# AMB3R-SLAM：基于分层后端的千米级 SLAM

Hengyi Wang　Lourdes Agapito
伦敦大学学院（UCL）计算机科学系
https://hengyiwang.github.io/projects/amber-slam

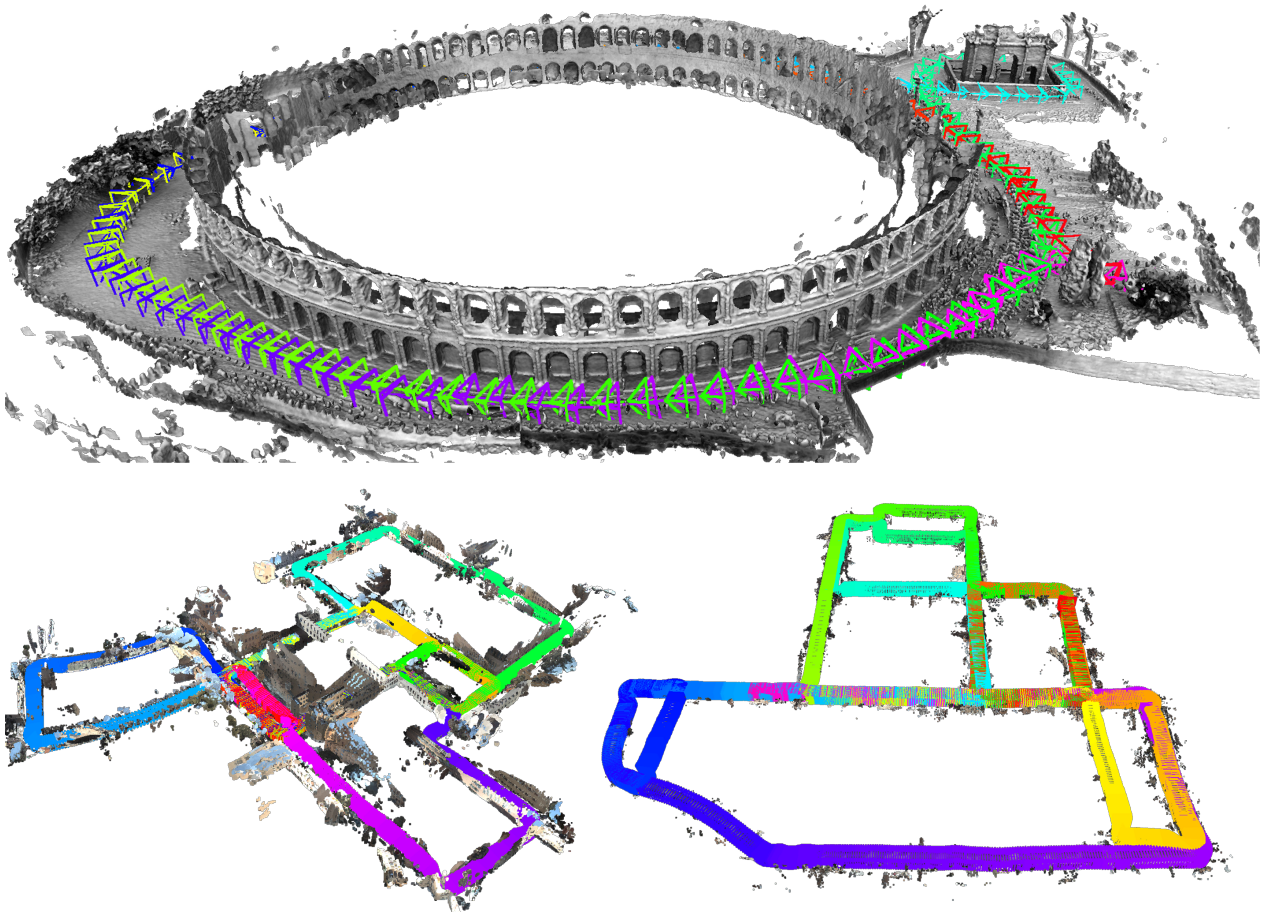


图 1. **概览**。 AMB3R-SLAM 是一个实时单目 SLAM 系统，能够在单块消费级 GPU 上重建超过 1 万帧、长达千米级的轨迹。AMB3R-SLAM 能够同时处理静态和动态场景，并可扩展为接收双目、RGB-D 或 LiDAR 作为额外输入。

## 摘要

我们提出了 *AMB3R-SLAM*，这是一个实时单目 *SLAM* 系统，能够在单块消费级 *GPU* 上重建超过 *1* 万帧、长达千米级的轨迹。我们的模型将用于低延迟在线跟踪的轻量级前端与一个分层后端相结合，后者逐步施加局部、中层和全局一致性约束。由于避免使用依赖静态世界假设的光束法平差（*bundle adjustment*），我们的系统无需额外处理即可自然地应对复杂的动态场景。此外，我们证明了该方法可以扩展为利用双目（*stereo*）、*RGB-D* 和 *LiDAR* 作为额外输入。*AMB3R-SLAM* 在 *9* 个数据集上取得了强劲的相机跟踪性能，在 *VBR* 和 *Oxford Spires* 上将此前最先进方法的绝对轨迹误差（*ATE*）降低了 *70%* 以上。在加入 *LiDAR* 输入后，我们的模型进一步将 *KITTI* 和 *VBR* 数据集上的 *ATE* 降低到亚米级。

## 1. 引言

视觉同时定位与建图（Simultaneous Localization and Mapping，SLAM）旨在从视频流中进行在线相机跟踪和稠密三维重建。近来，三维几何基础模型 [36, 72, 75] 在端到端稠密重建方面展现了卓越的能力。在一次性给定数十张图像的情况下，它们利用一个对所有视图进行注意力计算的 Transformer，以离线方式直接回归深度图和相机位姿。然而，视觉 SLAM 要求以低延迟、有界内存的方式持续进行流式推进。这使得将具有二次复杂度的三维基础模型适配到该任务上颇具挑战。

近期将几何基础模型扩展到流式视频的尝试，通常利用因果注意力或递归记忆压缩 [5, 69, 74, 89] 来代替稠密的双向注意力。虽然这些方法实现了恒定的单帧运行时间和有界内存，但它们牺牲了局部重建精度，而且因果方法的顺序链条不可避免地会导致漂移。相比之下，AMB3R-VO [70] 提出了一种与模型无关的视觉里程计流程，通过在一组紧凑的活动关键帧内保留完整的双向注意力，化解了流式效率与几何保真度之间的权衡。然而，AMB3R-VO 仍然只是一个视觉里程计流程；由于缺乏全局优化和长程约束，它在长轨迹上依然会累积漂移。

为了弥合前馈式视觉里程计与大规模 SLAM 之间的差距，我们提出了 AMB3R-SLAM，这是一个实时单目 SLAM 系统，能够在单块消费级 GPU 上重建超过 1 万帧、长达千米级的轨迹。与现有基于优化且使用稠密重建先验的 SLAM 系统 [41, 47, 86] 相比，我们的设计原则是依赖几何基础模型的前馈预测，并避免依赖事后估计（例如对应关系）的优化。通过这种方式，我们的系统能够继承从海量数据中学到的丰富几何先验，在各种具有挑战性的条件下保持鲁棒。

AMB3R-SLAM 由一个执行低延迟逐帧相机跟踪的前端和一个分层后端组成；后端通过带跨度为 2（span-2）边的稠密局部子图构建、稀疏长上下文建图以及回环检测，逐步保证局部、中层和全局一致性。由于避免了仅适用于静态世界的光束法平差，我们的方法无需额外处理即可自然地应对复杂的动态场景。此外，我们证明了该方法可以轻松扩展到双目、RGB-D 和 LiDAR 相机。我们在 9 个数据集上的实验结果表明，AMB3R-SLAM 取得了强劲的相机跟踪结果，并在 VBR [2] 和 Oxford Spires [63] 数据集上将此前最先进方法的绝对轨迹误差降低了 70% 以上。在加入 LiDAR 作为额外输入后，我们可以进一步将 VBR [2] 和 KITTI [20] 数据集上的跟踪误差降低到亚米级。

