T}_{i\to j} = \mathbf{T}_i^{-1}\mathbf{T}_j$ denote the relative transform from frame i to frame j.

VGGT-style training [65] supervises poses directly in one canonical coordinate system. Abstractly, with frame 1 used as the global anchor, the loss takes the form

$$\mathcal{L}_{\text{abs}} = \sum_{j=1}^{N} d_{\text{pose}}\Big(\hat{\mathbf{T}}_j, \mathbf{T}_j^*\Big)\,,$$

where all targets $\mathbf{T}_j^*$ are expressed in the same anchored world frame. This target is simple, but the network must learn a coordinate choice that is arbitrary from the scene's perspective and increasingly difficult to preserve as the trajectory grows.

π^3 [70] moves away from VGGT's fixed first-frame coordinate target: it predicts one pose per view and supervises them through all-pair relative-pose losses:

$$\mathcal{L}_{\pi^3} = \sum_{i \neq j} d_{\text{pose}}\Big(\hat{\mathbf{T}}_i^{-1}\hat{\mathbf{T}}_j, \mathbf{T}_{i\to j}^*\Big)\,.$$

This removes the need to match a fixed first-frame coordinate system at the loss level and provides $O(N^2)$ pairwise constraints. However, it is still not ideal for online reconstruction. The all-pair loss is defined over a complete set of views, whereas a causal stream must emit estimates from prefixes and cannot use future observations unless it revises past outputs. In addition, every pair contributes with the same weight, so difficult or weakly constrained pairs can consume the same gradient budget as reliable ones.

Our formulation predicts each relative transform directly and attaches separate learned confidences to rotation and translation:

$$\big(\hat{\mathbf{q}}_{i\to j}, \hat{\mathbf{t}}_{i\to j}, c_{i\to j}^{\text{R}}, c_{i\to j}^{\text{T}}\big) = \text{MLP}_{\text{rel}}([\mathbf{z}_i; \mathbf{z}_j])\,.$$

The pose loss uses confidence-weighted residuals. Omitting the per-frame focal-length term for clarity, we define

$$\mathcal{L}_{\text{rot}}(i,j) = c_{i\to j}^{\text{R}}\,\ell_{\text{rot}}^{L_1}(i,j) - \alpha \log c_{i\to j}^{\text{R}}, \qquad \mathcal{L}_{\text{trans}}(i,j) = c_{i\to j}^{\text{T}}\,\ell_{\text{trans}}^{L_1}(i,j) - \alpha \log c_{i\to j}^{\text{T}},$$

and average over a set of ordered pairs,

$$\mathcal{L}_{\text{cam}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \Big(\mathcal{L}_{\text{rot}}(i,j) + \mathcal{L}_{\text{trans}}(i,j)\Big).$$

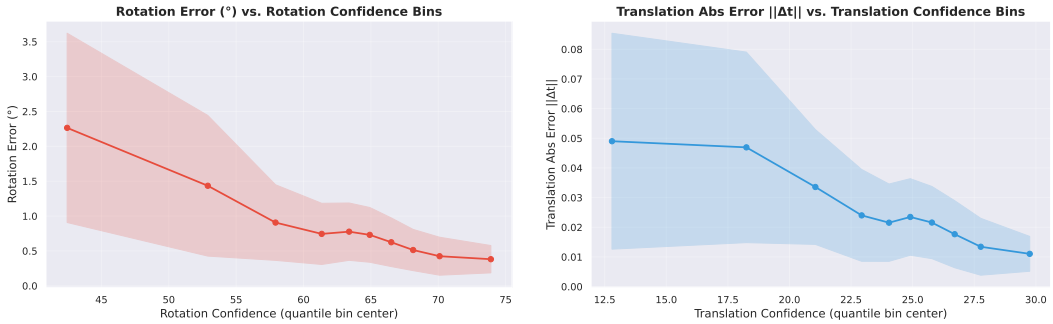
For the causal checkpoint, $\mathcal{P}$ is the past-to-current lower-triangular pair set described in App. F; the all-pair case corresponds to $\mathcal{P} = \{(i,j) : i \neq j\}$. Here c^{R} and c^{T} weight the rotation and translation residuals separately, while the $-\log c$ terms prevent the model from assigning uniformly low confidence. This gives the model a way to downweight poorly constrained pairs during training and then reuse the same signal as aggregation weights and online keyframe-bank utilities at inference (and as edge weights in the optional pose-graph refinement of App. C). The next diagnostic asks whether this training signal also behaves as geometric pair reliability.

