

VGGT- Ω

作者: Jianyuan Wang^{1,2}

Minghao Chen¹

Shangzhan Zhang¹

Nikita Karaev¹

Johannes Schönberger²

Patrick Labatut²

Piotr Bojanowski²

David Novotny

Andrea Vedaldi^{1,2}

Christian Rupprecht¹

¹ 牛津大学视觉几何组

²Meta AI

摘要

VGGT 等近期前馈式重建模型 (*feed-forward reconstruction model*) 的性能已可媲美传统的基于优化的重建方法, 同时还能提供适用于其他任务的几何感知特征。本文表明, 这些模型的质量随模型规模和数据规模以可预测的方式提升。为此, 我们提出 VGGT- Ω , 其在静态与动态场景中均显著提升了重建精度、效率和能力。为以前所未有的规模训练该模型, 我们引入提高训练效率的架构改进、支持动态场景的高质量数据标注流程, 以及自监督学习方案。我们采用具有多任务监督的单个稠密预测头, 并移除计算代价高昂的高分辨率卷积层, 从而简化 VGGT 的架构。此外, 我们使用寄存器将场景信息聚合为紧凑表示, 并提出寄存器注意力, 将帧间信息交换限制在这些寄存器上, 以部分替代全局注意力。由此, 训练期间 VGGT- Ω 仅使用其前代模型约 $\sim 30\%$ 的 GPU 内存, 使我们能够以比以往工作多 $15\times$ 的监督数据训练 VGGT- Ω , 并利用海量无标签视频数据。VGGT- Ω 在多个基准的静态与动态场景重建上均取得优异结果, e.g., 其在 *Sintel* 上的相机估计精度较此前最佳结果提升 77% 。我们还表明, 学习得到的寄存器能够改进视觉-语言-动作 (VLA) 模型并支持语言对齐, 这说明重建可作为空间理解的强大且可规模化的代理任务。项目主页: <http://vggt-omega.github.io/>

1. 引言

近期工作 [91, 101, 181, 186, 189] 表明, 前馈式重建模型在许多情况下能够比肩甚至超越传统的运动恢复结构 (Structure-from-Motion, SfM) 流程 [57, 140, 149]。此外, 此类模型学习的词元已作为有效的几何感知表示用于众多其他任务 [1, 9, 18, 51, 66, 72, 90, 93, 132, 135, 156, 172, 173, 196, 200, 204, 205, 214, 220, 222, 225, 227, 230]。这说明重建可作为代理任务, 用于学习一般空间理解所需的表示, 具有基础性价值。然而, 相较于规模作用已得到充分认识的基础模型 [80, 147, 215], 三维计算机视觉对此探索尚少。因此, 本文探究前馈式重建模型能否规模化, 以及规模扩展会带来哪些收益。

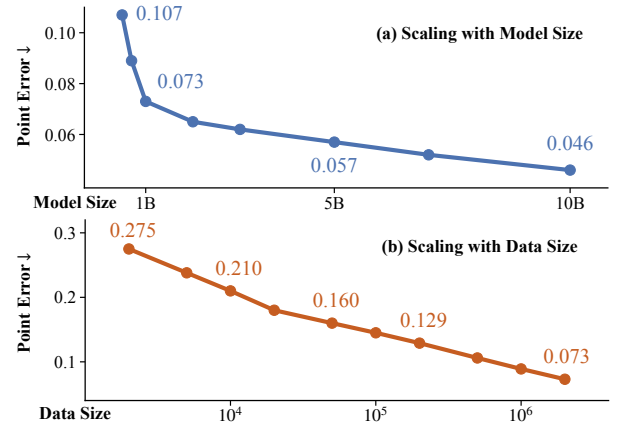


图 1. 数据与模型参数规模扩展带来的性能增益。随着模型规模从 0.2B 参数增至 10B 参数、数据规模从 2K 序列增至 2M 序列, 以三维点误差衡量的性能持续提升 (越低越好; 请注意坐标轴刻度不同)。所有模型均使用大致相同数量的词元进行训练, 并取六个数据集上的平均结果进行评估; 详情见章节 4.2。

为回答这一问题, 我们提出 VGGT- Ω , 将前馈式重建的数据规模以及可选的模型规模显著扩展至超越以往工作的水平。

与 VGGT [181] 相比, 新模型引入多项架构改进, 首先是寄存器的使用方式。近期工作 [32, 76, 117] 发现, 视觉 Transformer (Vision Transformer, ViT) 会自发利用少量图像词元承载全局信息, 并引入可学习寄存器, 以更直接、更高效地实现这一功能。VGGT 已具有逐帧寄存器, 而 VGGT- Ω 进一步提出寄存器注意力: 在部分全局注意力层中, 将帧间信息交换限制在寄存器上。更新后的寄存器随后在帧内注意力层中与其他词元进行局部交互, 从而形成聚合与再分发多帧信息的瓶颈。这一设计促使寄存器聚合整个场景的信息, 因此我们也称其为“场景”词元。

这一设计有两项优势。首先, 其他架构通常将寄存器视为辅助成分并在推理时丢弃 [32], 而本文表明寄存器承载了有用的全局信息。具体而言, 尽管没有显式监

督，寄存器仍可为视觉-语言-动作（Vision-Language-Action, VLA）模型和语言对齐提供有效特征。其次，寄存器注意力还能提高效率。全局注意力是 VGGT 的主要计算瓶颈，但其注意力图非常稀疏 [143, 174]。我们发现，寄存器注意力通过聚合全局信息，也能成为完整全局注意力的高效替代方案。具体而言，以寄存器注意力替换 25% 的全局注意力层不会造成可测量的性能下降，同时在训练期间为骨干网络节省约 23% 的 FLOPs 和 16% 的内存¹。

除寄存器外，我们还注意到，稠密预测头（*e.g.*，DPT）中的高分辨率卷积层虽然仅占模型参数的一小部分，却会消耗不成比例的 GPU 内存来存储前向激活。FSDP 或梯度检查点等技术无法消除存储这些激活的开销。因此，第二项改进是以单个 MLP 加像素重排算子，替代稠密预测器中内存开销最大的卷积层。无论从定量还是定性结果看，这一设计都能在不降低性能的前提下大幅减少内存占用。

最后，我们已在 VGGT 中证明，直接监督深度图、点图和跟踪特征的多任务训练是有效的。本文发现，实现这些收益并不需要额外的稠密预测头。第三项改进是继续使用多任务损失，但仅保留一个用于深度预测的稠密预测头和一个用于相机预测的稀疏预测头。

这三项改进可在训练期间节省 70% 的 GPU 内存，并小幅提高推理速度。

除效率外，我们发现训练数据的数量、多样性和质量对规模扩展至关重要。尤其是，处理动态内容十分关键，因为这能释放数量级更大的类互联网视频用于训练。因此，我们开发了一套高质量数据标注流程，可大规模生成刚性视频与动态视频的标注。该流程整合了基于视觉语言模型（VLM）的预筛选、VGGT、COLMAP、现代图像匹配模型和有监督几何后筛选。在约 40M 个内部类互联网视频上应用后，该筛选流程保留了 0.8M 个具有准确标注的序列，其中约三分之一包含动态内容。结合现有数据集（包括真实数据与合成数据），最终得到共 4M 个具有准确重建标注的多样化场景/序列，数量超过 VGGT 的 15 $\times$ 。

为进一步提高泛化能力，我们引入受 DINO 及相关动量教师-学生方法启发的自监督学习方案 [19, 60, 124, 147, 160]。教师模型和学生模型均由有监督的 VGGT- Ω 检查点初始化。两个模型在不同的数据增强与帧排列下处理相同的输入序列。学生模型经过训练，在对齐帧顺序后匹配教师模型的预测和特征分布；教师模型则通过学生模型的指数移动平均进行更新。我们使用该方案在 18M 个无标签视频上训练。

这些改进使我们得以研究前馈式重建模型的规模扩展特性。如图 1 所示，随着模型容量从 0.2B 参数增至 10B 参数、训练数据从数千个序列扩展至两百万个不同序列，以点误差衡量的重建精度持续呈现近似幂律的提升。

总体而言，VGGT- Ω 将前馈式重建性能提升到新的水平，在三个静态基准和三个动态基准上均以显著优

势取得最先进结果。特别是，它大幅超越 MegaSaM 等后优化方法以及 Depth Anything 3 等近期前馈式方法 [101]。在 Sintel 上，VGGT- Ω 的相机估计 AUC@3 $^\circ$ 达到 40.0，对比此前的 22.5 提升 77%；AUC@30 $^\circ$ 达到 79.1，对比 58.3 提升 35%。其深度估计 $\delta_{1.25}$ 达到 93.5，对比 74.1 提升 26%，同时速度比 MegaSaM 快 50 $\times$ 。最后，我们表明学习得到的寄存器可复用于重建之外的任务，既能改进 VLA 模型，也能支持语言对齐。

2. 相关工作

三维重建。 三维重建研究历史悠久，其开端是奠定多视图几何理论的早期奠基性工作 [43, 58, 123, 125]。后续研究推动了重大的实践进展，包括 COLMAP 等鲁棒的 SfM 系统 [140] 及其他流程 [2, 33, 46, 106, 149]。本文关注前馈式重建模型，*i.e.*，直接从一幅或多幅图像推断场景几何与相机位姿的神经网络。尽管近期 SfM 流程越来越多地引入关键点检测器 [38, 40, 170, 210]、特征匹配器 [21, 102, 138, 145] 等可学习组件，但与本文最相关的是直接学习几何估计的端到端可微 SfM 框架 [10, 158, 165, 166, 171, 178–180, 192, 232]。这些工作虽证明 SfM 可以进行端到端学习，但仍需组合经典 SfM 流程中的多个组成部分。DUST3R [186] 及其后续工作 MAST3R [39] 可直接从图像估计场景几何与相机参数（外参与内参）。然而，与立体视觉方法 [47, 54, 113, 121, 129, 193, 207, 208, 223] 类似，DUST3R 和 MAST3R 中的神经网络仅处理图像对，在处理更多视图时仍需后优化。能够联合处理多幅图像的方法带来了一项关键进展，它们不再需要跨视图优化 [41, 159, 175, 184, 197, 203, 221]。其中，VGGT [181] 是代表性方法：它仅依靠前馈推理便首次超越后优化方法（*e.g.*，使用束调整的方法），并催生了大量后续工作 [16, 24, 25, 35, 36, 64, 75, 79, 83, 87, 95, 100, 105, 110, 114, 115, 118, 131, 156, 183, 189, 199, 211, 213, 234]。若干工作通过词元合并、稀疏注意力或基于描述子的聚合，提高 VGGT 类模型的计算可扩展性，使多视图前馈式重建更加高效 [115, 143, 174]。另一些工作采用具有状态且容量受限的场景表示，以线性更新代价解决全局自注意力的二次复杂度问题。例如，部分方法采用推理时训练层 [26, 78, 219]。另一些方法 [22, 87, 157, 234] 以因果方式处理帧，同时通过缓存词元、键值对、局部窗口或地图来维持持久的几何状态。还有一些工作使用 VGGT 进行局部场景重建，再通过对齐、定位或 SLAM 算法融合结果 [35, 36, 114, 176]。

另有工作扩展 VGGT 类模型，使其能够接收额外的几何信息作为输入（*e.g.*，相机 [83]）、增加传感器（*e.g.*，LiDAR [183]），或对相机阵列进行建模 [95]。其他扩展工作预测法向量 [42]，或采用混合专家设计 [50, 191]。PI3 [189] 摆脱了对固定参考视图的依赖；Depth Anything 3 (DA3) [101] 则采用原生 DINO 作为骨干网络，二者均支持动态场景。部分研究探究了这些模型学到了什么，包括对应关系、极线结构，以及与人类三维形状感知之间的联系 [9, 11]。SelfEvo [68] 探索了利用时空上下文不对称性的自蒸馏方案。除重

¹以寄存器注意力替换全部全局注意力层可将 FLOPs 降至原来的 6%，但会导致明显的性能下降。

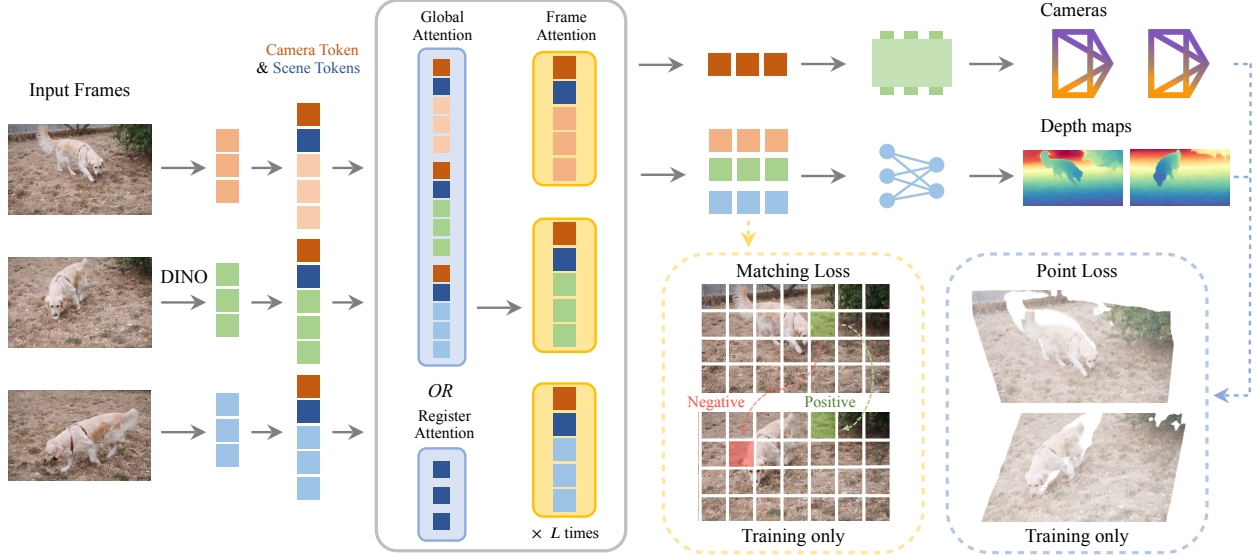


图 2. 架构概览。VGGT- Ω 在图像词元后附加相机词元和场景词元（寄存器），随后交替通过全局注意力（或寄存器注意力）层与帧内注意力层。我们用仅在训练时使用的损失替代 VGGT 中冗余的稠密预测头。

