VGGT：以视觉几何为基础的 Transformer

Jianyuan Wang[1,2] Minghao Chen[1,2] Nikita Karaev[1,2] Andrea Vedaldi[1,2]
Christian Rupprecht[1] David Novotny[2]

[1] 牛津大学视觉几何组 [2]Meta 人工智能实验室

图 1. **VGGT** 是一个仅含少量三维归纳偏置的大型前馈 Transformer，使用海量带三维标注的数据训练。它可接收多达数百张图像，并在不到一秒内同时预测所有图像的相机参数、点图、深度图和点轨迹；无需进一步处理，其性能通常已优于基于优化的替代方法。

摘要

本文提出 *VGGT*，这是一种前馈（*Feed-forward*）神经网络，可从场景的一张、少量乃至数百张视图中，直接推断场景的所有关键三维属性，包括相机参数、点图（*Point Map*）、深度图和三维点轨迹。这一方法推动了三维计算机视觉的发展；以往模型通常受限于单项任务，且专门为其设计。该方法简洁高效，可在一秒内完成图像重建，其性能仍优于需要视觉几何优化技术进行后处理的替代方案。该网络在多项三维任务上取得了当前最佳结果，包括相机参数估计、多视图深度估计、稠密点云重建和三维点跟踪。实验还表明，将预训练的 *VGGT* 用作特征骨干网络（*Feature Backbone*），可显著增强非刚性点跟踪和前馈新视角合成（*Feed-forward Novel View Synthesis*）等下游任务。代码和模型已公开发布于 *https://github.com/facebookresearch/vggt*。

1. 引言

本文研究如何利用前馈神经网络，从一组图像中估计场景的三维属性。传统三维重建采用视觉几何方法，并使用光束法平差（Bundle Adjustment，BA）[45] 等迭代优化技术。机器学习常作为重要补充，用于解决仅凭几何无法处理的任务，例如特征匹配和单目深度预测。二者的结合日益紧密；如今，VGGSfM [125] 等当前最佳的运动恢复结构（Structure-from-Motion，SfM）方法通过可微 BA，将机器学习与视觉几何端到端结合起来。即便如此，视觉几何在三维重建中仍然占据重要地位，增加了系统复杂度和计算成本。

随着网络能力不断增强，一个自然的问题是：能否最终由神经网络直接解决三维任务，几乎完全免除几何后处理？近期的 DUSt3R [129] 及其后续方法 MASt3R [62] 在这一方向展现出良好前景，但这些网络每次只能处理两张图像；要重建更多图像，仍须依赖后处理融合成对重建结果。

本文朝着消除三维几何后处理优化的目标更进一步。为此，本文提出以视觉几何为基础的 *Transformer*（*Visual Geometry Grounded Transformer*，*VGGT*），这是一种前馈神经网络，可从场景的一张、少量乃至数百张输入视图完成三维重建。VGGT 预测完整的三维属性，包括相机参数、深度图、点图和三维点轨迹。上

述预测通过单次前向传播即可在数秒内完成。即使不作进一步处理，其性能通常也优于基于优化的替代方案。这与 DUSt3R 、MASt3R 和 VGGSfM 有显著区别：后者仍需代价高昂的迭代后优化才能获得可用结果。

实验还表明，无须为三维重建设计专用网络。VGGT 基于相当标准的大型 Transformer [119]，除交替采用帧内注意力（Frame-wise Attention）与全局注意力（Global Attention）外，不包含特定的三维或其他归纳偏置；模型在大量公开的三维标注数据集上训练。因此，VGGT 沿用了自然语言处理和计算机视觉大模型的设计思路，例如 GPT [1, 29, 148]、CLIP [86]、DINO [10, 78] 和 Stable Diffusion [34]。这些模型已成为用途广泛的骨干网络，可通过微调解决新的特定任务。与之类似，本文表明 VGGT 计算的特征可显著增强动态视频点跟踪和新视角合成等下游任务。

近期已有多种大型三维神经网络，例如 DepthAnything [142]、MoGe [128] 和 LRM [49]。然而，这些模型仅关注单项三维任务，例如单目深度估计或新视角合成。相比之下，VGGT 使用共享骨干网络，联合预测所有关注的三维量。实验表明，尽管这些属性之间可能存在冗余，通过学习联合预测相互关联的三维属性仍能提高整体准确率。与此同时，在推理阶段，可由分别预测的深度与相机参数推导点图，相比直接采用专用点图头，准确率更高。

综上，本文贡献如下：(1) 提出 VGGT，一种大型前馈 Transformer；输入场景的一张、少量乃至数百张图像后，可在数秒内预测其全部关键三维属性，包括相机内参与外参、点图、深度图和三维点轨迹。(2) 证明 VGGT 的预测结果可直接使用：与采用缓慢后处理优化技术的当前最佳方法相比，其竞争力很强，且通常表现更优。(3) 进一步结合 BA 后处理后，VGGT 在各项任务上均取得当前最佳结果；即便与专攻部分三维任务的方法相比，也常能显著提高质量。

代码和模型已公开发布于 https://github.com/facebookresearch/vggt。我们相信，这将推动该方向的进一步研究，并为快速、可靠且通用的三维重建提供新的基础，从而惠及计算机视觉社区。

2. 相关工作

运动恢复结构 是经典的计算机视觉问题 [45, 77, 80]，其目标是从不同视点拍摄的静态场景图像集合中估计相机参数并重建稀疏点云。传统 SfM 流程 [2, 36, 70, 94, 103, 134] 包含图像匹配、三角化和光束法平差等多个阶段。COLMAP [94] 是基于传统流程最常用的框架。近年来，深度学习改进了 SfM 流程中的许多组件，其中关键点检测 [21, 31, 116, 149] 和图像匹配 [11, 67, 92, 99] 是两个主要研究方向。近期方法 [5, 102, 109, 112, 113, 118, 122, 125, 131, 160] 探索了端到端可微 SfM；其中，VGGSfM [125] 开始在具有挑战性的地标旅游照片场景上超越传统算法。

多视图立体 旨在从多张相互重叠的图像中稠密重建场景几何，通常假设相机参数已知，而这些参数往往由 SfM 估计。多视图立体（Multi-view Stereo，MVS）方法可分为三类：传统手工方法 [38, 39, 96, 130]、全局优化方法 [37, 74, 133, 147] 和基于学习的方法 [42, 72, 84, 145, 157]。与 SfM 类似，基于学习的 MVS 方法近年来也取得了长足进展。其中，DUSt3R [129] 和 MASt3R [62] 直接从一对视图估计对齐的稠密点云；该任务类似 MVS，但不需要相机参数。同期的一些工作 [111, 127, 141, 156] 探索用神经网络替代 DUSt3R 的测试时优化，但性能仅次于或相当于 DUSt3R 。相比之下，VGGT 大幅超越了 DUSt3R 和 MASt3R 。

任意点跟踪（Tracking-Any-Point） 最早由 Particle Video [91] 提出，并在深度学习时代由 PIPs [44] 重新带入研究视野，旨在跨包含动态运动的视频序列跟踪感兴趣点。给定一段视频和若干二维查询点，该任务需要预测这些点在其余所有帧中的二维对应位置。TAP-Vid [23] 为该任务提出了三个基准和一种简单基线，后者随后在 TAPIR [24] 中得到改进。CoTracker [55, 56] 利用不同点之间的相关性实现跨遮挡跟踪，而 DOT [60] 则实现了跨遮挡稠密跟踪。近期，TAPTR [63] 为该任务提出端到端 Transformer，LocoTrack [13] 则将常用的逐点特征扩展至邻近区域。上述方法均为专用点跟踪器。本文证明，将 VGGT 特征与现有点跟踪器结合，可获得当前最佳的跟踪性能。

3. 方法

本文提出 VGGT，一个以图像集为输入、输出多种三维量的大型 Transformer。我们首先在章节 3.1中介绍问题，在章节 3.2中介绍模型架构，在章节 3.3中介绍各预测头，最后在章节 3.4中说明训练设置。

3.1. 问题定义与符号

输入是由 N 幅 RGB 图像组成的序列 $(I_i)_{i=1}^N$，其中 $I_i \in \mathbb{R}^{3\times H\times W}$，这些图像观测同一个三维场景。VGGT 的 Transformer 是一个函数，它将该序列映射为相应的一组三维标注，每帧对应一组：

$$f\left((I_i)_{i=1}^N\right) = (\mathbf{g}_i, D_i, P_i, T_i)_{i=1}^N . \tag{1}$$

因此，该 Transformer 将每幅图像 I_i 映射为其相机参数 $\mathbf{g}_i \in \mathbb{R}^9$（内参与外参）、深度图 $D_i \in \mathbb{R}^{H\times W}$、点图（Point Map）$P_i \in \mathbb{R}^{3\times H\times W}$，以及用于点跟踪的 C 维特征网格 $T_i \in \mathbb{R}^{C\times H\times W}$。下文将说明这些量的定义。

对于**相机参数** $\mathbf{g}_i$，我们采用 [125] 中的参数化方式，令 $\mathbf{g} = [\mathbf{q}, \mathbf{t}, \mathbf{f}]$，即旋转四元数 $\mathbf{q} \in \mathbb{R}^4$、平移向量 $\mathbf{t} \in \mathbb{R}^3$ 和视场角 $\mathbf{f} \in \mathbb{R}^2$ 的拼接。我们假设相机主点位于图像中心，这也是运动恢复结构（Structure from Motion，SfM）框架中的常见设定 [95, 125]。