D.2 Learned Confidence as Pair Reliability

To check whether the predicted rotation and translation confidences act as learned pair-reliability estimates rather than only loss weights, we aggregate all image pairs from the first 20 ScanNet test scenes [14] and bin them by the corresponding predicted confidence(Fig. 8).

For both components, predicted confidence is monotone in the corresponding error: mean error decreases from low to high confidence bins, and the within-bin spread shrinks accordingly. The trend is sharpest at the low-confidence end, where unreliable pairs carry markedly larger residuals and wider variance. This is the property needed for the three downstream uses of confidence in our system: as a per-component loss weight in the Laplacian likelihood, as an edge weight in the PGO back-end, and as a frame-utility signal in the online keyframe bank.

The two panels also expose an asymmetry between components. Rotation confidence concentrates in a higher numerical range and reaches lower binned error than translation confidence, consistent with rotation being the easier learning target while translation remains more sensitive to baseline, scale, and parallax. A single shared pose confidence would have to average over this asymmetry, either over-trusting translation on rotation-easy pairs or under-using a reliable rotation edge when translation is ambiguous. Predicting the two confidences separately lets the model express this difference directly and lets each downstream consumer read the reliability of the component it actually depends on.



(a) Rotation error binned by rotation confidence. (b) Translation error binned by translation confidence.

Figure 8: **Learned confidence behaves as pair reliability.** Across all pairs from the first 20 ScanNet scenes, pairs are grouped into equal-mass confidence quantile bins; the x-axis is the bin center. The solid line shows the per-bin mean pose error and the shaded band shows the within-bin dispersion, so each polyline reports the average error and spread within that confidence quantile interval. Higher predicted confidence corresponds to lower mean error and tighter spread in the corresponding component.

Table 10: **Full-attention fine-tuning reference.** Pose ATE (lower is better). DA3-Large is the original backbone trained with absolute-pose supervision and substantially more data; R^3 uses the relative-pose objective. The comparison is included as a calibration reference, not as a controlled ablation.

Variant	Sintel	TUM-dyn.	ScanNet
DA3-Large [34]	0.140	0.013	0.039
R^3 (full-attention fine-tuned)	0.117	0.011	0.035

E Additional Ablations

Full-attention fine-tuning reference. We compare a full-attention fine-tuned version of our model against the original DA3-Large [34] backbone (Tab. 10). This is not a controlled ablation: DA3-Large is trained with substantially more data, whereas our variant uses the relative-pose objective from Sec. 5.1. We therefore treat DA3-Large as a backbone calibration reference rather than a matched baseline.

The full-attention fine-tuned variant improves over the original backbone on these pose metrics despite using a relative-pose target rather than DA3's direct camera-pose supervision.

Backbone vs. formulation. Tab. 11 separates the streaming gain into a backbone step and a formulation step. Replacing STream3R with a causal DA3 backbone trained with an absolute-pose objective improves ATE, and switching this model to our relative formulation improves it further on all three benchmarks.

Table 11: **Backbone vs. formulation in streaming.** ATE on Sintel / TUM-dynamics / ScanNet, lower is better. Tab. 7 is the controlled attribution study, and its streaming variants are trained separately from the middle row.

Configuration	Sintel	TUM-dyn.	ScanNet
STream3R [29] (absolute pose)	0.213	0.026	0.052
Causal DA3 + absolute-pose objective	0.155	0.020	0.049
Causal DA3 + relative formulation (R^3)	**0.115**	**0.018**	**0.038**

Table 12: **Aggregation strategy** (left) and **acceptance threshold** τ (right). Left: ATE on Sintel, TUM-dynamics, and ScanNet for top-k confidence-weighted aggregation, all-pair averaging, and an optional pose-graph optimization (PGO) on top. Right: ScanNet ATE at 100/300/500/800/1000 input frames for different τ. **Bold**/underlined marks best/second-best per column.