建本身外，前馈式重建模型学习的特征也被用作其他任务的几何感知表示。相关应用包括视频生成与新视图合成 [66, 72, 75, 109, 156]、视觉语言模型 [90, 132, 173, 196, 205, 214, 220, 222, 225, 227]、视觉-语言-动作模型 [1, 93, 135, 172]，以及检测、分割、匹配、占用预测和地点识别等感知任务 [18, 37, 51, 200, 204, 230]。

单目动态三维重建（即四维重建）旨在恢复随时间变化的场景几何。这一研究方向同样历史悠久，早期代表性工作包括 Bregler 等人 [12] 和 Torresani 等人 [168] 的研究。近期工作中，MegaSaM [99] 尤具影响力，它将前馈式深度预测与基于优化的非刚性重建相结合。ViPE [65] 在此方向上进一步推进。若干工作探索了减少优化的前馈式四维重建。MonST3R [218] 和 D²UST3R [55] 扩展 DUST3R，以处理动态三维内容。Align3R [112] 基于 DUST3R 推断相机，并沿时间维度对齐单目深度预测，但处理两个以上视图时仍依赖优化。CUT3R [184] 和 Point3R [197] 同时支持增量重建与动态场景。Geo4D [77] 通过微调视频生成器恢复四维几何。PI3 [189] 和 DA3 [101] 采用 VGGT 类模型并以动态数据训练；PAGE-4D [231] 则通过分离静态区域与运动区域的模块改造 VGGT。Human3R 专注于人场景重建 [27]。SpatialTracker [199]、St4RTrack [44]、DPM [153]、V-DPM [154]、Uni4D [206]、Any4D [82] 及相关方法统一了动态重建与点跟踪。D4RT [217] 提出统一的前馈式 Transformer，通过查询点级四维场景信息重建动态场景。

视觉 Transformer (ViT) 中的寄存器。 近期工作 [32, 117] 发现，ViT 中的少量词元不仅编码局部图块信息，还编码全局信息。这些离群词元具有较高范数，会破坏图块特征的空间连贯性。这促使研究者使用独立于图像词元的寄存器词元，以帮助保持图块表示之间的连贯性。后续研究进一步探究了这一现象的成因，以及寄

存器词元与图像词元之间的交互 [76, 88, 117, 144]。本文为每个输入图像帧加入寄存器词元，并使用它们跨图聚合与交换全局信息。由此，寄存器能够承载整个序列的信息。不同于以往工作通常将其丢弃的做法，我们证明寄存器可用于下游应用。

3. 方法

VGGT- Ω 以原始 VGGT 为基础，并从多个方面加以改进。首先介绍 VGGT- Ω 的新架构（章节 3.1）与训练流程（章节 3.2）。随后介绍利用无标注数据的自监督训练方案（章节 3.4），以及用于稳健标注数百万段视频的新数据流程（章节 3.5）。

3.1. 新的可扩展架构

如图 2 所示，VGGT- Ω 是一个前馈式 Transformer f ，将 N 张输入图像 $I_1, \dots, I_N \in \mathbb{R}^{3 \times H \times W}$ 映射为对应的相机和深度图：

$$((g_1, D_1), \dots, (g_N, D_N)) = f(I_1, \dots, I_N),$$

其中， $D_i \in \mathbb{R}^{H \times W}$ 是图像 I_i 的深度图， $g_i = (q_i, t_i, f_i) \in \mathbb{R}^9$ 由描述相应相机的旋转四元数 $q_i \in \mathbb{R}^4$ 、平移向量 $t_i \in \mathbb{R}^3$ 和视场角 $f_i \in \mathbb{R}^2$ 拼接而成。遵循常见做法 [140, 180, 181]，本文假设主点位于图像中心。因此，该问题定义与 VGGT [181] 相似，但模型不会直接预测点图（Point Map）或跟踪特征（尽管仍会监督这些量，详见章节 3.2）。网络 f 将每张图像编码为词元（章节 3.1.1），通过交替注意力（Alternating-Attention）聚合跨视图特征（章节 3.1.2），再用轻量级预测头将词元映射为最终预测（章节 3.1.3）。

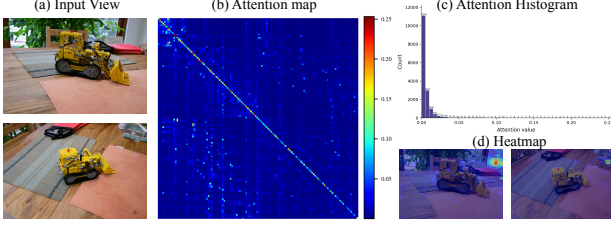


图 3. VGGT 全局注意力可视化。从 (b) 全局注意力矩阵、(c) 数值分布和 (d) 空间热力图可以看出，第 13 层的注意力十分稀疏。

3.1.1. 特征提取与词元化

使用以 DINOv3 初始化的视觉 Transformer [147] 对每张图像 I_i 进行词元化，得到 $z_i^F = \text{DINO}(I_i) \in \mathbb{R}^{H'W' \times C}$ ，其中，块大小为 r ， $H' = H/r$ ， $W' = W/r$ 。对于每张图像 I_i ，还附加一个相机词元 (Camera Token) $z_i^{\text{cam}} \in \mathbb{R}^{1 \times C}$ 和十六个寄存器 (场景词元, Scene Token) $z_i^{\text{scene}} \in \mathbb{R}^{16 \times C}$ 。相机词元用于预测相机参数，寄存器则聚合场景信息。与 [181] 相同，这些词元可采用两组可学习参数之一：若图像 I_i 为参考图像则采用一组，否则采用另一组。随后将其拼接成词元集合 $z = (z_1, \dots, z_N) \in \mathbb{R}^{N \times (H'W' + 17) \times C}$ 其中， $z_i = (z_i^F, z_i^{\text{cam}}, z_i^{\text{scene}})$ 是图像 I_i 的词元。

3.1.2. 寄存器注意力

VGGT 采用交替注意力 [181]，交替执行每张图像内部的帧内自注意力 (Frame-Wise Self-Attention) 与跨所有图像的全局自注意力 (Global Self-Attention)；两者在定义上均具有置换等变性。帧内注意力无需帧索引嵌入，否则会限制模型泛化到不同输入帧数的能力。因此，词元并未显式编码对应图像的身份 (仅标明某帧是否为参考帧)。全局注意力即对所有词元 z 应用标准注意力层，记为 $z' = \text{attn}(z)$ 。帧内注意力与之类似，但会对每张图像 I_i 的 z_i 独立应用，记为 $z' = \text{attn}_f(z) = (\text{attn}(z_1), \dots, \text{attn}(z_N))$ 。不同帧通过全局注意力交互，并在此形成多帧场景信息。但全局注意力需要关注所有帧的全部词元，计算成本随词元总数呈二次增长，因而开销很高。如图 3 所示，全局注意力图通常较为稀疏，这表明少量词元或许足以交换相应信息。这一观察与近期研究 [143, 174] 一致。因此，本文用寄存器注意力 (Register Attention) 替换 25% 的全局注意力层，并将自注意力限制在所有帧的寄存器之间。形式上，寄存器注意力仅更新寄存器： $z' = \text{attn}_{\text{scene}}(z)$ 其中， $(z_1^{\text{scene}'}, \dots, z_N^{\text{scene}'}) = \text{attn}(z_1^{\text{scene}}, \dots, z_N^{\text{scene}})$ ，i.e.，换言之，在这些模块中，只有寄存器参与跨帧自注意力。更新后的寄存器会在后续帧内注意力模块中与各帧的图像词元交互，将聚合的场景信息重新分配给图像词元。这既促使最终寄存器承载全局场景信息，又降低了全局注意力的成本。

3.1.3. 解码

注意力层输出的最终词元集合 $z' = (z'_1, \dots, z'_N)$ 被解码为深度图和相机参数。

深度。 在 VGGT 中，所有稠密解码器均由稠密预测 Transformer (Dense Prediction Transformer, DPT) 层 [134] 实现。然而，这些 DPT 预测头的末端卷积模块需要保留多个高分辨率特征图，内存开销很大。为降低这一成本，本文将工作分辨率高于输入分辨率 $1/4$ 的模块替换为轻量级上采样头：先通过一个多层感知机 (MLP)，再执行像素重排操作。MLP 输出 $2u^2$ 个通道 (实现中 $u = 4$)，像素重排操作将其从 $(H' \times W', 2u^2)$ 重排为 $(uH') \times (uW') \times 2$ ；两个输出通道分别对应深度和置信度。

本文还探索了完全无卷积的解码器，仅用 MLP 将词元映射为稠密预测。该设计在基准测试中表现良好，但预测深度图在视觉上会出现块状伪影，尤其是在包含天空、山脉等远距离结构的室外场景中；此类区域的深度无界，因而定义并不明确。因此，本文保留 DPT 前端计算成本较低的低分辨率卷积层。

相机。 先对相机词元和寄存器 $\{(z_i^{\text{cam}}, z_i^{\text{scene}})\}_{i=1}^N$ 应用轻量级 Transformer，再对每个更新后的相机词元应用 MLP，从而联合预测相机 $(g_1, \dots, g_N)$ 。与 VGGT 不同，本文的相机预测头只需单次前向传播即可预测相机参数，无需迭代细化。

3.2. 训练损失

在 VGGT 中，预测冗余的稠密输出 (e.g., 点图和轨迹) 虽然有益，但训练成本很高。VGGT- Ω 改为仅包含一个用于深度预测的稠密预测头 (Dense Prediction Head)。尽管模型不直接预测点图与轨迹，本文仍然通过相应损失监督这些量。实验发现，这种多任务监督 (Multi-Task Supervision) 与使用多个稠密预测头的性能几乎相同，同时可节省大量内存。因此，优化以下损失：

$$\mathcal{L} = \lambda_{\text{cam}} \mathcal{L}_{\text{cam}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}} + \lambda_{\text{point}} \mathcal{L}_{\text{point}} + \lambda_{\text{match}} \mathcal{L}_{\text{match}} \quad (1)$$

其中， λ_{cam} 、 λ_{depth} 、 λ_{point} 和 λ_{match} 为权重。

相机损失。 相机损失 $\mathcal{L}_{\text{cam}} = \sum_{i=1}^N \| \hat{g}_i - g_i \|$ 采用 ℓ_1 目标比较预测相机 $\hat{g}_i$ 与真实相机 g_i ；实验发现，该目标比 VGGT 使用的 Huber 损失更稳定。

深度损失。 遵循 VGGT，深度损失采用偶然不确定性 (Aleatoric Uncertainty) 和梯度一致性项。此外，还考虑相对于真实值的相对尺度。因此有： $\mathcal{L}_{\text{depth}} = \sum_{i=1}^N [\|c_i^D \odot (1 + D_i^{-1}) \odot e_i\| + \|c_i^D \odot \nabla e_i\|] - \alpha \sum_{i=1}^N \log c_i^D$ ，其中 $e_i = \hat{D}_i - D_i$ ， c_i^D 是预测的不确定性图， $\odot$ 表示逐元素乘积。

点损失。 点图为每个像素分配相应三维点在参考相机坐标系中的坐标。因此，可通过反投影从深度图和相机参数推导点图。相应地，点损失 $\mathcal{L}_{\text{point}}$ 与深度损失 $\mathcal{L}_{\text{depth}}$ 相同，只需将残差替换为 $e_i = \pi^{-1}(\hat{D}_i, \hat{g}_i) - P_i$ ，其中， π^{-1} 表示反投影， P_i 为点图。

匹配损失。 匹配损失 (Matching Loss) $\mathcal{L}_{\text{match}}$ 作用于最后一个注意力层输出的词元。该损失拉近正词元对 (对应同一三维位置) 的特征，并推远负词元对：



图 4. 定性结果。VGTT- Ω 能够处理静态与动态内容，图中叠加的车流和网球运动员轨迹即为例证。该方法还能泛化到困难场景，*e.g.*，水下珊瑚礁。各示例分别使用 64、4、9、16 和 32 帧输入。

$\mathcal{L}_{\text{match}} = \mathbb{E}_{\text{pos}}[-\log \sigma(s)] + \mathbb{E}_{\text{neg}}[-\log(1 - \sigma(s))]$ ，其中， s 是经 ℓ_2 归一化后词元间的余弦相似度， σ 是 Sigmoid 函数， $\mathbb{E}$ 表示分别对正、负样本对取平均，*i.e.*，即加权二元交叉熵。样本对的构建细节见补充材料。

3.3. 动态重建

VGTT- Ω 支持动态场景重建。除其他优势外，这使可用训练数据增加了数个数量级，因为大多数视频都包含运动。动态重建 (Dynamic Reconstruction) 需要用统计先验约束运动。经典的非刚性运动恢复结构 [12, 31, 161] 采用人工设计的低秩或局部刚性约束，但这类约束可能较为脆弱，泛化能力有限。VGTT- Ω 这类数据驱动模型有望从数据中学习更优的先验。然而，输出表示决定了相机运动与场景运动的耦合方式。例如，近期一些方法 [186] 使用点图表示场景，并从中恢复相机参数。这种方案在静态场景中效果良好，但在动态场景中，要么需要像 MonST3R [218] 一样分割运动像素，要么需要引入动态点图 [153, 154] 等扩展。另一种方案是预测深度图和射线图 [101]。然而，射线图会引入成本高昂的稠密输出，还可能使相机信息与逐像素外观变化相互纠缠。例如，固定相机拍摄舞者时，画面中存在大幅

运动，但相机参数保持不变。因此，如上所述，本文只预测深度图和相机参数，避免运动掩码等显式动态输出。

3.4. 自监督训练

受二维视觉中常见做法 [19, 60, 124, 147] 启发，本文采用教师-学生策略，利用无标注视频进行自监督学习 (Self-Supervised Learning)。具体而言，学生模型 (*Student*) 通过梯度下降更新，教师模型 (*Teacher*) 则仅通过学生模型参数的指数移动平均更新。两个网络均以在有监督数据上训练的 VGTT- Ω 检查点进行初始化。给定一个视频序列，将同一组帧输入两个网络，但分别施加独立的随机增强，包括颜色抖动与模糊、随机 90° 旋转、随机图像块掩蔽以及随机帧重排 (会影响参考帧的选择)。将两路数据恢复为相同顺序后，要求学生模型从两个方面与教师模型保持一致。 ℓ_2 特征匹配损失在多个层上对齐学生模型与教师模型的词元。回归损失则监督相机和深度。为防止坍塌，自监督期间会冻结相机和深度预测头。教师模型参数按 $\theta^T \leftarrow m \theta^T + (1 - m) \theta^S$ 更新，而不采用梯度下降。该蒸馏方案由此强制模型对外观变化和帧顺序保持不变