图像 I_i 的定义域记为 $\mathcal{I}(I_i) = \{1,\dots,H\} \times \{1,\dots,W\}$,，即所有像素位置的集合。**深度图** D_i 将每个像素位置 $\mathbf{y} \in \mathcal{I}(I_i)$ 与第 i 个相机所观测到的相应深度值 $D_i(\mathbf{y}) \in \mathbb{R}^+$ 关联起来。类似地，**点图** P_i 将每

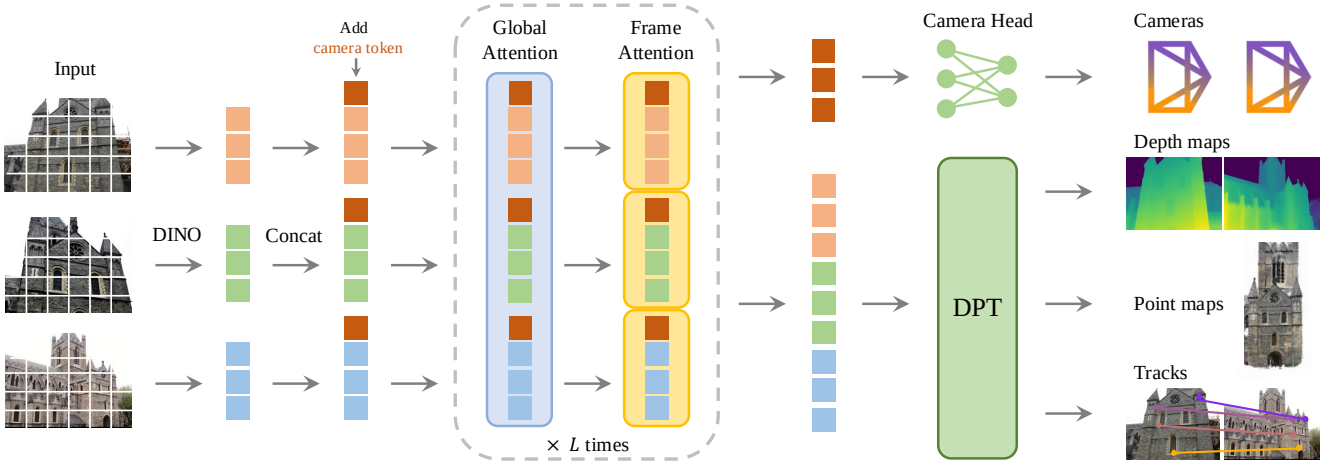

图 2. **架构概览**。模型首先使用 DINO 将输入图像划分为图像块并转换为词元，随后附加相机词元以预测相机。接着，模型交替使用帧内自注意力层和全局自注意力层。相机头最终预测相机外参与内参，而 DPT [87] 头负责生成各类稠密输出。

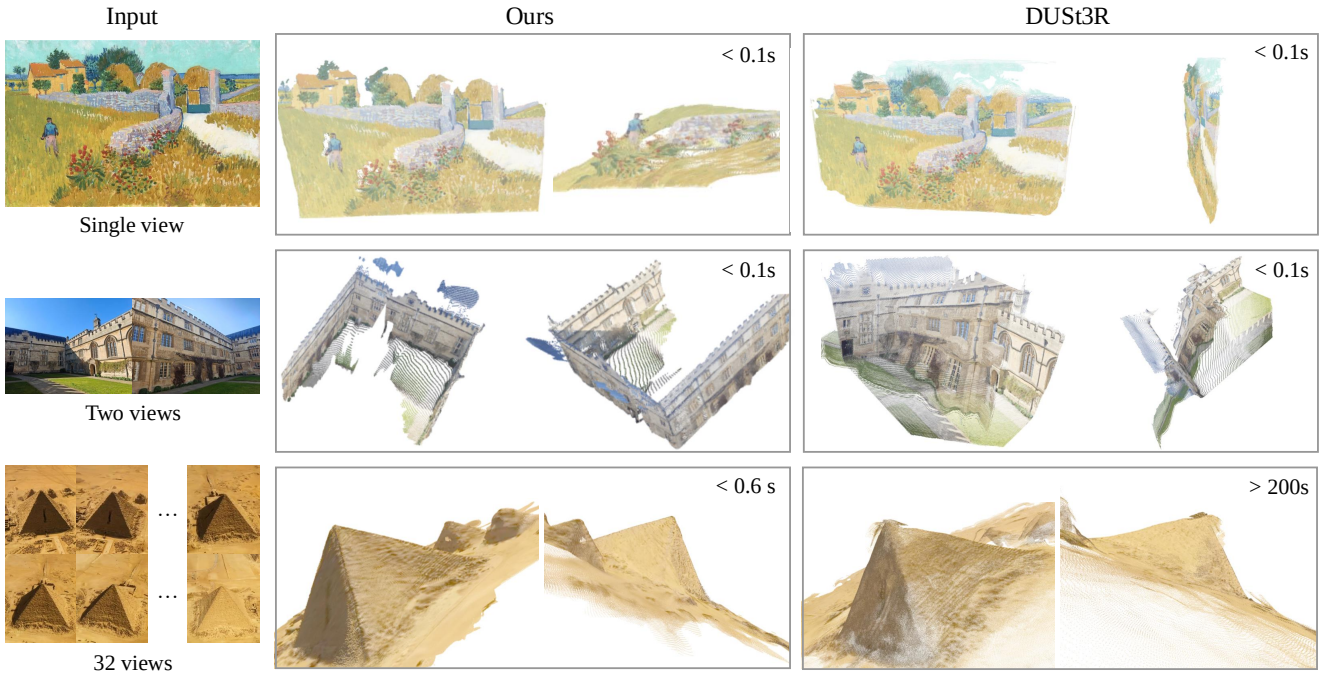

图 3. **在自然图像上，将本文预测的三维点与 DUSt3R 进行定性比较**。如第一行所示，本文方法成功预测了油画的几何结构，而 DUSt3R 预测出一个略有畸变的平面。第二行中，本文方法从两幅无重叠区域的图像中正确恢复出三维场景，而 DUSt3R 未能恢复。第三行给出了一个包含重复纹理的困难样例，本文方法仍能得到高质量预测。我们未展示输入超过 32 帧的样例，因为 DUSt3R 在超过该上限后会耗尽显存。

个像素与相应的三维场景点 $P_i(\mathbf{y}) \in \mathbb{R}^3$ 关联起来。关键在于，与 DUSt3R [129] 相同，这些点图具有视点不变性（Viewpoint Invariance），即三维点 $P_i(\mathbf{y})$ 定义在第一个相机 $\mathbf{g}_1$ 的坐标系中，我们将其作为世界参考坐标系。

最后，对于**关键点跟踪**，我们遵循 [25, 57] 等任意点跟踪（Track-Any-Point）方法。具体而言，给定查询图像 I_q 中的固定查询图像点 $\mathbf{y}_q$，网络输出一条轨迹 $\mathcal{T}^\star(\mathbf{y}_q) = (\mathbf{y}_i)_{i=1}^N$，它由所有图像 I_i 中对应的二维点 $\mathbf{y}_i \in \mathbb{R}^2$ 构成。

需要注意的是，上述 Transformer f 并不直接输出轨迹，而是输出用于跟踪的特征 $T_i \in \mathbb{R}^{C \times H \times W}$。跟踪任务交由章节 3.3中介绍的单独模块完成，该模块实现如下函数：$\mathcal{T}((\mathbf{y}_j)_{j=1}^M, (T_i)_{i=1}^N) = ((\hat{\mathbf{y}}_{j,i})_{i=1}^N)_{j=1}^M$. 它接收

