LONG3R: Long Sequence Streaming 3D Reconstruction

Zhuoguang Chen[1,2*]   Minghui Qin[2*]   Tianyuan Yuan[2,3*]   Zhe Liu[2]   Hang Zhao[2,1,3†]

[1]Shanghai Artificial Intelligence Laboratory   [2]IIIS, Tsinghua University   [3]Shanghai Qi Zhi Institute

Abstract

*Recent advancements in multi-view scene reconstruction have been significant, yet existing methods face limitations when processing streams of input images. These methods either rely on time-consuming offline optimization or are restricted to shorter sequences, hindering their applicability in real-time scenarios. In this work, we propose LONG3R (**LO**ng sequence streami**NG 3D R**econstruction), a novel model designed for streaming multi-view 3D scene reconstruction over longer sequences. Our model achieves real-time processing by operating recurrently, maintaining and updating memory with each new observation. We first employ a memory gating mechanism to filter relevant memory, which, together with a new observation, is fed into a dual-source refined decoder for coarse-to-fine interaction. To effectively capture long-sequence memory, we propose a 3D spatio-temporal memory that dynamically prunes redundant spatial information while adaptively adjusting resolution along the scene. To enhance our model's performance on long sequences while maintaining training efficiency, we employ a two-stage curriculum training strategy, each stage targeting specific capabilities. Experiments demonstrate that LONG3R outperforms state-of-the-art streaming methods, particularly for longer sequences, while maintaining real-time inference speed. Project page: https://zgchen33.github.io/LONG3R/.*

1. Introduction

Recovering dense geometry from a sequence of images is a fundamental task in 3D computer vision. It has widespread applications in robotics, autonomous driving, and indoor and outdoor scene reconstruction. Traditional approaches typically address this task through various methods, including Structure from Motion (SfM) [1, 9, 50, 53, 57, 71, 72], keypoint detection [33, 34, 46], bundle adjustment [2, 62, 73], Simultaneous Localization and Mapping (SLAM) [11, 28, 41], and Multi-View Stereo [22, 23, 41, 51]. While these methods have achieved notable success, they rely on hand-crafted heuristics, requiring significant engineering effort
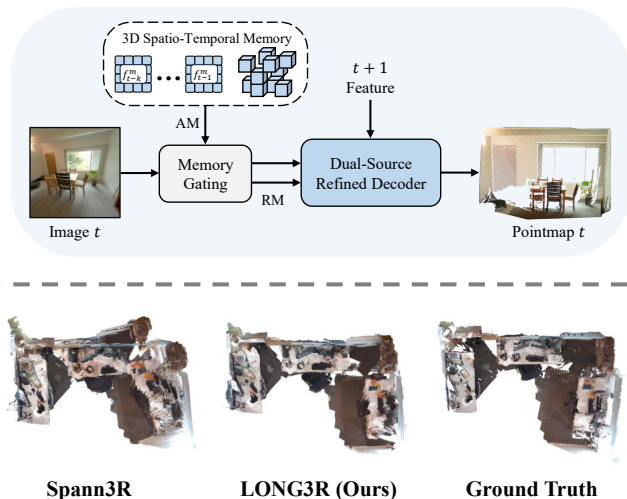

Figure 1. (Top): Overview of our LONG3R framework, which integrates a 3D spatio-temporal memory module and memory gating to refine features through a dual-source refined decoder. Here, **AM** denotes *All Memory* and **RM** denotes *Relevant Memory*. (Bottom): Qualitative comparison of 3D reconstructions from Spann3R, our method LONG3R, and the ground truth. LONG3R can achieve more accurate point prediction.

when assembled into a pipeline [49, 51].

Recently, a new class of methods, beginning with DUSt3R [68] and MASt3R [30], has tackled this problem using end-to-end neural networks. These models directly regress 3D representations (*i.e.*, pointmaps) from image pairs, offering a simple yet highly generalizable approach that has quickly gained traction in the community. A particularly promising direction extends this paradigm to online processing of streaming input images. For instance, Spann3R [63] introduced a recurrent model with memory to process streaming input images in real time, enabling various practical applications. However, despite its efficiency, Spann3R struggles with long input sequences due to three key issues: (1) its memory is only attended once per iteration, preventing effective reuse, (2) its memory becomes spatially redundant as images accumulate, and (3) its training strategy does not support adaptation to long sequences.

To address the aforementioned challenges, in this pa-

---

* Equal contribution. †Corresponding author.

per, we propose LONG3R (**LO**ng sequence streami**NG** **3D R**econstruction), a novel model designed for streaming multi-view 3D scene reconstruction over longer sequences, as illustrated in Fig. 1. We define long-sequence reconstruction as real-time processing of tens to hundreds of frames with near-constant memory requirements. Like Spann3R [63], our approach employs a recurrent network with a spatio-temporal 3D memory bank to process streaming image sequences. Given a new observation, our model retrieves relevant memories, interacts with the current view to predict its pointmap, and updates the memory accordingly. To enhance long-sequence processing, we introduce three key innovations:

1. Memory Gating & Dual-Source Decoder: We introduce a memory gating mechanism that selectively retains memories relevant to the current observation, followed by a Dual-Source Refined Decoder that enables coarse-to-fine interaction between observations and memories.
2. 3D Spatio-temporal Memory: We propose a dynamic 3D memory module that automatically prunes redundant memories and adapts resolution to the scene scale, balancing memory efficiency and reconstruction accuracy.
3. Two-stage Curriculum Training: We adopt a two-stage curriculum training strategy that progressively increases sequence length, enhancing the model's ability to handle increasingly complex memory interactions.

We conduct extensive experiments on multiple 3D datasets, comparing LONG3R with various state-of-the-art methods. Results demonstrate that our approach significantly improves long-sequence streaming reconstruction while maintaining real-time processing capability.