## 2. 相关工作

### 2.1. 几何基础模型

几何基础模型 [24, 32, 36, 72, 75] 旨在以前馈方式预测稠密几何与相机位姿，所遵循的范式要么是增量式范式 [4, 6, 7, 28, 31, 35, 38, 39, 42, 69, 74, 76, 78, 83, 93]，要么是全局范式 [12, 15, 19, 27, 62, 68, 70, 72, 76, 80, 81]。学到的数据驱动先验带来了前所未有的泛化能力，以及对各种挑战性条件的鲁棒性。通过从多样化的视频数据中学习，这也使动态场景重建 [16, 25, 33, 43, 64, 82, 90] 成为可能。对于大规模场景重建，若干工作 [11, 26, 79] 将视频划分为相互重叠的块，并通过全局优化将它们拼接建图。然而，这些方法虽然能够进行视频建图，但从构造上讲都是基于分块的离线方法。另有若干方法 [5, 70, 85, 89] 则直接利用几何基础模型本身来执行视觉里程计。然而，由于缺乏全局一致性，这些方法往往会产生漂移。

### 2.2. 视觉里程计与 SLAM

视觉里程计（Visual Odometry，VO）[14, 17, 49, 66, 70] 旨在估计相邻帧之间的 6 自由度相机自运动，从而实时跟踪局部轨迹；而视觉 SLAM [8, 29, 46] 则同时通过回环检测和全局优化构建全局一致的三维地图，以校正累积漂移。传统视觉 SLAM 大致可分为基于特征的方法 [46, 54] 和直接法 [13, 14]。除单目 SLAM 之外，多模态 SLAM 可以融合深度 [48, 77]、双目 [45] 或 LiDAR 传感器 [58, 59, 67, 87]，以提供鲁棒的度量尺度相机轨迹估计。随后，研究转向利用学习得到的稠密对应关系与可微光束法平差 [37, 65, 66]、神经隐式表示 [30, 52, 61, 71, 92] 或三维高斯溅射（3D Gaussian Splatting）[22, 44, 84] 来实现 SLAM。近来，几何基础模型 [32, 72] 的进展使 SLAM 能够在未标定的设置下运行 [41, 47, 86]。然而，这些工作仍难以扩展到长时程、大规模的环境。在本工作中，我们通过一个逐步保证局部、中层和全局一致性的分层后端，将未标定视觉 SLAM 的前沿推进到千米级序列。我们的系统还可以扩展为接收双目、RGB-D 和 LiDAR 作为额外输入，以实现度量尺度的相机跟踪与重建。

## 3. 方法

图 2 给出了我们方法的概览。AMB3R-SLAM 将用于低延迟相机跟踪的轻量级前端与一个逐步施加局部、中层和全局一致性的分层后端相结合：相互重叠的子图通过跨度为 2（Span-2）的边提供精确的局部重建与漂移校正，稀疏长上下文建图引入中程约束，而回环则施加全局一致性。我们进一步表明，AMB3R-SLAM 无需任何针对特定任务的适配即可自然地处理动态场景，并能融合包括双目、RGB-D 和 LiDAR 在内的额外模态。

### 3.1. 前端

前端在当前子图内对每一个新帧执行低延迟的在线相机位姿跟踪。为了保持高帧率，我们在前端采用了轻量级骨干网络，即具有 8000 万参数的 DA3-Small [36]。对于每个新帧，我们用锚定关键帧和最近的两帧构建一个紧凑的记忆，以估计相机位姿，并像 AMB3R-VO [70] 中那样通过鲁棒尺度求解器求解尺度因子。我们在表 1 中展示了其效率与精度。前端的置信度被用于后端中的视图选择，并且我们持续地利用后端估计对前端重新锚定，以重置累积的局部漂移。

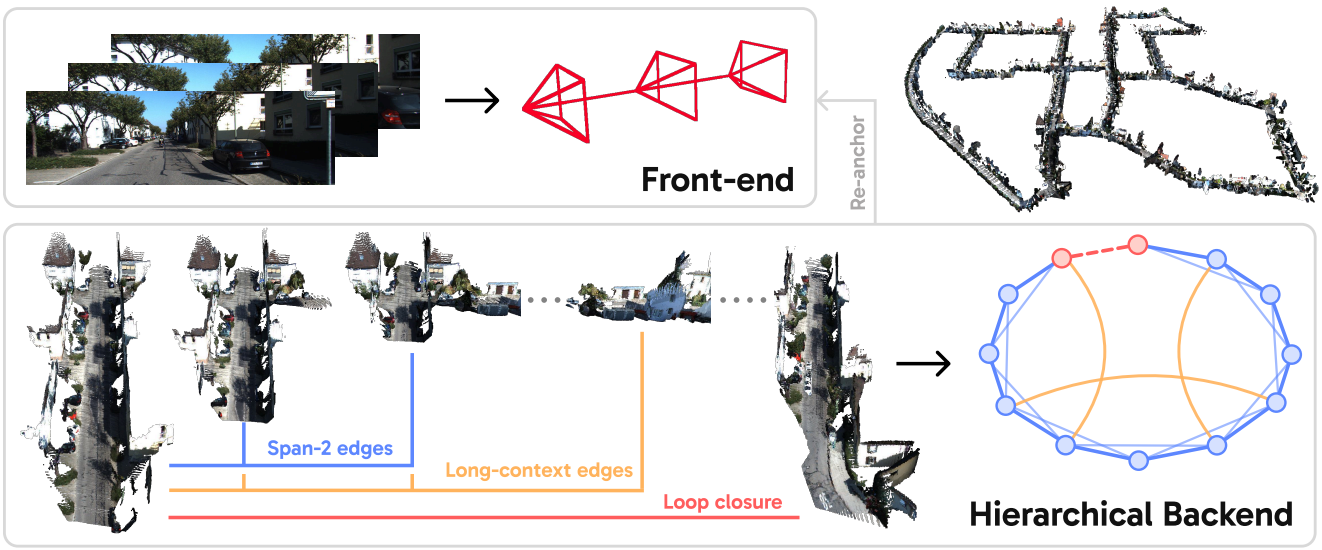



图 2. **AMB3R-SLAM 概览**。 AMB3R-SLAM 由一个执行低延迟相机位姿跟踪的前端和一个分层后端组成；后端通过包含跨度为 2 的边、长上下文边和回环边的位姿图优化，逐步保证局部、中层和全局一致性。

表 1. **前端漂移与效率评测**。前端漂移在每个建图窗口内、重新锚定之前进行评估。由于分辨率不同，运行速度（FPS）和峰值显存占用按数据集分别报告。

| 方法 | KITTI [20] | | | | TUM [60] | | | |
|---|---|---|---|---|---|---|---|---|
| | $R_{\text{err}}\downarrow$ | ATE%↓ | FPS | Mem [GB] | $R_{\text{err}}\downarrow$ | ATE%↓ | FPS | Mem [GB] |
| 双视图 | 2.61 | 12.09 | **105.6** | **0.30** | 5.87 | 8.84 | **70.3** | **0.45** |
| 本文（S） | 1.41 | 4.43 | 78.3 | 0.40 | 2.78 | 5.08 | 44.8 | 0.70 |
| 本文（B） | **1.40** | **2.87** | 48.0 | 1.12 | **1.60** | **4.33** | 22.9 | 1.72 |

### 3.2. 分层后端

**稠密局部建图**。虽然前端提供了低延迟跟踪，但其紧凑的上下文不可避免地会随时间出现跟踪漂移。为了在足够宽的基线上利用多视图信息，后端每隔 $\Delta$ 帧在一个包含 $n$ 个连续帧的滑动窗口上构建稠密子图 $\mathcal{M}_k$。我们设 $\Delta = n/3$，这保证了每个子图不仅与其相邻子图 $\mathcal{M}_{k+1}$ 重叠，还与 $\mathcal{M}_{k+2}$ 重叠，从而自然地支持 Span-2 边。为了限制后端计算开销，我们在每个子图内以步长 $d_{\text{stride}}$ 进行均匀的时间分组。后端只重建每组中置信度最高的帧，并对中间位姿进行插值，使模型能够以少量上下文帧感知宽基线。在这里，我们部署了一个大型基础模型 [36, 73]，并将每个子图的中间帧设为参考视图，以减少子图边界处的位姿漂移。重叠子图之间的坐标对齐沿用 AMB3R-VO [70] 的做法。

**稀疏长上下文建图**。局部子图只与其 Span-2 邻居相连。在没有物理回访的情况下，这种顺序结构不可避免地会在长轨迹上累积漂移。为了在不依赖回环的情况下抑制漂移，我们额外在多个子图上执行稀疏长上下文建图，为位姿图提供长程约束。每隔 $M$ 个子图，我们构建一个覆盖 $K$ 个连续子图的跨子图窗口。为了限制计算开销，我们在这 $K$ 个子图中采样一组稀疏的关键帧，得到一个中间重建 $\mathcal{C}$。我们从 $\mathcal{C}$ 中提取相距较远的子图 $\mathcal{M}_i$ 与 $\mathcal{M}_j$（$|i-j| \geq 3$）之间的相似变换 $G_{ij} \in \text{Sim}(3)$，以向位姿图注入长跨度边。由于模型在使用稀疏关键帧进行长上下文建图时性能可能下降，我们仅在满足以下条件时才将边 $G_{ij}$ 纳入图中：(i) $\mathcal{C}$ 的置信度保持在局部子图置信度的一定比例以内；(ii) $\mathcal{M}_i$ 与 $\mathcal{M}_j$ 之间的共视程度足够。