性, 使其能够从数百万段无标注视频中有效学习。

3.5. 训练数据

VGGT- Ω 的一项关键工作是扩展训练数据。为此, 我们将大量公开数据集 (章节 3.5.1) 与新开发的标注流程 (章节 3.5.2) 相结合; 该流程能够处理现成视频中的动态内容。

3.5.1. 数据来源

我们首先收集了以下公开数据集: Aria series [126], Bedlam [8], BEHAVIOR-1K [92], Co3Dv2 [136], uCo3D [108], DL3DV [103], Dynamic Replica [81], EDEN [89], EFM3D [151], HOT3D [6], Habitat [139], Hypersim [137], Mapfree [4], Mapillary Metropolis [116], MPSD [3], Megadepth [97], Megasynth [74], Mid-Air [45], Mvssynth [69], ParallelDomain-4D [61], Replica [150], SAIL-VOS [63], ScanNet Series [29, 209], TartanAirV2 [187], TartanGround [128], Taskonomy [212], UnrealStereo4K [169], Virtual KITTI [15], Waymo [155], 以及 WildRGBD [198]. 我们排除了 VGGT 使用的 Kubric [53] 和 PointOdyssey [228], 因为其背景几何是伪造的, 会产生无效深度。此外, 我们还使用若干内部数据集, 包括艺术家制作的物体资产、刚性和动态合成环境、真实设备采集数据及相关来源。对于非合成数据集 (*e.g.*, 由 SfM 流程标注的数据), 我们通过多视图一致性检查移除含噪深度值, 并丢弃有效深度像素过少的序列。这些数据集总计约含 3M 个序列, 每个序列包含 10 至 20,000 张图像。

3.5.2. 数据标注流程

为进一步扩展训练数据, 我们构建了一个大型内部视频集合, 包含约 40M 个互联网风格视频。首先评估每个视频是否适合重建, *e.g.*, 滤除带有大面积水印或突兀镜头切换的片段。通过检查的视频用于章节 3.4 所述的自监督训练。随后, 我们提出新流程, 为静态和动态场景视频标注相机参数与深度图。标注质量优先于数量, 因此会严格拒绝低质量数据, 并丢弃可能属于动态区域的深度标注。这可能排除极端相机运动或高度动态场景, 但合成数据集仍充分覆盖了这些情况。最终得到约 200K 个动态场景和 600K 个静态场景, 均具有高质量相机与深度标注。

VLM 预筛选。 我们提示视觉语言模型 (Vision-Language Model, VLM) 丢弃难以用多视图几何重建的视频。VLM 将 50% 的片段判定为重建难度过高, *e.g.*, 包含多个镜头、极端运动模糊或大面积叠层与水印。另有 40% 被判定为可重建, 但可能因视差不足、缺少非重复纹理等原因而精度较低。其余 10% 的视频进入下一阶段。在同一次处理中, VLM 还提取场景是否动态等元数据, 供后续阶段使用。

动态掩码提取。 我们使用 Grounding DINO [107] 检测人、汽车等潜在可移动物体类别的边界框, 并在匹配、跟踪与验证时排除这些区域。

特征匹配与跟踪。 我们组合以下方法提取跨帧匹配与轨迹: SIFT [111], SuperPoint & SuperGlue [38, 138], ALIKED & LightGlue [102, 226], 以及 VGGSfM Tracker [180]. 动态区域内的匹配会被丢弃。

重建与筛选。 当基于 RANSAC 的本质矩阵估计得到的内点过少时, 我们用原始 VGGT 初始化相机参数, 再依据上述对应关系运行 COLMAP [140], 迭代执行束调整与筛选。对于成功重建的序列, 我们丢弃未通过启发式检查的样本, *e.g.*, 图像注册率 < 99.5%、视场角不在 $[30^\circ, 120^\circ]$ 内, 或畸变率 > 0.1。这些标准会严格移除退化运动或极端变焦情况。随后使用基于图块的多视图立体方法 [141] 估计逐帧稠密深度图。

多视图一致性。 对于每一帧, 我们将深度图反投影到三维空间, 再把点重投影至其他视图, 并与对应深度比较。通过跨视图一致性检查的像素被标记为有效。我们丢弃有效深度像素少于 5% 的序列; 这通常 (但不总是) 意味着相机质量较低。

有监督几何筛选。 最终, 每个序列都获得相机、深度和有效掩码。我们用相机上方向向量一致性、视差角和轨迹平滑度等人工设计特征描述各序列。通过分别人工标注 500 个静态序列和 500 个动态序列, 训练分类器以移除低质量重建。该分类器集成了 XGBoost [23]、随机森林 [13] 和 CatBoost [130]。

4. 实验

本节进一步介绍框架细节, 将其与当前最先进的刚性和动态重建方法进行基准比较, 并对关键设计选择开展消融实验。

4.1. 实现细节

我们重点研究四种模型变体: 参数量分别为 200M、500M、1B 和 10B, 对应 12/12/24/16 个交替注意力块, 隐藏层维度分别为 384/768/1024/4096。视觉 Transformer 由 DINOv3 [147] 初始化, 训练期间不冻结。每个块包含一个全局注意力 (或寄存器注意力) 层和一个帧内注意力层。使用 AdamW 优化 240K 次迭代, 其中监督训练、自监督训练和最后一个监督训练阶段分别占 160K、50K 和 30K 次迭代。学习率在训练前 5% 进行线性预热, 随后 95% 采用余弦衰减; 监督训练和自监督训练的峰值学习率分别为 2×10^{-4} 和 1×10^{-4} 。每个批次的帧数从 [1, 24] 中均匀采样。图像增强包括: 在 [0.33, 1.33] 范围内随机改变宽高比, 同时将图像面积保持在约 512×512 像素, 并应用颜色抖动、灰度转换和随机图块掩码。训练使用 128 块 96GB H100 GPU, 并采用 bfloat16 混合精度、梯度检查点和 FSDP。

4.2. 基准评测

定量比较。 我们将 VGGT- Ω 与近期方法进行比较, 包括: (i) 前馈式重建模型; (ii) 基于优化的动态重建方法。评估使用三个静态数据集 (7 Scenes [146]、NRGBD [5] 和 ETH3D [142]) 以及三个动态数据集 (DyCheck [49]、Sintel [14] 和 TUM-Dynamic [152])。对于每个场景或

表 1. 静态与动态基准上的**相机位姿估计**。AUC 越高越好。前馈式模型 (DA3、PI3、VGGT) 在不同数据集上较为稳健, 但在 Sintel 的严格阈值下仍落后于动态优化方法 MegaSaM (*e.g.*, AUC@3° 为 16.2 *vs.* 22.5)。相比之下, 动态优化方法 (MegaSaM、MonST3R) 在宽基线场景中性能下降 (*e.g.*, ETH3D 上 AUC@30° 为 38.1 *vs.* 86.4)。本文方法在所有场景中显著刷新了最先进结果, *e.g.*, 将 Sintel 上的 AUC@3° 从 22.5 提升至 40.0, 相对提升 77%。

方法	静态场景						动态场景					
	7 Scenes		NRGBD		ETH3D		DyCheck		Sintel		TUM-Dynamic	
	指标	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@30°
MonST3R		9.0	68.3	13.9	79.7	1.7	14.3	11.5	45.4	4.3	45.8	7.7
MapAnything		5.8	61.4	35.2	88.9	13.2	51.0	6.1	60.3	2.9	31.6	4.3
MegaSaM		10.6	71.8	17.2	83.1	5.9	38.1	26.8	53.1	22.5	58.3	15.4
VGGT		10.9	74.4	81.7	97.7	18.8	62.1	21.0	78.7	15.0	50.0	16.6
PI3		13.3	77.0	83.8	98.2	35.3	79.6	23.3	81.0	14.8	53.5	16.1
DA3		18.7	78.2	86.4	98.4	46.1	87.0	32.1	83.9	16.2	52.7	20.8
本文-1B		29.6	83.1	89.7	98.8	49.8	88.5	38.4	87.3	35.3	73.0	30.2
本文-10B		36.4	88.2	92.5	99.1	56.3	90.4	43.7	90.9	40.0	79.1	36.4

表 2. 静态与动态基准上的**深度估计**。指标 $\delta_{1.25}$ 表示预测深度落在真实值指定倍率范围内的比例 (越高越好), AbsRel 为平均绝对相对误差 (越低越好)。

方法	静态场景						动态场景					
	7 Scenes		NRGBD		ETH3D		DyCheck		Sintel		TUM-Dynamic	
	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel
MonST3R	92.4	0.075	98.4	0.030	95.8	0.056	93.3	0.068	71.9	0.263	85.0	0.148
MapAnything	92.9	0.070	98.7	0.022	96.3	0.035	97.0	0.049	72.5	0.251	93.1	0.052
MegaSaM	93.8	0.065	96.2	0.057	94.8	0.083	97.4	0.042	74.1	0.207	92.9	0.083
VGGT	91.9	0.073	99.1	0.019	97.4	0.036	95.2	0.055	79.2	0.189	92.2	0.064
PI3	92.8	0.068	99.2	0.011	99.6	0.016	97.4	0.041	82.5	0.144	95.5	0.046
DA3	93.0	0.063	99.5	0.010	99.6	0.015	97.7	0.039	86.1	0.118	94.3	0.049
本文-1B	94.6	0.058	99.6	0.010	99.8	0.012	98.4	0.038	89.5	0.097	97.4	0.041
本文-10B	96.3	0.050	99.7	0.007	99.8	0.009	98.7	0.030	93.5	0.081	98.3	0.035

序列, 随机采样 10 帧。所有方法均使用其原始发布模型。对于 DA3, 我们使用其最大的 Giant 变体 (1B 参数)。

遵循 [181], 我们报告相机位姿估计的标准 AUC (越高越好), 其计算方式为: 相对旋转和平移误差低于给定角度阈值 (*e.g.*, 3°, 30°) 的图像对比例曲线下的面积。如表 1 所示, 前馈式模型在静态基准和较宽松阈值下通常表现较强; 基于优化且具备动态感知能力的 MegaSaM 在 Sintel 等困难动态序列上更具竞争力, 但在宽基线或低纹理场景中性能下降。相比之下, 我们的模型在静态和动态数据集以及严格和宽松阈值下均持续优于所有基线。

我们还使用绝对相对误差 (Absolute Relative Error, AbsRel; 越低越好) 和 $\delta_{1.25}$ (越高越好) 评估预测深度的精度; 后者衡量预测深度与真实深度之比处于指定阈值内的像素占比。如表 2 所示, 我们的模型在静态基准上优于基线, 并在 ETH3D 等现有方法已表现较强的数据集上进一步降低 AbsRel; 在动态场景中的提升更为明显, 深度误差更低且 $\delta_{1.25}$ 更高 (*e.g.*, 在 Sintel 上, $\delta_{1.25}$ 从 86.1 提升至 93.5, AbsRel 从 0.118 降至 0.081)。

规模更大的 10B 变体始终优于 1B 模型, 表明扩大重建模型规模能够直接提高相机和深度精度。

定性结果。 图 4 展示了模型在静态和动态场景中的结果, 包括交通、人类运动、自然景观和水下环境。

我们还在图 5 and 6 的困难案例上将 VGGT- Ω 与 DA3 和 MegaSaM 进行比较。在图 6 中, DA3 难以处理重复纹理, 在雪地缆车序列中估计出的相机运动

很小甚至为零; 面对强烈的相机滚转时, 其在无人机序列中多次重建出同一座塔。MegaSaM 在具有显著相机滚转的稀疏室内场景 (右上) 或无纹理墙面 (右下) 中也可能失效, 产生断裂结构或错位平面; 某些航拍序列 (左侧) 同样对其位姿估计构成挑战。相比之下, VGGT- Ω 的重建结果具有全局一致性, 这可能得益于从多样化数据中学到的强几何先验。

推理内存与速度。 下面将 VGGT- Ω 与 VGGT 和 DA3 的推理内存与速度进行比较 (请注意, 这与章节 3.1 讨论的训练效率有很大不同)。

进行比较前, 我们先修正原始 VGGT 实现。该模型在推理期间会缓存全部 24 层的中间张量。这些缓存张量虽在训练期间有用, 但测试时预测头只需四层的中间特征。因此, 我们不再缓存不必要的信息, 从而显著降低推理内存占用。

完成修正后, 我们在图 7 中比较 VGGT、DA3 和 VGGT- Ω 。为保证公平, 所有测量均在单块 80GB A100 GPU 上进行, 使用 PyTorch *scaled_dot_product_attention* 和 flash attention v2 后端。根据各模型的图块尺寸, 采用相近的输入分辨率: VGGT 和 DA3 使用 518×336 的分辨率和 14 像素的图块, VGGT- Ω 使用 512×336 的分辨率和 16 像素的图块。我们逐步增加输入帧数, 直至各方法耗尽内存; 图 7 中的叉号标出了首次内存不足的设置。

在内存效率方面, 修正后的 VGGT 与 VGGT- Ω 相近, 均可在单块 A100 GPU 上处理超过 1,000 帧。在速度方面, VGGT- Ω 更快, 主要原因是其采用图块尺寸为 16 的 DINOv3, 而非图块尺寸为 14 的 DINOv2,

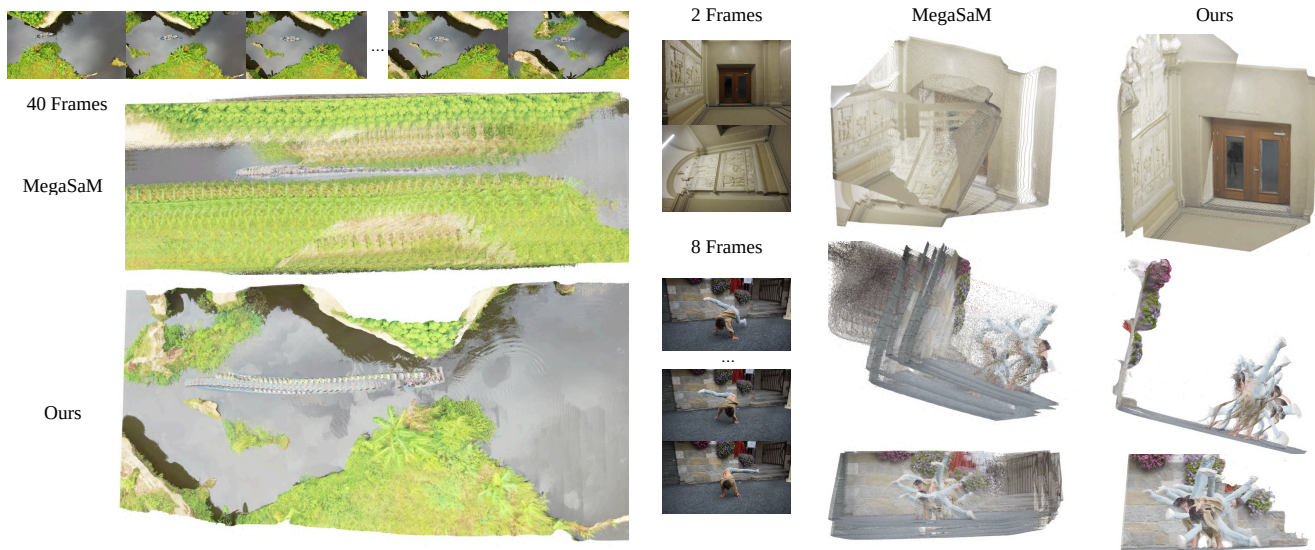


图 5. 与 MegaSaM 的定性比较。MegaSaM 经常出现几何不一致。相比之下，我们的方法在稀疏、动态和航拍场景中均能生成全局一致的重建结果。左列的航拍场景尤其明显：MegaSaM 出现严重的几何漂移和纹理拖影，导致重复图案和扭曲的全局布局。请注意，右上示例中的第二帧并非正立，使该案例尤其困难。

