STREAM3R：可扩展顺序 3D 重建与因果变换器

Yushi Lan[1*], **Yihang Luo**[1*], **Fangzhou Hong**[1], **Shangchen Zhou**[1],
Honghua Chen[1], **Zhaoyang Lyu**[2], **Shuai Yang**[3], **Bo Dai**[4], **Chen Change Loy**[1], **Xingang Pan**[1]
[1]S-Lab, Nanyang Technological University, Singapore
[2]Shanghai Artificial Intelligence Laboratory [3]WICT, Peking University [4]The University of Hong Kong
https://nirvanalan.github.io/projects/stream3r

从图像重建详细的三维几何是计算机视觉中的关键所在，并且是诸如自动驾驶、虚拟现实、机器人等一系列后续应用的前提条件。虽然传统的视觉几何方法如 SfM 和多视图立体通过手工设计解决了一系列子问题来应对这一挑战，但由 DUSt3R 引领的最新趋势展示了一种通过强大的 transformers 直接回归点云的有前途的新方法。这种模式以及其后续工作，包括 MASt3R、Fast3R 和 VGG-T，使得可以从多个输入图像（从单个到数百个）中重建三维几何，提供了一个更统一的三维重建解决方案。

虽然这些工作专注于处理固定图像集，但现实世界中的应用常常需要持续处理视频流输入并即时更新重建 [14]，例如当一个自主代理探索新的环境或处理一个长的视频序列时。处理视频流输入会带来显著的新挑战。例如，每次新图像到达时，天真地运行 Fast3R 或 VGG-T 会导致大量冗余计算，因为它们必须从头开始重建，而不能继承先前的结果。这些方法还因昂贵的全注意力操作而难以处理长视频。Spann3R [15] 扩展了 DUSt3R，并通过一种记忆设计 [16] 支持增量重建，但它仍然受到显著累积漂移的影响，并且在动态场景中失败。最相关的并行工作是 CUT3R [17]，它提出了一种 RNN 范式 [18] 来处理非结构化或视频流输入。然而，基于 RNN 的设计在与现代网络架构 [19] 扩展时表现不佳，并且由于其有限的内存大小而难以处理长距离依赖。

鉴于任务的流式特性，在这项工作中，我们有兴趣探究使用具有单向因果注意力的 transformer 来实现在线增量式 3D 重建。在具有因果注意力的 LLM 风格 transformer 中，每一步的预测通过 KVCache 重用以前的计算，这在许多语言和音频任务中已被证明是成功的 [20, 21]。我们观察到，这一特性对于解决来自流数据的在线 3D 重建也是非常理想的，因为每一步都应该在之前的重建基础上进行，同时整合来自接收帧的新内容。

随着神经网络变得更加强大，最近的关注已转向从大规模数据集中采用学习到的几何先验 [36, 70, 76, 37**?**] 来解决上述挑战。在这一方向上，VGGSfM [43] 首次引入了一种可微的束调整（BA）[**? ? ?**] 框架，通过端到端结合机器学习和视觉几何。然而，仍然需要昂贵的迭代优化。最近，DUSt3R [10] 及其后续工作 MASt3R [11] 通过引入点图回归的新颖范式来直接应对 3D 任务。具体地说，它将 3D 重建表述为一个密集预测任务 [66]，其中模型输出特征图的每个像素对应一个像素对齐的 3D 点 [**?** 46]。尽管这些方法实现了前所未有的性能，但它们是为成对输入而设计的，并且需要二次复杂度的后处理来融合更多的输入图像。因此，当有大量图像可用时，这种行为就成为一个需要解决的瓶颈。这个问题在最近的一些工作中得到了部分解决。具体来说，Spann3R [15] 扩展了 DUSt3R，并通过一种记忆设计 [16] 支持基于视频的增量重建。然而，它仍然导致显著的累积漂移，并且在动态场景中失败。Fast3R [12] 和 VGG-T [13] 通过用一堆自注意力块来替换 DUSt3R 的不对称交叉注意力解码器，以引入更多的输入视图。然而，昂贵的全注意力操作 [23] 仍然限制了输入视图的数量，并且无法处理流式输入 [55]，从而限制了其在实际应用中的可扩展性。最相关的同期工作是 CUT3R [17]，其提出了一种 RNN 范式 [18] 来处理非结构化或流式输入。然而，由于内存大小的限制，基于 RNN 的设计对于现代网络架构 [19] 是不可扩展的，并且在处理长程依赖时表现不佳。受此启发，我们提出了 STREAM3R，一个全面的框架，可以从非结构化或流式输入图像进行 3D 重建，并预测世界和局部坐标中的对应点图 [12]。不同于同时进行的工作 [12, 13]，这些工作通过用双向注意块替换 DUSt3R 的不对称解码器来解决此问题 [22, 23]，STREAM3R 遵循现代仅解码器 [24] 变换器的设计，其中输入帧被顺序

[*] Equal contribution.

Preprint. Under review.

处理并注册为因果注意 [25] 。通过这种方式，STREAM3R 自然兼容现代大语言模型（LLM）[20] 的训练和推理技术，如窗口注意 [26] 和 KVCache [24] ，i.e. ，已处理的观测令牌将被保存作为注册输入帧的参考。

我们对大量的 3D 数据进行端到端的训练，并在一系列下游应用中对所提出的方法进行基准测试。总结而言，我们的主要贡献如下：

1. 我们提出了 STREAM3R ，这是一种仅解码器的变换器框架，将密集的 3D 重建重新表述为带有因果注意力的顺序配准任务，从而实现对非结构化和流式输入的可扩展性。
2. STREAM3R 本质上与现代 LLM 风格的训练和推理技术兼容，允许在各帧之间实现高效和可扩展的上下文累积。
3. 我们的架构支持世界坐标和局部坐标的点图预测，并且通过基于 splatt 的渲染自然地推广到大规模的新视图合成场景。
4. 我们在各种 3D 数据上对模型进行端到端训练，并在标准基准测试中展示了具有竞争力或优越的性能，具有很强的泛化能力和快速的推理速度。