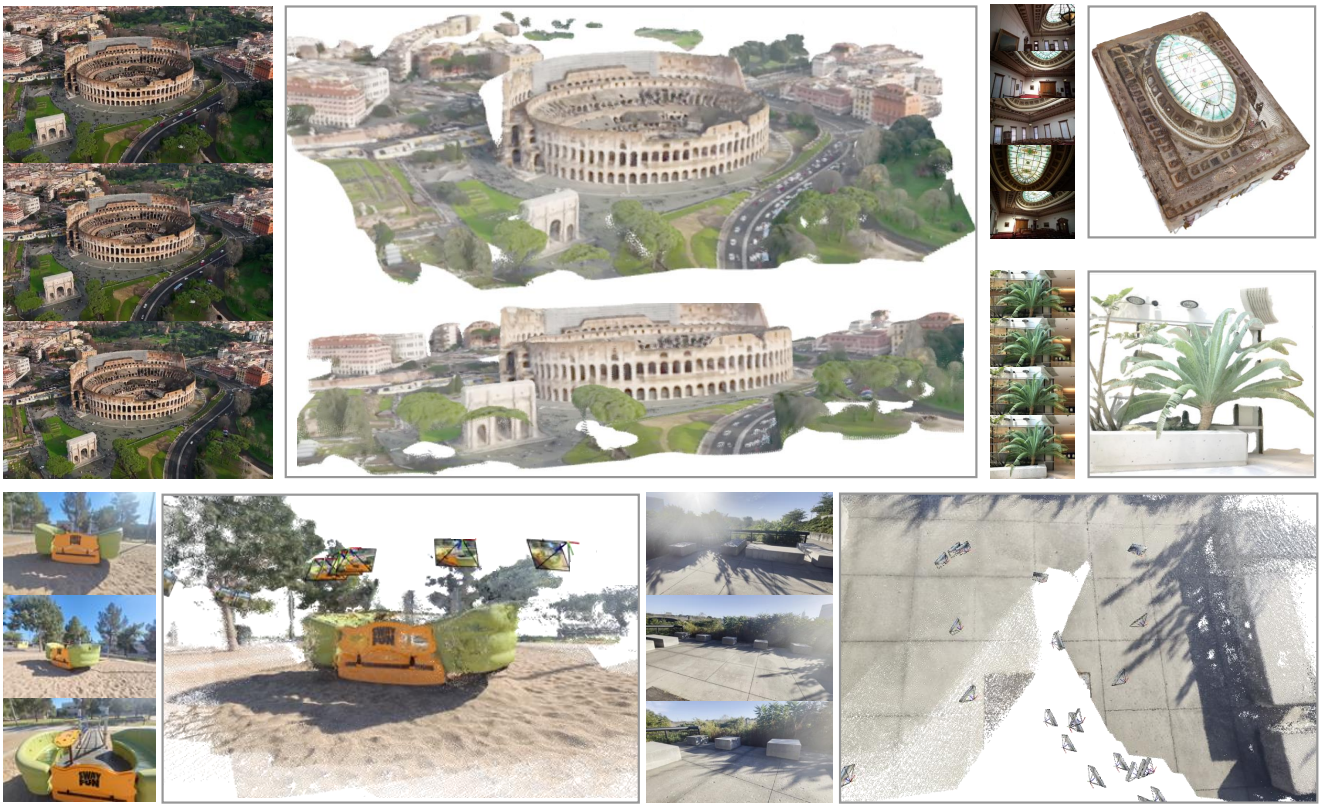

图 4. **点图估计的更多可视化结果**。相机视锥展示了估计的相机位姿。可通过我们的交互式演示获得更清晰的可视化效果。

查询点 $\mathbf{y}_q$ 和 Transformer f 输出的稠密跟踪特征 T_i，随后计算轨迹。两个网络 f 与 $\mathcal{T}$ 以端到端方式联合训练。

预测顺序。 除第一幅图像被选为参考帧外，输入序列中的图像顺序可以任意排列。网络架构被设计为对第一帧之外的所有帧满足置换等变性。

过完备预测。 值得指出的是，VGGT 预测的量并非全部相互独立。例如，正如 DUSt3R [129] 所示，相机参数 $\mathbf{g}$ 可以从不变点图 P 推断得到，例如通过求解透视-n-点（Perspective-n-Point，PnP）问题 [35, 61]。此外，深度图也可由点图和相机参数推导得到。然而，正如章节 4.5所示，即使上述各量之间存在闭式关系，在训练时要求 VGGT 显式预测所有这些量仍能显著提升性能。与此同时，在推理阶段，相比直接使用专门的点图分支，将独立估计的深度图与相机参数相结合可以生成更准确的三维点。

3.2. 特征骨干网络

遵循近期三维深度学习工作 [53, 129, 132]，我们设计了一个仅包含少量三维归纳偏置的简洁架构，使模型能够从大量带三维标注的数据中学习。具体而言，我们将模型 f 实现为大型 Transformer [119]。为此，首先通过 DINO [78] 将每幅输入图像 I 分块标记化（Patchify）为由 K 个标记组成的集合[1] $\mathrm{t}^I \in \mathbb{R}^{K\times C}$。随后，所有帧的图像标记集合，即 $\mathrm{t}^I = \cup_{i=1}^N \{\mathrm{t}_i^I\}$，经过帧内自注意力（Frame-Wise Self-Attention）层与全局自注意力（Global Self-Attention）层交替组成的主干网络处理。

交替注意力。 我们对标准 Transformer 设计略作调整，引入交替注意力（Alternating-Attention，AA），使 Transformer 交替关注帧内信息和全局信息。具体而言，帧内自注意力分别关注每一帧内的标记 t_k^I，而全局自注意力则同时关注所有帧中的标记 t^I。这种设计在跨图像整合信息与归一化每幅图像中各标记的激活之间取得了平衡。默认情况下，我们采用 $L = 24$ 层全局与帧内注意力。在章节 4中，我们证明 AA 架构能够显著提升性能。需要注意的是，该架构不含任何交叉注意力层，仅使用自注意力层。

3.3. 预测头

本节介绍 f 如何预测相机参数、深度图、点图和点轨迹。首先，对每幅输入图像 I_i，我们在对应的图像标记 t_i^I 中加入一个额外的相机标记 $\mathrm{t}_i^{\mathbf{g}} \in \mathbb{R}^{1\times C'}$ 和四个寄存器标记 [19] $\mathrm{t}_i^R \in \mathbb{R}^{4\times C'}$。随后，将 $(\mathrm{t}_i^I, \mathrm{t}_i^{\mathbf{g}}, \mathrm{t}_i^R j)_{i=1}^N$ 的拼接结果输入 AA Transformer，得到输出标记 $(\hat{\mathrm{t}}_i^I, \hat{\mathrm{t}}_i^{\mathbf{g}}, \hat{\mathrm{t}}_i^R)_{i=1}^N$。第一帧的相机标记和寄存器标记 $(\mathrm{t}_1^{\mathbf{g}} := \bar{\mathrm{t}}^{\mathbf{g}}, \mathrm{t}_1^R := \bar{\mathrm{t}}^R)$

[1] 标记数量取决于图像分辨率。

被设为一组可学习标记 $\bar{\mathrm{t}}^{\mathbf{g}}, \bar{\mathrm{t}}^{R}$，不同于其他所有帧同样可学习的标记（$\mathrm{t}_i^{\mathbf{g}} := \bar{\mathrm{t}}^{\mathbf{g}}, \mathrm{t}_i^{R} := \bar{\mathrm{t}}^{R}, i \in [2, \ldots, N]$）。这使模型能够区分第一帧与其余各帧，并在第一个相机的坐标系中表示三维预测结果。经过细化的相机标记和寄存器标记现在具有帧特异性——这是因为 AA Transformer 包含帧内自注意力层，使 Transformer 能够将相机标记和寄存器标记与同一图像中的对应标记进行匹配。按照常见做法，输出寄存器标记 $\hat{\mathrm{t}}_i^R$ 被丢弃，而 $\hat{\mathrm{t}}_i^I$、$\hat{\mathrm{t}}_i^{\mathbf{g}}$ 用于预测。

坐标系。 如上所述，我们在第一个相机 $\mathbf{g}_1$ 的坐标系中预测相机、点图和深度图。因此，第一个相机输出的相机外参被设为单位变换，即第一个旋转四元数为 $\mathbf{q}_1 = [0, 0, 0, 1]$，第一个平移向量为 $\mathbf{t}_1 = [0, 0, 0]$。前述特殊的相机标记和寄存器标记 $\mathrm{t}_1^{\mathbf{g}} := \bar{\mathrm{t}}^{\mathbf{g}}, \mathrm{t}_1^{R} := \bar{\mathrm{t}}^{R}$ 使 Transformer 能够识别第一个相机。

相机预测。 相机参数 $(\hat{\mathbf{g}}^i)_{i=1}^N$ 由输出相机标记 $(\hat{\mathrm{t}}_i^{\mathbf{g}})_{i=1}^N$ 预测，具体使用四个额外的自注意力层，之后接一个线性层。这些层构成预测相机内参与外参的相机头。

稠密预测。 输出图像标记 $\hat{\mathrm{t}}_i^I$ 用于预测稠密输出，即深度图 D_i、点图 P_i 和跟踪特征 T_i。更具体地，首先通过一个 DPT 层 [87] 将 $\hat{\mathrm{t}}_i^I$ 转换为稠密特征图 $F_i \in \mathbb{R}^{C'' \times H \times W}$。随后，分别用一个 3×3 卷积层将每个 F_i 映射为相应的深度图 D_i 和点图 P_i。此外，DPT 头还输出稠密特征 $T_i \in \mathbb{R}^{C \times H \times W}$，作为跟踪头的输入。我们还分别为每幅深度图和点图预测偶然不确定性（Aleatoric Uncertainty）[58, 76] $\Sigma_i^D \in \mathbb{R}_+^{H \times W}$ 和 $\Sigma_i^P \in \mathbb{R}_+^{H \times W}$。如章节 3.4所述，不确定性图用于计算损失；训练完成后，它们与模型对预测的置信度成正比。

跟踪。 为了实现跟踪模块 $\mathcal{T}$，我们采用以稠密跟踪特征 T_i 为输入的 CoTracker2 架构 [57]。更具体地，给定查询图像 I_q 中的查询点 $\mathbf{y}_j$（训练期间始终设 $q = 1$，但也可使用任何其他图像作为查询），跟踪头 $\mathcal{T}$ 预测所有图像 I_i 中的一组二维点 $\mathcal{T}((\mathbf{y}_j)_{j=1}^M, (T_i)_{i=1}^N) = ((\hat{\mathbf{y}}_{j,i})_{i=1}^N)_{j=1}^M$，这些点与 $\mathbf{y}$ 对应于同一个三维点。为此，首先在查询点 $\mathbf{y}_j$ 处对查询图像的特征图 T_q 进行双线性采样，以获得该点的特征。随后，将该特征与所有其他特征图 $T_i, i \neq q$ 进行相关性计算，得到一组相关图。这些相关图经过自注意力层处理，预测最终的二维点 $\hat{\mathbf{y}}_i$；这些点均与 $\mathbf{y}_j$ 对应。与 VGGSfM [125] 类似，我们的跟踪器不对输入帧作任何时序顺序假设，因此可应用于任意输入图像集，而不仅限于视频。