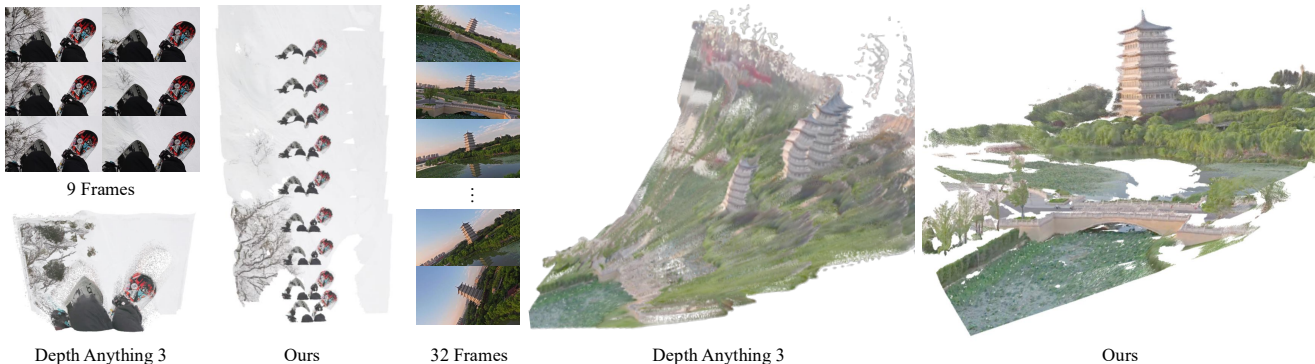


图 6. 与 Depth Anything 3 的定性比较。左：雪地上方的缆车序列，其中重复地形误导 DA3，使其估计出的相机运动很小甚至为零。我们的方法恢复了正确的相机轨迹。右：无人机朝塔飞行时相机发生滚转的序列。在这种强烈视角变化下，DA3 产生严重的重影和重复塔体结构，而我们的重建结果仍保持全局一致。

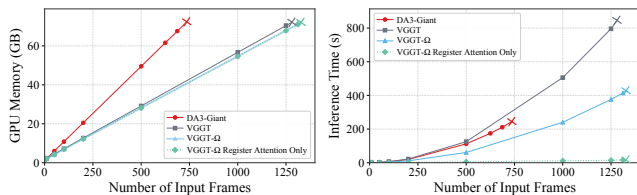


图 7. 内存与速度比较。我们在启用 flash attention v2 的 80GB A100 GPU 上比较 VGGT、DA3 和 VGGT- Ω 的推理内存占用与运行时间。对于 VGGT，我们首先修正了一个可提升其效率的实现细节（见正文）。叉号表示各方法首次出现内存不足的设置。

使图像词元数量减少约 25%。此外，VGGT- Ω 默认以

寄存器注意力替换 25% 的全局注意力层，从而减少 FLOPs 并获得 20–25% 的加速。总体而言，VGGT 和 VGGT- Ω 在内存与运行时间方面的规模扩展特性均优于 DA3；*e.g.*，在我们的测试中，DA3 在约 750 帧时耗尽内存，而 VGGT 和 VGGT- Ω 可处理约 1250 帧。一种更激进的 VGGT- Ω 变体将所有全局注意力层替换为寄存器注意力，可进一步提速，使 1000 帧的运行时间从 240.2 秒降至 11.7 秒，但会牺牲重建精度（章节 5）。

虽然这一现象可能出人意料，但需要强调的是，推理时移除全局注意力的主要收益是速度，而非内存。如图 7 所示，将所有全局注意力层替换为寄存器注意力可大幅提速，但推理峰值内存几乎不变。这是因为与全局注意力不同，寄存器注意力无需计算所有图像词元对之间的交互。不过，当 PyTorch

scaled dot product attention 使用 flash attention v2 时, 完整注意力矩阵并不会在内存中显式实例化。相反, 其内核以分块流式方式计算注意力, 在线更新 softmax 归一化并直接累积输出。峰值内存主要由尺寸与帧数和图像词元数量成正比的张量占用, 例如帧内注意力激活或前馈网络中间结果。这解释了图 7 中内存为何随帧数近似线性增长。

4.3. 消融实验

除非另有说明, 消融实验均采用 1B 模型变体: 在 2M 个序列上使用 64 块 GPU 进行 150k 步监督训练。为联合评估相机和深度精度, 我们利用相机将深度图反投影至三维空间, 并计算预测点与真实点之间的 ℓ_2 距离, 将其称为点误差 (*Point Error*)。之所以使用点误差而非 Chamfer 距离, 是因为无序点集上的最近邻匹配容易被墙面和地板等大面积表面主导。所有模型的训练词元数量大致相同。

模型与数据规模。 如图 1 所示, 扩大模型规模或数据规模均能持续改善性能。训练序列数量以 $10\times$ 的幅度增加时, 点误差从 0.275 单调降至 0.073。总体而言, 两条曲线的形状表明, 幂律可能刻画此类模型的规模扩展规律。

寄存器注意力。 仅使用全局注意力层的变体达到 0.071 的点误差。以寄存器注意力替换 25% 的全局注意力层后, 性能与原模型几乎相同 (0.073)。

多任务学习。 移除点损失和匹配损失后, 点误差从 0.073 增至 0.078。作为参考, 采用 VGGT 原始的多头多任务设置可达到 0.070, 但需要多个稠密预测头, 难以进行规模扩展。

自监督训练。 将 10% 的监督训练步骤替换为自监督训练后, 点误差从 0.073 略降至 0.070。性能提升源于使用更加多样的无标签数据进行训练。我们还观察到分布外泛化能力有所提升。

标注质量。 为验证标注流程生成的伪真值质量, 我们在提供合成相机与深度真值的 Sintel 上将其与 MegaSaM [99] 进行比较。为保证公平, 只评估同时满足我们的筛选标准并通过 MegaSaM 验证的序列和像素, 排除 8/23 个序列及所有动态区域。在该协议下, 我们的流程在相机位姿上达到 96.4% AUC@30°, 在深度上达到 99.3% $\delta_{1.25}$; MegaSaM 的对应结果分别为 62.1% 和 77.2%, 验证了所得标注的高质量。生成伪标签的目标并非最大化产出量, 而是仅保留极可能正确的序列和像素, 因为实践表明, 规模较小但精度较高的伪真值集合比规模更大但噪声更多的集合更有益。因此, 标注流程有意采用保守策略: 只要序列存在轻微歧义, 或像素无法得到可靠验证, 我们就将其丢弃, 以免引入噪声监督。

4.4. 寄存器的应用

机器人。 近期工作探索了利用重建模型提升视觉-语言-动作 (VLA) 系统空间理解能力的方法 [1, 93]。我

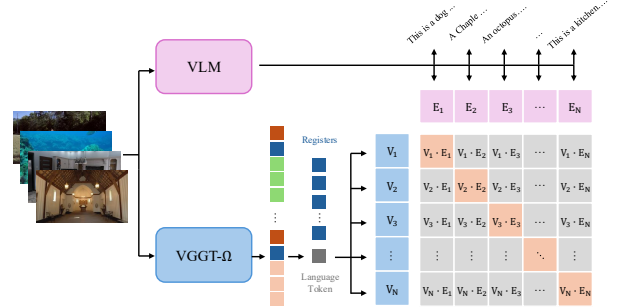


图 8. **语言对齐。** 我们通过语言对齐验证寄存器是否包含高层信息。对于每个序列, VLM 描述场景内容、粗略布局和外观, 并对生成的文本词元进行均值池化, 形成语言嵌入。在 VGGT-Ω 一侧, 一个小型自注意力堆栈以寄存器和可学习语言词元作为输入, 再对输出语言词元进行投影, 形成寄存器衍生嵌入。在全局批次的所有序列—描述配对上, 使用对称 InfoNCE 损失优化这两个嵌入。

们评估 VGGT-Ω 能否作为 VLA 的即插即用几何编码器。给定输入图像后, 从 VGGT-Ω 提取寄存器 (场景词元), 并将其与 OpenVLA-OFT 的原始输入词元拼接 [85]。随后按照标准协议训练 OpenVLA-OFT, 同时始终冻结 VGGT-Ω。如表 3 所示, 具备几何感知能力的寄存器在所有 LIBERO 任务上均持续提升性能 [104]。

表 3. **LIBERO 基准上的性能。** 我们冻结预训练模型, 将场景词元作为额外输入馈送至 OpenVLA-OFT, 并报告成功率 (Success Rate, SR; 越高越好)。显著的性能提升验证了场景词元聚合空间信息的有效性。

方法	空间 SR (%)	物体 SR (%)	目标 SR (%)	长程 SR (%)	平均 SR (%)
Diffusion Policy	78.3	92.5	68.3	50.5	72.4
TraceVLA	84.6	85.2	75.1	54.1	74.8
Octo	78.9	85.7	84.6	51.1	75.1
OpenVLA	84.7	88.4	79.2	53.7	76.5
Dita	84.2	96.3	85.4	63.8	82.4
CoT-VLA	87.5	91.6	87.6	69.0	83.9
π_0 -FAST	96.4	96.8	88.6	60.2	85.5
π_0	96.8	98.8	95.8	85.2	94.2
UniVLA	96.5	96.8	95.6	92.0	95.2
OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1
OpenVLA-OFT + Our Frozen Scene Tokens	99.3	99.2	99.0	96.7	98.5

语言对齐。 为进一步验证寄存器是否包含高层信息, 我们研究其能否与自然语言对齐。该过程如图 8 所示, 遵循 CLIP 式对比对齐的思想 [133, 216]。为尽可能保持简单, 我们使用训练充分的 VGGT-Ω, 冻结语言编码器, 仅微调本模型。

具体而言, 我们分别使用视觉语言模型 (Vision-Language Model, VLM) 和 VGGT-Ω 为给定图像序列提取全局描述符, 随后将二者对齐。提取该全局描述符时, VLM 观察所有输入视角, 并根据提示将场景内容、粗略布局和外观描述为一个连贯的整体场景。随后, 对

VLM 生成的文本词元隐藏状态进行均值池化和 ℓ_2 归一化。为从 VGGT- Ω 获得全局描述符，一个小型自注意力堆栈以寄存器和一个随机初始化的新可学习语言词元作为输入。对输出语言词元进行投影和 ℓ_2 归一化，得到 VGGT- Ω 中源自寄存器的嵌入。

需要注意的是，引入的语言词元只能读取寄存器，从不直接观察图像图块词元。因此，成功对齐说明寄存器本身携带可与语言匹配的场景级信息。

我们使用对称 InfoNCE 损失，最大化匹配的寄存器衍生嵌入与 VLM 语言嵌入之间的余弦相似度，同时最小化不匹配样本对的相似度。模型在与主要重建训练相同的图像序列上进行微调。与分布式 CLIP 训练相同，嵌入会跨 GPU 聚合，使全局批次提供批内负样本。VLM 保持冻结；与机器人实验中固定 VGGT- Ω 不同，此处以 1×10^{-5} 的恒定学习率端到端微调 VGGT- Ω 。

仅以较小学习率迭代 10K 次后，模型便能很好地迁移至语言检索任务。评估时，我们构建了由 100 个手工筛选网络视频组成的基准，涵盖烹饪、祈祷和赛车等多种场景。

给定完成对齐的 VGGT- Ω 所产生的寄存器衍生嵌入，我们使用余弦相似度为每个候选视频检索配对的语言嵌入，并报告 top- K 准确率。使用对齐期间采用的 VLM 嵌入时，模型的 top-1 准确率达到 76.8%，top-3 准确率达到 97.0%。换言之，几乎所有视频的正确语言描述都位列前三名候选结果。为检验这种对齐能否迁移到特定训练目标之外，我们在不进行任何额外训练的情况下，将 VLM 嵌入替换为纯文本 LLM 嵌入。具体而言，提示 VLM 生成视频描述，并仅将该描述输入 Qwen3 LLM [201]。在这种零样本迁移设置下，模型仍取得 47.5% 的 top-1 准确率和 77.8% 的 top-3 准确率。

总体而言，这些结果表明寄存器携带高层信息，可能具有很强的语义性，并且能够与语言空间对齐。与此同时，对齐微调后几何任务性能并未下降。补充材料提供了文本描述示例和提示词。这可能说明，强几何模型学到的表示能够自然地与语言等其他模态的表示对齐。这一观察与柏拉图表示假说 (Platonic Representation Hypothesis) [56, 71] 一致：在不同模态上训练且能力足够强的模型，往往会趋向共享表示空间。

一般应用。 寄存器为从模型中提取序列级和帧级表示提供了一种通用机制。通过引入面向特定任务的可学习词元，该设计可自然扩展到其他预测任务。对于序列级预测，可将可学习词元与寄存器词元拼接，并将其解码为所需输出，类似于上文讨论的语言词元。对于帧级预测，可在各帧复制同一词元，将其与相应寄存器拼接，并用于生成逐帧估计。我们已在若干此类任务上开展初步实验，包括度量尺度估计、重力方向估计和人员存在检测，并观察到较有潜力的结果。这些可能性留待社区进一步探索。从这一角度看，相机词元也可视为类似寄存器的可学习词元，主要区别在于它接受直接的相机监督。



图 9. **运动感知表征。**我们用 k -means 对经 PCA 降维的中间图像词元进行聚类，结果始终能将移动的舞者与静态人群和背景分开，红色表示高响应区域。这表明 VGGT- Ω 无需显式运动监督，即可从重建训练中学得运动感知表征。