早期的 3D 重建管道，如 Structure-from-Motion (SfM) 和 SLAM，通过几何推理从图像集合中估计稀疏的几何结构和相机位置。更近期的方法如 NeRF 和 Gaussian Splatting 将重点转向使用连续体积表示进行高保真度的新视图合成。然而，这些方法通常针对每个场景进行训练，没有学习到先验知识，导致收敛速度慢以及对稀疏或被遮挡的输入泛化能力差——这一限制有时被称为白板假设。相比之下，我们采用数据驱动的方法，从大规模 3D 数据集中学习几何先验，以实现从非结构化或流式输入中进行快速且可泛化的重建。

近期的研究利用大规模数据来学习深度估计、位姿加深度估计和光束调整的先验知识。尽管这些方法提高了泛化能力，大多数主要集中在单目深度或双视图设置，限制了在未知内参情况下重建完整几何体的能力。VGGSfM 通过将神经特征匹配与经典优化相结合，引入了可微分光束调整，但仍然是迭代的，计算量大，阻碍了可扩展性。在多视图立体领域，类似 MVSNeRF 和 MVSNet 的方法将神经网络集成到 MVS 管道中，但通常需要已知的相机位姿，并且仍然很大程度上依赖于手工设计的组件来有效整合三维几何信息。

近年来，基于点图的表示形式 [10, 11, 46, 47, 48, 49, 50] 已经成为密集的三维几何预测的统一格式，与神经网络的输出结构很好地对齐。与体素 [51] 、网格 [52] 或隐式场 [53, 31] 相比，点图能够进行前馈推断和实时渲染，还可以直接支持基于光栅化渲染 [34] 、SLAM [54, 55] 和少样本合成 [56] 等应用。DUSt3R [10] 和类似的后续工作如 MASt3R [11] 将立体 3D 重建重新表述为密集点图回归，从图像对中联合估计深度、姿态和内参。然而，它们的成对设计从根本上限制了可扩展性——在处理多视图场景时，需要二次融合操作和复杂的全局对齐过程。我们的方法在克服这些可扩展性限制的同时，还保持了点图表示的优点。

从单目视频重建动态场景的密集几何结构是一个重要但对传统方法具有挑战性的任务。最近的方法利用深度先验来解决这一挑战。具体来说，Robust-CVD 和 MegaSAM 需要对每个视频进行耗时的优化。MonST3R 基于 DUSt3R，通过在动态数据集上微调 DUSt3R 为动态场景输出点图。然而，它仍然需要基于滑动窗口的每视频全局对齐作为后处理。相比之下，我们的方法直接从单目视频实现前馈的四维重建，支持在线预测，而无需为每个视频进行昂贵的优化或后处理对齐。

Reconstruction Methods from Streaming Inputs. 流式方法为单目 SLAM 流水线 [14, 55, 60] 所代表的 3D 重建问题提供了一种更具可扩展性的替代解决方案。受现有基于学习的在线 3D 重建方法 [61, 62, 63] 的启发，最近 Spann3R [15] 为 DUSt3R 引入了基于内存的扩展，而 Fast3R [12] 和 VGG-T [13] 则用基于 Transformer 的注意力堆栈替换了非对称解码器，以直接实现多视角融合。尽管有这些进展，这些方法仍然主要依赖于全局全注意力机制，这限制了其随着序列长度增加的实时可扩展性。CUT3R [17] 采用 RNN 风格的架构来增量地处理非结构化输入，但受限于内存容量有限和与现代硬件加速技术的兼容性差 [19] 。我们的方法从根本上重新构想了点图预测为仅解码器 Transformer 任务，利用 KVCache 和窗口化注意力等技术实现高效的因果推理 [26, 24] 。这种架构设计使我们能够有效地扩展到长序列，同时保持与现代 LLM 风格的训练基础设施和优化技术的完全兼容，克服了以前方法的局限性。

1 预备知识：DUSt3R

我们对 DUSt3R [10] 进行重新表述，以接受图像流作为输入。在 DUSt3R 中，每个输入的图像 $\boldsymbol{I}_t$ 最初被分割成一组 K 标记，$\boldsymbol{F}_t = \text{Encoder}(\boldsymbol{I}_t)$ ，其中 $\boldsymbol{F}_t \in \mathbb{R}^{K \times C}$ 和 Encoder 是一个权

重共享的 ViT [64] 。具体来说，DUSt3R 被设计为一次摄入两个输入图像，i.e. ， $t \in \{1, 2\}$ 。编码后的图像产生两组标记：

$$\boldsymbol{F}_1 = \text{Encoder}(\boldsymbol{I}_1), \quad \boldsymbol{F}_2 = \text{Encoder}(\boldsymbol{I}_2). \tag{1}$$

。之后，解码器网络 Decoder_t 通过一系列带有交叉注意力层的变换块对它们进行推理：

$$G_1^i = \text{DecoderBlock}_1^i(G_1^{i-1}, G_2^{i-1}), \quad G_2^i = \text{DecoderBlock}_2^i(G_2^{i-1}, G_1^{i-1}), \tag{2}$$

，其中 i 从 1 到 B ，代表 B 块总数中解码器的块索引。$G_1^0 \coloneqq \boldsymbol{F}_1$ 和 $G_2^0 \coloneqq \boldsymbol{F}_2$ 。最后，每个分支的对应回归头预测一个带有相关置信图的点图：

$$\hat{\boldsymbol{X}}_{1,1}, \hat{\boldsymbol{C}}_{1,1} = \text{Head}_1(G_1^0, \ldots, G_1^B), \quad \hat{\boldsymbol{X}}_{2,1}, \hat{\boldsymbol{C}}_{2,1} = \text{Head}_2(G_2^0, \ldots, G_2^B). \tag{3}$$

。请注意，DUSt3R 是为两个视图输入设计的，并且需要一个昂贵且不可扩展的全局对齐过程以加入更多输入视图。

2 方法

我们介绍了 STREAM3R ，这是一种转换器，可以将未经校准的流图像作为输入，并生成一系列 3D 属性作为输出。输入可以是非结构化图像集合或视频。不像现有方法 [13, 12] 需要在整个输入序列上采用高昂的双向注意力或者使用固定大小的内存缓存 [15, 17] ， STREAM3R 则缓存过去帧的特征作为上下文，并通过对累计观测进行因果注意力来顺序处理进入的帧。该设计不仅加快了训练速度和收敛速度，还与现代 LLM 的架构原则一致，使我们能够利用该领域的进步。我们首先在第 2.1 节介绍问题的表述，第 **??** 节介绍架构，第 2.2 节介绍训练目标，第 3 节介绍实现细节。所提出方法的概述见图 **??** 。