2. Related Work

Traditional 3D Reconstruction. 3D reconstruction from visual data modalities has been a long-standing research challenge in computer vision and graphics. This field has evolved through paradigm shifts from Structure-from-Motion (SfM) frameworks [6, 16] and SLAM systems [11, 18, 28, 39, 41] to advanced scene representation methods like Neural Radiance Fields (NeRF) [7, 21, 37, 38, 66, 81] and 3D Gaussian Splatting [26, 27, 78, 82, 84, 86]. Contemporary reconstruction frameworks expose fundamental limitations of classical geometric approaches when confronted with sparse observations, ill-posed problems, or long unconstrained sequences. Traditional methods based on geometric optimization—such as explicit feature matching and iterative bundle adjustment—are inherently sensitive to observation sparsity, computationally redundant, and notably slow due to the absence of learned priors for regularizing underconstrained scenarios. In contrast, our streaming 3D reconstruction method, LONG3R, directly regresses 3D pointmaps from images using learned priors, resulting in significantly faster reconstruction.

Learning-Based 3D Reconstruction. Replacing traditional handcrafted components with learning-based approaches, including learning-based priors [8, 12, 17, 47, 56, 60, 61, 77, 84, 90], depth estimation [5, 24, 25, 31, 42, 42–44, 76], and end-to-end system optimization [58, 61, 65, 79], is becoming a prevailing trend for scalable scene representation. This trend elevates data-driven methodologies to prominence within 3D reconstruction frameworks, with pointmap representations [13, 19, 29, 32, 35, 36, 63, 67, 68, 75, 87] catalyzing the evolution of this field. Drawing inspiration from the CroCo [69, 70] cross-view completion paradigm, DUSt3R [68] pioneers a geometry-agnostic pointmap prediction mechanism without prior calibration. MASt3R [30] implements a coarse-to-fine feature matching strategy, which improves the predictions of the metric-scale point map. Metric scale dense pointmaps predictions [30] can serve as a front-end to improve initialization and triangulation processes in SFM/SLAM pipelines [15, 40]. SLAM3R [32] and Reloc3r [13] propose to leverage a feedforward network for initial mapping and localization estimation, instead of relying on computationally intensive bundle adjustment-based backends. MV-DUSt3R [59] incorporates multi-view decoder blocks for cross-view information exchange and cross-reference view blocks to enhance robustness against reference view selection. For dynamic scene reconstruction, both MonST3R [85] and CUT3R [67] reconstruct temporally coherent 3D representations from unconstrained monocular video sequences.

Streaming Reconstruction. Current streaming scene reconstruction methods using pairwise image matching struggle with long sequences. Accumulative scale drift and error propagation in pose estimation degrade 3D reconstruction consistency. Traditional monocular SLAM systems [18, 20, 61, 90] mitigate these issues through optimized tracking, optical flow-integrated bundle adjustment, and loop closure detection. Their reliance on predefined intrinsic parameters limits generalizability, and reconstruction quality heavily depends on depth estimation accuracy. Spann3R [63] refines point map predictions using a hybrid memory feature bank with attention interaction, enabling feedforward scene reconstruction without iterative optimization. CUT3R [67] uses persistent state tokens with transformer-based recurrent state updates for online reconstruction from streaming sequences, yet suffers from limited extreme viewpoint extrapolation capabilities due to deterministic inference and potential drift accumulation in extended sequences lacking global alignment. Our model utilizes spatio-temporal contextual information during training and inference phases to reduce cumulative errors caused by the lack of loop closure detection and post-optimization.
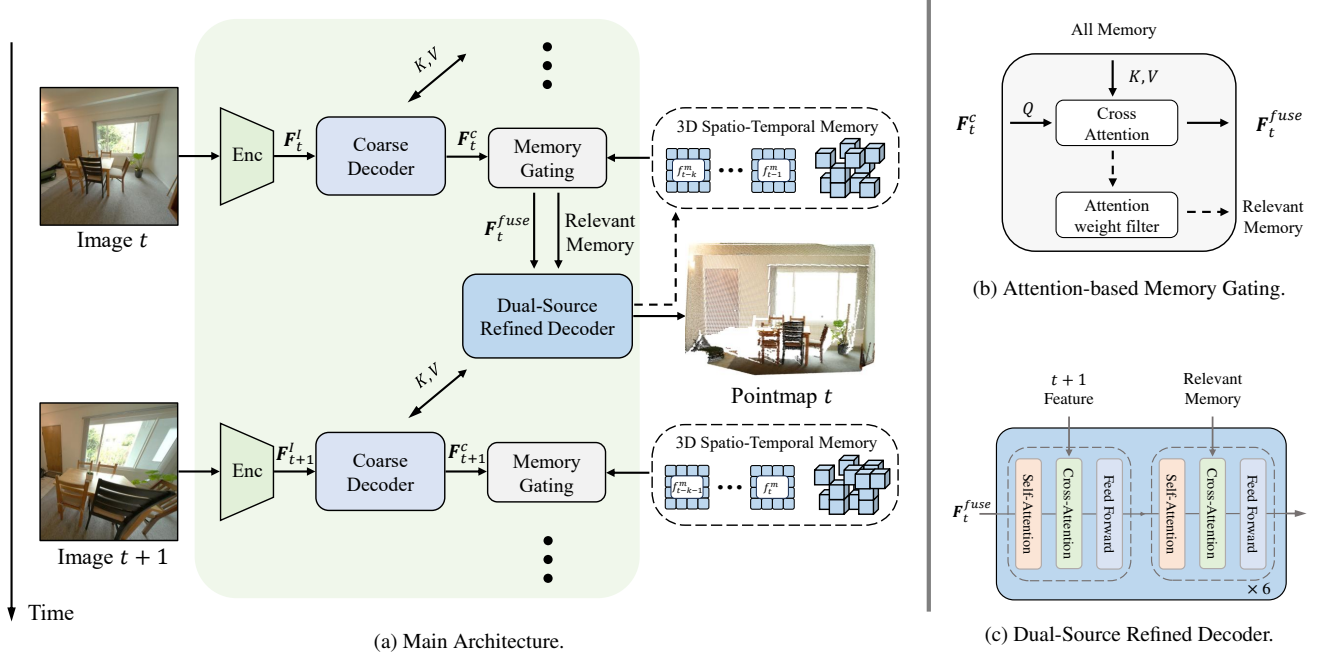
Figure 2. **Method Overview**. (a) Illustrates the overall architecture, where image features F_t^I first interact with F_{t-1}^I in the Coarse Decoder to generate F_t^c, after which a memory-gating module filters irrelevant entries from the spatio-temporal memory F_{mem}. The Dual-Source Refined Decoder subsequently interacts with both the filtered memory and features from $t+1$, ultimately generating the pointmap t. (b) Details the attention-based memory gating module, which selects relevant information from the memory. (c) Illustrates the dual-source refined decoder, which alternately interacts with the next-frame features and relevant memory features through multiple self- and cross-attention layers to optimize memory information utilization and maintain alignment with the subsequent frame.