5. 进一步观察

在开发 VGGT- Ω 的过程中，我们得到若干尚未严格验证的经验观察；这些发现仍值得与有意开发类似模型的研究者分享。

模型融合：信息存储在哪里？ 为更好理解模型行为，我们将模型融合 (Model Souping) [194] 用作探针。具体而言，无需进一步训练，直接对特定权重子集取平均，将公开 VGGT 检查点融合进 VGGT- Ω 。尽管两者架构不同（例如 VGGT 使用图块大小为 14 的 DINOv2，而 VGGT- Ω 使用图块大小为 16 的 DINOv3 和寄存器注意力），平均后的模型仍能产生视觉上合理的重建，这一点颇为意外。

融合不同权重子集，可以揭示不同类型信息在模型中的存储位置。首先，深度和视场角信息主要存储在注意力块的前馈网络 (FFN) 中，尤其是帧内注意力块，其次是全局注意力块。这与语言模型领域的观察一致 [30, 52, 119]。例如，章节 B 指出，模型会记忆训练数据中特殊的含噪信号，如在特定场景中把人体视为地面的一部分。将 VGGT 帧内注意力块的 FFN 权重以 50%-50% 的比例平均融合进 VGGT- Ω 后，这些错误得以消除；融合 Q/K/V 投影权重则没有同样效果。相机外参信息似乎在更高层编码，并非仅由 FFN 权重控制。最后，对可变输入帧数的泛化能力也与帧内注意力块密切相关。

运动感知。 由于模型能重建包含大幅非相机运动的动态序列，我们进一步考察其是否学会定位运动区域。为此，我们分析中间词元的特征分布：在时空维度上归一化中间图像词元，用 PCA 降维，再用 k -means 聚类。整个过程均不使用标签、光流或学习式探针。

如图 9 所示，一个 PCA 簇始终跨帧跟踪移动的舞者，而静态人群和背景留在其他簇中。这表明，无需显式运动监督，运动判别能力会作为重建目标的副产物自动涌现。此外，无论训练还是推理，模型从未获得输入帧的时间顺序。

我们还考察了该信号随层深的演化。浅层（如图 9 中的第 4 层）产生最干净的运动分割，能隔离舞者且背景响应极少；中层（如第 13 层）保留较弱但仍可辨的运动信号；在最深层（如第 23 层），相同聚类过程会突出场景中的所有人物，说明表征逐渐更具全局性和语义性。

预测归一化。 VGGT 曾讨论是否应在线将预测归一化到单位坐标空间，这里进一步分析该问题。与 VGGT 的观察一致，模型收敛后，使用与不使用预测归一化的定量性能没有差异。归一化的主要优势体现在定性结果上：最终点云在空间中分布更均匀。缺点是优化更不稳定，学习曲线更陡峭，也更依赖谨慎的超参数调节以避免梯度爆炸。若必须归一化预测，建议像 PI3 一样从预训练模型初始化，以稳定训练。

DINO 初始化。 如前所述，VGGT- Ω 由 DINOv3 初始化。我们发现，不采用 DINO 初始化（即从头训练）仍可收敛到相近性能，但需要多 4-8 $\times$ 的训练迭代。DINOv3 显著降低了优化难度并提高训练效率。

无效区域预测。 我们参照 [83, 185] 试验了预测无效区域的额外分支，希望帮助模型忽略天空等区域，避免点云中的天空像素黏附于前景物体。尽管模型能准确预测无效掩码，天空像素仍出现在深度估计前景中，可能因为这些区域缺少监督信号。因此，为尽量减少稠密预测头数量，最终模型未加入该预测器。

仅使用寄存器注意力。 我们还测试了完全不使用完整全局注意力、仅使用寄存器注意力的方案。此时，一帧中的图像词元无法关注其他帧的图像词元，所有帧间信息交换均依赖寄存器。尽管 FLOPs 降至原模型的 6%，性能也降至原始 VGGT 的水平。对于端侧应用，这一权衡值得探索。

辅助输入。 理论上，引入时间顺序、相机参数、深度图或尺度因子等辅助输入可进一步提高性能。然而，经验表明，在预训练阶段引入这些先验往往有害，即使在不同训练迭代中随机使用或掩蔽亦然。相反，初步实验显示，仅在微调阶段提供条件辅助输入非常有效，既能提高任务特定性能，又不破坏学得表征的完整性。全面探索超出本文范围，但这是一个有前景的未来方向。

合成与真实数据混合。 与 VGGT [181] 的观察一致，合成数据和真实数据在训练中作用互补。粗略而言，合成数据更直接地贡献于准确率，真实数据则改善泛化能力；后者的主要作用是让模型适应真实世界外观，并接触更多样的相机轨迹。实践中，建议每个训练轮次采样约 80% 合成数据和 20% 真实数据。若能获得足够干净的合成标注，例如解决章节 B 所述问题后，将合成数据比例提高至约 90% 可能更佳。

如何微调 VGGT/VGGT- Ω 。 在新数据域上微调 VGGT/VGGT- Ω 以完成重建任务时，我们建议采用两项实践策略。第一，使用完整学习率调度，即线性预热后接余弦衰减。峰值学习率应保持较小，迭代次数不必很多，但完整走完调度通常优于恒定学习率。第二，偶然不确定性（置信度）损失在微调期间可能不稳定，因此建议初期移除该损失，尤其当微调数据集很小时。若下游任务需要置信度估计，可单独微调不确定性头。对于重建之外的任务，建议提高预热比例（如从 5% 提至 10-15%），并使用完整调度周期。

自监督训练。 原则上，自监督训练是扩展训练数据的有力工具。目前它有助于提高模型泛化能力，尤其针对分布外数据，但对大多数基准影响很小；进一步改进并不容易，并行工作也得到类似观察 [67]。我们投入大量时间探索其他自监督方案，包括与 RayZer [73] 和 E-RayZer [224] 类似的新视图合成变体、生成词元而非像素、NeRF 表征 [7] 或高斯泼溅 [84]。还尝试了掩蔽图像词元、跨帧区分物体、引入时间顺序及相关变体。在我们的实现中，只有学生-教师方法有效。例如，E-RayZer 等方法可能适用于静态场景，却难以处理动态场景。这一结果有违直觉，因为二维自监督训练很大程度上依赖某种不变性（如 DINO 中的图像增强），重建理应自然编码此类不变性。我们原本预期自监督重建可从头训练，但成功方案仍需要预训练模型。因此，尽管章节 3.4 中的教师-学生自监督训练并非取得基准结果所必需，且会使训练流程更复杂，我们仍将其分享给社区。自监督重建仍是开放问题，最终可能需要整合进后文讨论的统一模型或全能模型。

稠密预测头。 原则上，强潜在表征应允许仅用 MLP 解码稠密预测任务，无需依赖卷积层；JiT [96] 和 LagerNVS [156] 等近期结果支持这一观点。这与保持框架简洁、可扩展的目标一致，因此我们曾对此展开探索。然而，纯 MLP 预测头总会在深度图中产生可见的图块/块状伪影，尽管其定量指标通常优于卷积替代方案，速度更快、显存效率更高，训练梯度也更稳定。这些伪影会引入本应平滑的几何不连续，对人眼尤其明显。我们尝试了 mipmap 式监督 [7]（引入局部深度方差）、概率建模（以多个独立通道的概率混合估计深度）及相关变体，但均无法可靠消除伪影。有趣的是，该问题在室外场景中远比室内常见，尤其当场景含有远处物体时。我们推测这与预测目标的数值分布有关：JiT 研究的图像生成输出位于边界明确的数值空间，而深度等几何量的范围实际上无界。因此我们采用折中方案：保留少量卷积层，后接 MLP 层。浅层卷积在 16×16 或 32×32 等低特征分辨率上运行，计算开销很小，却贡献了观察到的大部分空间平滑度提升。为保持与先前方法的架构可比性，我们使用 DPT 的浅层；实际上，任何特征金字塔式卷积架构都能缓解图块伪影。尽管 VGGT- Ω 采用此折中方案，我们仍认为纯 MLP 稠密解码头是有前景且重要的未来方向。

6. 讨论

下面讨论 VGGT- Ω 设计中的若干取舍。这些取舍在一定程度上源于我们对前馈式模型如何为研究社区创造最大价值的思考。

优先保持简洁性。 我们发现，若干架构改动可以进一步提升性能。例如，尽管本文模型只需单次前向传播即可预测相机参数，但采用原始 VGGT [181] 的迭代细化设计可取得更好的结果。将原始 RGB 值注入稠密预测头也能改善深度输出。结合这些及其他技术后，在多个数据集上，AUC@3° 可进一步提高 4%-6%， $\delta_{1.25}$ 可提高约 2%；我们预计，进一步针对特定任务进行修

改还能带来更多增益。即便如此，我们仍有意优先保证模型整体的简洁性，以及特征骨干网络（聚合器）所提取表征的质量。这一选择源于如下观察：骨干网络经过充分训练后，训练一个新的或改进后的预测头通常只需 5–10K 次迭代。保持架构简洁还能得到更干净的基础模型，便于研究社区在此基础上继续开发。

前馈式重建的优势。 与传统重建流程相比，前馈式重建具有三项重要优势：(a) 效率高，(b) 鲁棒性强，(c) 表征能力强。首先，其速度明显更快，在某些设置下可达到数十倍乃至数百倍；采用流式处理或多 GPU 并行后，差距还会进一步扩大。其次，相比束调整等高度依赖显式几何优化的方法，前馈式重建的鲁棒性强得多。例如，它可以处理视差很小或完全没有视差的情况，而基于三角测量的方法难以应对此类情况。如实验所示，前馈式重建也易于扩展到动态场景。第三，也是最重要的一点，前馈式重建能够提取通用的几何感知表征。三维视觉领域缺少强大且开箱即用的通用特征提取器；在二维视觉领域，已有较早的 VGG16 [148] 和 ResNet [59]，更不用说现代二维视觉或视觉-语言基础模型。本文通过多个示例表明，前馈式重建是为三维任务学习此类表征的一项前景良好的代理任务。

COLMAP [140] 等基于优化的重建方法仍有其优势。在条件良好的情况下，束调整能以极高精度估计相机参数，将角度误差降至百分之几度。对于使用 NeRF [120] 或高斯泼溅 [84] 进行新视角合成等应用，这种精度仍然很有价值。前馈式重建与优化并不冲突；相反，它可为束调整等后续优化过程提供良好的初始化。尽管如此，鉴于 VGGT 提出后前馈式重建在短期内取得的快速进展，我们仍对其未来的精度提升和这一范式的适用性保持乐观。

大模型时代三维/四维技术的作用。 空间理解正日益受到视觉-语言、视觉-语言-动作、视觉-动作及世界模型领域的关注 [17, 20, 28, 62, 86, 94, 122, 127, 163, 235]。这是开发具身智能体的自然结果，因为它们需要在三维物理世界中运行。短期来看，重建模型可作为外部工具，用于获取深度和相机参数等显式三维信息。如本文所述，重建模型还可提供隐式编码空间信息的“结构化”词元。长期来看，重建可能成为未来统一模型（“全能模型”）的组成部分，*i.e.*，即使用语言、视觉等多种模态进行预训练，并面向理解与生成等多种能力的大模型 [34, 162, 164, 167, 177, 188, 195, 229]。这一目标有望通过在模型训练中加入面向重建的数据来实现，而无需大幅修改架构或训练框架。例如，可将相机参数作为文本进行自回归预测，并将深度等逐像素量建模为图像生成任务，近期工作 [202] 已对此展开探索。

下一代重建模型乃至更广泛的感知系统，或将建立在统一模型之上。 首先，最大的收益很可能来自数据，而数据正是重建等感知任务目前的主要瓶颈。大规模文本与视频语料包含对物理世界丰富的隐式描述。将几何任务与这些模态联合训练，便可利用这一规模庞大、覆盖更广的监督来源。其次，大多数感知问题在孤立处理时都存在严重的欠约束。统一模型会自然施加跨任

务一致性，使一个领域中的歧义（*e.g.*，深度估计中的无纹理区域）能够借助另一领域的先验（*e.g.*，语义上下文）来消解。第三，从范式层面也有理由朝这一方向发展：近期观察可能表明，相比仅面向感知的视觉模型，生成式视觉模型更易扩展，并且生成模型似乎能够在一定程度上迁移到感知任务 [48]。我们预计，更强、更鲁棒的感知模型将来自多模态、多任务训练，而不是孤立的重建或感知目标。

7. 结论

本文提出了 VGGT- Ω ，一种在静态和动态基准上均取得优异结果的前馈式重建模型。我们从架构、数据和训练三个方面改进原始 VGGT：引入寄存器注意力，采用带多任务监督的单一稠密预测头，构建处理动态内容的大规模标注流程，并设计利用海量无标签视频的自监督训练方案。这些设计使模型得以在前所未有的规模上训练。实验表明，VGGT- Ω 的性能会随模型容量和数据规模可预测地提升。除几何信息外，学得的寄存器还承载有用的全局信息，可改进 VLA 模型并支持语言对齐。我们希望 VGGT- Ω 能成为社区进一步研究的有用基础。

致谢。 Jianyuan Wang 获得 Facebook AI Research 的支持。Christian Rupprecht 部分获得 ERC 启动基金“Volute”（编号 101222037）的支持。许多人为本工作提供了帮助。完整致谢请参见项目主页。

参考文献

- [1] Ali Abouzeid, Malak Mansour, Qinbo Sun, Zezhou Sun, and Dezheng Song. Geoaware-VLA: Implicit geometry aware vision-language-action model. *arXiv*, 2025. 1, 3, 9
- [2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10), 2011. 2
- [3] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kotschieder. Mapillary planet-scale depth dataset. In *Proc. ECCV*, 2020. 6
- [4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocation: Metric pose relative to a single image. In *Proc. ECCV*, 2022. 6
- [5] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proc. CVPR*, 2022. 6
- [6] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Intro-