**回环检测**。尽管稀疏长上下文建图缓解了漂移，千米级轨迹仍然需要显式的回环来施加全局一致性。我们采用 DBoW2 [18] 在历史子图 $\mathcal{M}_i$ 与当前子图 $\mathcal{M}_j$ 之间提出回环候选。对于每个候选，我们从两个子图中采样视图，并用我们的基础模型对其进行联合重建，计算三维体素占据重叠度以进行局部几何验证。我们滤除旋转和平移校正量不合理的回环边，以防止在重复性环境中出现感知混淆（perceptual aliasing）。通过验证的回环边随后进入式 1 中的位姿图优化。

**位姿图优化**。我们在关键帧节点 $\{S_i \in \text{Sim}(3)\}$ 上构建一个 Sim(3) 位姿图，用于全局轨迹优化。在 Span-2 边、长跨度边和回环边 $(i,j) \in \mathcal{E}$ 上的目标函数表述为：

$$\{S_i^\star\} = \arg\min_{\{S_i\}} \sum_{(i,j)\in\mathcal{E}} \rho\Big(\mathbf{r}_{ij}^T \mathbf{W}_{ij}^T \mathbf{W}_{ij} \mathbf{r}_{ij}\Big), \tag{1}$$

其中 $\mathbf{r}_{ij} = \log_{\text{Sim}(3)}\left(G_{ij}^{-1} S_i^{-1} S_j\right) = [\boldsymbol{\nu}_{ij}^T, \boldsymbol{\phi}_{ij}^T, \sigma_{ij}]^T \in \mathfrak{sim}(3) \cong \mathbb{R}^7$ 将误差分解为平移（$\boldsymbol{\nu} \in \mathbb{R}^3$）、旋转（$\boldsymbol{\phi} \in \mathbb{R}^3$）和对数尺度（$\sigma \in \mathbb{R}$）。为了在不同距离下平衡平移、旋转与尺度，我们将白化矩阵定义为：

$$\mathbf{W}_{ij} = \sqrt{w_{ij}}\, \text{diag}\left(\frac{\mathbf{I}_3}{d_{ij}}, \mathbf{I}_3, 1\right), \tag{2}$$

其中 $w_{ij}$ 为边权重，$d_{ij} = \max(\|\mathbf{t}_{ij}\|, d_{\min})$ 按边的基线长度缩放平移误差，较小的下限 $d_{\min}$ 用于避免除以零。这衡量的是相对漂移而非绝对漂移，从而防止长程

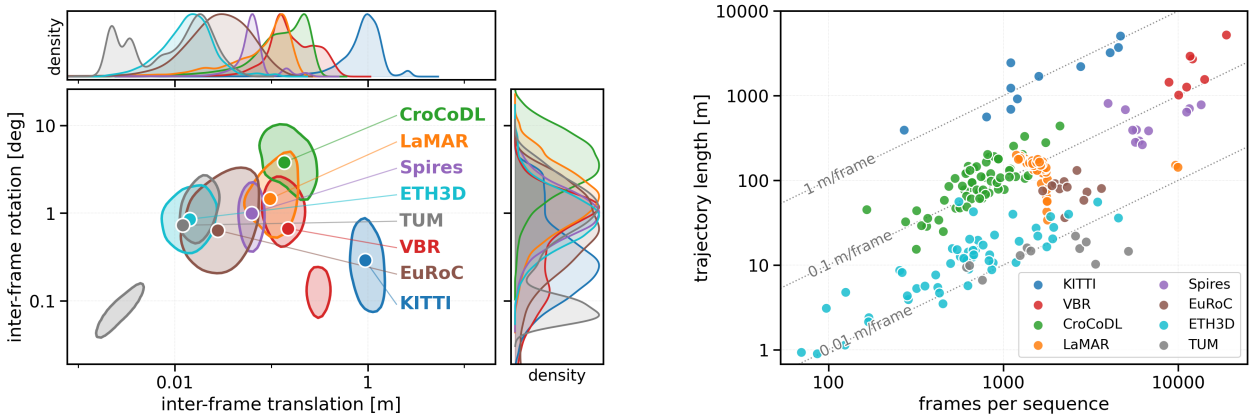



图 3. **评测数据集统计**。我们可视化了相邻帧之间旋转与平移的统计量，以及每个序列的帧数和轨迹长度。对于 LaMAR，我们按其真值（gt）帧步长对统计量进行了校准。

平移误差主导优化。我们采用 Huber 损失 $\rho$ 来抑制外点。得到的 Sim(3) 校正量会被传播到所有帧，在重叠窗口上使用加权位姿平均 [70]。

### 3.3. 对动态场景的鲁棒性

传统视觉 SLAM 系统严重依赖光束法平差，而它本质上假设世界是静态的，在存在动态物体时很容易失效。相比之下，我们的框架在设计上就消除了静态场景假设。首先，前馈式基础模型 [36, 73] 在多样化的动态数据上训练，自然地学会了将刚性的相机自运动与运动物体的刚性或非刚性运动解耦。其次，我们的分层后端所优化的 Sim(3) 位姿图严格基于相机之间的相对位姿，而非跨越运动像素的 2D–3D 重投影误差。因此，我们的系统无需显式的运动分割，即可直接处理复杂的动态环境。

### 3.4. 扩展到双目与 RGB-D

单目 SLAM 存在尺度模糊问题，在缺乏回环的情况下会导致连续子图间的累积漂移。在给定双目或 RGB-D 输入时，我们通过将子图校准到度量尺度来约束这种尺度漂移。我们直接从 RGB-D 传感器获得度量深度 $d_t^s$，或对已标定的双目图像对通过半全局匹配（semi-global matching）[21] 获得。对于每个子图 $\mathcal{M}_k$，我们通过下式估计度量尺度因子 $\hat{s}_k$：

$$\hat{s}_k = \mathrm{median}_{t\in\mathcal{M}_k,\,\mathbf{u}\in\mathcal{V}_{k,t}} \left( \frac{d_t^s(\mathbf{u})}{d_{k,t}(\mathbf{u})} \right), \tag{3}$$

其中 $\mathcal{V}_{k,t}$ 表示有效像素。我们舍弃深度内点比例较低的尺度估计，以避免非线性的形状畸变。一旦子图被校准到度量尺度，它们的相对尺度就固定为 $s_{ij} = 1.0$。因此，式 1 中的位姿图优化从 Sim(3) 退化为 SE(3)，从而防止沿轨迹的累积尺度漂移。

### 3.5. 扩展到 LiDAR

双目或结构光传感器由于基线三角测量和信号衰减，其深度误差随距离按 $O(z^2)$ 增长；与之不同，LiDAR 通过直接的脉冲飞行时间（time-of-flight）测距，在数百米范围内具有几乎与距离无关的 $O(1)$ 误差，精度可达厘米级。当 LiDAR 可用时，我们改为利用基于 ICP 的方法 [67] 来获得位姿图中的相对位姿。当检测到回环时，我们利用几何基础模型的鲁棒性来估计初始回环边，再进行基于 ICP 的点云精化。由于每条回环边的配准质量可能不同，我们利用其内点均方根误差（RMSE）$\varepsilon_{ij}$ 对回环边进行加权：

$$w_{ij} = \bar{w} \left( \frac{\bar{\varepsilon}}{\varepsilon_{ij}} \right)^2, \tag{4}$$

其中 $\bar{w}$ 为回环边权重，$\varepsilon_{ij}$ 为 ICP 配准的内点 RMSE，$\bar{\varepsilon}$ 为所有候选回环的 RMSE 中位数。通过这种方式，高质量的回环边被赋予更大的权重，而潜在的错误对齐则被抑制，从而确保位姿图优化的鲁棒性。

## 4. 实验

### 4.1. 实验设置

**基准**。我们在 9 个数据集上评测 AMB3R-SLAM，其轨迹统计如图 3 所示。对于千米级室外跟踪，我们使用 KITTI [20]（0.3k–4.7k 帧，0.4–5.1 km）、VBR [2]（8.8k–18.8k 帧，1.0–5.2 km）和 Oxford Spires [63]（4.0k–13.5k 帧，0.26–0.80 km）。对于包含室内外混合穿行的建筑级采集，我们使用 LaMAR [55] 和 CroCoDL [1]。对于房间级序列，我们使用 TUM RGB-D [60]、ETH3D-SLAM [57] 和 EuRoC [3]。对于动态场景，我们使用 Bonn 动态数据集 [50]。这些数据集合在一起，覆盖了从几米范围的桌面运动到 5 km 连续城市驾驶的各种情形。