2.1 问题定义和符号表示

STREAM3R 是一个回归模型，它顺序地接收一串 N RGB 图像 $(\boldsymbol{I})_t^N$ ，其中每个图像 $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$ 属于同一个三维场景。流输入依次被转换为对应每一帧的一组三维注释：技术上，STREAM3R 被实现为一个因果变换器，将每个图像 $\boldsymbol{I}_t$ 映射到其对应的局部坐标的点图 $\hat{\boldsymbol{X}}_t^{\text{local}} \in \mathbb{R}^{3\times H\times W}$ 及其在全局坐标 $\hat{\boldsymbol{X}}_t^{\text{global}} \in \mathbb{R}^{3\times H\times W}$ 中的点图，该坐标由第一个输入帧 $\boldsymbol{I}_0$ 指定，以及其相对相机位姿 $\hat{\boldsymbol{P}}_t \in \mathbb{R}^9$ ，包括内参和外参。之后我们将说明这些三维属性是如何预测的。

正如在第 1 节中提到的，给定流媒体输入，对于每一个当前图像 $\boldsymbol{I}_t$ ，我们的方法首先将其划分为特征 $\boldsymbol{F}_t = \text{Encoder}(\boldsymbol{I}_t)$ 。主要区别在于解码器的一侧：我们并不在整个序列上执行双向注意 [12] 或像在 RNN 中那样与一个可学习状态交互 [17] ，而是从 LLMs [20, 24, 65] 中汲取灵感，与先前的观察进行高效因果注意。具体来说，在每个解码器块中执行逐帧自注意后，当前特征 G_t^{i-1} 将与先前观测到的对应于同一层的帧的特征进行交叉注意：

$$G_t^i = \text{DecoderBlock}^i\left(G_t^{i-1},\ G_0^{i-1} \oplus G_1^{i-1} \oplus \cdots \oplus G_{t-1}^{i-1}\right). \tag{4}$$

这种交互确保了高效的信息传递以处理长上下文依赖关系。请注意，这个操作很容易实现，并在推理期间经过 KV 缓存的优化，以实现高效计算 [24, 20] 。为了实现这一目标，需要进行几个网络架构的修改。在 DUSt3R 中，解码器遵循对称设计，两个独立的解码器 Decoder_1 、Decoder_2 用于处理两个输入视图。为了扩展到任意数量的输入，我们移除对称设计，仅保留一个解码器 Decoder 来处理所有输入帧。具体而言，解码器中的每个块包含一个 SelfAttn 块用于逐帧注意力，以及一个 CrossAttn 块用于因果地关注所有先前观测的特征。请注意，由于缺乏历史背景，我们处理前两个帧遵循 DUSt3R 的惯例。其后的所有传入帧都遵循方程 (4) 中的因果操作。请注意，为了表明经典世界空间，我们在如图 **??** 所示，以元素方式向第一个帧 $\boldsymbol{F}_1 = \boldsymbol{F}_1 + [\text{reg}]$ 的 tokens 添加一个可学习的寄存器 token [reg] 。通过这种方式，模型能够在不引入 N 独立解码器的情况下学习输出全局点。与现有工作 [12] 不同的是，为了简化，我们没有对其他帧施加位置嵌入。

在解码操作之后，可以相应地预测每帧对应的 3D 属性。基于现有工作，我们预测两组点图 $\hat{\boldsymbol{X}}_t^{\text{local}}, \hat{\boldsymbol{X}}_t^{\text{global}}$ 及其对应的置信度图 $\hat{\boldsymbol{C}}_t^{\text{local}}, \hat{\boldsymbol{C}}_t^{\text{global}}$ 。具体而言，局部点图 $\hat{\boldsymbol{X}}_t^{\text{local}}$ 是在查看相机的坐标系中定义，而全局点图 $\hat{\boldsymbol{X}}_t^{\text{global}}$ 是在第一幅图像 $\boldsymbol{I}_1$ 的坐标系中定义。我们使用两个 DPT 头来进行点图预测：这个冗余的预测已经被证明能够简化训练，并促进在具有部分标注的 3D 数据集和单视图深度数据集上的训练。

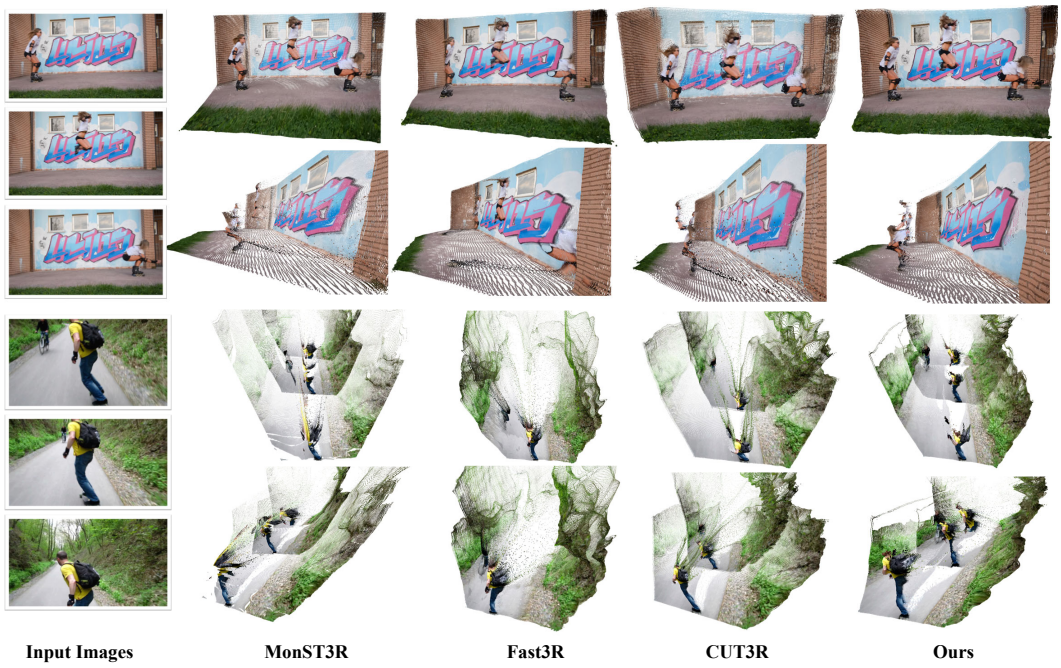


Figure 1: 对自然图片的定性结果。我们将我们的方法与 MonST3R、Fast3R 和 CUT3R 方法进行比较，结果表明，本文方法在视觉质量上更胜一筹。