Variant	TUM	Sintel	ScanNet	Avg.
Top-1	0.0199	0.1272	0.0401	0.0624
Top-5	0.0188	0.1177	0.0392	0.0586
Top-10	0.0183	0.1153	0.0387	0.0574
All-avg	<u>0.0179</u>	<u>0.1152</u>	<u>0.0384</u>	<u>0.0572</u>
+ PGO	**0.0179**	**0.1142**	**0.0381**	**0.0567**

τ	100	300	500	800	1k
0.96	<u>0.066</u>	0.200	0.270	0.334	0.357
0.97	**0.061**	0.180	0.250	0.306	0.332
0.98	**0.061**	**0.171**	**0.245**	**0.295**	**0.318**
0.99	0.068	<u>0.172</u>	<u>0.246</u>	<u>0.300</u>	<u>0.329</u>

E.1 Aggregation and Selection Heuristics

We further investigate sensitivity to the aggregation strategy and the keyframe admission threshold τ, which together govern the density and reliability of the reconstructed trajectory.

Influence of neighborhood size K. We evaluate the impact of the number of reference frames K used during pose fusion (Tab. 12, left). Accuracy improves as K grows, and all-pair averaging gives the best closed-form ATE across the three benchmarks; top-K aggregation is therefore best understood as an efficiency cap that trades a small amount of accuracy for lower fusion cost. Adding an optional confidence-weighted pose-graph optimization on top of the all-pair aggregate yields a further small but consistent improvement, indicating that the same confidence also serves as a useful per-pair weight in a global solve.

Keyframe admission threshold τ. The threshold τ controls the feature novelty required for a frame to enter the keyframe bank. Since admission requires the maximum token similarity to fall below τ, a lower threshold is more selective and a higher threshold is more permissive. Empirical results on ScanNet (Tab. 12, right) identify $\tau = 0.98$ as the best balance across varying sequence lengths: $\tau = 0.96$ can starve the bank in long-horizon streams, while $\tau = 0.99$ admits more redundant frames and noisy edges.

Pose-distance admission. Token novelty can be complemented by camera distance, which MUSt3R [6] and AMB3R [62] use for keyframe admission. In our preliminary experiments, this rule gives a slight improvement on DL3DV-Benchmark, suggesting that explicit geometric cues can further improve spatial coverage of the keyframe bank.

F Training Details

Architecture and trainable parameters. We build on the pretrained DA3-Large [34] backbone and keep most of it frozen: only the global-attention layers, the focal-length head, and the newly introduced pairwise relative-pose/confidence head are updated, yielding roughly 110M trainable parameters out of the 372M-parameter inference model. The focal-length head is a small MLP that maps each camera token $\mathbf{z}_i$ to a single scalar $\hat{f}_i$ and is supervised by a plain L_1 residual against the ground-truth focal length. Training uses frame-causal global attention. The same checkpoint is used for both streaming and full-context inference; the latter removes the causal mask at test time so global attention runs bidirectionally over the full clip, then applies the lightweight pose-graph refinement described in App. C.

Implementation details. The model is trained on sequences of 4–32 views. Most batches use clips of up to 16 views, and the relative-pose stage extends the curriculum to 32 views in the second half of training. All training is done on 48 GB GPUs. To fit long streams in this memory budget, we use gradient checkpointing, `bfloat16` mixed precision, and a FlexAttention [16] implementation of frame-causal attention, which lets us pack about 200 views per 48 GB GPU without the depth teacher and 96 views per 48 GB GPU when the DA3 teacher is enabled. For pose supervision, we evaluate the relative-pose head on the causal lower-triangular set of past-to-current ordered pairs after the DA3 forward pass. This provides $O(N^2)$ supervised relative edges during training without a comparable

Table 13: **Training datasets.** Counts denote the number of scenes or sequences used for training after dataset-specific filtering.