**基线方法**。我们主要与最先进的流式 VO/SLAM 系统进行比较：LoGeR [89]、LingBot-Map [5]、MASt3R-SLAM [47]、VGGT-SLAM2 [40] 和 AMB3R-VO [70]。离线方法 [9, 10, 51, 53, 79]（包括基于分块的方法 [11, 36, 76]）仅作参考展示。在动态场景上，我们还额外与专为运动物体设计的方法 [34, 91] 进行比较。为了将我

表 2. **KITTI 上的 ATE [m]↓ 比较。**

| 类别 | 方法 | 00<br>4.5k/3.7km | 01<br>1.1k/2.5km | 02<br>4.7k/5.1km | 03<br>0.8k/0.6km | 04<br>0.3k/0.4km | 05<br>2.8k/2.2km | 06<br>1.1k/1.2km | 07<br>1.1k/0.7km | 08<br>4.1k/3.2km | 09<br>1.6k/1.7km | 10<br>1.2k/0.9km | 平均 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 离线 | Scal3R [79] | 4.30 | 45.29 | 42.06 | 3.36 | 1.74 | 3.30 | 2.49 | 2.03 | 36.69 | 12.32 | 6.46 | 14.55 |
| | Glob3R [10] | 2.77 | 53.12 | 30.10 | 1.53 | 0.75 | 2.80 | 2.67 | 1.77 | 32.76 | 5.51 | 11.51 | 13.21 |
| 分块 | VGGT-Long [11] | 8.64 | 61.21 | 52.72 | 8.78 | 4.20 | 9.88 | 4.67 | 2.66 | 72.98 | 31.84 | 27.71 | 25.94 |
| | $\pi$3-Long [76] | 5.55 | 114.83 | 50.29 | 1.63 | 1.11 | 3.48 | 2.88 | 3.92 | 24.25 | 7.38 | 17.61 | 21.17 |
| | DA3-Long [36] | 5.13 | 78.76 | 35.64 | 5.38 | 3.18 | 3.04 | 2.83 | 2.32 | 26.55 | 8.86 | 13.42 | 16.83 |
| 在线 | DROID-SLAM [65] | 92.10 | 344.60 | 107.61 | 2.38 | 1.00 | 118.50 | 62.47 | 21.78 | 161.60 | 72.32 | 118.70 | 100.28 |
| | VGGT-SLAM2 [40] | TL | 163.65 | TL | 50.04 | 19.38 | 159.58 | 46.35 | 57.80 | TL | 167.96 | 76.99 | 92.72 |
| | DA3-SLAM | 91.07 | 263.25 | 136.08 | 3.72 | 6.79 | 50.41 | 115.39 | 33.34 | 28.67 | 67.67 | 14.36 | 73.71 |
| | MASt3R-SLAM [47] | OOM | 530.37 | OOM | 18.87 | 88.99 | 159.43 | 92.00 | OOM | 263.75 | TL | 153.07 | 186.64 |
| | AMB3R-VO [70] | 167.38 | 276.12 | 157.07 | 22.06 | 6.91 | 149.11 | 55.94 | 37.25 | 85.68 | 134.73 | 52.15 | 104.04 |
| | LingBot-Map [5] | 27.17 | 70.94 | 112.02 | 2.02 | 1.36 | 26.14 | 16.61 | 10.48 | 23.82 | 17.84 | 6.48 | 28.63 |
| | LoGeR [89] | 30.47 | 47.91 | 36.32 | 5.38 | 1.95 | 26.34 | 6.60 | 5.55 | 24.41 | 10.12 | 10.11 | 18.65 |
| | **本文** | 3.67 | 55.91 | 31.55 | 4.80 | 2.04 | 2.68 | 3.33 | 1.73 | 23.56 | 9.03 | 5.96 | 13.11 |
| | **本文（双目）** | 2.67 | 56.04 | 29.89 | 2.29 | 0.74 | 1.71 | 3.69 | 1.46 | 23.46 | 5.63 | 6.05 | 12.15 |
| LiDAR | PIN-SLAM [52] | 0.87 | 5.23 | 2.34 | 0.44 | **0.11** | **0.26** | **0.15** | **0.27** | **1.85** | 1.20 | **0.71** | 1.22 |
| | ↪ SE(3) | 0.88 | 5.43 | 2.71 | 0.72 | **0.12** | **0.31** | 0.45 | **0.30** | **1.88** | 1.23 | **0.74** | 1.34 |
| | **本文（LiDAR）** | **0.86** | **2.95** | **1.53** | **0.27** | 0.16 | 0.28 | 0.19 | 0.28 | 2.03 | **1.12** | 0.74 | **0.95** |
| | ↪ SE(3) | **0.86** | **3.07** | **1.70** | **0.45** | 0.28 | **0.31** | **0.29** | **0.30** | 2.03 | **1.13** | 0.76 | **1.01** |

表 3. **Oxford Spires 上的 ATE [m]↓ 比较。**

| 方法 | keb_02<br>6.0k/292m | keb_03<br>5.7k/282m | keb_04<br>13.5k/779m | keb_05<br>11.5k/701m | obs_01<br>5.7k/396m | obs_02<br>5.5k/391m | ble_05<br>6.8k/384m | chr_02<br>11.2k/637m | chr_03<br>6.2k/262m | chr_05<br>4.0k/801m | bod_02<br>5.0k/685m | 平均 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGGT-SLAM2 [40] | 20.56 | 16.20 | 24.71 | 30.02 | 8.97 | 7.42 | **3.56** | 30.09 | 4.90 | 28.13 | 63.24 | 21.62 |
| DA3-SLAM | 2.86 | 3.88 | 4.34 | 7.97 | 18.69 | 3.87 | 14.84 | 8.89 | 6.93 | 15.48 | 44.78 | 12.05 |
| LoGeR [89] | 6.14 | 4.50 | 5.35 | 13.37 | 7.35 | 8.76 | 5.51 | 9.33 | 1.46 | 21.79 | 10.92 | 8.59 |
| LingBot-Map [5] | 4.30 | 2.15 | 4.37 | 6.65 | 4.33 | 3.53 | 6.99 | 13.34 | 2.20 | 36.05 | 10.91 | 8.62 |
| **本文 (Ω)** | **0.75** | **0.34** | **2.03** | **0.74** | **0.40** | **0.75** | 5.60 | **0.72** | **0.19** | **1.27** | 10.15 | **2.08** |
| **本文** | 0.80 | 0.55 | 2.06 | 1.15 | 1.78 | 1.74 | 6.74 | 2.23 | 0.25 | 2.09 | **6.11** | 2.32 |

表 4. **VBR 上的 ATE [m]↓ 比较。**

| | 方法 | col_0<br>8.8k/1.5km | cam_0<br>12.0k/2.7km | cam_1<br>11.7k/3.0km | pin_0<br>11.1k/1.3km | spa_0<br>14.1k/1.6km | dia_0<br>10.0k/1.0km | cia_1<br>18.8k/5.2km | 平均 |
|---|---|---|---|---|---|---|---|---|---|
| 离线 | VGGT-Long [11] | 45.73 | 132.68 | 115.65 | 64.40 | 58.49 | 33.66 | 187.97 | 91.23 |
| | $\pi$3-Long [76] | 81.63 | 86.54 | 71.49 | 49.85 | 54.27 | 28.07 | 118.81 | 70.09 |
| | DA3-Long [36] | 65.78 | 88.12 | 102.28 | 41.08 | 55.93 | 4.15 | 11.21 | 52.65 |
| 在线 | VGGT-SLAM [41] | 102.92 | 110.77 | 89.69 | 72.98 | 62.67 | 35.62 | 144.02 | 88.38 |
| | DA3-SLAM | 97.19 | 93.57 | 91.61 | 81.82 | 65.44 | 36.61 | 192.95 | 94.17 |
| | LoGeR [89] | 49.38 | 22.44 | 34.90 | 11.08 | 27.05 | 32.88 | 44.51 | 31.75 |
| | LingBot-Map [5] | 15.02 | 20.34 | 9.94 | 30.60 | 23.59 | 22.75 | 64.34 | 26.65 |
| | **本文 (Ω)** | **2.52** | 26.43 | **3.72** | **3.95** | **5.10** | **2.66** | **5.66** | **7.15** |
| | **本文** | 3.95 | **9.29** | 6.73 | 10.75 | 5.94 | 3.36 | 11.92 | 7.42 |
| LiDAR | PIN-SLAM [52] | 1.69 | **0.74** | **0.14** | 0.46 | 0.31 | 1.10 | 1.24 | 0.81 |
| | ↪ SE(3) | 1.70 | **0.74** | **0.17** | 0.50 | 0.32 | 1.10 | 1.25 | 0.83 |
| | **本文（LiDAR）** | **0.22** | 1.01 | 0.28 | **0.18** | **0.10** | **0.25** | **0.50** | **0.36** |
| | ↪ SE(3) | **0.22** | 1.02 | 0.28 | **0.37** | **0.10** | **0.25** | **0.52** | **0.39** |