2.2 训练目标

STREAM3R 使用 DUSt3R 中引入的点图损失的一般形式进行训练。给定一系列来自视频或图像集合的随机采样图像 N，我们训练模型以生成由 $\mathcal{X} = \{\hat{\mathcal{X}}^{\text{local}}, \hat{\mathcal{X}}^{\text{global}}\}$ 表示的点图预测，其中 $\hat{\mathcal{X}}^{\text{local}} = \{\hat{\boldsymbol{X}}_t^{\text{local}}\}_{t=1}^N$ 和 $\hat{\mathcal{X}}^{\text{global}} = \{\hat{\boldsymbol{X}}_t^{\text{global}}\}_{t=1}^N$。对应的置信度得分被表示为 $\hat{\mathcal{C}}$。根据 DUSt3R [10]，我们对点图应用置信度感知回归损失：

$$\mathcal{L}_{\text{conf}} = \sum_{(\hat{\boldsymbol{x}},\hat{c})\in(\hat{\mathcal{X}},\hat{\mathcal{C}})} \left(\hat{c} \cdot \left\| \frac{\hat{\boldsymbol{x}}}{\hat{s}} - \frac{\boldsymbol{x}}{s} \right\|_2 - \alpha \log \hat{c} \right), \tag{5}$$

，其中 $\hat{s}$ 和 s 是 $\hat{\mathcal{X}}$ 和 $\mathcal{X}$ 的尺度归一化因子，以实现尺度不变监督 [69]。我们还设置 $\hat{s} := s$ 用于度量尺度数据集，如 MASt3R [11]，以实现度量尺度点图预测。对于相机预测损失，我们将姿态 $\hat{\boldsymbol{P}}_t$ 参数化为四元数 $\hat{\boldsymbol{q}}_t$、平移 $\hat{\boldsymbol{\tau}}_t$ 和焦距 $\hat{\mathrm{f}}_t$，并最小化预测和真实值之间的 L2 范数：

$$\mathcal{L}_{\text{pose}} = \sum_{t=1}^{N} \left(\|\hat{\boldsymbol{q}}_t - \boldsymbol{q}_t\|_2 + \left\| \frac{\hat{\boldsymbol{\tau}}_t}{\hat{s}} - \frac{\boldsymbol{\tau}_t}{s} \right\|_2 + \left\| \hat{\boldsymbol{f}}_t - \boldsymbol{f}_t \right\|_2 \right). \tag{6}$$

。

3 实验

我们的方法在一个大型且多样化的 3D 数据集上进行训练，包括 e.g.、Co3Dv2 [37]、ScanNet++ [70]、ScanNet [71]、HyperSim [72]、Dynamic Replica [73]、DL3DV [36]、BlendedMVS [74]、Aria Synthetic Environments [75]、TartanAir [76]、MapFree [77]、MegaDepth [78] 和 ARKitScenes [79]。完整版的数据集详情请查阅补充材料。

我们提供两个版本的 STREAM3R，其中 STREAM3R $^{\alpha}$ 是受 DUSt3R [10] 预训练权重启发并进行微调的，STREAM3R $^{\beta}$ 则从旗舰 VGG-T [13] 模型初始化。对于 STREAM3R $^{\alpha}$，我们继承了 24 层的 CroCo ViT [80, 81] 作为编码器，并通过仅保留首个解码器 $\text{Decoder} = \text{Decoder}_1$，对其 12 层的解码网络进行改造。DPT-L [66] 头被用来将解码的 tokens 分别映射为局部和全局点图。对于 STREAM3R $^{\beta}$，我们用 CausalAttn 替换 VGG-T 全局注意力中的 SelfAttn 层，并对所有参数进行微调。为了实现高效的内存使用和稳定的训练，我们在每个 transformer 层注入了 QK-Norm [82]，并利用 FlashAttention [19] 进行 BFloat16 混合精度训练。

Table 1: 单帧深度评估。我们报告了在 Sintel、Bonn、KITTI 和 NYU-v2（静态）数据集上的性能。在每个类别中，最佳和第二佳结果分别以粗体和 underlined 显示。我们的方法在与现有方法的比较中，取得了更好或相当的性能。

Method	Sintel		Bonn		KITTI		NYU-v2	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
VGG-T	**0.271**	**67.7**	0.053	97.3	**0.076**	93.3	**0.060**	**94.8**
Fast3R	0.502	52.8	0.192	77.3	0.129	81.2	0.099	88.9
DUSt3R	0.424	58.7	0.141	82.5	0.112	86.3	0.080	90.7
MASt3R	0.340	60.4	0.142	82.0	0.079	**94.7**	0.129	84.9
MonST3R	0.358	54.8	0.076	93.9	0.100	89.3	0.102	88.0
Spann3R	0.470	53.9	0.118	85.9	0.128	84.6	0.122	84.9
CUT3R	0.428	55.4	0.063	96.2	0.092	91.3	0.086	90.9
STREAM3R $_{\alpha}$	0.350	59.0	0.075	93.4	0.088	91.3	0.091	89.9
STREAM3R $_{\beta}$	0.228	70.7	0.061	**96.7**	0.063	95.5	**0.057**	**95.7**

我们的模型使用 AdamW 优化器在批量大小为 64 的情况下进行训练，学习率为 1e-4，共进行 $400K$ 次迭代。对于每个批次，我们从一个随机的训练场景中随机采样 $4-10$ 帧。输入帧被裁剪成不同的分辨率，从 224×224 到 512×384 以提高通用性。训练在七天内端到端运行在 8 个 NVIDIA A100 GPU 上。还采用了梯度检查点来优化内存使用。