3. Method

Our proposed streaming 3D reconstruction method, LONG3R, is depicted in Fig. 2. It starts by encoding features from consecutive frames, which are then processed by a Coarse Decoder to generate a rough 3D structure. A memory gating mechanism then filters the spatio-temporal memory to retain relevant memory. The relevant memory, along with $t+1$ context, is used by a Dual-Source Refined Decoder to produce accurate reconstructions.

3.1. Feature Encoding and Coarse Decoding

Following previous studies [63, 68], the input image I_t is processed by a ViT encoder that partitions it into patches and linearly projects them into visual feature tokens:

$$F_t^I = \text{Encoder}(I_t). \quad (1)$$

These tokens are then forwarded to a Coarse Decoder, implemented as a generic transformer composed of B PairwiseBlocks. Each PairwiseBlock, which comprises self-attention, cross-attention, and an MLP, is used as follows:

$$F_{t,i}^c = \text{PairwiseBlock}_i^c\left(F_{t,i-1}^c, F_{t-1,i-1}^r\right), \quad (2)$$

for $i = 1, 2, \ldots, B$, with $F_{t,0}^c = F_t^I$. Here, $F_{t-1,i-1}^r$ denotes the refined tokens produced by the corresponding

blocks of the Refined Decoder for the $(t-1)$-th frame. By efficiently interacting with temporal features from the previous frame, the Coarse Decoder generates coarse representations that serve as the basis for subsequent processing.

3.2. Attention-based Memory Gating

As depicted in Fig. 2b, the memory gating mechanism serves two primary functions: aggregating information from all memory entries and filtering out irrelevant memory tokens to reduce the computational load of the subsequent Refined Decoder. For brevity, we denote the final Pairwise-Block output $F_{t,B}^c$ by F_t^c, which then attends to the memory keys F_{mem}^K and values F_{mem}^V via cross-attention:

$$W_t = \text{Softmax}\left(\frac{F_t^c\left(F_{\text{mem}}^K\right)^\top}{\sqrt{C}}\right), \quad (3)$$

$$F_t^{\text{fuse}} = W_t F_{\text{mem}}^V, \quad (4)$$

where $W_t \in \mathbb{R}^{P \times S}$ denotes the attention weights for each token in the current query relative to all memory keys, with P as the number of tokens in the current frame and S as the number of memory tokens.

To filter out memory irrelevant to the current observation, we employ the attention weights $W_t \in \mathbb{R}^{P \times S}$ and

an attention threshold τ (5×10^{-4}). Specifically, for each memory index $s \in \{1, \ldots, S\}$, if there exists at least one token $p \in \{1, \ldots, P\}$ such that $W_t(p, s) > \tau$, the s-th memory feature is retained; otherwise, it is discarded. Thus, the relevant memory $\boldsymbol{F}_{\text{r\_mem}}$ is given by:

$$\delta(s) = \begin{cases} 1, & \text{if } \max_p W_t(p, s) > \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

$$\boldsymbol{F}_{\text{r\_mem}} = \{\, \boldsymbol{F}_{\text{mem}}(s) \mid \delta(s) = 1 \,\}. \quad (6)$$

Here, $\delta(s)$ is an indicator function. Notably, $\boldsymbol{F}_{\text{mem}}$ comprises two components, $\boldsymbol{F}_{\text{mem}}^K$ and $\boldsymbol{F}_{\text{mem}}^V$, both of which undergo the same above filtering process. This mechanism ensures that only memory elements with sufficient relevance, as determined by the attention weights, contribute to the following refining process.

3.3. Dual-Source Refined Decoder

To maximize the utilization of memory information and maintain alignment with the subsequent frame, we propose the Dual-Source Refined Decoder, as illustrated in Fig. 2c. Unlike the Coarse Decoder, which consists solely of generic PairwiseBlocks, the Dual-Source Refined Decoder alternates between two types of blocks: a PairwiseBlock and a MemoryBlock. This design allows the current-frame tokens to fully exploit and aggregate both spatio-temporal memory tokens and features from the next frame.

Let $\boldsymbol{F}_{t,i}^r$ denote the refined feature representation at block i. The overall operation can be formulated as:

$$\boldsymbol{F}_{t,i}^r = \begin{cases} \text{PairwiseBlock}(\boldsymbol{F}_{t,i-1}^r, \boldsymbol{F}_{t+1,i-1}^c), & i \text{ odd}, \\ \text{MemoryBlock}(\boldsymbol{F}_{t,i-1}^r, \boldsymbol{F}_{\text{r\_mem}}), & i \text{ even}, \end{cases} \quad (7)$$

for $i = 1, 2, \ldots, B$, with $\boldsymbol{F}_{t,0}^r = \boldsymbol{F}_t^{\text{fuse}}$. The Pairwise-Block (applied for odd i) facilitates feature interactions between the refined current-frame features and the coarse tokens from the next frame, while the MemoryBlock (applied for even i) integrates these refined features with the relevant memory tokens $\boldsymbol{F}_{\text{r\_mem}}$, thereby enhancing long-range spatio-temporal dependencies. This alternating structure enables the decoder to construct a robust, context-aware feature representation by leveraging both immediate and historical information.

Following the decoder, an explicit 3D reconstruction prediction is generated from its outputs using a DPT head.