Dataset	Count	Type
AriaSyntheticENV [47]	2,000	Synthetic
ARKitScenes [3]	4,388	LiDAR/RGB-D
CO3Dv2 [48]	30,616	COLMAP
DL3DV [36]	6,379	COLMAP
HyperSim [49]	344	Synthetic
MapFree [1]	921	COLMAP
MVS-Synth [23]	121	Synthetic
OmniWorld-Game [81]	484	Synthetic
ScanNet [14]	1201	LiDAR/RGB-D
ScanNet++ [76]	230	LiDAR/RGB-D
Spring [40]	47	Synthetic
TartanAir [68]	355	Synthetic
Dynamic Replica [26]	524	Synthetic
Virtual KITTI 2 [5]	50	Synthetic
WildRGBD [72]	23,050	LiDAR/RGB-D

backbone cost, because the all-pair computation is only an MLP over compact camera tokens and is negligible relative to image-token attention. At inference time, streaming does not materialize all historical pairs: each incoming frame is paired only with the bounded active context $\mathcal{C}_t$, so per-frame cost is bounded by the keyframe bank. Confidence logits are passed through a softplus nonlinearity to ensure positive confidence values.

Optimization schedule. The model is optimized with AdamW in two stages that update the same parameter groups listed above. *Stage 1* adapts the DA3-Large backbone into a frame-causal absolute-pose checkpoint for 15k iterations at a constant learning rate of $1\mathrm{e}{-4}$ with a dynamic batch size of 4–16 views. *Stage 2* initializes from this checkpoint and trains with the relative-pose objective for 25k iterations; it updates the global-attention layers, the focal-length head, and the pairwise relative-pose/confidence head, not the relative-pose head alone. Its learning rate starts at $1\mathrm{e}{-4}$ and decays with a cosine schedule to $1\mathrm{e}{-5}$, with gradient accumulation of 2 and a dynamic batch size that begins at 4–16 views and is extended to 4–32 views in the second half of the schedule.

Training datasets. We train on a diverse collection of 15 publicly available datasets spanning synthetic scenes, LiDAR/RGB-D captures, and COLMAP reconstructions. Table 13 reports the number of training scenes or sequences retained after dataset-specific filtering. For AriaSyntheticENV, we use the first 2,000 scenes, and for OmniWorld we use the OmniWorld-Game subset.

G Robustness Details

Evaluation protocol. We follow the Robust-VGGT [22] controlled-distractor protocol with two modifications for streaming evaluation. (i) We preserve the dataset's original frame order: distractor frames are interleaved into the clean stream rather than reshuffled, so R^3 is exposed to distractors at arbitrary points along the trajectory. (ii) We keep the first three frames clean as a calibration prefix; this gives the streaming model a short, distractor-free initialization before the gating decision begins. For each trial we sample N_c clean images from a single scene and N_n distractor images drawn uniformly at random from other scenes in the same dataset, with disjoint pools and scene-agnostic selection. Three settings, **Small / Medium / Large**, differ only in N_n. On RobustNeRF [50] we use $N_c{=}30$ and $N_n{\in}\{10, 30, 50\}$. ETH3D [54] has fewer frames per scene, so we use $N_c{=}14$ and $N_n{\in}\{5, 14, 30\}$. Each setting is repeated for 10 trials with different seeds and we report the mean. All methods are evaluated on identical sampled sets, and the same protocol is used for both the pose and depth/point-map evaluations.

Balanced filtering score. The main robustness table reports the distractor rejection success rate, matching Robust-VGGT. To guard against the degenerate strategy of rejecting too many frames, we

additionally compute a balanced filtering score,

$$\mathrm{BFS} = \tfrac{1}{2}\,\mathrm{DistractorReject} + \tfrac{1}{2}\,\mathrm{CleanAccept}.$$

Under the Small/Medium/Large settings of Tab. 6, the R^3 confidence gate obtains balanced filtering scores of $0.84/0.89/0.93$ on ETH3D and $0.999/0.992/0.996$ on RobustNeRF, indicating that high distractor rejection does not come from uniformly rejecting clean frames.