们系统本身的贡献与其骨干网络的贡献区分开，我们进一步将 DepthAnything 3 [36] 集成到 VGGT-SLAM 流程中，并将这一受控基线称为 DA3-SLAM。

**评价指标**。遵循先前工作 [5, 47, 70, 89, 91]，我们在 KITTI、VBR、Oxford Spires、TUM 和 Bonn 上报告绝对轨迹误差（ATE RMSE）。遵循 VidMap [53]，我们在 ETH3D-SLAM 和 EuRoC 上报告平移误差在 5 cm 到 10 m 阈值范围内的 AUC。在 LaMAR 和 CroCoDL 上，以及在 VBR 上的消融实验中，我们报告 VidMap [53] 的窗口化 AUC（W-AUC），它以给定长度的轨迹片段为单位进行评估，阈值设为该长度的 5%。

**实现细节**。我们的前端使用 DA3-Small（0.08B），后端使用一个 Giant 规模的模型 [36]；**Ours (Ω)** 表示使用 VGGT-Ω [73] 作为后端的同一流程，我们将其纳入以表明我们的系统与底层几何基础模型无关。除非另有说明，**Ours** 表示单目系统。

## 4.2. 评测

**大规模室外场景**。在千米尺度上，AMB3R-SLAM 是唯一能够匹敌乃至超越离线全局重建方法的在线方法（表 2–3）。在 KITTI 上（表 2），我们的单目版本取得了 13.11 m 的平均 ATE，比领先的在线竞争者（LoGeR，18.65 m）低 30%，比最好的基于分块的方法（DA3-Long，16.83 m）低 22%，并略优于离线 SfM 方法（Glob3R [10]，13.21 m）。在更多样化的室外轨迹上评测时，我们的优势变得更大。在 VBR 上（表 4），其城市序列最长可达 18.8k 帧、5.2 km，AMB3R-SLAM 将平均 ATE 从 26.65 m（LingBot-Map）和 31.75 m（LoGeR）降低到 7.42 m（使用 Ω 时为 7.15 m），误差降低了 72%，同时也显著优于所有基于分块的方法。在 Oxford Spires 上（表 3 和图 4），AMB3R-SLAM 将平均 ATE 从 8.59 m（LoGeR）降低到 2.32 m（使用 Ω 时为 2.08 m），降幅达 73%。我们的方法在包含具有挑战性的“室内—室外—室内”过渡的 Christ Church 5 序列上展现出显著更好的相机跟踪性能，体现了其在不同工况下的鲁棒性。

**建筑级室内外采集**。遵循 VidMap [53]，我们在表 5 中于 LaMAR [55] 和 CroCoDL [1] 上进行评测，二者的轨迹平均都超过 100 m，且包含具有挑战性的条件和分布外环境。与现有的 SLAM 系统（如 DA3-SLAM）相比，我们方法的优势随窗口长度增加而扩大。在 CroCoDL（一个在分布外的灾害现场建筑中采集视频的数据集）

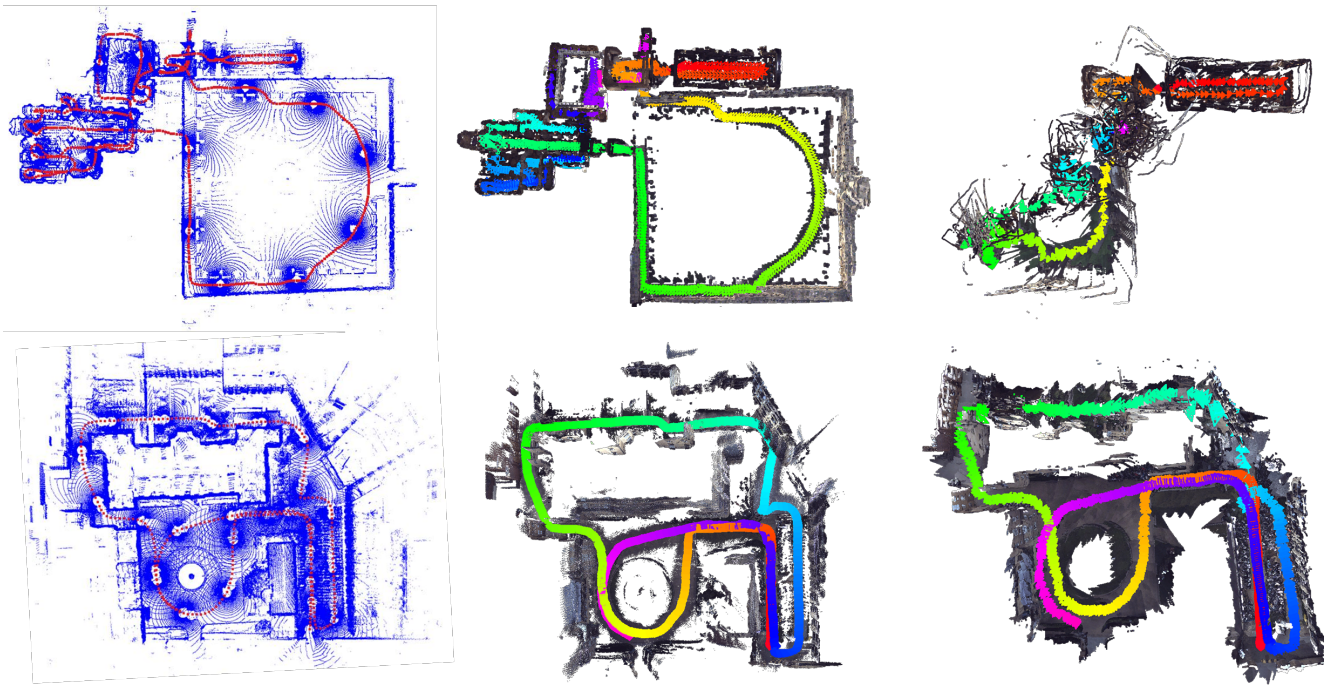



图 4. **Oxford Spires 数据集上的定性展示。**

表 5. **在 LaMAR 和 CroCoDL 上的评测（未标定）**。我们报告平移误差的窗口化 AUC [53]。

| | 方法 | LaMAR [55] | | | | | CroCoDL [1] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 10 m | 25 m | 50 m | 100 m | Full | 10 m | 25 m | 50 m | 100 m | Full |
| 离线 | GLOMAP [51] | 76.4 | 69.4 | 60.1 | 53.6 | 62.3 | 77.8 | 66.6 | 60.0 | 65.9 | 76.4 |
| | DA3-Long [36] | 86.7 | 86.7 | 84.6 | 78.1 | 76.5 | 86.2 | 88.1 | 88.5 | 88.8 | 88.9 |
| | VidMap [53] | **91.9** | **92.3** | **89.9** | **89.3** | **88.5** | **94.0** | **95.0** | **95.3** | **93.0** | **95.5** |
| 在线 | LoGeR [89] | 65.9 | 64.5 | 60.3 | 60.1 | 59.8 | 50.0 | 56.7 | 57.9 | 63.7 | 63.8 |
| | LingBot-Map [5] | 69.2 | 75.6 | 75.0 | 72.7 | 69.6 | 54.4 | 66.3 | 69.9 | 72.2 | 72.0 |
| | VGGT-SLAM2 [40] | 74.0 | 69.2 | 60.3 | 49.4 | 56.9 | 62.8 | 64.3 | 66.0 | 62.5 | 74.9 |
| | DA3-SLAM | 82.5 | 76.4 | 67.9 | 60.8 | 62.5 | 77.1 | 72.5 | 65.9 | 57.6 | 70.9 |
| | MASt3R-SLAM [47] | 60.8 | 63.6 | 55.6 | 47.3 | 51.5 | 51.8 | 57.3 | 60.2 | 65.2 | 70.0 |
| | MegaSaM [34] | 77.5 | 68.4 | 59.9 | 51.4 | 52.7 | 50.2 | 50.2 | 43.5 | 35.2 | 50.1 |
| | ViPE [23] | 86.6 | 83.8 | 80.2 | 78.1 | 76.6 | 66.7 | 68.2 | 68.6 | 70.7 | 71.4 |
| | **本文 (Ω)** | **88.9** | **89.3** | **87.7** | **85.5** | **84.7** | 88.6 | 90.1 | **91.0** | **92.2** | **91.7** |
| | **本文** | 88.7 | 87.2 | 83.8 | 78.0 | 75.7 | **90.9** | **90.8** | 90.1 | 91.8 | 91.2 |