3.4. 3D Spatio-Temporal Memory

Our memory mechanism handles long sequences by concurrently maintaining short-term temporal memory and long-term 3D spatial memory. The memory consists of historical tokens generated by the Dual-Source Refined Decoder. With a fixed storage capacity, our long-term 3D
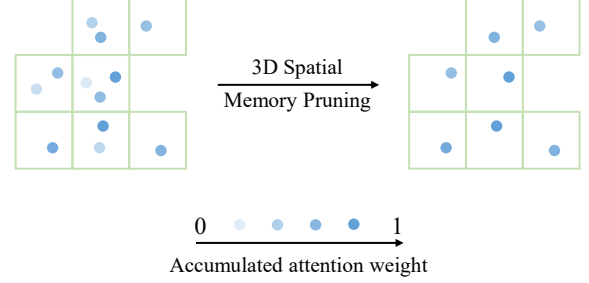

Figure 3. **3D Spatial Memory Pruning**. Memory tokens are grouped into voxels based on 3D positions (illustrated here in a 2D simplified view), with only the token having the highest accumulated attention weight retained per voxel. Darker blue dots indicate higher attention weights.

spatial memory maintains an overall spatial representation while avoiding redundant tokens. This memory design efficiently captures essential spatio-temporal features without overwhelming memory resources.

The short-term temporal memory stores historical tokens from the time window $[t - K, t - 1]$, where K denotes the window length. It stores key features $f^K \in \mathbb{R}^{(K \cdot P) \times C}$ and value features $f^V \in \mathbb{R}^{(K \cdot P) \times C}$, ensuring the effective utilization of time-dependent information.

The long-term 3D spatial memory mitigates GPU memory constraints and improves inference speed by managing tokens within $[1, t - K]$ while limiting their number. Inspired by the occupancy mechanism, we employ voxels as storage units, with each voxel retaining a single token. This sparsification balances the number of memory units with the scene's spatial size. However, since different scenes require varying voxel sizes and the model's optimization is metric-invariant, predefined voxel sizes are unsuitable. To address this, we introduce an adaptive voxel size strategy.

Adaptive Voxel Size. Since the memory stores patch-based tokens, we first compute a unique 3D position P for each patch using the point map predicted in each frame via a weighted average. Each token then calculates the 3D Euclidean distance to its eight neighboring tokens in the image plane, with the average distance defined as $d_i = 0.125 \sum_{j \in \mathcal{N}(i)} \|P_i - P_j\|_2$, where P_i is the 3D position of token i, $\mathcal{N}(i)$ represents the set of its eight neighboring tokens, and $\| \cdot \|_2$ is the Euclidean norm.

The optimal image voxel size v_{img} is determined as the minimum d_i across all tokens to balance memory usage and storage efficiency. The scene voxel size v_{scene} is computed as the average of image voxel sizes across all frames:

$$v_{\text{img}} = \min_i d_i \quad \text{and} \quad v_{\text{scene}} = \frac{1}{t-1} \sum_{j=1}^{t-1} v_{\text{img},j}, \quad (8)$$

where t denotes the sequence index of the current frame.

Method	7Scenes						NRGBD						FPS
	Acc↓		Comp↓		NC↑		Acc↓		Comp↓		NC↑		
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	
F-Recon [74]	12.43	7.62	5.54	**2.31**	**61.89**	**68.85**	28.55	20.59	15.05	6.31	**65.47**	**75.77**	≪1
Dust3R [68]	3.01	**1.47**	**5.11**	2.79	58.83	63.73	3.94	**2.48**	**5.31**	**3.58**	62.62	72.29	≤3
MASt3R [29]	**2.82**	1.56	5.26	3.24	58.22	62.46	**3.85**	2.54	5.50	3.62	60.92	68.67	≤3
MV-DUSt3R [59]	**2.92**	1.24	**2.49**	**0.78**	66.42	76.07	3.76	1.99	**2.55**	0.92	81.16	95.39	∼15
MV-DUSt3R+ [59]	2.93	**1.07**	8.63	0.95	66.38	**76.18**	3.47	**1.60**	3.69	**0.85**	84.33	97.27	∼3
CUT3R [67]	7.73	3.57	7.75	1.83	65.74	73.98	12.48	5.57	6.34	2.35	75.84	90.05	∼23
Spann3R [63]	3.42	1.48	2.41	0.85	66.35	76.25	6.91	3.15	**2.91**	**1.10**	77.75	93.71	∼22
Ours	**2.57**	**1.14**	**2.08**	**0.73**	**66.55**	**76.43**	6.66	2.54	3.11	1.21	77.56	93.08	∼22

Table 1. **Quantitative results on 7Scenes [52] and NRGBD [3] datasets.** All models are using 224×224 image inputs.
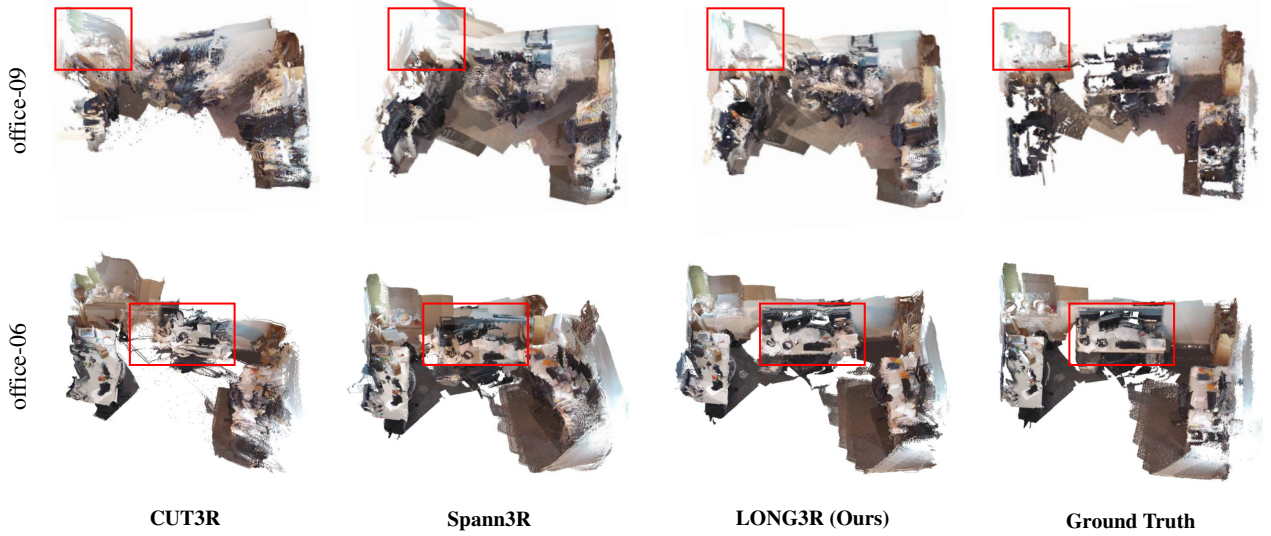