H Long-Sequence Pose Evaluation on DL3DV-Benchmark

We provide the protocol details for the DL3DV-Benchmark [36] evaluation reported in Tab. 5 of the main paper. The evaluation is purely camera-trajectory based. We report Absolute Trajectory Error (ATE) after Sim(3) alignment, together with per-trajectory rotation RMSE. Aggregates are mean values over a random subset of 25 scenes drawn from the 140 DL3DV-Benchmark scenes, with per-scene sequence lengths ranging from 304 to 439 frames. All methods process every frame in the sequence. R^3 runs the full sequence in a *single pass with no reset*; in our DL3DV run, TTT3R [10] uses a 100-frame reset interval, which gives better results than no-reset inference, and LongStream [13] uses the reset behavior of its released default configuration.

The reported numbers are consistent with the long-sequence pose curves in Sec. 4.3: the keyframe-bank front-end with confidence-weighted relative-pose aggregation accumulates drift more slowly than recurrent and KV-cache baselines at the horizons covered by DL3DV-Benchmark.

I Visual Odometry on TUM RGB-D fr1

We evaluate on the standard TUM RGB-D fr1 visual-odometry sequences [57], processing every frame at the original frame rate. As shown in Tab. 14, R^3 achieves the lowest mean ATE and is best on $5/9$ sequences. TTT3R and LongStream rely on resets in their default configurations; with resets disabled, TTT3R degrades from 0.148 to 0.305 mean ATE and LongStream runs out of memory.

Table 14: **Visual odometry on TUM RGB-D fr1.** Sim(3)-aligned ATE RMSE (m), lower is better, with every frame processed at the original frame rate. The mean is computed from unrounded per-sequence values. †Evaluated with the reset mechanism of its released default configuration; R^3 runs in a single pass without reset.

Method	360	desk	desk2	floor	plant	room	rpy	teddy	xyz	Mean
TTT3R† [10]	**0.106**	0.107	0.159	**0.252**	0.108	**0.264**	0.053	0.198	0.083	0.148
LongStream† [13]	0.125	0.129	0.206	0.267	0.188	0.349	0.040	0.168	0.040	0.168
InfiniteVGGT [77]	0.140	0.066	0.126	0.747	0.174	0.801	**0.035**	0.117	0.024	0.248
$\mathbf{R}^3$	0.134	**0.063**	**0.068**	0.271	**0.079**	0.311	0.045	**0.070**	**0.018**	**0.117**

J Video Depth Estimation

We additionally evaluate video depth estimation against the DA3-Large backbone on Sintel [4], Bonn [45], and KITTI [21]. Following the standard protocol, we report Absolute Relative error (Abs Rel, lower is better) and $\delta < 1.25$ accuracy (higher is better) under per-sequence median scale alignment.

Our model matches the DA3-Large teacher on Bonn, improves Abs Rel on Sintel, and trails slightly on KITTI, showing that joint relative-pose and depth training preserves the backbone's depth quality despite the added pose objective.

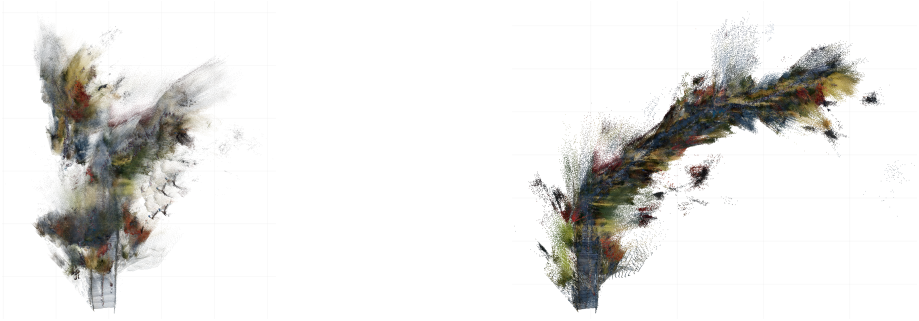
K Limitations and Future Work

R^3 is a 372M-parameter model trained on six 48 GB GPUs, and we have not explored larger backbones or substantially more data. The keyframe bank still relies on a few hand-set thresholds (τ,

Table 15: **Video depth estimation on Sintel, Bonn, and KITTI.** Abs Rel is lower-is-better; $\delta < 1.25$ is higher-is-better. **Bold** marks the best result per column.