- ducing HOT3D: an egocentric dataset for 3D hand and object tracking. *arXiv*, 2406.09598, 2024. 6
- [7] Jonathan T. Barron, Ben Mildenhall, Matthew Tan-cik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proc. ICCV*, 2021. 11
 - [8] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: a synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proc. CVPR*, 2023. 6
 - [9] Tyler Bonnen, Jitendra Malik, and Angjoo Kanazawa. Human-level 3D shape perception emerges from multi-view learning. *arXiv*, 2026. 1, 2
 - [10] Eric Brachmann, Jamie Wynn, Shuai Chen, Tom-maso Cavallari, Áron Monszpart, Daniyar Tur-mukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collec-tions via incremental learning of a relocalizer. In *Proc. ECCV*, 2024. 2
 - [11] Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understand-ing and learned data priors in VGGT. *arXiv*, 2025. 2
 - [12] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3D shape from image streams. In *Proc. CVPR*, 2000. 3, 5
 - [13] Leo Breiman. Random forests. *Machine learning*, 45 (1), 2001. 6
 - [14] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. ECCV*, 2012. 6
 - [15] Yann Cabon, Naila Murray, and Martin Humen-berger. Virtual KITTI 2. *arXiv*, 2001.10773, 2020. 6
 - [16] Yann Cabon, Lucas Stof, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *Proc. CVPR*, 2025. 2
 - [17] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric depth from vision language mod-els. In *Proc. ICLR*, 2026. 12
 - [18] Yang Cao, Feize Wu, Dave Zhenyu Chen, Yingji Zhong, Lanqing Hong, and Dan Xu. VGGT-Det: Mining VGGT internal priors for sensor-geometry-free multi-view indoor 3D object detection. *arXiv*, 2026. 1, 3
 - [19] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 2, 5
 - [20] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spa-tialVLM: Endowing vision-language models with spa-tial reasoning capabilities. In *Proc. CVPR*, 2024. 12
 - [21] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proc. CVPR*, 2021. 2
 - [22] Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. Geometric context transformer for streaming 3D re-construction. *arXiv*, 2026. 2
 - [23] Tianqi Chen and Carlos Guestrin. Xgboost: A scal-able tree boosting system. In *Proc. Knowledge Dis-covery and Data Mining*, 2016. 6
 - [24] Tianrun Chen, Yuanqi Hu, Yidong Han, Hanjie Xu, Deyi Ji, Qi Zhu, Chunyan Yu, Xin Zhang, Cheng Chen, Chaotao Ding, Ying Zang, Xuanfu Li, et al. HD-VGGT: High-resolution visual geometry transformer. *arXiv*, 2026. 2
 - [25] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. *arXiv*, 2025. 2
 - [26] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *Proc. ICLR*, 2026. 2
 - [27] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Ev-eryone everywhere all at once. *arXiv*, 2025. 3
 - [28] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Proc. NeurIPS*, 2024. 12
 - [29] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of in-door scenes. In *Proc. CVPR*, 2017. 6
 - [30] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pre-trained transformers. In *Proc. ACL*, 2022. 10
 - [31] Yuchao Dai, Hongdong Li, and Mingyi He. A sim-ple prior-free method for non-rigid structure-from-motion factorization. *IJCV*, 107, 2014. 5
 - [32] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *Proc. ICLR*, 2024. 1, 3
 - [33] Frank Dellaert. Factor graphs and GTSAM: A hands-on introduction, 2012. 2
 - [34] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretrain-ing. *arXiv*, 2505.14683, 2025. 12
 - [35] 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. In *Proc. 3DV*, 2026. 2
 - [36] 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*, 2025. 2
- [37] Tianchen Deng, Xun Chen, Ziming Li, Hongming Shen, Danwei Wang, Javier Civera, and Hesheng Wang. Unipr-3d: Towards universal visual place recognition with visual geometry grounded transformer. *arXiv*, 2025. 3
- [38] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: self-supervised interest point detection and description. In *Proc. CVPR Workshop*, 2018. 2, 6
- [39] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MAST3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. In *Proc. 3DV*, 2025. 2
- [40] 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 *Proc. CVPR*, 2019. 2
- [41] Sven Elfle, Qunjie Zhou, and Laura Leal-Taixé. Light3R-SfM: towards feed-forward structure-from-motion. In *Proc. CVPR*, 2025. 2
- [42] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, and Chengfei Lyu. Dens3R: A foundation model for 3D geometry prediction. *Proc. ICLR*, 2026. 2
- [43] Olivier D. Faugeras and Stephen J. Maybank. Motion from point matches: Multiplicity of solutions. *IJCV*, 4(3), 1990. 2
- [44] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: simultaneous 4D reconstruction and tracking in the world. In *Proc. ICCV*, 2025. 3
- [45] Michael Fonder and Marc Van Droogenbroeck. Mid-Air: A multi-modal dataset for extremely low altitude drone flights. In *Proc. CVPR Workshop*, 2019. 6
- [46] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building Rome on a cloudless day. In *Proc. ECCV*, 2010. 2
- [47] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-Neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. In *Proc. NeurIPS*, 2022. 2
- [48] Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannan, et al. Image generators are generalist vision learners. *arXiv*, 2026. 12
- [49] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Proc. NeurIPS*, 2022. 6
- [50] 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. *arXiv*, 2025. 2
- [51] Yulu Gao, Bohao Zhang, Zongheng Tang, Jitong Liao, Wenjun Wu, and Si Liu. VGGT-Segmentor: Geometry-enhanced cross-view segmentation. *arXiv*, 2026. 1, 3
- [52] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proc. Empirical Methods in Natural Language Processing*, 2021. 10
- [53] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, et al. Kubric: a scalable dataset generator. In *Proc. CVPR*, 2022. 6
- [54] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proc. CVPR*, 2020. 2
- [55] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. D²USt3R: Enhancing 3D reconstruction with 4d pointmaps for dynamic scenes. *arXiv*, 2504.06264, 2025. 3
- [56] Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pre-training. In *Proc. NeurIPS Workshop*, 2025. 10
- [57] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 1
- [58] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004. 2
- [59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 12
- [60] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, 2020. 2, 5
- [61] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. *arXiv*, 2405.14868, 2024. 6
- [62] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv*, 2023. 12
- [63] Yuan-Ting Hu, Jiahong Wang, Raymond A Yeh, and Alexander G Schwing. SAIL-VOS 3D: A synthetic

- dataset and baselines for object detection and 3d mesh reconstruction from video data. In *Proc. CVPR*, 2021. 6
- [64] Guichen Huang, Ruoyu Wang, Xiangjun Gao, Che Sun, Yuwei Wu, Shenghua Gao, and Yunde Jia. LongSplat: Online generalizable 3D Gaussian splatting from long sequence images. In *Proc. ICCV*, 2025. 2
- [65] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, et al. ViPE: video pose engine for 3D geometric perception. *arXiv*, 2508.10934, 2025. 3, 24
- [66] Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3R: 3D scene generation meets feed-forward reconstruction. In *Proc. CVPR*, 2026. 1, 3
- [67] Nan Huang, Pengcheng Yu, Weijia Zeng, James M Rehg, Angjoo Kanazawa, Haiwen Feng, and Qianqian Wang. Self-improving 4d perception via self-distillation. *arXiv preprint arXiv:2604.08532*, 2026. 11
- [68] Nan Huang, Pengcheng Yu, Weijia Zeng, James M Rehg, Angjoo Kanazawa, Haiwen Feng, and Qianqian Wang. Self-improving 4D perception via self-distillation. *arXiv*, 2026. 2
- [69] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *Proc. CVPR*, 2018. 6
- [70] Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. PointWorld: Scaling 3D world models for in-the-wild robotic manipulation. *arXiv*, 2026. 24
- [71] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proc. ICML*, 2024. 10
- [72] Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. *arXiv*, 2026. 1, 3
- [73] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. RayZer: a self-supervised large view synthesis model. In *Proc. ICCV*, 2025. 11
- [74] Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, Jiuxiang Gu, Qixing Huang, et al. MegaSynth: Scaling up 3d scene reconstruction with synthesized data. In *Proc. CVPR*, 2025. 6
- [75] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: feed-forward 3D Gaussian Splatting from unconstrained views. *arXiv*, 2505.23716, 2025. 2, 3
- [76] Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don’t need trained registers. *arXiv*, 2025. 1, 3
- [77] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025. 3
- [78] Haian Jin, Rundui Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T. Barron, Noah Snaveley, and Aleksander Holynski. ZipMap: linear-time stateful 3D reconstruction via test-time training. In *Proc. CVPR*, 2026. 2
- [79] Dongki Jung, Jaehoon Choi, Yonghan Lee, Sungmin Eum, Heesung Kwon, and Dinesh Manocha. MoRe: Monocular geometry refinement via graph optimization for cross-view consistency. In *Proc. WACV*, 2026. 2
- [80] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv*, 2020. 1
- [81] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6
- [82] Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4d: Unified feed-forward metric 4d reconstruction. *arXiv preprint arXiv:2512.10935*, 2025. 3
- [83] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, et al. MapAnything: universal feed-forward metric 3D reconstruction. In *Proc. 3DV*, 2026. 2, 11
- [84] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *Proc. SIGGRAPH*, 42(4), 2023. 11, 12
- [85] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proc. RSS*, 2025. 9
- [86] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, et al. OpenVLA: an open-source vision-language-action model. In *Proc. CoRL*, 2025. 12
- [87] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. SStream3R:

- Scalable sequential 3D reconstruction with causal transformer. In *Proc. ICLR*, 2026. 2
- [88] Alexander Lappe and Martin A Giese. Register and [CLS] tokens induce a decoupling of local and global features in large ViTs. In *Proc. NeurIPS*, 2025. 3
- [89] Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. Eden: Multimodal synthetic dataset of enclosed garden scenes. In *Proc. WACV*, pages 1579–1589, 2021. 6
- [90] Seonho Lee, Jiho Choi, Inha Kang, Jiwook Kim, Jun-sung Park, and Hyunjung Shim. 3D-aware vision-language models fine-tuning with geometric distillation. In *Proc. EMNLP*, pages 10628–10647, 2025. 1, 3
- [91] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MAST3R. In *Proc. ECCV*, 2024. 1
- [92] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, et al. BEHAVIOR-1K: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv*, 2024. 6
- [93] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv*, 2025. 1, 3, 9
- [94] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. In *Robotics: Science and Systems*, 2026. 12
- [95] Samuel Li, Pujith Kachana, Prajwal Chidananda, Saurabh Nair, Yasutaka Furukawa, and Matthew Brown. Rig3R: Rig-aware conditioning for learned 3D reconstruction. In *Proc. NeurIPS*, 2025. 2
- [96] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv*, 2026. 11
- [97] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*, 2018. 6
- [98] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, et al. Sekai: A video dataset towards world exploration. *arXiv*, 2025. 24
- [99] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 9, 24
- [100] Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. WinT3R: Window-based streaming reconstruction with camera token pool. In *Proc. ICLR*, 2025. 2
- [101] 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*, 2511.10647, 2025. 1, 2, 3, 5
- [102] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: local feature matching at light speed. In *Proc. ICCV*, 2023. 2, 6
- [103] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, et al. DL3DV-10K: a large-scale scene dataset for deep learning-based 3d vision. In *Proc. CVPR*, 2024. 6
- [104] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In *Proc. NeurIPS*, 2023. 9
- [105] Changkun Liu, Bin Tan, Zeran Ke, Shangzhan Zhang, Jiachen Liu, Ming Qian, Nan Xue, Yujun Shen, and Tristan Braud. PLANA3R: Zero-shot metric planar 3D reconstruction via feed-forward planar splatting. *arXiv*, 2025. 2
- [106] Shaohui Liu, Yidan Gao, Tianyi Zhang, Rémi Pautrat, Johannes L Schönberger, Viktor Larsson, and Marc Pollefeys. Robust incremental structure-from-motion with hybrid features. In *Proc. ECCV*, 2024. 2
- [107] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *Proc. ECCV*, 2024. 6
- [108] Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. uCO3D uncommon objects in 3D. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6
- [109] Xuanyi Liu, Deyi Ji, Liquan Liu, Lanyun Zhu, Xuhang Chen, Qianxiong Xu, Peng Shu, Huan Yu, Jie Jiang, Feng Gao, and Siwei Ma. CamGeo: Sparse camera-conditioned image-to-video generation with 3D geometry priors. In *Proc. ICML*, 2026. 3
- [110] Xuanyi Liu, Chunan Yu, Deyi Ji, Qi Zhu, Lingyun Sun, Xuanfu Li, Jin Ma, Tianrun Chen, and Lanyun Zhu. StreamCacheVGGT: Streaming visual geometry transformers with robust scoring and hybrid cache compression. *arXiv*, 2026. 2
- [111] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1999. 6
- [112] 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. In *Proc. CVPR*, 2025. 3

- [113] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar RAFT. In *Proc. ECCV*, 2022. 2
- [114] Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the $sl(4)$ manifold. In *Proc. NeurIPS*, 2025. 2
- [115] Soroush Mahdi, Fardin Ayar, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. Evict3R: Training-free token eviction for memory-bounded streaming visual geometry transformers. *arXiv*, 2025. 2
- [116] Mapillary. Metropolis dataset, 2025. 6
- [117] Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy V. Vo. Revisiting [CLS] and patch token interaction in vision transformers. *arXiv*, 2026. 1, 3
- [118] Kirill Mazur, Marwan Taher, and Andrew J. Davison. 4D Primitive-Mâché: Glueing primitives for persistent 4D scene reconstruction. In *Proc. CVPR*, 2026. 2
- [119] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Proc. NeurIPS*, 2022. 10
- [120] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tan-cik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 12
- [121] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *Proc. CVPR*, 2020. 2
- [122] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, et al. Cosmos world foundation model platform for physical ai. *arXiv*, 2501.03575, 2025. 12
- [123] John Oliensis. A critique of structure-from-motion algorithms. *CVIU*, 80(2), 2000. 2
- [124] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 5
- [125] Onur Özyesil, V. Voroninski, R. Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26, 2017. 2
- [126] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for ego-centric 3D machine perception. In *Proc. ICCV*, 2023. 6
- [127] Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models, 2025. 12
- [128] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. TartanGround: A large-scale dataset for ground robot perception and navigation. In *Proc. IROS*, 2025. 6
- [129] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proc. CVPR*, 2022. 2
- [130] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Proc. NeurIPS*, 2018. 6
- [131] Dexin Qi, Tao Tao, Zhihong Zhang, and Xuesong Mei. Fupad: Scalable pose estimation by fusing patch-wise vgg with dense bundle adjustment. In *Proc. ICIRA*, 2025. 2
- [132] Kangan Qian, ChuChu Xie, Yang Zhong, Jingrui Pang, Siwen Jiao, Sicong Jiang, Zilin Huang, Yunlong Wang, Kun Jiang, Mengmeng Yang, Hao Ye, Guanghao Zhang, et al. Xembodied: A foundation model with enhanced geometric and physical cues for large-scale embodied environments. *arXiv*, 2026. 1, 3
- [133] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 9
- [134] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. ICCV*, 2021. 4
- [135] Zhifeng Rao, Wenlong Chen, Lei Xie, Xia Hua, Dongfu Yin, Zhen Tian, and F. Richard Yu. AugVLA-3D: Depth-driven feature augmentation for vision-language-action models. In *arXiv*, 2026. 1, 3
- [136] 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 *Proc. ICCV*, 2021. 6
- [137] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proc. ICCV*, 2021. 6
- [138] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: learning feature matching with graph neural networks. In *Proc. CVPR*, 2020. 2, 6
- [139] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proc. ICCV*, 2019. 6
- [140] Johannes Lutz Schönberger and Jan-Michael Frahm.