3.4. 训练

训练损失。 我们使用多任务损失对 VGGT 模型 f 进行端到端训练：

$$\mathcal{L} = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pmap}} + \lambda \mathcal{L}_{\text{track}}. \tag{2}$$

实验发现，相机损失（$\mathcal{L}_{\text{camera}}$）、深度损失（$\mathcal{L}_{\text{depth}}$）和点图损失（$\mathcal{L}_{\text{pmap}}$）的数值范围相近，无需相互加权。跟踪损失 $\mathcal{L}_{\text{track}}$ 的权重因子设为 $\lambda = 0.05$。下面依次介绍各损失项。

相机损失 $\mathcal{L}_{\text{camera}}$ 监督相机预测 $\hat{\mathbf{g}}$：$\mathcal{L}_{\text{camera}} = \sum_{i=1}^N \|\hat{\mathbf{g}}_i - \mathbf{g}_i\|_\epsilon$，该损失使用 Huber 损失 $|\cdot|_\epsilon$ 比较预测相机 $\hat{\mathbf{g}}_i$ 与真值 $\mathbf{g}_i$。

深度损失 $\mathcal{L}_{\text{depth}}$ 遵循 DUSt3R [129]，采用偶然不确定性损失 [59, 75]，根据预测的不确定性图 $\hat{\Sigma}_i^D$ 对预测深度 $\hat{D}_i$ 与真值深度 D_i 之间的差异加权。与 DUSt3R 不同，我们还加入了单目深度估计中广泛使用的梯度项。因此，深度损失为 $\mathcal{L}_{\text{depth}} = \sum_{i=1}^N \|\Sigma_i^D \odot (\hat{D}_i - D_i)\| + \|\Sigma_i^D \odot (\nabla \hat{D}_i - \nabla D_i)\| - \alpha \log \Sigma_i^D$，其中 $\odot$ 表示沿通道广播的逐元素乘积。点图损失的定义与之类似，但使用点图不确定性 Σ_i^P：$\mathcal{L}_{\text{pmap}} = \sum_{i=1}^N \|\Sigma_i^P \odot (\hat{P}_i - P_i)\| + \|\Sigma_i^P \odot (\nabla \hat{P}_i - \nabla P_i)\| - \alpha \log \Sigma_i^P$。

最后，跟踪损失定义为 $\mathcal{L}_{\text{track}} = \sum_{j=1}^M \sum_{i=1}^N \|\mathbf{y}_{j,i} - \hat{\mathbf{y}}_{j,i}\|$。其中，外层求和遍历查询图像 I_q 中的所有真值查询点 $\mathbf{y}_j$，$\mathbf{y}_{j,i}$ 是 $\mathbf{y}_j$ 在图像 I_i 中的真值对应点，而 $\hat{\mathbf{y}}_{j,i}$ 是应用跟踪模块 $\mathcal{T}((\mathbf{y}_j)_{j=1}^M, (T_i)_{i=1}^N)$ 得到的相应预测。此外，遵循 CoTracker2 [57]，我们采用可见性损失（二元交叉熵）估计一个点在给定帧中是否可见。

真值坐标归一化。 如果缩放一个场景或改变其全局参考坐标系，场景图像完全不受影响，这意味着任何此类变体都是三维重建的合理结果。我们通过归一化数据消除这种歧义，从中确定一个规范形式，并要求 Transformer 输出这一特定变体。遵循 [129]，首先在第一个相机 $\mathbf{g}_1$ 的坐标系中表示所有量。然后，计算点图 P 中所有三维点到原点的平均欧氏距离，并使用该尺度归一化相机平移 $\mathbf{t}$、点图 P 和深度图 D。与 [129] 不同的是，我们不会对 Transformer 的输出预测执行这种归一化；相反，我们要求模型从训练数据中学习所选用的归一化方式。

实现细节。 默认情况下，我们分别采用 $L = 24$ 层全局注意力和帧内注意力。模型总计约含 12 亿个参数。我们使用 AdamW 优化器优化训练损失 (2)，共训练 160K 次迭代。采用余弦学习率调度器，峰值学习率为 0.0002，并进行 8K 次迭代的预热。对于每个批次，从随机训练场景中随机采样 2–24 帧。输入帧、深度图和点图均被缩放至最大尺寸为 518 像素。宽高比在 0.33 至 1.0 之间随机取值。我们还对输入帧随机应用颜色抖动、高斯模糊和灰度增强。训练使用 64 块 A100 GPU，历时九天。采用阈值为 1.0 的梯度范数裁剪，以保证训练稳定性。同时使用 bfloat16 精度和梯度检查点，以提高 GPU 显存与计算效率。

训练数据。 模型使用大规模、多样化的数据集集合进行训练，包括：Co3Dv2 [88]、BlendMVS [146]、DL3DV [69]、MegaDepth [64]、Kubric [41]、WildRGB [135]、ScanNet [18]、HyperSim [89]、Mapillary [71]、Habitat [107]、Replica [104]、MVS-Synth [50]、PointOdyssey [159]、Virtual KITTI [7]、Aria Synthetic Environments [82]、Aria Digital Twin [82]，以及一个由艺术家创作资产组成、类似 Objaverse [20] 的合成数据集。这些数据集覆盖室内和室外等不同领域，同时包含合成与真实世界场景。数据集中的三维标注来自多种来

方法	Re10K （未见） AUC@30 ↑	CO3Dv2 AUC@30 ↑	时间
Colmap+SPSG [92]	45.2	25.3	∼ 15s
PixSfM [66]	49.4	30.1	> 20s
PoseDiff [124]	48.0	66.5	∼ 7s
DUSt3R [129]	67.7	76.7	∼ 7s
MASt3R [62]	76.4	81.8	∼ 9s
VGGSfM v2 [125]	78.9	83.4	∼ 10s
MV-DUSt3R [111] ‡	71.3	69.5	∼ 0.6s
CUT3R [127] ‡	75.3	82.8	∼ 0.6s
FLARE [156] ‡	78.8	83.3	∼ 0.5s
Fast3R [141] ‡	72.7	82.5	∼ **0.2s**
本文方法（前馈）	85.3	88.2	∼ **0.2s**
本文方法（含 BA）	**93.5**	**91.8**	∼ 1.8s

表 1. **在 RealEstate10K [161] 和 CO3Dv2 [88] 上的相机位姿估计**，使用 10 个随机帧。所有指标均为越高越好。所有方法均未在 Re10K 数据集上训练。运行时间使用一张 H100 GPU 测得。以 ‡ 标记的方法为同期工作。

已知真值相机	方法	准确度 ↓	完整度 ↓	总体误差 ↓
✓	Gipuma [40]	**0.283**	0.873	0.578
✓	MVSNet [144]	0.396	0.527	0.462
✓	CIDER [139]	0.417	0.437	0.427
✓	PatchmatchNet [121]	0.427	0.377	0.417
✓	MASt3R [62]	0.403	0.344	0.374
✓	GeoMVSNet [157]	0.331	**0.259**	**0.295**
✗	DUSt3R [129]	2.677	0.805	1.741
✗	本文方法	**0.389**	**0.374**	**0.382**

表 2. **在 DTU [51] 数据集上的稠密 MVS 估计**。表格上半部分为已知真值相机的方法，下半部分为未知真值相机的方法。

方法	准确度 ↓	完整度 ↓	总体误差 ↓	时间
DUSt3R	1.167	0.842	1.005	∼ 7s
MASt3R	0.968	0.684	0.826	∼ 9s
本文方法（点图）	0.901	0.518	0.709	∼ 0.2s
本文方法（深度 + 相机）	**0.873**	**0.482**	**0.677**	∼ 0.2s

表 3. **在 ETH3D [97] 上的点图估计**。DUSt3R 和 MASt3R 使用全局对齐，而本文方法采用前馈方式，因而速度快得多。本文方法（点图）一行表示直接使用点图头的结果，本文方法（深度 + 相机）表示结合深度图头与相机头构建点云。

源，例如传感器直接采集、合成引擎或 SfM 技术 [95]。我们的数据集组合在规模和多样性上与 MASt3R [30] 大致相当。

4. 实验

本节在多项任务上将本文方法与当前最先进的方法进行比较，以验证其有效性。

方法	AUC@5 ↑	AUC@10 ↑	AUC@20 ↑
SuperGlue [92]	16.2	33.8	51.8
LoFTR [105]	22.1	40.8	57.6
DKM [32]	29.4	50.7	68.3
CasMTR [9]	27.1	47.0	64.4
Roma [33]	31.8	53.4	70.9
本文方法	**33.9**	**55.2**	**73.4**

表 4. **在 ScanNet-1500 [18, 92] 上的双视图匹配比较**。尽管本文跟踪头并非专为双视图设置设计，其性能仍超过当前最先进的双视图匹配方法 Roma。采用 AUC 衡量（越高越好）。