Figure 4. **Qualitative comparisons.** We present a comparison of reconstruction results on Office-06 and Office-09 from the 7Scenes [52] dataset with the Spann3R [63] and CUT3R [67] methods. In comparison with other online reconstruction methods, our approach achieves superior spatial consistency (e.g., the regions enclosed by the red bounding boxes) while preserving real-time performance.

Given the model's streaming architecture, v_{scene} undergoes continuous online updates during inference, enabling adaptive scene-specific adjustments across temporal sequences.

3D Spatial Memory Pruning. Once the scene voxel size is determined, tokens with similar 3D positions are grouped into the same voxel, while tokens that are farther apart are assigned to different voxels. The cumulative attention weight of each token is tracked, and only the token with the highest weight within each voxel is retained, as illustrated in Fig. 3. This mechanism effectively balances memory size to avoid storing similar memories while preserving the spatial representation of the scene.

3.5. Training

Loss Function. Following the approach in [63, 68], we employ a confidence-aware loss $\mathcal{L}_{\text{conf}}$ for 3D regression and a scale loss $\mathcal{L}_{\text{scale}}$, which encourages the predicted point cloud to have an average distance smaller than that of the

ground truth. Overall, the final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{scale}}. \quad (9)$$

Two-stage Curriculum Training. To enable our model to better handle long sequences, we adopt a two-stage training strategy. In the first stage, the model is trained by randomly sampling 5 frames per video sequence. This initial training phase allows the model to develop a preliminary understanding, enabling the encoder to be frozen while fine-tuning subsequent modules in the second stage. In the second stage, the ViT encoder remains frozen while the other modules are fine-tuned, allowing the model to be trained with an increased number of frames for addressing long sequences. Specifically, we initially sample 10 frames and subsequently 32 frames, enabling the model to gradually adapt to longer sequences. This phased approach enhances long-sequence reconstruction capabilities by exploiting spatio-temporal feature correlations across progres-

| Method | Replica$_{100}$ | | | | | | Replica$_{200}$ | | | | | | FPS |
| | Acc↓ | | Comp↓ | | NC↑ | | Acc↓ | | Comp↓ | | NC↑ | | |
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	
Dust3R [68]	6.34	3.99	6.44	3.68	61.67	69.27	**4.99**	**2.76**	**4.63**	**2.59**	**62.26**	**70.76**	≤3
MASt3R [29]	**5.10**	**2.96**	**6.00**	**3.43**	**61.81**	**69.52**	5.26	3.23	7.31	3.75	58.03	62.81	≤3
MV-DUSt3R [59]	10.41	6.48	4.34	1.22	73.76	88.36	17.02	11.70	**5.10**	**1.36**	66.74	78.24	∼7
MV-DUSt3R+ [59]	**5.28**	**3.26**	**2.56**	**0.89**	**79.07**	**93.63**	**11.79**	**8.37**	5.64	1.53	**70.66**	**83.86**	∼1
CUT3R [67]	20.44	14.64	5.67	2.32	69.63	84.31	28.3	20.68	6.61	1.88	63.95	73.85	∼23
Spann3R[63]	14.08	8.88	4.67	1.61	72.46	88.98	16.29	10.17	4.02	1.16	68.56	82.80	∼21
Ours	**11.46**	**7.55**	**3.68**	**1.24**	**73.29**	**89.86**	**11.93**	**7.42**	**2.73**	**0.87**	**68.67**	**82.92**	∼21

Table 2. **Quantitative results on Replica [54] datasets.** All models are using 224×224 image inputs. Replica$_{100}$ and Replica$_{200}$ represent sequence lengths of 100 and 200 frames, respectively.

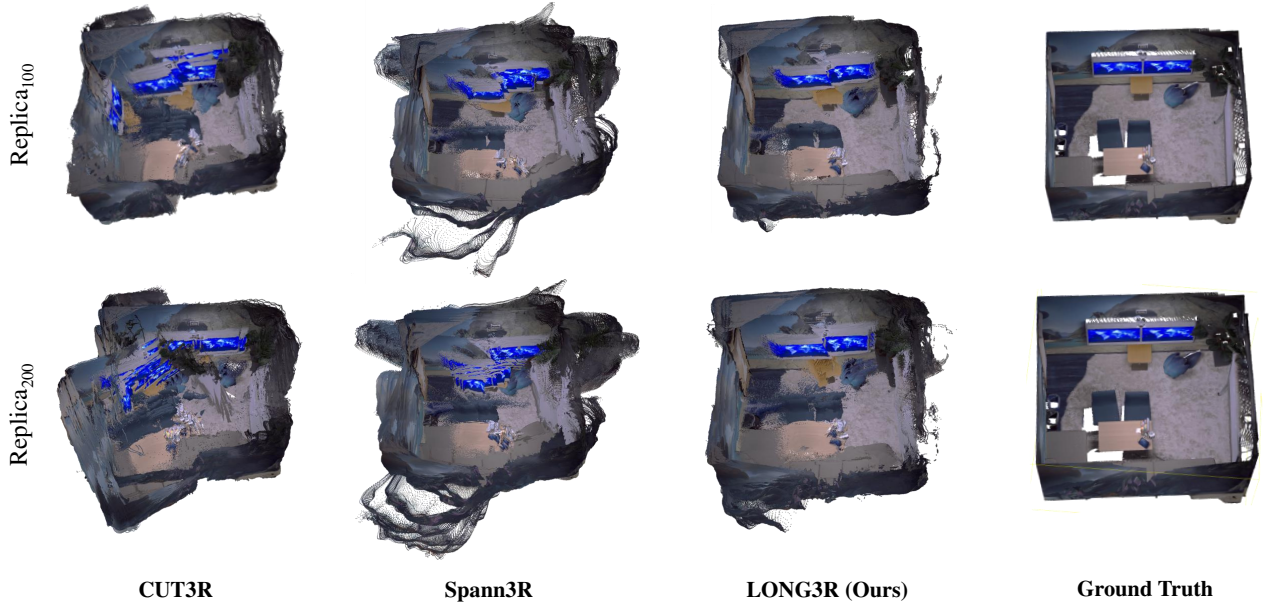**CUT3R** **Spann3R** **LONG3R (Ours)** **Ground Truth**