- Structure-from-motion revisited. In *Proc. CVPR*, 2016. 1, 2, 3, 6, 12, 24
- [141] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. ECCV*, 2016. 6
- [142] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR*, 2017. 6
- [143] You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: training-free acceleration of visual geometry transformer. In *Proc. ICLR*, 2026. 2, 4
- [144] Cheng Shi, Yizhou Yu, and Sibe Yang. Vision transformers need more than registers. *arXiv*, 2026. 3
- [145] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. ClusterGNN: cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proc. CVPR*, 2022. 2
- [146] 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 *Proc. CVPR*, 2013. 6
- [147] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, et al. DINOv3. *arXiv preprint*, 2025. 1, 2, 4, 5, 6
- [148] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 12
- [149] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. on Graphics (TOG)*, 2006. 1, 2
- [150] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv*, 2019. 6
- [151] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. EFM3D: A benchmark for measuring progress towards 3D egocentric foundation models. *arXiv*, 2024. 6
- [152] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proc. IROS*, 2012. 6
- [153] Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic Point Maps: A versatile representation for dynamic 3D reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025. 3, 5
- [154] Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: Video reconstruction with dynamic point maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3, 5
- [155] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. CVPR*, 2020. 6
- [156] Stanislaw Szymanowicz, Minghao Chen, Jianyuan Wang, Christian Rupprecht, and Andrea Vedaldi. LagerNVS: Latent geometry for fully neural real-time novel view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 1, 2, 3, 11
- [157] Marwan Taher, Ignacio Alzugaray, Kirill Mazur, Xin Kong, and Andrew J. Davison. KV-Tracker: Real-time pose tracking with transformers. *arXiv*, 2025. 2
- [158] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment network. In *Proc. ICLR*, 2019. 2
- [159] 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 *Proc. CVPR*, 2025. 2
- [160] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 2017. 2
- [161] Jonathan Taylor, Allan D Jepson, and Kiriakos N Kutulakos. Non-rigid structure from locally-rigid motion. In *Proc. CVPR*, 2010. 5
- [162] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv*, 2024. 12
- [163] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, et al. Gemini robotics: Bringing ai into the physical world. *arXiv*, 2025. 12
- [164] Meituan LongCat Team. LongCat-Next: Lexicalizing modalities as discrete tokens. *arXiv*, 2026. 12
- [165] Zachary Teed and Jia Deng. DeepV2D: video to depth with differentiable structure from motion. In *Proc. ICLR*, 2020. 2
- [166] Zachary Teed and Jia Deng. DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Proc. NeurIPS*, 2021. 2
- [167] Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaëys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, et al. Beyond language modeling: An exploration of multimodal pre-training. *arXiv*, 2603.03276, 2026. 12

- [168] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 30(5), 2008. 3
- [169] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo mixture density networks. In *Proc. CVPR*, 2021. 6
- [170] MJ Tyszkiewicz, P Fua, and E Trulls. DISK: learning local features with policy gradient. In *Proc. NeurIPS*, 2020. 2
- [171] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: depth and motion network for learning monocular stereo. In *Proc. CVPR*, 2017. 2
- [172] An Dinh Vuong, Minh Nhat Vu, and Ian Reid. Improving robotic manipulation with efficient geometry-aware vision encoder. In *arXiv*, 2025. 1, 3
- [173] Chongyu Wang, Ting Huang, Chunyu Sun, Xinyu Ning, Di Wang, and Hao Tang. Let geometry GUIDE: Layer-wise unrolling of geometric priors in multi-modal LLMs. *arXiv*, 2026. 1, 3
- [174] Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster VGGT with block-sparse global attention. *arXiv*, 2025. 2, 4
- [175] Hengyi Wang and Lourdes Agapito. Spann3R: 3D reconstruction with spatial memory. In *Proc. 3DV*, 2024. 2
- [176] Hengyi Wang and Lourdes Agapito. AMB3R: accurate feed-forward metric-scale 3D reconstruction with backend. *arXiv*, 2511.20343, 2025. 2
- [177] Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He, Dou Hong, Qiwen Liu, Shuohuan Wang, et al. Ernie 5.0 technical report. *arXiv*, 2026. 12
- [178] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *Proc. CVPR*, 2021. 2
- [179] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: solving pose estimation via diffusion-aided bundle adjustment. In *Proc. ICCV*, 2023.
- [180] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGsFM: visual geometry grounded deep structure from motion. In *Proc. CVPR*, 2024. 2, 3, 6
- [181] 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)*, 2025. 1, 2, 3, 4, 7, 11, 22
- [182] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiaoxiao Long, Hao Zhu, et al. SpatialVID: a large-scale video dataset with spatial annotations. *arXiv*, 2509.09676, 2025. 24
- [183] Lijie Wang, Lianjie Guo, Ziyi Xu, Qianhao Wang, Fei Gao, and Xieyuanli Chen. LiDAR-VGGT: Cross-modal coarse-to-fine fusion for globally consistent and metric-scale dense mapping. *IEEE Robotics and Automation Letters (RA-L)*, 2026. 2
- [184] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proc. CVPR*, 2025. 2, 3
- [185] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv*, 2410.19115, 2024. 11
- [186] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUST3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 1, 2, 5, 22
- [187] Wenshan Wang, DeLong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: a dataset to push the limits of visual SLAM. In *Proc. IROS*, 2020. 6
- [188] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yuezhe Wang, Zhen Li, Qiyang Yu, Yingli Zhao, Yulong Ao, et al. Emu3: Next-token prediction is all you need. *arXiv*, 2024. 12
- [189] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3 : Permutation-equivariant visual geometry learning. *arXiv*, 2507.13347, 2025. 1, 2, 3
- [190] Yunnan Wang, Kecheng Zheng, Jianyuan Wang, Minghao Chen, David Novotny, Christian Rupprecht, Yinghao Xu, Xing Zhu, Wenjun Zeng, Xin Jin, and Yujun Shen. SceneScribe-1M: A large-scale video dataset with comprehensive geometric and semantic annotations. In *Proc. CVPR*, 2026. 24
- [191] Zichen Wang, Ang Cao, Liam J Wang, and Jeong Joon Park. MoE3D: A mixture-of-experts module for 3D reconstruction. *arXiv*, 2026. 2
- [192] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. DeepSfM: structure from motion via deep bundle adjustment. In *Proc. ECCV*, 2020. 2
- [193] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proc. ICCV*, 2021. 2
- [194] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proc. ICML*, 2022. 10

- [195] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proc. CVPR*, 2025. 12
- [196] Changli Wu, Haodong Wang, Jiayi Ji, Yutian Yao, Chunsai Du, Jihua Kang, Yanwei Fu, and Lijuan Cao. MVGGT: Multimodal visual geometry grounded transformer for multiview 3D referring expression segmentation. *arXiv*, 2026. 1, 3
- [197] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Proc. NeurIPS*, 2025. 2, 3
- [198] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. RGBD objects in the wild: Scaling real-world 3D object learning from RGB-D videos. In *Proc. CVPR*, 2024. 6
- [199] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D point tracking made easy. In *Proc. ICCV*, 2025. 2, 3
- [200] Jingyi Xu, Zhangshuo Qi, Zhongmiao Yan, Xuyu Gao, Qianyun Jiao, Songpengcheng Xia, Xieyuanli Chen, and Ling Pei. VGGT-MPR: VGGT-enhanced multimodal place recognition in autonomous driving environments. *arXiv*, 2026. 1, 3
- [201] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, et al. Qwen3 technical report. *arXiv*, 2025. 10
- [202] Ceyuan Yang, Zhijie Lin, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Chaorui Deng, Kunchang Li, Zihan Ding, Yuwei Guo, Fuyun Wang, Fangqi Zhu, et al. Context unrolling in omni models. *arXiv*, 2026. 12
- [203] 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. *Proc. CVPR*, 2025. 2
- [204] Songlin Yang, Tianyi Wei, Yushi Lan, Zeqi Xiao, Anyi Rao, and Xingang Pan. Dense semantic matching with VGGT prior. *arXiv*, 2025. 1, 3
- [205] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, et al. Cambrian-S: Towards spatial supersensing in video. In *Proc. ICLR*, 2026. 1, 3
- [206] David Yifan Yao, Albert J. Zhai, and Shenlong Wang. Uni4D: unifying visual foundation models for 4D modeling from a single video. *arXiv*, 2503.21761, 2025. 3
- [207] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proc. ECCV*, 2018. 2
- [208] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Proc. NeurIPS*, 2020. 2
- [209] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: a high-fidelity dataset of 3d indoor scenes. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 6
- [210] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: learned invariant feature transform. In *Proc. ECCV*, 2016. 2
- [211] Yuheng Yuan, QiuHong Shen, Shizun Wang, Xingyi Yang, and Xinchao Wang. Test3R: Learning to reconstruct 3D at test time. In *Proc. NeurIPS*, 2025. 2
- [212] Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. CVPR*, 2018. 6
- [213] Ying Zang, Yidong Han, Chaotao Ding, Yuanqi Hu, Deyi Ji, Qi Zhu, Xuanfu Li, Jin Ma, Lingyun Sun, Tianrun Chen, and Lanyun Zhu. Robust 4D visual geometry transformer with uncertainty-aware priors. *arXiv*, 2026. 2
- [214] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, Xing Wei, and Ning Guo. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv*, 2025. 1, 3
- [215] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proc. CVPR*, pages 12104–12113, 2022. 1
- [216] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. CVPR*, 2023. 9
- [217] Chuhan Zhang, Guillaume Le Moing, Skanda Kopula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, et al. Efficiently reconstructing dynamic scenes one d4rt at a time. *arXiv*, 2512.08924, 2025. 3
- [218] 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*, 2410.03825, 2024. 3, 5
- [219] 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*, 2026. 2
- [220] Jiang Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. *arXiv*, 2026. 1, 3
- [221] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun

- Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proc. CVPR*, 2025. 2
- [222] Shihua Zhang, QiuHong Shen, Shizun Wang, Tianbo Pan, and Xinchao Wang. Make geometry matter for spatial reasoning. *arXiv*, 2026. 1, 3
- [223] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. GeoMVSNet: Learning multi-view stereo with geometry perception. In *Proc. CVPR*, 2023. 2
- [224] Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-RayZer: Self-supervised 3D reconstruction as spatial visual pre-training. In *Proc. CVPR*, 2026. 11
- [225] Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv*, 2025. 1, 3
- [226] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. ALIKED: a lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Trans. on Instrumentation and Measurement*, 72, 2023. 6
- [227] Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3D world: Enhancing mllms with 3D vision geometry priors. *arXiv*, 2025. 1, 3
- [228] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proc. CVPR*, 2023. 6
- [229] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *Proc. ICLR*, 2025. 12
- [230] Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse gaussian occupancy prediction. *arXiv*, 2026. 1, 3
- [231] Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. PAGE-4D: disentangled pose and geometry estimation for 4D perception. *arXiv*, 2510.17568, 2025. 3
- [232] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. CVPR*, 2017. 2
- [233] Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, et al. OmniWorld: A multi-domain and multi-modal dataset for 4D world modeling. In *Proc. ICLR*, 2026. 24
- [234] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *Proc. ICLR*, 2026. 2
- [235] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proc. CoRL*, 2023. 12

VG GT- Ω

补充材料

本补充材料在章节 A 中给出更多实现细节，在章节 B 中讨论常见数据问题，并在章节 C 中分析局限性。

A. 补充细节

本节详细介绍训练设置与架构（章节 A.1）、正负样本对的构造（章节 A.2）、VLM 提示词（章节 A.3）、标注筛选标准（章节 A.4）和语言对齐流程（章节 A.5）。

A.1. 训练与架构

训练时，Eq. (1) 中的损失权重设为 $\lambda_{\text{cam}} = 5.0$ 、 $\lambda_{\text{depth}} = 1.0$ 、 $\lambda_{\text{point}} = 0.5$ 和 $\lambda_{\text{match}} = 0.1$ 。遵循 [181, 186]，我们将每个场景的真实值归一化到单位空间。具体而言，先将所有量变换到第一台相机的坐标系，计算所有三维点到原点的平均距离，再用该值缩放深度图和平移向量。与 [181] 相同，此归一化仅用于真实值，不用于预测值。为稳定训练，我们采用阈值为 1.0 的梯度范数裁剪，并在注意力层中使用 QKNorm。每次迭代从 [1, 24] 随机采样输入帧数，并相应设置批大小以充分利用 GPU 显存。每帧以 0.05 的概率独立添加矩形掩码，其高和宽从 [32, 128] 像素均匀采样。掩码区域内的像素设为黑色，对应深度值标记为无效。颜色抖动采用 PyTorch 实现，亮度、对比度、饱和度和色相参数分别为 0.5、0.5、0.5 和 0.1。自监督训练期间，教师模型权重以衰减率 0.999 的指数移动平均更新。相机预测头沿用 [181] 的实现：焦距使用 ReLU 激活，四元数和平移参数不使用激活函数。对于具有时间顺序的序列（e.g.，视频），我们从局部时间窗口采样帧，而不依据共视关系采样。由此可能生成极具挑战的重建子集：图像语义相关，却几乎没有视觉重叠。实验发现，这些样本能提高模型泛化能力。

A.2. 构造正负样本对

下面补充主文（章节 3.2）中的匹配损失，详述正负图元对的构造方法。

给定由深度图反投影得到的逐像素三维点，我们先在查询帧中采样有效像素，并将各采样像素归入一个查询图块。随后利用已知相机内外参，将对应三维点投影至所有其他帧，仅保留同时满足以下条件的投影：(i) 落在图像内；(ii) 与目标帧深度图保持深度一致（相对容差为 1%，并排除宽 4 像素的窄图像边界）。

对每个目标帧，我们统计各查询图块有多少三维点落入各目标图块。由此为每个查询图块得到目标图块上的软对应图，以投影重叠率表示。我们选择跨视图有效投影充足的查询图块，并随机采样不超过固定数量的图块；采样概率与其对应总数成正比。对于每个选中的查询图块，所有重叠率 $>10\%$ 的目标图块均构成正图元对。