4.1. 相机位姿估计

首先在 CO3Dv2 [88] 和 RealEstate10K [161] 数据集上评估本文方法的相机位姿估计（Camera Pose Estimation）性能，结果如 表 1 所示。遵循 [124]，我们从每个场景中随机选择 10 张图像，并采用结合 RRA 与 RTA 的标准指标 AUC@30 进行评估。相对旋转准确率（Relative Rotation Accuracy，RRA）与相对平移准确率（Relative Translation Accuracy，RTA）分别计算每个图像对的旋转和平移相对角度误差。随后对这些角度误差设置阈值，以确定准确率分数。AUC 是不同阈值下 RRA 与 RTA 较小值所形成的准确率–阈值曲线下面积。表 1 中的（可学习）方法均在 Co3Dv2 上训练，且未在 RealEstate10K 上训练。在两个数据集的所有指标上，本文的前馈模型始终优于其他方法，包括采用高计算开销后优化步骤的方法，例如 DUSt3R/MASt3R 的全局对齐和 VGGSfM 的光束法平差（Bundle Adjustment，BA）；这些方法通常需要超过 10 秒。相比之下，VGGT 仅以纯前馈方式运行便取得更优性能，在相同硬件上只需 0.2 秒。与同期工作 [111, 127, 141, 156]（以 ‡ 标记）相比，本文方法具有显著性能优势，速度则与其中最快的 Fast3R [141] 相近。在 RealEstate10K 数据集上，这一性能优势更加明显，因为 表 1 中的方法均未在该数据集上训练。这验证了 VGGT 出色的泛化能力。

结果还表明，将 VGGT 与 BA 等视觉几何优化方法结合可进一步提升性能。具体而言，使用 BA 优化预测的相机位姿与轨迹可进一步提高准确率。本文方法直接预测接近准确的点图/深度图，可为 BA 提供良好的初始化。因此，无需像 [125] 那样在 BA 中执行三角测量和迭代细化，使本文方法快得多（即使采用 BA 也只需约 2 秒）。由此可见，尽管 VGGT 的前馈模式优于此前所有方案（无论其是否为前馈方法），后优化仍能带来收益，性能还有进一步提升的空间。

4.2. 多视图深度估计

遵循 MASt3R [62]，我们进一步在 DTU [51] 数据集上评估多视图深度估计结果。报告的标准 DTU 指标包括准确度（Accuracy，从预测到真值的最小欧氏距离）、完整度（Completeness，从真值到预测的最小欧氏距离），以及二者平均值总体误差（Overall，即 Chamfer 距离）。

图 5. **刚性点跟踪与动态点跟踪的可视化**。上：VGGT 的跟踪模块 $\mathcal{T}$ 为描述静态场景的无序输入图像集合输出关键点轨迹。下：我们微调 VGGT 的骨干网络，以增强处理时序输入的动态点跟踪器 CoTracker [56]。

在 表 2 中，DUSt3R 和本文的 VGGT 是仅有的两个不使用真值相机信息的方法。MASt3R 使用真值相机对匹配点进行三角测量，从而得到深度图。GeoMVSNet 等深度多视图立体视觉方法则使用真值相机构建代价体。

本文方法显著优于 DUSt3R，将总体误差从 1.741 降至 0.382。更重要的是，其结果与测试时已知真值相机的方法相当。这一显著性能提升很可能得益于模型的多图像训练方案：该方案使模型原生具备多视图三角测量推理能力，而不依赖特设的对齐过程。相比之下，DUSt3R 仅对多个成对相机三角测量结果取平均。

4.3. 点图估计

我们还在 ETH3D [97] 数据集上比较本文方法、DUSt3R 和 MASt3R 所预测点云的准确性。每个场景随机采样 10 帧。采用 Umeyama [117] 算法将预测点云与真值对齐。使用官方掩码过滤无效点后报告结果。点图估计采用准确度、完整度和总体误差（Chamfer 距离）作为指标。如 表 3 所示，尽管 DUSt3R 和 MASt3R 执行了高开销优化（全局对齐——每个场景约 10 秒），本文方法仅以简单的前馈方式、每次重建只需 0.2 秒，仍显著优于二者。

与直接使用估计点图相比，深度头和相机头的预测结果（即使用预测相机参数将预测深度图反投影到三维空间）具有更高准确率。这说明，将复杂任务（点图估计）分解为较简单的子问题（深度图与相机预测）能够带来收益，尽管训练期间相机、深度图和点图受到联合监督。

图 3 给出了本文方法与 DUSt3R 在自然场景上的定性比较，更多示例见 图 4。VGGT 能够输出高质量预测并具有良好泛化能力，在油画、无重叠帧，以及沙漠等具有重复或均匀纹理的场景等困难域外样例上表现出色。

4.4. 图像匹配

双视图图像匹配是计算机视觉中得到广泛研究的课题 [68, 93, 105]。它是仅限两个视图的刚性点跟踪特例，因此即使本文模型并非专门面向该任务，它仍是衡量跟踪准确率的合适评估基准。我们在 ScanNet 数据集 [18] 上遵循标准协议 [33, 93]，结果见 表 4。对于每个图像对，首先提取匹配并据此估计本质矩阵，再将其分解得到相对相机位姿。最终指标为以 AUC 衡量的相对位姿准确率。评估时，使用 ALIKED [158] 检测关键点，并将其作为查询点 $\mathbf{y}_q$。随后将这些点输入跟踪分支 $\mathcal{T}$，以在第二帧中寻找对应点。评估超参数（例如匹配数量、RANSAC 阈值）采用 Roma [33] 的设置。尽管未针对双视图匹配进行显式训练，表 4 表明 VGGT 在所有基线中取得最高准确率。

ETH3D 数据集	准确度 ↓	完整度 ↓	总体误差 ↓
交叉注意力	1.287	0.835	1.061
仅全局自注意力	<u>1.032</u>	<u>0.621</u>	<u>0.827</u>
交替注意力	**0.901**	**0.518**	**0.709**

表 5. **在 ETH3D 上对 Transformer 骨干网络进行消融实验**。将本文的交替注意力架构与两个变体进行比较：一个仅使用全局自注意力，另一个使用交叉注意力。

含 $\mathcal{L}_{\text{camera}}$	含 $\mathcal{L}_{\text{depth}}$	含 $\mathcal{L}_{\text{track}}$	准确度 ↓	完整度 ↓	总体误差 ↓
✗	✓	✓	1.042	0.627	0.834
✓	✗	✓	<u>0.920</u>	<u>0.534</u>	<u>0.727</u>
✓	✓	✗	0.976	0.603	0.790
✓	✓	✓	**0.901**	**0.518**	**0.709**

表 6. **多任务学习的消融实验**。结果表明，同时训练相机、深度和轨迹估计时，在 ETH3D 上取得最高的点图估计准确率。

4.5. 消融实验

特征骨干网络。首先将本文提出的交替注意力（Alternating-Attention）设计与两种替代注意力架构进行比较，以验证其有效性：(a) 仅使用全局自注意力；(b) 交叉注意力。为确保公平比较，所有模型变体的参数量相同，均使用总计 $2L$ 个注意力层。在交叉注意力变体中，每一帧分别关注其余所有帧的标记，最大限度融合跨帧信息，但运行时间显著增加，输入帧数增多时尤为明显。隐藏维度和注意力头数量等超参数保持一

致。消融实验选择点图估计准确率作为评估指标，因为它能反映模型对场景几何与相机参数的联合理解能力。表 5 的结果表明，交替注意力架构以明显优势超过两个基线变体。其他初步探索实验也一致表明，使用交叉注意力的架构通常不如仅使用自注意力的架构。

多任务学习。 我们还验证了训练单个网络同时学习多种三维量的收益，尽管这些输出可能存在重叠（例如深度图与相机参数可共同生成点图）。如 表 6 所示，训练时移除相机、深度或轨迹估计都会使点图估计准确率明显下降。相机参数估计能显著提高点图准确率，而深度估计仅带来小幅提升。

4.6. 面向下游任务的微调

下面说明 VGGT 预训练特征提取器可复用于下游任务。具体展示其在前馈新视角合成（Novel View Synthesis）和动态点跟踪（Dynamic Point Tracking）中的应用。

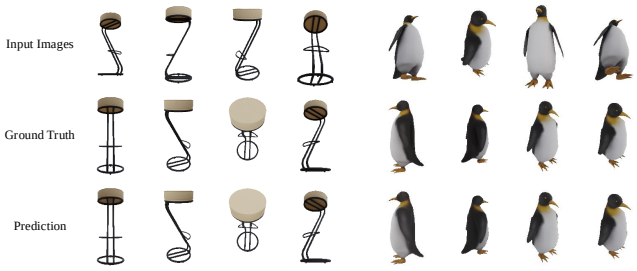

图 6. **新视角合成的定性示例。**上排为输入图像，中排为目标视点的真值图像，下排为本文方法合成的图像。

方法	已知输入相机	尺寸	PSNR ↑	SSIM ↑	LPIPS ↓
LGM [110]	✓	256	21.44	0.832	0.122
GS-LRM [154]	✓	256	29.59	0.944	0.051
LVSM [53]	✓	256	31.71	0.957	0.027
本文方法-NVS*	✗	224	30.41	0.949	0.033