Baselines. 我们将我们的方法与一组基线进行比较，这些基线旨在以一对视图作为输入：DUSt3R [10]，MASt3R [11]，和 MonST3R [49]。此外，我们包括了与最新方法的对比：Spann3R [15]，CUT3R [17]，SLAM3R [55]，和 Fast3R [12]，这些方法专门设计用于处理数量变化的输入图像。我们还包括了旗舰 3D 几何模型 VGG-T [13] 以供参考。需要注意的是 Fast3R 和 VGG-T 是双向注意力方法，我们将它们与需要全局优化（GA）的方法归为一组。我们将其他最新的方法归为流式方法，这些方法支持处理顺序输入。需要注意的是，除 VGG-T 和 STREAM3R $^{\beta}$ 外的所有方法，我们都用 512 的最大维度进行推理。对基于 VGG-T 的方法，由于 DINO-V2 tokenizer [83] 的要求，我们用 518 的最大维度进行推理。关于 FPS，我们在 A100 GPU 上以 FP32 进行推理速度的基准测试。更多同时方法的对比 [84, 38] 包括在补充材料中。

3.1 单目和视频深度估计

继先前的方法后，我们首先在 Sintel、Bonn、KITTI 和 NYU-v2 数据集上评估单目深度估计，这些数据集涵盖动态和静态、室内和室外、真实和合成数据。这些数据集未用于训练，适合跨领域零样本性能的基准测试。我们的评估包括绝对相对误差（Abs Rel）和在真实深度 1.25 倍范围内的内点百分比，遵循现有方法的惯例。每帧中位数缩放按照 DUSt3R 进行。我们在表中包含了定量结果。正如可以看到的那样，与流式方法相比，我们的方法达到最新水平，甚至在 Sintel、KITTI 和 NYU-2 上与 VGG-T 相比表现最佳。还请注意，与 CUT3R 相比，我们的方法使用了更少的数据集和计算资源。具体来说，CUT3R 采用四个阶段的课程训练，共进行 $100+35+40+10=185$ 个周期，而由于计算资源的限制，我们的方法仅在部分使用 CUT3R 数据集的情况下进行端到端训练，仅需 7 个周期。

我们还在视频深度任务上对我们的模型进行了基准测试，该任务通过将输出深度图与使用给定每序列尺度的真实深度图对齐来评估每帧深度质量和帧间深度一致性。像 MASt3R、CUT3R 和我们的方法等度量点图方法也在不对齐的情况下报告。两种方法的定量结果包含在表 5 中。在每序列尺度对齐的情况下，我们的方法超越了基于优化的基线 DUSt3R-GA [10] 和 MASt3R-GA [11]（静态场景假设），甚至 MonST3R-GA [49]（动态场景，依赖光流 [89]）。与流媒体前沿方法 CUT3R 相比，我们在所有三个基准测试上都实现了更高的准确性，同时运行速度快 40%。STREAM3R 还优于全注意力模型 Fast3R [12]、流媒体方法 Spann3R [15] 和 Sintel 上的旗舰模型 VGG-T。值得注意的是，使用滑动窗口注意力 [26] 保持恒定缓存的 STREAM3R $^{\beta}$ -W，尽管仅访问了五帧过去帧，却在 Bonn 和 KITTI 上超过了 STREAM3R $^{\beta}$。

3.2 3D 重建

我们还在 7-scenes [90] 数据集上对场景级别的 3D 重建进行基准测试，并按照现有方法 [15, 17, 10] 的惯例，使用准确性 (Acc)、完整性 (Comp) 和法线一致性 (NC) 指标。按照 CUT3R 的方法，我们通过评估稀疏采样的图像来评估模型在图像集合上表现的性能，这些图像之间的重叠最少或没有，每个场景有 i.e.，3 到 5 帧。定量结果包含在表中 3。与基于

Table 2: 视频深度评估。我们在 Sintel、Bonn 和 KITTI 数据集上评估尺度不变和度量深度精度。需要全局对齐的方法标记为“GA”。“类型”列指示该方法是基于优化的（“Optim”）、流式处理（“Stream”）还是全注意力（“FA”）。我们还报告在 KITTI 数据集上使用 512 × 144 分辨率在 A100 GPU 上所有方法的推理速度 FPS，除了 Spann3R ，它支持 Stream 224 × 224 输入。我们的方法实现了比 CUT3R 更好的性能，同时提供更快的推理。对于 STREAM3R $^{\beta}$ -W[5]，我们表示在 STREAM3R $^{\beta}$ 上使用滑动窗口注意力，窗口大小为 5。注意 STREAM3R $^{\beta}$ -W[5] 在所有基于流的方式中实现了最快的 FPS。

Alignment	Method	Type	Sintel		Bonn		KITTI		FPS
			Abs Rel ↓	δ <1.25 ↑	Abs Rel ↓	δ <1.25 ↑	Abs Rel ↓	δ <1.25 ↑	
Per-sequence scale	VGG-T [13]	FA	**0.297**	**68.8**	0.055	97.1	0.073	96.5	7.32
	Fast3R [12]	FA	0.653	44.9	0.193	77.5	0.140	83.4	**47.23**
	DUSt3R-GA [10]	Optim	0.656	45.2	0.155	83.3	0.144	81.3	0.76
	MASt3R-GA [11]	Optim	0.641	43.9	0.252	70.1	0.183	74.5	0.31
	MonST3R-GA [49]	Optim	0.378	55.8	0.067	96.3	0.168	74.4	0.35
	Spann3R [15]	Stream	0.622	42.6	0.144	81.3	0.198	73.7	13.55
	CUT3R [17]	Stream	0.421	47.9	0.078	93.7	0.118	88.1	16.58
	STREAM3R $^{\alpha}$	Stream	0.478	51.1	0.075	94.1	0.116	89.6	23.48
	STREAM3R $^{\beta}$	Stream	0.264	70.5	0.069	95.2	**0.080**	94.7	12.95
	STREAM3R $^{\beta}$ -W[5]	Stream	0.279	68.6	**0.064**	**96.7**	0.083	**95.2**	32.93
Metric scale	MASt3R-GA [11]	Optim	1.022	14.3	0.272	70.6	0.467	15.2	0.31
	CUT3R	Stream	1.029	23.8	0.103	88.5	0.122	85.5	16.58
	STREAM3R $^{\alpha}$	Stream	1.041	21.0	0.084	94.4	0.234	57.6	23.48

Table 3: 7-Scenes 上的 3D 重建评估 [90] 。尽管在流式传输设置中操作，我们的方法提供了具有竞争力的性能，达到甚至超过基于离线优化并利用全局对齐的方法。STREAM3R $^{\beta}$ -FA 表示在我们训练的模型中采用全注意力进行 3D 重建。