Figure 5. **Qualitative comparisons.** We present a comparison of reconstruction results on Office-0 from the Replica [54] dataset with the Spann3R [63] and CUT3R [67] methods. During streaming observations within the same scene, existing methods suffer from increasing ambiguity due to error accumulation. In contrast, our approach effectively suppresses cumulative drift and maintains spatial consistency.

sively expanded temporal contexts, thereby optimizing the model's ability to capture and utilize memory-related patterns in sequential data processing.

4. Experiments

To comprehensively assess overall performance and component effectiveness, we evaluate our method on 3D reconstruction (Sec. 4.2) and camera pose estimation (Sec. 4.3), along with ablation and analysis presented in Sec. 4.4.

4.1. Setup

Training Datasets. Following Spann3R [63], we train our model with a mixture of 6 datasets: Habitat [48], ARKitScenes [4], BlendedMVS [80], ScanNet++ [83], Co3D-v2 [45], ScanNet [10]. These datasets integrate real-world

and synthetic data, encompassing metric-scale measurements and normalized-scale samples.

Baselines. For online streaming reconstruction methods, we evaluate LONG3R with Spann3R [63] and CUT3R [67] as primary baselines. To ensure a comprehensive analysis, we further compare our method with DUSt3R [68] and MASt3R [29], which involve additional post-processing stages, as well as offline approaches MV-DUSt3R and MV-DUSt3R+ [59]. All evaluations are conducted on a single NVIDIA RTX 3090 GPU with 24 GB of VRAM, and all input images are resized to 224×224 for a fair comparison.

Implementation details. We use ViT-Large [14] as LONG3R's encoder, initialized with DUSt3R's encoder weights. Both training stages utilize images at a resolution of 224×224. Experiments are conducted with 10-frame

Method	7Scenes			TUM			ScanNet		
	ATE	RPE$_t$	RPE$_r$	ATE	RPE$_t$	RPE$_r$	ATE	RPE$_t$	RPE$_r$
Spann3R [63]	12.64	6.15	1.88	5.66	**2.13**	**0.59**	9.83	2.30	0.66
CUT3R [67]	12.40	7.65	2.34	6.25	2.55	0.69	14.27	3.58	0.92
Ours	**8.72**	**5.03**	**1.67**	**5.40**	2.36	0.60	**6.44**	**2.14**	**0.61**

Table 3. **Evaluation of Camera Pose Estimation (cm/°).** All three models using 224 × 224 image inputs.

Method	7Scenes				NRGBD				FPS
	Acc↓		Comp↓		Acc↓		Comp↓		
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	
w/o Gating	2.53	**1.12**	2.12	0.74	6.72	3.14	**2.91**	**1.20**	18.0
w/ Gating	2.57	1.14	**2.08**	**0.73**	**6.66**	**3.11**	2.92	1.21	21.4

Table 4. **Attention-based Memory Gating ablation study on 7Scenes [52] and NRGBD [3].**

short-term memory and 3000-token long-term memory. In the first stage, we employ the AdamW optimizer with a learning rate of 1.12×10^{-4} and a batch size of 10 per GPU, training for 120 epochs. In the second stage, we fine-tune the model using the AdamW optimizer with a learning rate of 1×10^{-5} on 10-view and 32-view sequences, training for 12 epochs each. The first stage runs on 16 A100 GPUs for 28 hours, while fine-tuning requires about 20 hours on the same hardware configuration.

4.2. 3D Reconstruction

We evaluate scene-level reconstruction on three unseen datasets: 7Scenes [52], NRGBD [3], and Replica [54] to demonstrate generalization capability in long-sequence reconstruction using *accuracy*, *completion* and *normal consistency* as in previous works [3, 63, 64, 89] and we report accuracy and completeness in centimeters. For the Replica [54] dataset, we consider two settings: uniformly sampling 100 and 200 frames, denoted Replica$_{100}$ and Replica$_{200}$, respectively.

Quantitative comparisons. Tab. 1 and Tab. 2 present the quantitative comparison of reconstruction metrics between our method and existing 3D reconstruction approaches. The learning-based approaches evaluated are categorized into three classes: optimization methods with post-processing refinement like DUSt3R [68] and MASt3R [29], offline methods like MV-DUSt3R [59], and streaming online reconstruction methods like Spann3R [63] and CUT3R [67]. As demonstrated in Tab. 1, our method achieves the highest reconstruction accuracy compared to online streaming reconstruction methods, while attaining a precision comparable to post-optimization and offline approaches on the 7Scenes [52] and NRGBD [3] datasets. Furthermore, our real-time performance offers a crucial advantage over the slower post-optimization and offline methods.