表 7. **在 GSO [28] 数据集上的视图合成定量比较。**将 VGGT 微调用于前馈新视角合成后，即使不知道输入图像的相机外参与内参，也能取得有竞争力的性能。注意，* 表示使用小规模训练集（仅 20%）。

前馈新视角合成 近年来发展迅速 [8, 43, 49, 53, 108, 126, 140, 155]。多数现有方法将已知相机参数的图像作为输入，并预测新相机视点所对应的目标图像。本文遵循 LVSM [53]，不依赖显式 3D 表示，而是修改 VGGT 以直接输出目标图像。但我们不假设输入帧的相机参数已知。

训练与评估严格遵循 LVSM 的协议，例如使用 4 个输入视图，并以 Plücker 射线表示目标视点。我们对 VGGT 进行简单修改。与此前相同，输入图像由 DINO 转换为标记。对于目标视图，则使用卷积层将其 Plücker 射线图像编码为标记。表示输入图像和目标视图的这些标记被拼接起来，并由 AA Transformer 处理。随后使用 DPT 头回归目标视图的 RGB 颜色。需要强调的是，本文不输入源图像的 Plücker 射线。因此，模型无法获知这些输入帧的相机参数。

LVSM 在 Objaverse 数据集 [20] 上训练。本文使用类似的内部数据集，其规模约为 Objaverse 的 20%。训练和评估的更多细节见 [53]。如 表 7 所示，尽管无需输入相机参数且训练数据少于 LVSM，本文模型在 GSO 数据集 [28] 上仍取得有竞争力的结果。预计采用更大的训练数据集可取得更好结果。定性示例见 图 6。

方法	Kinetics			RGB-S			DAVIS		
	AJ	$\delta^{\text{vis}}_{\text{avg}}$	OA	AJ	$\delta^{\text{vis}}_{\text{avg}}$	OA	AJ	$\delta^{\text{vis}}_{\text{avg}}$	OA
TAPTR [63]	49.0	64.4	85.2	60.8	76.2	87.0	63.0	76.1	91.1
LocoTrack [13]	52.9	66.8	85.3	69.7	83.2	89.5	62.9	75.3	87.2
BootsTAPIR [26]	54.6	68.4	86.5	70.8	83.0	89.9	61.4	73.6	88.7
CoTracker [56]	49.6	64.3	83.3	67.4	78.9	85.2	61.8	76.1	88.3
CoTracker + 本文方法	**57.2**	**69.0**	**88.9**	**72.1**	**84.0**	**91.6**	**64.7**	**77.5**	**91.4**

表 8. **在 TAP-Vid 基准上的动态点跟踪结果。**尽管本文模型并非为动态场景设计，仅使用本文预训练权重微调 CoTracker 即可显著提高性能，说明所学特征具有良好的鲁棒性和有效性。

动态点跟踪 近年来已成为竞争激烈的任务 [25, 44, 57, 136]，也是所学特征的另一项下游应用。遵循标准实践，报告以下点跟踪指标：遮挡准确率（Occlusion Accuracy，OA），即遮挡预测的二分类准确率；$\delta^{\text{vis}}_{\text{avg}}$，即在给定像素阈值内被准确跟踪的可见点平均比例；以及平均 Jaccard（Average Jaccard，AJ），用于联合衡量跟踪与遮挡预测的准确率。

我们将当前最先进的 CoTracker2 模型 [57] 的骨干网络替换为本文的预训练特征骨干网络。这是因为 VGGT 在无序图像集合上训练，而非时序视频。本文骨干网络预测跟踪特征 T_i，以其替代特征提取器的输出；这些特征随后进入 CoTracker2 架构的其余部分，并最终用于预测轨迹。整个修改后的跟踪器在 Kubric [41] 上微调。如 表 8 所示，集成预训练 VGGT 后，CoTracker 在 TAP-Vid 基准 [23] 上的性能显著提升。例如，在 TAP-Vid RGB-S 数据集上，VGGT 的跟踪特征将 $\delta^{\text{vis}}_{\text{avg}}$ 指标从 78.9 提升至 84.0。尽管 TAP-Vid 基准包含来自不同数据源、具有快速动态运动的视频，本文模型仍取得优异性能，表明其特征即使在并非显式设计所面向的场景中也具有良好泛化能力。

5. 讨论

局限性。 尽管本文方法对多样的自然场景表现出强泛化能力，但仍存在若干局限。首先，当前模型不支持鱼眼图像或全景图像；输入图像发生极端旋转时，重建性能也会下降。此外，模型虽能处理存在轻微非刚性运动的场景，却无法应对大幅非刚性形变。

本文方法的重要优势在于灵活且易于适配。只需在针对性数据集上微调模型，并对架构做少量修改，即可

输入帧数	1	2	4	8	10	20	50	100	200
时间 (s)	0.04	0.05	0.07	0.11	0.14	0.31	1.04	3.12	8.75
显存 (GB)	1.88	2.07	2.45	3.23	3.63	5.58	11.41	21.15	40.63

表 9. **不同输入帧数下的运行时间与 GPU 峰值显存占用**。运行时间以秒为单位，GPU 显存占用以 GB 为单位。

直接解决这些局限。相比之下，现有方法通常需要在测试时优化过程中进行大量重新设计，才能适应此类特殊场景；这种适应能力使本文方法与其明显不同。

运行时间与显存。 如 表 9 所示，我们评估了特征骨干网络处理不同数量输入帧时的推理运行时间和 GPU 峰值显存占用。测量在单张 NVIDIA H100 GPU 上使用 Flash Attention v3 [98] 完成。图像分辨率为 336×518。

由于用户可能根据具体需求与可用资源选择不同分支组合，此处重点分析特征骨干网络的开销。相机头较为轻量，其运行时间通常约为特征骨干网络的 5%，GPU 显存占用约为后者的 2%。每个 DPT 头处理一帧平均耗时 0.03 秒，占用 0.2 GB GPU 显存。

GPU 显存充足时，可在单次前向传播中高效处理多帧。在本文模型中，帧间关系仅由特征骨干网络处理，而 DPT 头对每一帧独立进行预测。因此，GPU 资源受限的用户可逐帧预测，具体取舍由用户决定。

当标记数量很大时，全局自注意力的朴素实现会占用大量显存。采用大语言模型（Large Language Model，LLM）部署中的技术可节省资源或加速计算。例如，Fast3R [141] 采用张量并行（Tensor Parallelism），通过多张 GPU 加速推理，这一技术可直接用于本文模型。

分块标记化。 如 章节 3.2 所述，我们探索了两种将图像分块标记化（Patchifying）为标记的方法：使用 14×14 卷积层或预训练 DINOv2 模型。实验结果表明，DINOv2 性能更好，且训练稳定得多，尤其是在初始阶段。DINOv2 对学习率或动量等超参数的变化也不太敏感。因此，本文模型默认使用 DINOv2 进行分块标记化。

可微光束法平差。 我们还探索了像 VGGSfM [125] 一样使用可微光束法平差（Differentiable Bundle Adjustment，BA）。小规模初步实验表明，可微 BA 具有良好潜力。但其瓶颈在于训练期间的计算成本。在 PyTorch 中使用 Theseus [85] 启用可微 BA，通常会使每个训练步骤慢约 4 倍，对于大规模训练而言开销很高。定制框架以加速训练可能是一种解决方案，但超出本文范围。因此，本文没有加入可微 BA；不过，它可在缺乏显式 3D 标注的场景中提供有效监督信号，是大规模无监督训练中值得探索的方向。

单视图重建。 DUSt3R 和 MASt3R 等系统必须复制图像以构成图像对，而本文模型架构原生支持单张图像输入。此时，全局注意力会直接转变为逐帧注意力。尽管模型未针对单视图重建进行显式训练，却取得了出人意料的良好结果。部分示例见 图 3 和 图 7。我们强烈建议试用演示，以获得更直观的可视化效果。

预测归一化。 如 章节 3.4 所述，本文方法使用 3D 点的平均欧氏距离对真值进行归一化。虽然 DUSt3R 等方法也对网络预测应用这种归一化，但实验表明，它既非模型收敛所必需，也无益于最终性能。此外，这种做法往往会在训练阶段引入额外的不稳定性。

6. 结论

本文提出以视觉几何为基础的 Transformer（VGGT），一种前馈神经网络，可直接估计数百个输入视图中的所有关键三维场景属性。该方法在多项 3D 任务上取得当前最先进的结果，包括相机参数估计、多视图深度估计、稠密点云重建和 3D 点跟踪。本文简单且以神经网络为先的方法不同于传统的视觉几何方法；后者依赖优化和后处理才能针对特定任务获得准确结果。本文方法简单高效，非常适合实时应用，这也是其相较于优化类方法的另一项优势。

附录

附录包含以下内容：

- 章节 A给出关键术语的形式化定义。
- 章节 B提供完整的实现细节，包括架构和训练超参数。
- 章节 C给出补充实验与讨论。
- 章节 D展示单视图重建的定性示例。
- 章节 E进一步综述相关工作。

A. 形式化定义

本节补充形式化定义，为方法部分奠定更严谨的基础。

相机外参相对于*世界参考系*定义，本文将第一台相机的坐标系作为世界参考系。为此，引入两个函数。第一个函数 $\gamma(\mathbf{g}, \mathbf{p}) = \mathbf{p}'$ 将 $\mathbf{g}$ 编码的刚性变换应用于世界参考系中的点 $\mathbf{p}$，得到相机参考系中的对应点 $\mathbf{p}'$。第二个函数 $\pi(\mathbf{g}, \mathbf{p}) = \mathbf{y}$ 进一步应用透视投影，将三维点 $\mathbf{p}$ 映射为二维图像点 $\mathbf{y}$。相机 $\mathbf{g}$ 观测到的该点深度记为 $\pi^{\mathrm{D}}(\mathbf{g}, \mathbf{p}) = d \in \mathbb{R}^{+}$.