Method	Type	Acc ↓		Comp ↓		NC ↑		FPS
		Mean	Med.	Mean	Med.	Mean	Med.	
VGG-T [13]	FA	0.087	0.039	0.091	0.039	0.787	0.890	12.0
Fast3R [12]	FA	0.164	0.108	0.163	0.080	0.686	0.775	**30.92**
DUSt3R-GA [10]	Optim	0.146	0.077	0.181	0.067	0.736	0.839	0.68
MASt3R-GA [11]	Optim	0.185	0.081	0.180	0.069	0.701	0.792	0.34
MonST3R-GA [49]	Optim	0.248	0.185	0.266	0.167	0.672	0.759	0.39
STREAM3R $^{\beta}$ -FA	Stream	0.091	0.043	0.075	0.042	0.769	0.879	12.0
Spann3R [15]	Stream	0.298	0.226	0.205	0.112	0.650	0.730	12.97
SLAM3R [55]	Stream	0.287	0.155	0.226	0.066	0.644	0.720	**38.40**
CUT3R [17]	Stream	0.126	0.047	0.154	0.031	0.727	0.834	17.00
STREAM3R $^{\alpha}$	Stream	0.148	0.077	0.177	0.058	0.700	0.801	26.4
STREAM3R $^{\beta}$	Stream	**0.122**	**0.044**	0.110	**0.038**	**0.746**	**0.856**	20.12

优化的方法和包括 Spann3R、Fast3R、CUT3R 和 SLAM3R 在内的强大基线相比，我们的方法实现了更好的性能。与 CUT3R 相比，我们的方法在推理过程中速度快了超过 50% 倍，同时表现更佳。尽管 SLAM3R 在推理速度上最快，但其重建准确性明显低于我们的方法。这种性能差距可以部分归因于 SLAM3R 是在较低的输入分辨率 224 × 224 下进行训练和评估的。我们还包括了 STREAM3R $^{\beta}$ -FA 进行比较，这表示将 STREAM3R $^{\beta}$ 中的因果注意力替换为完全注意力（FA）。有趣的是，STREAM3R $^{\beta}$ -FA 表现出与 VGG-T 可比的性能，甚至在完整性指标上取得更好的结果。这突出了我们提出的方法的有效性和通用性。NRGBD [91] 基准上的比较结果包含在补充材料中。

3.3 相机姿态估计

根据 CUT3R [17] ，我们评估在 Sintel [85] 、TUM-dynamics [92] 和 ScanNet [71] 数据集上的相机位姿估计精度。Sintel 和 TUM-dynamics 都具有大量的动态运动，对传统的 SfM 和 SLAM 流程构成了重大挑战。我们在与真实值进行 Sim(3) 对齐后，报告绝对平移误差（ATE）、相对平移误差（RPE $_{\text{trans}}$ ）和相对旋转误差（RPE $_{\text{rot}}$ ），遵循 [30, 49, 17] 中的协议。我们的方法可以在不需要相机校准的情况下运行，类似于比较的基线方法 [30] 。虽然许多先前的方法 [59, 94] 通过测试时优化来解决这个问题——共同估计每个序列的内参和密集深度——我们专注于纯在线处理。表 4 报告了在线（Onl）和优化（Optim）类别的结果，后者包括 DUSt3R [10] （将所有帧对齐到第一个帧而不进行全局对齐）。尽管基于优化

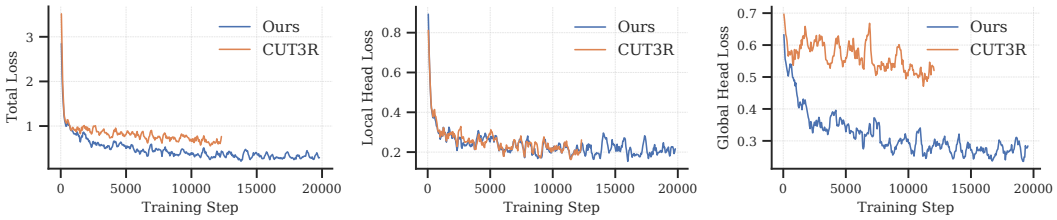
Table 4: 在 Sintel [85]、TUM-dynamic [92] 和 ScanNet [71] 数据集上的相机姿态评估。我们的方法在大多数基准测试上与 CUT3R 实现了相当的性能。

Method	Type	Sintel			TUM-dynamics			ScanNet		
		ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓
Particle-SfM [93]	Optim	0.129	0.031	**0.535**	-	-	-	0.136	0.023	0.836
Robust-CVD [59]	Optim	0.360	0.154	3.443	0.153	0.026	3.528	0.227	0.064	7.374
CasualSAM [94]	Optim	0.141	0.035	0.615	0.071	0.010	1.712	0.158	0.034	1.618
DUSt3R-GA [10]	Optim	0.417	0.250	5.796	0.083	0.017	3.567	0.081	0.028	0.784
MASt3R-GA [11]	Optim	0.185	0.060	1.496	**0.038**	0.012	**0.448**	0.078	0.020	**0.475**
MonST3R-GA [49]	Optim	**0.111**	0.044	0.869	0.098	0.019	0.935	**0.077**	**0.018**	0.529
DUSt3R [10]	Onl	0.290	0.132	7.869	0.140	0.106	3.286	0.246	0.108	8.210
Spann3R [15]	Onl	0.329	0.110	4.471	0.056	0.021	0.591	0.096	0.023	0.661
CUT3R [17]	Onl	**0.213**	**0.066**	**0.621**	0.046	0.015	0.473	0.099	0.022	**0.600**
STREAM3R $^{\beta}$	Onl	**0.213**	0.076	0.868	**0.026**	**0.013**	**0.330**	**0.052**	**0.021**	0.850

Table 5: 关于视频深度估计的消融实验。当评估以相同迭代次数训练的检查点时，我们提出的架构在视频深度估计任务上始终表现优于基于 RNN 的 CUT3R。

Method	Sintel		BONN		KITTI	
	Abs Rel ↓	δ <1.25 ↑	Abs Rel ↓	δ <1.25 ↑	Abs Rel ↓	δ <1.25 ↑
CUT3R	0.598	40.7	0.102	90.7	0.157	77.4
STREAM3R $_{\alpha}$	0.535	47.0	0.083	94.2	0.141	81.8