表 6. **TUM RGB-D 上的 ATE [cm]↓ 比较。** † 表示仅在关键帧位姿上进行评估的方法。

| | 方法 | 360 | desk | desk2 | floor | plant | room | rpy | teddy | xyz | 平均 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 离线 | Scal3R [79] | 6.6 | 5.5 | 2.9 | 18.9 | 4.5 | 11.0 | 3.2 | 9.5 | 5.0 | 7.4 |
| | Glob3R [10] | 8.1 | 1.8 | 2.6 | 3.0 | **1.8** | 5.0 | 2.2 | 3.7 | 0.9 | 3.2 |
| | SAILRecon [9] | 7.0 | 2.4 | 4.2 | 10.7 | 3.1 | 11.3 | 2.0 | 3.7 | 1.2 | 5.1 |
| 分块 | VGGT-Long [11] | 11.8 | 5.8 | 11.1 | 11.8 | 7.1 | 15.5 | 14.0 | 12.0 | 9.9 | 11.0 |
| | π3-Long [76] | 11.5 | 4.7 | 5.2 | 16.0 | 8.5 | 11.4 | 14.3 | 8.1 | 5.2 | 9.4 |
| | DA3-Long [36] | 5.9 | 3.4 | 4.2 | 10.7 | 6.0 | 10.5 | 20.6 | 12.6 | 4.4 | 8.7 |
| 在线 | DROID-SLAM† [65] | 20.2 | 3.2 | 9.1 | 6.4 | 4.5 | 91.8 | 5.6 | 4.5 | 1.2 | 15.8 |
| | MASt3R-SLAM† [47] | 7.0 | 3.5 | 5.5 | 5.6 | 3.5 | 11.8 | 4.1 | 11.4 | 2.0 | 6.0 |
| | LoGeR [89] | 10.6 | 3.2 | 4.6 | 10.3 | 4.6 | 9.8 | 3.0 | 8.4 | 2.2 | 6.3 |
| | MUSt3R [4] | 7.8 | 5.1 | 7.1 | 5.0 | 4.0 | 9.9 | 4.3 | 4.2 | 1.3 | 5.4 |
| | VGGT-SLAM2† [40] | 5.0 | 2.5 | 2.9 | 10.2 | 2.6 | 6.3 | 2.6 | 3.8 | 1.4 | 4.1 |
| | LingBot-Map [5] | 6.3 | 2.8 | 4.4 | 5.9 | 4.4 | 9.2 | 2.4 | 3.6 | 1.1 | 4.4 |
| | AMB3R-VO [70] | 4.6 | 1.9 | 2.8 | 3.2 | 2.9 | 5.8 | 2.3 | 3.7 | 1.1 | 3.2 |
| | **本文 (单目)** | **3.6** | **1.4** | **2.2** | **2.6** | 2.3 | 2.9 | **1.8** | 2.9 | **0.7** | 2.3 |
| | **本文 (RGB-D)** | **3.6** | **1.4** | **2.2** | 2.7 | 2.0 | **2.8** | **1.8** | **2.8** | 0.8 | **2.2** |

上，我们的方法以显著优势超越所有在线方法，并超过了离线的 GLOMAP（76.4%）和 DA3-Long（88.9%）。这些结果证明了我们的系统在严重的域外条件下防止漂移的有效性。

**室内场景**。房间级序列测试的是相反的工况：此时累积漂移很小，跟踪质量主要由局部精度决定。在这种情况下，我们的方法同样始终领先。在 TUM RGB-D 上（表 6），AMB3R-SLAM 取得了 2.3 cm 的平均 ATE，比此前最好的在线方法 AMB3R-VO（3.2 cm）误差降低 28%。值得注意的是，尽管为保持实时性能，我们的基础模型使用了明显更小的帧预算，我们仍然超越了所有离线和基于分块的方法。这表明我们的系统并没有以牺牲局部保真度为代价换取全局一致性。在 ETH3D-SLAM 上（表 7），AMB3R-SLAM 将 AUC@5 cm 从 38.7%（ViPE）提升到 44.0%，并在每个阈值下都保持领先。在 EuRoC 上，快速的无人机运动导致所有先前的在线系统在严格阈值下崩溃，而我们的方法依然鲁棒，甚至超过了经过大量优化的离线方法 VidMap（11.5）。

**动态场景**。我们还展示了在 Bonn 动态数据集上的性能（表 8）。AMB3R-SLAM 取得了 1.3 cm 的平均 ATE，并在全部 8 个序列上都是最精确的方法，相比 WildGS-SLAM（2.3 cm）误差降低 43%，相比 MegaSaM（3.5 cm）降低 63%。值得注意的是，这些结果使用的是与静态场景完全相同的流程，没有任何针对特定任务的适配。

表 7. **在 ETH3D-SLAM 和 EuRoC 上的评测（未标定）**。我们报告平移误差的 AUC [53]。

| | 方法 | ETH3D-SLAM [57] | | | | | EuRoC [3] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 5 cm | 10 cm | 50 cm | 1 m | 10 m | 5 cm | 10 cm | 50 cm | 1 m | 10 m |
| 离线 | GLOMAP [51] | 20.4 | 30.9 | 58.7 | 67.4 | 77.8 | 0.4 | 1.8 | 16.1 | 29.7 | 70.7 |
| 离线 | DA3-Long [36] | 19.0 | 37.2 | 73.6 | 84.4 | 98.0 | 0.1 | 0.8 | 31.4 | 57.6 | 95.4 |
| 离线 | VidMap [53] | **51.4** | **68.3** | **90.1** | **94.0** | **99.0** | **11.5** | **31.6** | **79.1** | **89.4** | **99.0** |
| 在线 | LoGeR [89] | 6.4 | 16.9 | 57.7 | 73.8 | 95.9 | 0.1 | 0.5 | 18.2 | 45.1 | 92.6 |
| 在线 | LingBot-Map [5] | 13.1 | 27.8 | 68.3 | 81.3 | 97.2 | 0.1 | 0.7 | 32.0 | 55.5 | 91.7 |
| 在线 | VGGT-SLAM2 [40] | 26.4 | 44.1 | 75.7 | 84.3 | 97.1 | 1.3 | 7.8 | 55.4 | 74.2 | 96.9 |
| 在线 | MASt3R-SLAM [47] | 8.9 | 21.6 | 65.3 | 77.0 | 96.4 | 1.6 | 5.6 | 51.0 | 71.9 | 97.0 |
| 在线 | MegaSaM [34] | 38.4 | 53.9 | 79.2 | 86.7 | 98.0 | 0.8 | 5.7 | 50.3 | 73.2 | 97.3 |
| 在线 | ViPE [23] | 38.7 | 54.9 | 78.3 | 85.7 | 97.6 | 1.3 | 6.2 | 55.4 | 75.0 | 94.0 |
| 在线 | **本文（单目）** | 44.0 | 60.6 | 84.7 | 90.4 | 98.8 | **20.3** | **44.0** | **85.5** | **92.8** | **99.3** |
| 在线 | **本文（RGB-D）** | **44.8** | **61.8** | **85.3** | **90.7** | **98.9** | – | – | – | – | – |

表 8. **Bonn 动态数据集上的 ATE [cm]↓ 比较**。

| 方法 | bal_1 | bal_2 | crd_1 | crd_2 | per_1 | per_2 | mov_1 | mov_2 | 平均 |
|---|---|---|---|---|---|---|---|---|---|
| DSO [14] | 7.3 | 21.8 | 10.1 | 7.6 | 30.6 | 26.5 | 4.7 | 11.2 | 15.0 |
| DROID-SLAM [65] | 7.5 | 4.1 | 5.2 | 6.5 | 4.3 | 5.4 | 2.3 | 4.0 | 4.9 |
| MonoGS [44] | 15.3 | 17.3 | 11.3 | 7.3 | 26.4 | 35.2 | 22.2 | 47.2 | 22.8 |
| DynaMoN [56] | 6.8 | 3.8 | 6.1 | 5.6 | 2.4 | 3.5 | 1.4 | 2.6 | 4.0 |
| MonST3R [88] | 5.4 | 7.2 | 5.4 | 6.9 | 11.9 | 11.1 | 3.3 | 7.4 | 7.3 |
| MegaSaM [34] | 3.7 | 2.6 | 1.6 | 7.2 | 4.1 | 4.0 | 1.4 | 3.4 | 3.5 |
| WildGS-SLAM [91] | 2.8 | 2.4 | 1.5 | 2.3 | 3.1 | 2.7 | 1.6 | 2.2 | 2.3 |
| **本文 (Ω)** | 1.7 | 2.0 | 1.3 | 1.3 | 3.6 | 2.2 | 1.6 | 2.7 | 2.1 |
| **本文** | **1.3** | **1.4** | **1.1** | **1.2** | **1.8** | **1.3** | **1.0** | **1.1** | **1.3** |