To systematically assess the robustness and geometric fidelity of reconstruction approaches under streaming observations, we design a controlled experiment on the Replica [54] dataset. We sample 100 and 200 frames sequentially from the same scene as complete observations. The predicted pointmaps are aggregated as the reconstruction results for metric computation, reflecting the impact of sequence length on reconstruction accuracy. As shown in Tab. 2, post-optimization methods achieve higher accu-
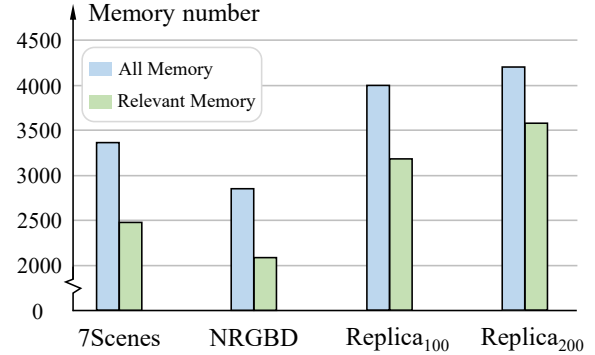

Figure 6. **Memory number comparison.** This figure illustrates the number of memory tokens before (all memory) and after (relevant memory) passing through the memory gating module.

racy due to global alignment effectively mitigating drift, albeit at the cost of significantly lower FPS, which limits their applicability in real-time scenarios. Other optimization-free baselines exhibit reduced robustness with denser observational inputs due to their inherent lack of holistic spatial consistency awareness during training and inference. In contrast, our approach maintains competitive reconstruction accuracy compared to online methods while avoiding catastrophic performance degradation on extended sequences.

Qualitative comparisons. We qualitatively compare our method with online reconstruction baselines on the 7Scenes [52] and Replica [54] datasets. Detailed comparisons of reconstruction results are presented in Fig. 4 and Fig. 5. As shown in Fig. 4, our method achieves superior spatial consistency while maintaining real-time performance compared to Spann3R [63] and CUT3R [67]. Fig. 5 presents results from long-sequence streaming observation experiments. With increasing observation sequences, Spann3R [63] and CUT3R [67] exhibit amplified ambiguity in geometric predictions (e.g., erroneous surfaces of televisions), whereas our approach maintains consistent reconstruction quality via 3D spatiotemporal context aggregation.

4.3. Camera Pose Estimation

We evaluate camera pose estimation on 7Scenes [52], TUM [55], and ScanNet [10] using Absolute Translation Error (ATE), Relative Translation Error (RPE$_t$), and Relative Rotation Error (RPE$_r$), following [8, 85, 88].

(a) Interleaved Dual-Source Refined Decoder.   (b) Concatenated Dual-Source Refined Decoder.
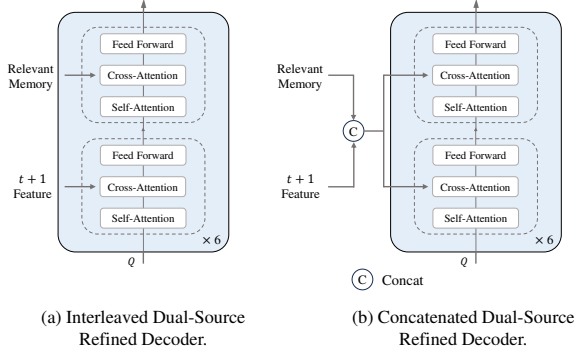
Figure 7. **Dual-Source Decoder Comparison.** This figure compares the interleaved and concatenated architectures of the Dual-Source Refined Decoder.

As shown in Tab. 3, our method significantly outperforms Spann3R and CUT3R on static scene datasets, namely 7Scenes and ScanNet. Despite being trained exclusively on static scenes, our method remains competitive with Spann3R and CUT3R on the TUM Dynamics dataset, which features dynamic human motion.

4.4. Ablation and analysis

Attention-based Memory Gating. We analyze the impact of the memory gating mechanism on the 7Scenes [52] and NRGBD [3] datasets. The experimental results are summarized in Tab. 4, and Fig. 6. The memory gating mechanism removes memory features irrelevant to the current frame, exemplified by a 27% reduction on 7Scenes, and achieves an optimal balance between reconstruction accuracy and computational efficiency in streaming reconstruction. We evaluated the FPS with (21.4 FPS) and without (18.0 FPS) memory gating, resulting in a 20% boost.

Dual-Source Refined Decoder. We conduct an ablation study on the Replica [54] dataset to compare the effectiveness of various dual-source block designs, with Fig. 7 depicting the design variations. Due to memory constraints associated with the concatenation method, the experiments reported in Tab. 5 use a sequence length of 24 frames instead of 32 during the second stage of training. As shown in Tab. 5, compared with the concatenation approach, our proposed interleaved attention blocks yield superior scene reconstruction accuracy under both sampling settings in Replica while reducing computational complexity. These performance gains primarily stem from mitigating information loss caused by feature space misalignment between memory features and the coarse features of frame $t+1$. Our interleaved attention blocks address this issue by employing alternating cross-attention, which progressively aligns feature spaces and improves computational efficiency.

Method	Replica$_{100}$				Replica$_{200}$			
	Acc↓		Comp↓		Acc↓		Comp↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Concat.	14.83	10.26	4.59	1.81	29.52	21.04	8.88	4.05
Interleaved.	**12.06**	**7.67**	**3.68**	**1.23**	**13.34**	**8.41**	**3.15**	**0.94**

Table 5. **Refined Decoder ablation study on Replica [54].**

Method	7Scenes				Replica$_{200}$			
	Acc↓		Comp↓		Acc↓		Comp↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
w/o 3D Spa. Mem.	5.76	2.96	3.30	1.22	65.75	47.63	13.24	3.49
w/ Spann3R Mem.	2.64	1.16	2.10	0.74	12.41	7.87	3.07	0.88
LONG3R (ours)	**2.57**	**1.14**	**2.08**	**0.73**	**11.93**	**7.42**	**2.74**	**0.87**

Table 6. **Memory frameworks ablation study on 7Scenes [52] and Replica [54].** w/o 3D Spa. Mem.: without 3D spatial memory, only with temporal memory.