的系统总体上仍能实现最低的误差，我们的方法在流处理方法中建立了最强性能，并显著超越了 CUT3R [17] 在 TUM-dynamics 和 ScanNet 上的表现，展示了在动态环境中特别的鲁棒性。在此，我们对 STREAM3R 进行详细的消融分析，以展示其设计的有效性。由于训练模型所需的计算资源庞大，我们仅在 224 × 224 分辨率图像上训练消融模型。所有数据集都被用于训练模型。请注意，为了公正的比较，我们使用预训练的 MASt3R [11] 检查点初始化所有模型，并使用相同的超参数和计算资源来训练模型。我们展示了仅解码器结构的



(a) Overall Training Curve (b) Training Curve of Local Branch (c) Training Curve of Global Branch

Figure 2: 我们提出的 STREAM3R 的消融研究。与基于 RNN 的架构 [17] 相比，我们的仅解码器网络在三维点图预测任务中，尤其是在全局分支中，具有更好的收敛性和更快的训练速度。

transformer 在时序 3D 点图预测中相对于 RNN 设计的有效性。主要的基准是 CUT3R [17]，它利用 RNN 设计来实现这一点。为了公平比较，我们使用相同的数据集和预训练模型权重初始化来重新训练 CUT3R 和我们的方法。训练曲线包含在图 2a 中，其中两个模型使用相同的超参数和计算资源进行训练。可以观察到，STREAM3R 比 CUT3R 收敛更快，并在给定时间内执行了 60% 更多的训练步骤。这可能听起来违反直觉，因为 STREAM3R 针对的是比 CUT3R 的常量状态内存更长的上下文。然而，由于 CUT3R 架构在每次状态读取交互后都需要状态更新操作，而 STREAM3R 直接关注于已有观测的缓存特征。我们还在图 2b 中注意到，两个架构中 $\text{Head}_{\text{local}}$ 的收敛情况相似，而对于 $\text{Head}_{\text{global}}$，我们提出的架构显示出明显更快的收敛速度，如图 2c 所示。这表明，由于有限的内存容量，使用单一状态使模型更难以注册输入帧。

在定量分析中，我们在视频深度估计（表 5）和 3D 重建（表 6）上对消融模型进行基准测试，分别评估 $\text{Head}_{\text{local}}$ 和 $\text{Head}_{\text{global}}$。为了公平比较，我们评估了经过相同次数迭代训练的检查点。可以看出，我们提出的架构在这两个任务上始终表现更好。

Table 6: 在 7-Scenes 上的 3D 重建消融实验。我们提出的架构在相同配置下训练时，在 3D 重建任务中始终比基于 RNN 的 CUT3R 取得更好的性能。请注意，我们的架构训练速度更快。

Method	Acc ↓		Comp ↓		NC ↑	
	Mean	Med.	Mean	Med.	Mean	Med.
CUT3R	0.480	0.365	0.330	0.148	0.555	0.583
STREAM3R $_{\alpha}$	**0.328**	**0.261**	**0.255**	**0.095**	**0.605**	**0.659**

我们的方法存在一些局限性。首先，朴素的自回归建模自然会遭遇误差累积和漂移。一些抗漂移采样策略可以被提出以缓解这个问题。其次，目前这仍然是一个具有确定性输出的回归模型。进一步将其扩展为一个自回归生成模型将进一步解锁一系列的下游应用。最后，由于该模型遵循现代大型语言模型的类似设计，可以引入更多的训练技术，如 MLA，以进一步提高训练效率和性能。

我们引入了一个仅解码器的变换器框架，用于从非结构化或流式图像输入中进行密集的 3D 重建。通过将重建重新表述为具有因果注意力的顺序配准任务，该框架克服了先前工作在可扩展性上的瓶颈，并自然地与大型语言模型风格的训练和推理流程相一致。我们的设计允许跨帧高效整合几何上下文，支持双坐标点图预测，并且能够在大规模场景上进行新视图合成而不需要全局后处理。通过在标准基准上的广泛实验，我们展示了该模型在单目/视频深度估计和 3D 重建任务中取得了竞争性或更优的性能，并显著提高了推理效率。通过将几何学习与可扩展序列建模相结合，我们希望这项工作为更通用的实时 3D 理解系统铺平道路。

我们的方法存在一些局限性。首先，XMATHXDQ 天真的因果建模自然会遭受误差积累和漂移 [95] 。可以提出一些推断策略来缓解这个问题。其次，目前 STREAM3R 仍然是一个具有确定性输出的回归模型。进一步扩展为自回归生成模型 [25, 95] 将进一步解锁一系列下游应用。最后，既然 STREAM3R 遵循现代 LLMs 的类似设计，可以引入更多的训练技术，如 MLA [65] ，以进一步提升训练效率和性能。

References

[1] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016.

[2] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021.

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

[5] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *CVPR*, 2023.

[6] Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, and Yinda Zhang. Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. In *ECCV*, 2024.

[7] Muhammad Zubair Irshad, Mauro Comi, Yen-Chen Lin, Nick Heppert, Abhinav Valada, Rares Ambrus, Zsolt Kira, and Jonathan Tremblay. Neural fields in robotics: A survey, 2024.

[8] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *ECCV*, 2018.

[9] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *CVPR*, 2019.

[10] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024.

[11] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

[12] 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 *CVPR*, June 2025.

[13] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025.

[14] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *TPAMI*, 29(6):1052–1067, 2007.

[15] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.

[16] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022.

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

[18] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization, 2015.

[19] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[21] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2023.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[23] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[24] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[25] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS*, 37:24081–24125, 2025.

[26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[27] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[28] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018.

[29] 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.

[30] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *NeurIPS*, pages 16558–16569, 2021.

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[32] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

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

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[35] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024.

[36] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024.

[37] 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 *ICCV*, 2021.

[38] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[39] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024.

[40] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025.

[41] 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. *arXiv preprint*, 2024.

[42] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. 2024.

[43] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, pages 21686–21697, 2024.

[44] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *CVPR*, pages 9043–9053, 2023.

[45] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024.

[46] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.

[47] Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. *arXiv preprint arXiv:2412.09573*, 2024.

[48] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *arXiv*, 2023.

[49] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024.

[50] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *ECCV*, 2024.

[51] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias NieBner, Gordon Wetzstein, and Michael Zollhofer. DeepVoxels: Learning persistent 3D feature embeddings. In *CVPR*, pages 2432–2441. IEEE.