这是因为我们的系统对相机相对位姿进行图优化，而非光束法平差，从而无需滤除运动物体或对复杂的动态对应关系进行建模。

**额外模态**。除单目跟踪外，AMB3R-SLAM 还可以通过固定位姿图的尺度（$s_{ij} = 1.0$）来融合双目和 RGB-D 传感器，将优化从 Sim(3) 限制为 SE(3)，从而缓解尺度漂移。在 KITTI 上（表 2），引入双目传感器将平均 ATE 降低到 12.15 m。在室内场景中，RGB-D 输入将 TUM（表 6）的 ATE 提升至 2.2 cm，并将 ETH3D-SLAM（表 7）的 AUC@5 cm 提升至 44.8%。

由于深度传感器的量程限制和噪声，这些提升相对有限。相比之下，加入 LiDAR 作为额外输入后，AMB3R-SLAM 的性能大幅提升。在 KITTI 上（表 2），AMB3R-SLAM（LiDAR）在 Sim(3) 下将平均 ATE 降低到 0.95 m（SE(3) 下为 1.01 m），优于基于神经隐式表示的最先进 LiDAR SLAM 方法 PIN-SLAM（1.22 m）。在 KITTI-01 上——这是一条 2.5 km 的高速公路序列，视觉特征有限且相机运动迅速，单目跟踪在此漂移严重——LiDAR 输入将 ATE 从 55.91 m 降低到 2.95 m。在相机运动更加多样的 VBR 上（表 4），AMB3R-SLAM（LiDAR）将平均 ATE 降低到 0.36 m（SE(3) 下为 0.39 m）。

### 4.3. 分析

**计算分析**。如表 9 所示，AMB3R-SLAM 在单块 NVIDIA RTX 4090 上实时运行，单目模式下达到 10.2–17.6 FPS，使用 LiDAR 时达到 31.3–47.8 FPS，同时峰值显存保持在约 10–14 GB，且不随序列长度增长。几何基础模型的推理占据了单帧延迟的绝大部分，而回环检索和 Sim(3) 图优化的开销可以忽略不计。由于优化的是子图之间的相对位姿，而非在数百万个三维点

表 9. **AMB3R-SLAM 的运行时间与显存**。我们还报告了建图、回环检索和位姿图优化的开销分解。

| 方法 | 数据集 | FPS↑ | Mem. [GB]↓ | 时间 [ms/帧]↓ | | |
|---|---|---|---|---|---|---|
| | | | | 建图 | 检索 | PGO |
| 单目 | KITTI | 17.6 | 10.3 | 56.5 | 0.7 | 0.17 |
| 单目 | VBR | 10.2 | 14.0 | 97.5 | 1.0 | 0.12 |
| LiDAR | KITTI | 47.8 | 10.5 | 20.8 | 0.7 | 0.05 |
| LiDAR | VBR | 31.3 | 14.1 | 44.0 | 1.0 | 0.04 |

表 10. **AMB3R-SLAM（单目）在 VBR 上的消融实验**。我们报告全部 7 个序列上的 W-AUC [%]↑。

| 变体 | W-AUC [%]↑ | | | | | |
|---|---|---|---|---|---|---|
| | 10 m | 50 m | 100 m | 500 m | 1000 m | 全程 |
| w/o LC | **89.38** | 86.31 | 80.94 | 55.39 | 47.82 | 42.02 |
| w/o 白化 | 88.11 | 84.65 | 80.92 | 77.47 | 82.79 | 86.97 |
| w/o SLCM | 89.15 | 86.71 | 84.28 | 84.36 | 88.35 | **90.46** |
| **本文（完整）** | 89.14 | **87.81** | **85.57** | **86.00** | **90.17** | 90.39 |

表 11. **AMB3R-SLAM（LiDAR）在 KITTI 上的消融实验**。我们报告在 3 个含回环的序列上平均的 ATE [m]↓。

| 变体 | ATE | ATE (SE(3)) |
|---|---|---|
| w/o GFM init | 1.39 | 1.45 |
| **本文（完整）** | **1.17** | **1.23** |

上进行光束法平差，我们的分层后端可以轻松扩展到数千米的序列。

**消融实验**。我们在表 10 中于 VBR 上评测了后端的各个组件。在没有回环检测（*w/o LC*）的情况下，AMB3R-SLAM 在 100 m 以内仍能保持可靠的相机跟踪，但最终会随时间漂移，W-AUC 在 500 m 处下降到 55.39%，在完整轨迹上下降到 42.02%。去掉白化矩阵（*w/o whitening*）会使所有窗口上的精度下降，因为未经归一化的平移残差会在较长基线上主导位姿图优化。没有稀疏长上下文建图（*w/o SLCM*）时，相机跟踪在回环稀疏或缺失的 50–1000 m 范围内性能下降。这表明我们的长上下文边提供了有效的中程约束，能在回环触发之前限制漂移。

在表 11 中，我们消融了几何基础模型在 LiDAR 流程中的作用。由于结合 ICP 方法的 LiDAR 本身已经相当精确，几何基础模型在这里充当了良好的先验，用于在回环时初始化配准。通过避免在视角变化下陷入局部几何极小值，这将平均 ATE 从 1.45 m 降低到 1.23 m。

**局限性与未来工作**。目前，AMB3R-SLAM 通过点云而非统一、紧凑的表面表示来表达三维结构。因此，在感知上存在歧义的区域，我们的地图可能出现重复表面或重影伪影。引入紧凑的表示（例如神经隐式场）可能有助于维持全局一致的地图。此外，虽然 AMB3R-SLAM 在系统层面通过图优化有效地融合了双目、RGB-D 和 LiDAR，但其底层的几何基础模型仅基于 RGB 输入运行，并未以这些额外模态为条件。探索具有原生多模态条件输入的基础模型，是进一步提升几何保真度和跟

踪精度的一个有前景的方向。

## 5. 结论

我们提出了 AMB3R-SLAM，这是一个实时单目 SLAM 系统，能够在单块消费级 GPU 上对超过 1 万帧的序列进行千米级跟踪。通过将轻量级前端与分层后端相结合，我们的系统逐步施加局部、中层和全局一致性，有效地扩展到大规模重建。由于我们的系统不依赖基于静态世界假设的光束法平差，我们表明它无需额外处理即可应对复杂的动态环境。此外，我们还表明该系统可以扩展到包括双目、RGB-D 和 LiDAR 在内的额外模态，以进一步提升其在各种挑战性环境中的鲁棒性。

**致谢**。本文所述研究得到了 Cisco Research 的资助研究奖，以及 UKRI 资助（编号 EP/S021566/1）的 UCL 基础人工智能博士培训中心（UCL Centre for Doctoral Training in Foundational AI）的支持。

## 参考文献

[1] Hermann Blum, Alessandro Mercurio, Joshua O'Reilly, Tim Engelbracht, Mihai Dusmanu, Marc Pollefeys, and Zuria Bauer. Crocodl: Cross-device collaborative dataset for localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27424–27434, 2025. 4, 5, 6

[2] Leonardo Brizi, Emanuele Giacomini, Luca Di Giammarino, Simone Ferrari, Omar Salem, Lorenzo De Rebotti, and Giorgio Grisetti. Vbr: A vision benchmark in rome. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 15868–15874. IEEE, 2024. 2, 4

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016. 4, 7

[4] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1050–1060, 2025. 2, 6

[5] Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026. 2, 4, 5, 6, 7

[6] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 2

[7] Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. *arXiv preprint arXiv:2507.18255*, 2025. 2

[8] 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 (PAMI)*, 29(6):1052–1067, 2007. 2

[9] Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. Sail-recon: Large sfm by augmenting scene regression with localization. *arXiv preprint arXiv:2508.17972*, 2025. 4, 6

[10] Junyuan Deng, Heng Li, Kejie Qiu, Lingteng Qiu, Rui Peng, Weichao Shen, Weihao Yuan, Siyu Zhu, Zilong Dong, and Ping Tan. Glob3r: Global structure-from-motion with 3d foundation models. *arXiv preprint arXiv:2607.09225*, 2026. 4, 5, 6