本文将场景建模为一组正则曲面 $S_i \subset \mathbb{R}^3$。由于场景会随时间变化 [151]，因此令其依赖于第 i 张输入图像。像素位置 $\mathbf{y} \in \mathcal{I}(I_i)$ 处的深度定义为场景中所有投影至 $\mathbf{y}$ 的三维点 $\mathbf{p}$ 的最小深度，即 $D_i(\mathbf{y}) = \min\{\pi^D(\mathbf{g}_i, \mathbf{p}) : \mathbf{p} \in S_i \ \wedge \ \pi(\mathbf{g}_i, \mathbf{p}) = \mathbf{y}\}$. 于是，像素位置 $\mathbf{y}$ 处的点为 $P_i(\mathbf{y}) = \gamma(\mathbf{g}, \mathbf{p})$，其中 $\mathbf{p} \in S_i$ 是使上式取最小值的三维点，即 $\mathbf{p} \in S_i \ \wedge \ \pi(\mathbf{g}_i, \mathbf{p}) = \mathbf{y} \ \wedge \ \pi^D(\mathbf{g}_i, \mathbf{p}) = D_i(\mathbf{y})$.

B. 实现细节

架构。如正文所述，VGGT 由 24 个注意力块组成，每个块包含一个帧内自注意力层和一个全局自注意力层。遵循 DINOv2 [78] 采用的 ViT-L 模型，每个注意力层的特征维度设为 1024，并使用 16 个注意力头。注意力层采用 PyTorch 官方实现（即 *torch.nn.MultiheadAttention*），并启用 Flash Attention。为稳定训练，每个注意力层还采用 QKNorm [48] 和 LayerScale [115]。LayerScale 的初始值设为 0.01。图像词元化采用 DINOv2 [78]，并加入位置嵌入。与 [143] 一致，将第 4、11、17 和 23 个块的词元输入 DPT [87] 进行上采样。

训练。 构建训练批次时，首先随机选择一个训练数据集（与 [129] 一致，各数据集权重不同但大致接近），再从中均匀随机采样一个场景。训练期间，每个场景选择 2 至 24 帧，同时保持每个批次共 48 帧不变。训练使用各数据集对应的训练集。包含少于 24 帧的训练序列会被排除。首先对 RGB 帧、深度图和点图进行等比例缩放，使长边为 518 像素。随后围绕主点裁剪短边，使其尺寸介于 168 至 518 像素，并保持为 14 像素图像块尺寸的整数倍。同一场景中的各帧会独立应用幅度较大的颜色增强，以提高模型对不同光照条件的鲁棒性。遵循 [33, 105, 125] 构建真实轨迹：将深度图反投影至三维空间，再把点重投影到目标帧，并保留重投影深度与目标深度图相符的对应关系。批次采样时，排除与查询帧相似度较低的帧。在少数不存在有效对应关系的情况下，省略跟踪损失。

C. 补充实验

IMC 上的相机位姿估计 本文还在图像匹配挑战赛（Image Matching Challenge，IMC） [54] 上评估；这是一个专注于地标旅游照片数据的相机位姿估计基准。直到近期，该基准仍由经典增量式 SfM 方法 [94] 主导。

基线。 评估模型的两个版本：VGGT 和 VGGT + BA。VGGT 直接输出相机位姿估计，而 VGGT + BA 通过额外的光束法平差阶段细化估计结果。对比对象包括 [66, 94] 等经典增量式 SfM 方法和近期提出的深度学习方法。具体而言，近期的 VGGSfM [125] 是首个在具有挑战性的地标旅游照片数据集上超越增量式 SfM 的端到端训练深度学习方法。

除 VGGSfM 外，还与近期广受关注的 DUSt3R [129] 和 MASt3R [62] 进行比较。需要指出的是，DUSt3R 和 MASt3R 使用了 MegaDepth 数据集的很大一部分进行训练，仅排除场景 0015 和 0022。其训练采用的 MegaDepth 场景与 IMC 基准存在一定重叠；虽然图像并不相同，但两个数据集中包含相同场景。例如，MegaDepth 的场景 0024 对应大英博物馆，而 IMC 基准同样包含大英博物馆场景。为进行同等条件下的比较，本文采用与 DUSt3R 和 MASt3R 相同的训练划分。正文中，为确保在 ScanNet-1500 上公平比较，训练时排除了对应的 ScanNet 场景。

结果。 表 10给出了评估结果。尽管地标旅游照片数据历来是 SfM 方法的重点，VGGT 的前馈性能仍与当前最佳的 VGGSfMv2 相当：二者 AUC@10 分别为 71.26 和 76.82；与此同时，前者速度显著更快（每个场景 0.2 秒，对比 10 秒）。VGGT 在所有准确率阈值下均显著超越 MASt3R [62] 和 DUSt3R [129]，且速度快得多。这是因为 MASt3R 和 DUSt3R 的前馈预测一次只能处理一对帧，因此需要代价高昂的全局对齐（Global Alignment）步骤。加入光束法平差后，VGGT + BA 的性能进一步大幅提升，在 IMC 上取得当前最佳结果：AUC@10 从 71.26 提升至 84.91，AUC@3 从 39.23 提升至 66.37。本文模型直接预测三维点，可作为 BA 的初始化。因此无需像 [125] 那样执行三角化和 BA 迭代细化。由此，VGGT + BA 比 [125] 快得多。

D. 定性示例

单视图重建的定性示例进一步见图 7。

E. 相关工作

本节进一步讨论相关工作。

视觉 Transformer。 Transformer 架构最初针对语言处理任务提出 [6, 22, 120]，随后由 ViT [27] 引入计算机视觉领域，并得到广泛采用。凭借结构简洁、容量大、灵活性强以及捕获长程依赖的能力，视觉 Transformer 及其变体此后成为各类计算机视觉任务架构设计的主流选择 [4, 12, 83, 137]。

图 7. **基于点图估计的单视图重建**。 DUSt3R 需要将一张图像复制为图像对，而本文模型可直接从单张输入图像预测点图。模型对未见过的真实世界图像表现出很强的泛化能力。

方法	测试时优化	AUC@3°	AUC@5°	AUC@10°	运行时间
COLMAP (SIFT+NN) [94]	✓	23.58	32.66	44.79	>10s
PixSfM (SIFT + NN) [66]	✓	25.54	34.80	46.73	>20s
PixSfM (LoFTR) [66]	✓	44.06	56.16	69.61	>20s
PixSfM (SP + SG) [66]	✓	45.19	57.22	70.47	>20s
DFSfM (LoFTR) [47]	✓	46.55	58.74	72.19	>10s
DUSt3R [129]	✓	13.46	21.24	35.62	∼ 7s
MASt3R [62]	✓	30.25	46.79	57.42	∼ 9s
VGGSfM [125]	✓	45.23	58.89	73.92	∼ 6s
VGGSfMv2 [125]	✓	59.32	67.78	76.82	∼ 10s
VGGT（本文）	✗	39.23	52.74	71.26	**0.2s**
VGGT + BA（本文）	✓	**66.37**	**75.16**	**84.91**	1.8s

表 10. **IMC [54] 上的相机位姿估计**。本文方法在具有挑战性的地标旅游照片数据上取得当前最佳性能，优于在最新 CVPR'24 IMC 挑战赛相机位姿（旋转和平移）估计任务中排名第一的 VGGSfMv2 [125]。

DeiT [114] 证明，采用强数据增强策略可在 ImageNet 等数据集上有效训练视觉 Transformer。DINO [10] 揭示了视觉 Transformer 通过自监督学习获得的特征所具有的独特性质。CaiT [115] 引入层缩放以应对更深层视觉 Transformer 的训练难题，有效缓解了梯度相关问题。此外，QKNorm [48, 150] 等技术也被用于稳定训练过程。文献 [138] 还探索了目标跟踪中帧内注意力模块与全局注意力模块之间的动态关系，但其采用交叉注意力。

相机位姿估计。 从多视图图像估计相机位姿是三维计算机视觉中的关键问题。过去几十年间，无论是增量式方法 [2, 36, 94, 103, 134] 还是全局方法 [3, 14–17, 52, 73, 79, 81, 90, 106]，运动恢复结构（SfM）都已成为主流方案 [46]。近期，一系列方法将相机位姿估计视为回归问题 [65, 100, 109, 112, 113, 118, 122, 123, 131, 152, 153, 160]，并在稀疏视图设置下取得了良好结果。Ace-Zero [5] 进一步提出回归三维场景坐标，而 FlowMap [101] 则聚焦深度图，二者均将其作为相机预测的中间表示。VGGSfM [125] 则将经典 SfM 流程简化为可微框架，并展现出很高的性能，尤其是在地标旅游照片数据集上。与此同时，DUSt3R [62, 129] 提出学习像素对齐点图的方法，从而可通过简单对齐恢复相机位姿。这一范式转变引起广泛关注，因为点图这种过参数化表示可与三维高斯泼溅等多种下游应用无缝结合。

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[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):105–112, 2011. 2, 11

[3] Mica Arie-Nachimson, Shahar Z Kovalsky, Ira Kemelmacher-Shlizerman, Amit Singer, and Ronen Basri. Global motion estimation from point matches. In *2012 Second international conference on 3D imaging, modeling, processing, visualization & transmission*, pages 81–88. IEEE, 2012. 11

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 10

[5] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2, 11

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 10

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

[8] Ang Cao, Justin Johnson, Andrea Vedaldi, and David Novotny. Lightplane: Highly-scalable components for neural 3Dfields. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2025. 8

[9] Chenjie Cao and Yanwei Fu. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12129–12139, 2023. 6

[10] 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, 11

[11] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2021. 2

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 10

[13] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. *Proc. ECCV*, 2024. 2, 8

[14] David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2841–2853, 2012. 11

[15] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1212–1221, 2017.