[52] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. In *ICCV*, pages 9785–9795, 2019.

[53] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019.

[54] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. *arXiv preprint*, 2024.

[55] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. *arXiv preprint arXiv:2412.09401*, 2024.

[56] Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025.

[57] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024.

[58] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *NeurIPS*, 2024.

[59] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021.

[60] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *3DV*, March 2024.

[61] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-r2n2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, volume 9912 of *Lecture Notes in Computer Science*, pages 628–644. Springer, 2016.

[62] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. PixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021.

[63] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*, 2021.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[65] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

[66] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

[67] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction, 2025.

[68] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, pages 21013–21022, June 2022.

[69] 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, 2024.

[70] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, pages 12–22, 2023.

[71] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

[72] 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 *ICCV*, 2021.

[73] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023.

[74] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *CVPR*, 2020.

[75] 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 egocentric 3d machine perception. In *ICCV*, pages 20133–20143, October 2023.

[76] 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 *IROS*, 2020.

[77] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, 2022.

[78] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *ICCV*, pages 2041–2050, 2018.

[79] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data, 2022.

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

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

[82] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023.

[83] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

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

[85] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *ECCV*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[86] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. *arXiv*, 2019.

[87] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer-Verlag Berlin, October 2012.

[88] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[89] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.

[90] 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 *CVPR*, June 2013.

[91] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *CVPR*, pages 6290–6301, June 2022.

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

[93] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *ECCV*, 2022.

[94] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *ECCV*, pages 20–37. Springer, 2022.

[95] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025.

[96] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024.

[97] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, June 2020.

[98] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, June 2023.

[99] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019.

[100] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation, 2021.

[101] Anastasiia Kornilova, Marsel Faizullin, Konstantin Pakulev, Andrey Sadkov, Denis Kukushkin, Azat Akhmetyanov, Timur Akhtyamov, Hekmat Taherinejad, and Gonzalo Ferrer. Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis, 2022.

[102] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. *arXiv preprint arXiv:2406.09414*, 2024.

[103] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *ICCV*, pages 9466–9476, October 2023.

[104] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors, 2024.

[105] Ioan Andrei Bârsan, Peidong Liu, Marc Pollefeys, and Andreas Geiger. Robust dense mapping for large-scale dynamic environments. In *ICRA*, pages 7510–7517, 2018.

A 数据集详情

我们在包括静态和动态场景及物体的多种场景类型的 29 数据集上训练我们的模型。具体而言，我们主要遵循 CUT3R [17] 的数据分割，最高采样比的主要 15 数据集包括：Co3Dv2 [37] 、ScanNet++ [70] 、ScanNet [71] 、HyperSim [72] 、Dynamic Replica [73] 、DL3DV [36] 、BlendedMVS [74] 、Aria Synthetic Environments [75] 、TartanAir [76] 、MapFree [77] 、MegaDepth [78] 、WildRGBD [96] 、Waymo [97] 、Bedlam [98] 和 ARKitScenes [79] 。由于 3D Ken Burns [99] 、IRS [100] 和 SmartPortraits [101] 这些数据集要么是单视图，要么未能成功下载，因此我们未将它们用于训练。我们改编 CUT3R [17] 、DUSt3R [10] 和 Spann3R [15] 提供的官方脚本进行数据集处理。对于训练 STREAM3R $^{\beta}$ ，我们移除了 VGG-T 中的所有单视图数据集，剩下 19 数据集用于训练。移除单视图数据集后，我们没有发现性能下降。有关更多数据集详情，请参阅 CUT3R 的表 6。

B 更多实现细节

More Training Details. 我们的方法在所有数据集上进行端到端训练，使用混合的 12 种不同分辨率，范围从 224×224 到 512×384 。在数据增强方面，我们通过在序列中的所有帧应用相同的颜色抖动来执行序列级颜色抖动。

我们遵循 DUSt3R 并使用 CroCoNet 的预训练 ViT 进行编码器和解码器设计。我们直接使用 DPT 头进行 $\text{Head}_{\text{global}}$ 和 $\text{Head}_{\text{local}}$ 的实现。我们在每次注意力操作之前对 ViT 编码器的查询和键特征应用 RoPE，但为了对任意数量的输入视图进行泛化，我们在 ViT 解码器中忽略它。对于消融研究，我们在相同的数据集上训练我们的模型，但分辨率为 224×224 。

对于滑动窗口注意版本 STREAM3R $^{\beta}$ -W[5]，我们总是包含第一个帧的标记以保持规范坐标空间不变。我们将窗口大小设置为 $W = 5$ ，因为它在性能和速度之间达到了平衡，其他窗口大小也能稳定工作。对于全注意力版本 STREAM3R $^{\beta}$ -FA，我们直接使用因果训练的模型 STREAM3R $^{\beta}$ ，并在 SelfAttn 中去除了因果掩码。这与 CUT3R 中的“重访”操作相似。

我们在主论文中进一步扩展了视频深度比较，并包含了更广泛的基线方法，包括单帧深度方法（Marigold [39] 和 DepthAnything-V2 [102] ）、视频深度方法（NVDS [103] 、DepthCrafter [40] 和 ChronoDepth [104] ），以及最近的联合深度与姿势估计方法，如 Robust-CVD [105] 、CausalSAM [94] 、DUSt3R [10] 、MASt3R [11] 、MonST3R [49] 和 Spann3R [15] 。扩展结果显示在表 **??** 中。STREAM3R 在每序列比例 & 位移设置下，始终优于其基于 RNN 的对应方法 CUT3R，甚至在 KITTI 数据集上实现了最新的性能记录，同时在 FPS 方面也是最快的。

我们进一步在表 **??** 中列出了 NRGBD 基准 [91] 上的比较结果。在这里，我们还加入了与同时期研究工作 StreamVGGT [84] 的比较，该方法将 VGG-T 微调为流式版本，类似于我们的方法。我们还包括了 VGG-T[streaming]，这意味着在流式设置中使用 VGG-T 的方法是用因果注意力替换 VGG-T 中的全局注意力。可以看出，我们的方法明显优于所有基于优化和在线的方法，包括官方的 VGG-T 模型。在流式设置中直接使用 VGG-T 会显著降低性能，这强调了在因果约束下进行微调的必要性。