	Sintel		Bonn		KITTI	
Method	Abs Rel↓	$\delta < 1.25$↑	Abs Rel↓	$\delta < 1.25$↑	Abs Rel↓	$\delta < 1.25$↑
DA3-Large [34]	0.460	**0.518**	0.106	**0.890**	**0.122**	**0.870**
$\mathbf{R}^3$	**0.410**	0.479	**0.102**	0.880	0.130	0.845

$\Delta_{\max}$, $M_{\max}$, outlier-gate constants) that generalize across our benchmarks but may need re-tuning on very different domains. On very long streams R^3-Stream can still drift without a periodic reset (Fig. 9), and because it cannot revisit past tokens, distant viewpoints, occlusions, or long sequences can leave residual inconsistencies in the geometry (Fig. 10).



(a) Without reset. (b) With periodic reset.

Figure 9: Bird's-eye view of a long streaming sequence. Without reset, the trajectory eventually drifts and the camera is lost; a periodic reset restores a clean track.

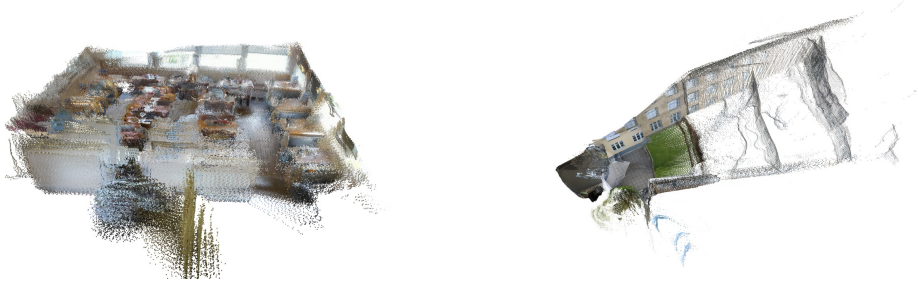


Figure 10: Causal streaming reconstructions. Distant viewpoints, occlusions, and long sequences can leave residual inconsistencies in the recovered geometry.

Natural next steps include scaling the backbone and training data, replacing the lightweight pose-only refinement with bundle adjustment or joint depth–pose optimization, learning the keyframe-bank thresholds and a reset policy in place of the current scalars, and a global refinement or limited bidirectional re-attention pass to address causal inconsistencies in the streaming setting.

L Broader Impacts

Improved feed-forward 3D reconstruction can benefit robotics, AR/VR, accessibility, and mapping applications. Potential negative uses include unauthorized 3D capture of private spaces or surveillance scenarios. The foundational nature of the work and the absence of person-specific modeling limit direct paths to misuse.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction describe a relative-pose regression head with learned rotation and translation confidences, a confidence-weighted pose-graph back-end, and a single causal checkpoint usable in both streaming and full-context modes; these claims are substantiated by the method and experiments sections.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: A dedicated Limitations and Future Work section is included in the appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper does not include formal theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The architecture, losses, training data, optimizer, and inference procedure are described in the method section and appendix in sufficient detail to reproduce the main results, and code and checkpoints will be released.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We will release the training and evaluation code together with trained checkpoints and reproduction instructions on the project page (`https://kevinxu02.github.io/r3-site/`). All datasets used in this work are publicly available.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Datasets, splits, evaluation protocols, hyperparameters, and optimizer settings are specified in the experiments section and appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Inference is deterministic and evaluation is performed on full benchmark splits, so the reported numbers are exactly reproducible from a fixed checkpoint without run-to-run variance.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Training is performed on six 48GB GPUs as reported in the paper; the appendix provides further details on training time and inference cost.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: See Appendix L.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The released model is a 3D reconstruction network and does not generate language, images, or scraped data, so it does not pose a high risk of misuse warranting additional safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All existing datasets and the DA3 backbone used in this work are cited and used under their original research licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The code and trained checkpoints to be released will include documentation covering training, evaluation, and intended use, and will be released on the project page.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.