[16] Zhaopeng Cui and Ping Tan. Global structure-from-motion by similarity averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 864–872, 2015.

[17] Zhaopeng Cui, Nianjuan Jiang, Chengzhou Tang, and Ping Tan. Linear global translation estimation with feature tracks. *arXiv preprint arXiv:1503.01832*, 2015. 11

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5, 6, 7

[19] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 4

[20] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 5, 8

[21] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 10

[23] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *arXiv*, 2022. 2, 8

[24] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. *arXiv*, 2306.08637, 2023. 2

[25] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: tracking any point with per-frame initialization and temporal refinement. In *Proc. CVPR*, 2023. 3, 8

[26] Carl Doersch, Yi Yang, Dilara Gokay, Pauline Luc, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ross Goroshin, João Carreira, and Andrew Zisserman. Bootstap: Bootstrapped training for tracking-any-point. *arXiv preprint arXiv:2402.00847*, 2024. 8

[27] 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 16×16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021. 10

[28] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 8

[29] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, and Kevin Stone. The Llama 3 herd of models. *arXiv*, 2407.21783, 2024. 2

[30] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. *arXiv*, 2409.19152, 2024. 6

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

[32] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6

[33] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 6, 7, 10

[34] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. CVPR*, 2021. 2

[35] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[36] 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 *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 368–381. Springer, 2010. 2, 11

[37] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2

[38] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[39] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2015. 2

[40] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. 6

[41] 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, Dmitry Lagun,

Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *Proc. CVPR*, 2022. 5, 8

[42] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2

[43] Junlin Han, Jianyuan Wang, Andrea Vedaldi, Philip Torr, and Filippos Kokkinos. Flex3d: Feed-forward 3d generation with flexible reconstruction model and input view curation. *arXiv preprint arXiv:2410.00890*, 2024. 8

[44] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *Proc. ECCV*, 2022. 2, 8

[45] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 1, 2

[46] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004. 11

[47] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *arxiv*, 2023. 11

[48] Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245*, 2020. 10, 11

[49] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *Proc. ICLR*, 2024. 2, 8

[50] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[51] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 6

[52] Nianjuan Jiang, Zhaopeng Cui, and Ping Tan. A global linear method for camera pose registration. In *Proceedings of the IEEE international conference on computer vision*, pages 481–488, 2013. 11

[53] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. LVSM: a large view synthesis model with minimal 3D inductive bias. *arXiv*, 2410.17242, 2024. 4, 8

[54] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 10, 11

[55] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 2

[56] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *Proc. ECCV*, 2024. 2, 7, 8

[57] Nikita Karaev, Ignacio Rocco, Ben Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3, 5, 8

[58] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *Proc. ICRA*. IEEE, 2016. 5

[59] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Proc. NeurIPS*, 2017. 5

[60] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. In *CVPR*, 2024. 2

[61] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009. 4

[62] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1, 2, 6, 10, 11

[63] Hongyang Li, Hao Zhang, Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, and Lei Zhang. Taptr: Tracking any point with transformers as detection. *arXiv preprint arXiv:2403.13042*, 2024. 2, 8

[64] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 5

[65] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 11

[66] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. *arXiv.cs*, abs/2108.08291, 2021. 6, 10, 11

[67] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at

light speed. *arXiv preprint arXiv:2306.13643*, 2023. 2

[68] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: local feature matching at light speed. In *Proc. ICCV*, 2023. 7

[69] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5

[70] 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 *European Conference on Computer Vision*, pages 249–269. Springer, 2025. 2

[71] Manuel Lopez-Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota BulÃ², Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5

[72] Zeyu Ma, Zachary Teed, and Jia Deng. Multi-view stereo with cascaded epipolar raft. In *European Conference on Computer Vision*, pages 734–750. Springer, 2022. 2

[73] Pierre Moulon, Pascal Monasse, and Renaud Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of the IEEE international conference on computer vision*, pages 3248–3255, 2013. 11

[74] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3504–3515, 2020. 2

[75] David Novotný, Diane Larlus, and Andrea Vedaldi. Learning 3D object categories by looking around them. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 5

[76] David Novotný, Diane Larlus, and Andrea Vedaldi. Capturing the geometry of object categories from video supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 5

[77] John Oliensis. A critique of structure-from-motion algorithms. *Computer Vision and Image Understanding*, 80(2):172–214, 2000. 2

[78] 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, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 4, 10

[79] Onur Ozyesil and Amit Singer. Robust camera location estimation by convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2674–2683, 2015. 11

[80] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017. 2

[81] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 11

[82] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 5

[83] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 10

[84] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8645–8654, 2022. 2

[85] Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky TQ Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, et al. Theseus: A library for differentiable nonlinear optimization. *Advances in Neural Information Processing Systems*, 35:3801–3818, 2022. 9

[86] 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*, pages 8748–8763, 2021. 2

[87] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3, 5, 10

[88] 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. 5, 6

[89] 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 *International Conference on Computer Vision (ICCV) 2021*, 2021. 5

[90] Rother. Linear multiview reconstruction of points, lines, planes and cameras using a reference plane. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1210–1217. IEEE, 2003. 11

[91] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80, 2008. 2

[92] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 6

[93] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: learning feature matching with graph neural networks. In *Proc. CVPR*, 2020. 7

[94] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 10, 11

[95] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 2, 6

[96] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 2

[97] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 6, 7

[98] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024. 9

[99] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12517–12526, 2022. 2

[100] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21349–21359, 2023. 11

[101] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024. 11

[102] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. FlowMap: high-quality camera poses, intrinsics, and depth via gradient descent. *arXiv*, 2404.15259, 2024. 2

[103] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2, 11

[104] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[105] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 6, 7, 10

[106] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE international conference on computer vision*, pages 801–809, 2015. 11

[107] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[108] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10208–10217, 2024. 8

[109] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018. 2, 11

[110] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 8

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

[112] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 2, 11

[113] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34: 16558–16569, 2021. 2, 11

[114] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 11

[115] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 10, 11

[116] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2

[117] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4), 1991. 7

[118] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 2, 11

[119] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017. 2, 4

[120] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 10

[121] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pages 14194–14203, 2021. 6

[122] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2021. 2, 11

[123] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 11

[124] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: solving pose estimation via diffusion-aided bundle adjustment. In *Proc. ICCV*, 2023. 6

[125] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: visual geometry grounded deep structure from motion. In *Proc. CVPR*, 2024. 1, 2, 5, 6, 9, 10, 11

[126] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. PF-LRM: pose-free large reconstruction model for joint pose and shape prediction. *arXiv.cs*, abs/2311.12024, 2023. 8

[127] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state, 2025. 2, 6

[128] 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. 2

[129] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 1, 2, 3, 4, 5, 6, 10, 11

[130] Yuesong Wang, Zhaojie Zeng, Tao Guan, Wei Yang, Zhuo Chen, Wenkai Liu, Luoyuan Xu, and Yawei Luo. Adaptive patch deformation for textureless-resilient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1621–1630, 2023. 2

[131] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 230–247. Springer, 2020. 2, 11

[132] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. MeshLRM: large reconstruction model for high-quality mesh. *arXiv*, 2404.12385, 2024. 4

[133] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5610–5619, 2021. 2

[134] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 2, 11

[135] 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. 5

[136] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 8

[137] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 10

[138] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8760, 2022. 11

[139] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *AAAI*, 2020. 6

[140] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation. *arXiv*, 2403.14621, 2024. 8

[141] 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. *arXiv preprint arXiv:2501.13928*, 2025. 2, 6, 9

[142] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proc. CVPR*, 2024. 2

[143] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 10

[144] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 6

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

[146] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 5

[147] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2

[148] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G., Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H. Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv.cs*, abs/2305.10435, 2023. 2

[149] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In *Proc. ECCV*, 2016. 2

[150] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023. 11

[151] 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. 10

[152] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, pages 592–611. Springer, 2022. 11

[153] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. 11

[154] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 8

[155] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: large reconstruction model for 3D Gaussian splatting. *arXiv*, 2404.19702, 2024. 8

[156] 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, 2025. 2, 6

[157] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *CVPR*, 2023. 2, 6

[158] 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 Transactions on Instrumentation and Measurement*, 72:1–16, 2023. 7

[159] 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 *ICCV*, 2023. 5

[160] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2, 11

[161] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 6