由于部分序列是动态的，不满足正样本标准并不自动意味着有效负样本。因此，我们随机采样图块作为查

询和目标候选，并施加两项约束以构造负样本对：(i) 几何约束，要求候选目标图块中心与查询图块诱导的极线相距足够远（i.e.，具有较大的极线/Sampson 距离）；(ii) 外观约束，要求两个图块的平均 RGB 值具有足够大的 ℓ_2 距离。同时满足两项约束的图块视为负样本。随后对负样本下采样，得到均衡的正负图块对，用于监督最后一层词元的匹配损失。

A.3. VLM 提示词

我们使用 VLM 丢弃不适合多视图几何的视频。VLM 提示词如下：

“你是计算机视觉、摄影测量和多视图几何 (MVG) 专家。”

“分析视频片段，判断其是否适合高质量三维重建。”

“必须依据以下严格标准对视频分类并提取场景元数据。”

步骤 1: 检查硬拒绝因素

“评估以下项目。若存在任意一项，则分类为 ‘REJECT_HARD’：”

1. 不连续与剪辑：视频是否包含切镜、叠化、划变、淡入淡出或蒙太奇剪辑？（必须是单个连续镜头。）
2. 非物理/二维内容：是否为屏幕录制、静态图像幻灯片、动画或卡通？（必须是真实世界影像。）
3. 严重视觉失效：严重运动模糊（边缘丢失）、严重滚动快门（晃动/果冻效应）或数据损坏/画面故障。
4. 严重遮挡：是否存在大面积叠加层（大段文字/UI）、水印，或手指遮住镜头等物理遮挡？
5. 非针孔投影：影像是否为 360° 等距柱状投影，或未经标定且畸变严重的鱼眼画面？

步骤 2: 检查几何与纹理适用性

“若视频通过步骤 1，则评估重建质量。若存在任意一项，则分类为 ‘REJECT_SOFT’：”

1. 视差不足：相机是否静止？是否只有旋转（平移拍摄/俯仰）或变焦而无物理平移？（深度恢复需要平移。）
2. 纹理问题：场景是否缺少非重复纹理？（如空白白墙、只有纯蓝天空、暗部阴影或网格瓷砖等高度重复图案。）
3. 镜面与透明物体：主要特征是否具有反射性（镜子、水、玻璃）或透明？（这会产生“幽灵”几何。）
4. 对焦与景深问题：是否存在严重的反复对焦（脉动），或景深过浅导致背景完全虚化（散景）？
5. 动态主体占优：移动主体（如说话人头部特写）是否占据画面 $>90\%$ ，而可见静态背景很少？

步骤 3: 提取元数据

“判断场景动态性：”

1. 动态：存在移动物体（行人、行驶车辆、被强风吹动的树木）。
2. 静态：场景为刚性，只有相机移动。

输出说明：

“严格按以下模式返回 JSON 对象。不得输出 Markdown 格式，也不得在 JSON 外添加解释。”

```
{
  "classification":
    "ACCEPT" | "REJECT_HARD" | "REJECT_SOFT",
  "reason": "Brief descriptions of the primary
    flaw or 'Good candidate'",
  "scene_dynamics": "Static" | "Dynamic"
}
```

“标签定义：”

REJECT_HARD: 因剪辑、合成内容、严重伪影或遮挡而不可用。
REJECT_SOFT: 技术上可用, 但可能因纯旋转、反射、缺少纹理或对焦问题而失败。
ACCEPT: 高质量候选; 单镜头、平移/视差清晰、纹理良好、几何刚性。

A.4. 标注筛选标准

为训练章节 3.5.2 所述的集成分类器 (XGBoost、随机森林和 CatBoost), 我们从每个重建结果中提取若干几何特征。这些特征用于检测共线运动退化、尺度不一致和相机轨迹含噪等常见失败模式。下面给出流程中最重要特征的定义。

轨迹平滑度。 我们用平移向量 $\mathbf{t}$ 的加速度量化估计相机轨迹的平滑度。平滑度分数 S_{trans} 定义为加速度模长平方的均值:

$$S_{\text{trans}} = \frac{1}{N-2} \sum_{i=1}^{N-2} \|\mathbf{t}_{i+1} - 2\mathbf{t}_i + \mathbf{t}_{i-1}\|^2 \quad (2)$$

对于旋转, 我们利用旋转向量的二阶差分计算类似指标 S_{rot} 。 S_{trans} 或 S_{rot} 较高表示轨迹抖动, 通常与 SfM 收敛不佳有关。

视差角分析。 为确保多视图立体具有足够基线, 我们分析重建稀疏点云的视差角。对于稀疏点子集 $\mathcal{P}$, 找出可见点 $p \in \mathcal{P}$ 的相机集合 $\mathcal{C}_p$, 并计算任意相机对 $c_j, c_k \in \mathcal{C}_p$ 在点 p 处张成的最大夹角。最终特征取所有采样点最大夹角的中位数。中位视差较低意味着纯旋转运动退化或距离极远。

点云 PCA 形状 (线性度与平面度)。 为检测“掠过式”直线重建 (会造成圆柱歧义) 等几何退化, 我们对归一化点云坐标执行主成分分析 (Principal Component Analysis, PCA)。令 $v_1 \geq v_2 \geq v_3$ 为点云协方差矩阵的特征值。定义线性度为 $(v_1 - v_2)/v_1$, 平面度为 $(v_2 - v_3)/v_1$, 散射度为 v_3/v_1 。线性度过高的重建会被标记为线性退化, 通常由分类器丢弃。

深度图完整度。 该指标评估稠密重建阶段的密度。对每一帧, 计算具有有效有限深度值的像素占总图像分辨率的比例, 最终特征为所有帧完整度的平均值。数值极低通常表示多视图立体阶段失败, 常由无纹理表面、强镜面反射或劣质相机导致。

点云噪声水平。 我们通过统计离群点检测估计信噪比。对每个点, 计算其到最近邻的平均距离; 若该距离超过全局均值两个标准差以上, 则将该点归为噪声。该特征取噪声点比例, 用于检测受漂浮伪影和离群点困扰的重建。

该流程只能为刚性像素提供深度值, 因此上述指标假设动态像素已被掩蔽。我们假设模型可从合成数据中学会估计动态像素深度。

A.5. 语言对齐

在章节 4.4 中, 我们使用 VLM 为每个训练序列生成序列级语言嵌入。这里给出确切的 VLM 提示词及代表性生成描述。生成文本词元的隐状态经均值池化和投影后形成语言嵌入, 提示词和图像词元不参与池化。新生成词元最多为 128 个, VLM 参数保持冻结。对齐微调期间按章节 4.4 所述优化 VGGT- Ω 侧。无论输入 VLM 还是本文模型, 图像均以无序集合形式提供。

语言对齐提示词

这些图像从多个视角展示同一场景。
请把完整场景描述为一个连贯场景, 而非多张独立图像。

采用以下格式:

场景: < 场景类别或环境 >。

内容: < 主要物体及粗略空间布局 >。

外观: < 材质、颜色或风格等稳定视觉属性 >。

每个字段用一个短句。

应提及跨多个视图一致的信息, 即使它并非在每个视图中都可见。

重点描述共享场景内容、主要物体、粗略布局和稳定外观。

仅当动态物体在多个视图中显著且可辨认时才提及。

不要描述运动、动作或特定帧变化。

不要逐张描述图像。

不要提及相机运动、视角顺序、图像质量、模糊、曝光、光照伪影或不确定性。

描述应简洁、客观且语义密集。

代表性描述

场景: 一间光线昏暗、泛红色调的房间, 内有门、窗和带图案纹理的墙。

内容: 房间左侧有一扇关着的门, 右侧有一扇窗, 另有一面带纹理图案的墙。

外观: 墙面带花卉或锦缎图案, 地板颜色较深, 整体光照呈红色。

场景: 带蜿蜒道路的雪林景观。

内容: 积雪覆盖的森林中有高大松树、蜿蜒道路和陡坡。

外观: 场景对比度较高, 上方为乌云, 明亮白雪覆盖地面和树木。

场景: 室外公共区域。

内容: 铺装地面上有一张深绿色塑料野餐桌, 桌面为六边形、桌腿为黑色, 并配有相应长凳。

外观: 桌椅均由统一的深绿色哑光塑料制成, 地面铺设浅灰色方砖。

场景: 具有现代建筑的城市广场。

内容: 中央喷泉设有大型圆形雕塑, 周围是铺装步道和建筑。

外观: 喷泉主体为银色金属结构, 底座较深; 周边铺有浅灰色石砖, 并有采用玻璃与混凝土立面的现代建筑。

B. 数据质量

数据是训练大型“基础”模型的关键要素之一, 但仅有大量数据并不足够; 数据质量同样会显著影响模型行为。对于三维重建, 我们发现含噪数据会在自监督和有监督训练中引入特定失败模式。

在自监督训练中, 难度不合理的视频 (如电视素材中常见的不连续镜头切换) 会引起剧烈损失尖峰, 甚至导致梯度爆炸和性能下降。因此, 自监督训练仅使用通

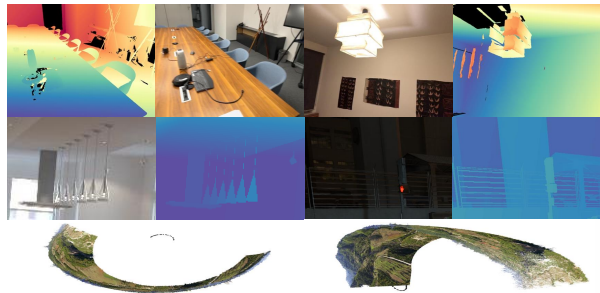


图 10. 常见数据问题。顶行：ScanNet++ 样本。中行：Hypersim 等合成数据集样本。底行：穹顶效应。

过章节 3.5.2 所述 VLM 预筛选的视频；该检查能可靠移除这类样本。

在有监督阶段，含噪数据的影响更隐蔽，却同样重要。不同类型的标注错误会在推理时转化为特定失败模式；更糟的是，这些模式不影响标准基准中的大多数图像，因而可能无法从定量结果中发现。只有当新测试样本与训练期间见过的错误标注样本相似时，问题才会显现。这表明模型虽学会了三维重建的一般原理，也会记忆特殊噪声。下面讨论最具代表性的案例。

传感器。 部分数据集使用 LiDAR 或深度相机等传感器标注真实场景，常见失败模式是前景—背景泄漏。例如，图 10 顶行展示的 ScanNet++ 样本中，椅背深度对应背景地板或墙壁，可能源于传感器采集未对齐或存在孔洞。悬挂灯具周围也出现虚假深度伪影：渲染深度含有与可见灯具几何不符的碎片化前景结构。这些伪影可能由过曝或半透明物体附近不可靠的传感器采集或重建造成，因此我们在训练后期排除此类数据集。

细薄结构。 合成数据通常提供更准确的几何标注，但部分合成数据集中的细薄结构仍可能不准。栅栏杆等物体往往只占数个像素，并位于陡峭深度不连续处。尽管其在 RGB 图像中清晰可见，渲染深度却可能不完整、过度平滑或未对齐。因此，标签可能将细薄结构赋为其后方墙壁、窗户、地板或远处背景的深度，而非真实物体表面，如图 10 中行所示。推理时，模型可能忽略细薄物体，仅恢复其较粗或较低部分，或使其在深度图中被邻近大表面淹没。这是现有模型的普遍问题，我们推测更高分辨率输入或更优合成数据可缓解该问题。

虚假背景。 合成数据的另一常见问题是虚假背景。例如，在 Kubric、PointOdyssey 和 BEDLAM 中，背景深度可能对应 HDRI 渲染所用的代理穹顶或地面几何，而非背景外观暗示的真实三维结构。这种设计虽适合渲染，却会提供与场景语义不一致的深度值。对于 BEDLAM 等真实前景与合成背景边界清晰的数据集，我们通过阈值化最大有效前景深度，在训练时滤除背景；其余此类数据则不用于训练。

穹顶效应。 COLMAP [140]、MegaSaM [99] 和 ViPE [65] 等采用束调整的方法可生成伪真值，但也可能产生图 10 底行所示的穹顶效应等退化解。这通常源于图像集合的全局约束较弱，*e.g.*，近乎平行的观察

方向、环路闭合不足、三角化夹角较小，或相机自标定不准，特别是互联网视频中常见的径向畸变误差。此时束调整仍可获得较低重投影误差，却会把恢复的相机和场景几何弯曲成全局不一致的曲面。这类重建局部看似合理，但大尺度几何错误，不适合作为监督信号。可采用章节 3.5 所述有监督几何筛选，或将标注深度与现有单目深度模型、前馈式重建模型的预测比较，以滤除这类样本。

融入墙体的人体。 在有行人穿行的近静态街景视频中，我们反复观察到一种失败模式：几乎所有现有前馈式重建模型都会把人体吸收到静态背景中，将其估计为邻近墙体或建筑的一部分。这似乎与现有训练数据集中含糊的边界像素有关。例如，广泛使用的多视图数据集 MegaDepth 通过 COLMAP 标注旅游摄影图像，而这些图像常含有人物。在人像边界处，图块匹配立体算法可能把部分像素分配给周围静态建筑，使监督信号把人体局部视为背景。排除 MegaDepth 或重新估计其深度可显著减少此伪影。

数据歧义。 不同数据集间的标注不一致会导致令人困惑的模型行为。例如，在 Aria Synthetic Environments 和 HyperSim 等合成数据集中，室外场景常被渲染成贴于窗户上的二维纹理，因此真实深度对应窗面本身；而在其他数据集中，深度值正确表示透过窗户可见的真实物体。训练语料中这种固有的真实值不匹配会混淆模型，偶尔产生异常预测行为。

近期多个数据集提出了互联网视频标注的替代流程，包括 SceneScribe-1M [190]、PointWorld [70]、Sekai [98]、SpatialVID [182] 和 OmniWorld [233]，也值得进一步探索。

C. 局限性

失败案例。 尽管 VGGT- Ω 总体表现优异，但在特定场景中仍面临困难。例如，强运动模糊会显著降低性能；视角突然变化（如数秒内从 10° 变为 160° ）或相机畸变严重时，重建质量也常会下降。此外，模型在训练早期接触过一些含噪数据（*e.g.*，ScanNet++），因此面对包含大量显示器的办公室等场景时，预测有时不稳定。这些局限主要源于训练数据分布，未来将引入更多样、更具挑战性的序列加以缓解。

被掩蔽的敏感内容。 受隐私和许可限制，训练数据中的人脸、商标等部分内容被掩蔽或模糊处理。因此，模型偶尔会在这些区域产生异常伪影或不稳定预测。例如，身穿黑色衣物的人体深度预测有时不够平滑。这些伪影可能导致视觉畸变或观感欠佳的输出。