3D Spatio-Temporal Memory. We conduct an ablation study in two parts to evaluate our 3D spatio-temporal memory design. In the first part, we compare the complete design with a variant that excludes the long-term 3D spatial memory component on the Replica [54] dataset. The results detailed in Tab. 6 reveal that omitting the long-term 3D spatial memory significantly degrades reconstruction performance, highlighting its essential role in long-sequence streaming reconstruction. In the second part, we compare our proposed design with the Spann3R memory [63] framework under consistent network architectures. Our findings demonstrate that our 3D spatio-temporal memory achieves superior scene-level reconstruction accuracy through reducing spatially redundant memory while preserving coherent geometric understanding across sequential frames.

5. Conclusion

In this paper, we propose LONG3R, a novel framework for long-sequence streaming 3D reconstruction that overcomes key limitations of existing methods. It combines a memory gating mechanism, a Dual-Source Refined Decoder, and a dynamic 3D Spatio-Temporal Memory Module to improve memory efficiency and reduce redundancy. A two-stage curriculum training strategy further helps the model adapt to long sequences. Experiments on multiple datasets demonstrate that LONG3R achieves state-of-the-art while maintaining real-time performance. Future work will explore scaling the framework to more complex environments.

Limitations. Since our predictions are defined relative to the first frame, our model may produce blurry results if the viewpoint deviates significantly. Due to the lack of dynamic training data, the current model struggles to handle highly dynamic scenes with large object motions.

Acknowledgements

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 1

[2] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 29–42. Springer, 2010. 1

[3] 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*, pages 6290–6301, 2022. 5, 7, 8

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 6

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016. 2

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 2

[8] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19844–19853, 2024. 2, 7

[9] David Crandall, Andrew Owens, Noah Snavely, and Dan Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition.*, pages 3001–3008. IEEE, 2011. 1

[10] 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*, pages 5828–5839, 2017. 6, 7

[11] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 1, 2

[12] 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*, pages 224–236, 2018. 2

[13] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. *arXiv preprint arXiv:2412.08376*, 2024. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[15] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. 2

[16] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006. 2

[17] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2

[18] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 2

[19] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *arXiv preprint arXiv:2412.06777*, 2024. 2

[20] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on robotics*, 33(2):249–265, 2016. 2

[21] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022. 2

[22] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 1

[23] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2015. 1

[24] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular

depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2

[25] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[26] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 2

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[28] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007. 1, 2

[29] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2, 5, 6, 7

[30] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1, 2

[31] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

[32] 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. *arXiv preprint arXiv:2412.09401*, 2024. 2

[33] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 1

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1

[35] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. *arXiv preprint arXiv:2412.03079*, 2024. 2

[36] Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. Lora3d: Low-rank self-calibration of 3d geometric foundation models. *arXiv preprint arXiv:2412.07746*, 2024. 2

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[39] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[40] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. *arXiv preprint arXiv:2412.12392*, 2024. 2

[41] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011. 1, 2

[42] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 2

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[44] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2

[45] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 6

[46] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1

[47] 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*, pages 4938–4947, 2020. 2

[48] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 6

[49] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[51] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1

[52] 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*, pages 2930–2937, 2013. 5, 7, 8

[53] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, pages 835–846. ACM, 2006. 1

[54] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6, 7, 8

[55] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 7

[56] 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*, pages 8922–8931, 2021. 2

[57] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE international conference on computer vision*, pages 801–809, 2015. 1

[58] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018. 2

[59] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *arXiv preprint arXiv:2412.06974*, 2024. 2, 5, 6, 7

[60] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6243–6252, 2017. 2

[61] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Neural Information Processing Systems*, 2021. 2

[62] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 1

[63] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 1, 2, 3, 5, 6, 7, 8

[64] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Coslam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13293–13302, 2023. 7

[65] 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*, pages 21686–21697, 2024. 2

[66] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[67] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2, 5, 6, 7

[68] 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*, pages 20697–20709, 2024. 1, 2, 3, 5, 6, 7

[69] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2023. 2

[70] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 2

[71] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *European conference on computer vision*, pages 61–75. Springer, 2014. 1

[72] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 1

[73] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. Multicore bundle adjustment. In *CVPR 2011*, pages 3057–3064. IEEE, 2011. 1

[74] Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, and Feng Zhao. Frozenrecon: Pose-free 3d scene reconstruction with frozen depth models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9310–9320, 2023. 5

[75] 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. 2

[76] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[77] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1281–1292, 2020. 2

[78] Runyi Yang, Zhenxin Zhu, Zhou Jiang, Baijun Ye, Xiaoxue Chen, Yifei Zhang, Yuantao Chen, Jian Zhao, and Hao Zhao. Spectrally pruned gaussian fields with neural compensation, 2024. 2

[79] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

[80] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 6

[81] Baijun Ye, Caiyun Liu, Xiaoyu Ye, Yuantao Chen, Yuhai Wang, Zike Yan, Yongliang Shi, Hao Zhao, and Guyue Zhou. Blending distributed nerfs with tri-stage robust pose optimization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7975–7981. IEEE, 2024. 2

[82] Baijun Ye, Minghui Qin, Saining Zhang, Moonjun Goon, Shaoting Zhu, Zebang Shen, Luan Zhang, Lu Zhang, Hao Zhao, and Hang Zhao. Gs-occ3d: Scaling vision-only occupancy reconstruction for autonomous driving with gaussian splatting. *arXiv preprint arXiv:2507.19451*, 2025. 2

[83] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 6

[84] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19447–19456, 2024. 2

[85] 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. *arXiv preprint arxiv:2410.03825*, 2024. 2, 7

[86] Saining Zhang, Baijun Ye, Xiaoxue Chen, Yuantao Chen, Zongzheng Zhang, Cheng Peng, Yongliang Shi, and Hao Zhao. Drone-assisted road gaussian splatting with cross-view uncertainty. *arXiv preprint arXiv:2408.15242*, 2024. 2

[87] 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. *arXiv preprint arXiv:2502.12138*, 2025. 2

[88] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pages 523–542. Springer, 2022. 7

[89] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12786–12796, 2022. 7

[90] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *2024 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2024. 2