[11] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025. 2, 4, 5, 6

[12] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16774–16784, 2025. 2

[13] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 834–849, 2014. 2

[14] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(3):611–625, 2017. 2, 7

[15] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, et al. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025. 2

[16] 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

[17] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 15–22. IEEE, 2014. 2

[18] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on robotics*, 28(5):1188–1197, 2012. 3

[19] Jingnan Gao, Zhe Wang, Xianze Fang, Xingyu Ren, Zhuo Chen, Shengqi Liu, Yuhao Cheng, Jiangjing Lyu, Xiaokang Yang, and Yichao Yan. More: 3d visual geometry reconstruction meets mixture-of-experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14680–14691, 2026. 2

[20] 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)*, pages 3354–3361, 2012. 2, 3, 4

[21] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2): 328–341, 2008. 4

[22] Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21584–21593, 2024. 2

[23] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 6, 7

[24] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1071–1081, 2025. 2

[25] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025. 2

[26] Bharath Raj Nagoor Kani and Noah Snavely. G3t up! gravity aligned coordinate frames simplify pointmap processing. *arXiv preprint arXiv:2605.27372*, 2026. 2

[27] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2

[28] Ramil Khafizov, Artem Komarichev, Ruslan Rakhimov, Peter Wonka, and Evgeny Burnaev. G-cut3r: Guided 3d reconstruction with camera and depth prior integration. *arXiv preprint arXiv:2508.11379*, 2025. 2

[29] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 225–234. IEEE, 2007. 2

[30] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 952–961, 2023. 2

[31] 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. 2

[32] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–91. Springer, 2024. 2

[33] Haodong Li, Chen Wang, Jiahui Lei, Kostas Daniilidis, and Lingjie Liu. Stereodiff: Stereo-diffusion synergy for video depth estimation. *arXiv preprint arXiv:2506.20756*, 2025. 2

[34] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 4, 6, 7

[35] 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. *arXiv preprint arXiv:2509.05296*, 2025. 2

[36] Haotong Lin, Sili Chen, Junhao 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. 2, 3, 4, 5, 6, 7

[37] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 424–440. Springer, 2024. 2

[38] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16651–16662, 2025. 2

[39] Jiahao Ma, Lei Wang, David Ahmedt-Aristizabal, Chuong Nguyen, et al. Puzzles: Unbounded video-depth augmentation for scalable end-to-end 3d reconstruction. *arXiv preprint arXiv:2506.23863*, 2025. 2

[40] Dominic Maggio and Luca Carlone. Vggt-slam 2.0: Real-time dense feed-forward scene reconstruction. *arXiv preprint arXiv:2601.19887*, 2026. 4, 5, 6, 7

[41] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 2, 5

[42] Soroush Mahdi, Fardin Ayar, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. Evict3r: Training-free token eviction for memory-bounded streaming visual geometry transformers. *arXiv preprint arXiv:2509.17650*, 2025. 2

[43] Jinjie Mai, Wenxuan Zhu, Haozhe Liu, Bing Li, Cheng Zheng, Jürgen Schmidhuber, and Bernard Ghanem. Can video diffusion model reconstruct 4d geometry? *arXiv preprint arXiv:2503.21082*, 2025. 2

[44] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE Conference on Computer Vision*

*and Pattern Recognition (CVPR)*, pages 18039–18048, 2024. 2, 7

[45] Raul 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. 2

[46] 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

[47] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 2, 4, 5, 6, 7

[48] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011. 2

[49] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–I. Ieee, 2004. 2

[50] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 4

[51] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 4, 6, 7

[52] Yue Pan, Xingguang Zhong, Louis Wiesmann, Thorbjörn Posewsky, Jens Behley, and Cyrill Stachniss. Pin-slam: Lidar slam using a point-based implicit neural representation for achieving global map consistency. *IEEE Transactions on Robotics*, 40:4045–4064, 2024. 2, 5

[53] Zador Pataki, Paul-Edouard Sarlin, and Marc Pollefeys. VidMap: Exploiting Temporal Structure for Video-Based Structure-from-Motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. 4, 5, 6, 7

[54] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2564–2571, 2011. 2

[55] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. LaMAR: Benchmarking Localization and Mapping for Augmented Reality. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4, 5, 6

[56] Nicolas Schischka, Hannah Schieber, Mert Asim Karaoglu, Melih Gorgulu, Florian Grötzner, Alexander Ladikos, Nassir Navab, Daniel Roth, and Benjamin Busam. Dynamon: Motion-aware fast and robust camera localization for dynamic neural radiance fields. *IEEE Robotics and Automation Letters*, 10(1): 548–555, 2024. 7

[57] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 134–144, 2019. 4, 7

[58] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 5135–5142. IEEE, 2020. 2

[59] Tixiao Shan, Brendan Englot, Carlo Ratti, and Daniela Rus. Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 5692–5698. IEEE, 2021. 2

[60] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 573–580, 2012. 3, 4

[61] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6229–6238, 2021. 2

[62] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5283–5293, 2025. 2

[63] Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The oxford spires dataset: Benchmarking large-scale lidar-visual localisation, reconstruction and radiance field methods. *International Journal of Robotics Research*, 2025. 2, 4

[64] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 2

[65] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 2, 5, 6, 7

[66] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:39033–39051, 2023. 2

[67] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023. 2, 4

[68] Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster vggt with block-sparse global attention. *arXiv preprint arXiv:2509.07120*, 2025. 2

[69] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025. 2

[70] Hengyi Wang and Lourdes Agapito. Amb3r: Accurate feed-forward metric-scale 3d reconstruction with backend. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14612–14625, 2026. 2, 3, 4, 5, 6

[71] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13293–13302, 2023. 2

[72] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 2

[73] Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026. 3, 4, 5

[74] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10510–10522, 2025. 2

[75] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2

[76] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 2, 4, 5, 6

[77] Thomas Whelan, Stefan Leutenegger, Renato F Salas-Moreno, Ben Glocker, and Andrew J Davison. Elasticfusion: Dense slam without a pose graph. In *Robotics: science and systems*. Rome, Italy, 2015. 2

[78] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025. 2

[79] Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, et al. Scal3r: Scalable test-time training for large-scale 3d reconstruction. *arXiv preprint arXiv:2604.08542*, 2026. 2, 4, 5, 6

[80] Congrong Xu, Huachen Gao, Xingyu Chen, Yuliang Xiu, Jun Gao, and Anpei Chen. $R^3$: 3d reconstruction via relative regression. *arXiv preprint arXiv:2605.26519*, 2026. 2

[81] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21924–21935, 2025. 2

[82] Zhu Yu, Jingnan Gao, Runmin Zhang, Lingteng Qiu, Zhengyi Zhao, Rui Peng, Yichao Yan, Kejie Qiu, Siyu Zhu, Zilong Dong, et al. Towards consistent video geometry estimation. *arXiv preprint arXiv:2605.30060*, 2026. 2

[83] Yijun Yuan, Zhuoguang Chen, Kenan Li, Weibang Wang, and Hang Zhao. Slam-former: Putting slam into one transformer. *arXiv preprint arXiv:2509.16909*, 2025. 2

[84] Vladimir Yugay, Yue Li, Theo Gevers, and Martin R Oswald. Gaussian-slam: Photo-realistic dense slam with gaussian splatting. *arXiv preprint arXiv:2312.10070*, 2023. 2

[85] Vladimir Yugay, Duy-Kien Nguyen, Theo Gevers, Cees GM Snoek, and Martin R Oswald. Fvo: Fast visual odometry with transformers. *arXiv preprint arXiv:2510.03348*, 2025. 2

[86] Ganlin Zhang, Shenhan Qian, Xi Wang, and Daniel Cremers. Vista-slam: Visual slam with symmetric two-view association. *arXiv preprint arXiv:2509.01584*, 2025. 2

[87] Ji Zhang, Sanjiv Singh, et al. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and systems*, pages 1–9. Berkeley, CA, 2014. 2

[88] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 82863–82886, 2025. 7

[89] 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. 2, 4, 5, 6, 7

[90] Songyan Zhang, Yongtao Ge, Jinyuan Tian, Guangkai Xu, Hao Chen, Chen Lv, and Chunhua Shen. Pomato: Marrying pointmap matching with temporal

motion for dynamic 3d reconstruction. *arXiv preprint arXiv:2504.05692*, 2025. 2

[91] Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni. Wildgs-slam: Monocular gaussian splatting slam in dynamic environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471. IEEE, 2025. 4, 5, 7

[92] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12786–12796, 2022. 2

[93] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2