arXiv:2611.32297v1 [cs.CV] 24 Nov 2026

MaskedVGGT：掩码多视图变换器是可扩展的三维学习器

Zichen Zou[1,2], Xiaosong Jia[1†], Wen-Ming Chen[1,2], ZuxuanWu[1,2], Yu-Gang Jiang[1]可信具身人工智能研究院（TEAI），复旦大学[2]上海创新研究院

†通讯作者

摘要

多视图前馈模型展现出很强的可扩展性，但它们依赖代价高昂的几何标注，这限制了数据扩展。现有三维自监督方法通过跨视图补全（例如 CroCo、MuM、Muskie）或新视图合成（例如 RayZer、E-RayZer）减少了这种依赖，通常会联合学习图像表征与跨视图交互。在这项工作中，我们认为，**从头开始学习单张图像表征对于三维预训练方法而言并无必要**，因为这样做无法利用二维基础模型（例如 DINO、CLIP 和前沿视觉语言模型）的持续进步。为此，我们提出 MaskedVGGT：冻结一个强大的预训练二维编码器，仅对新增的随机初始化多视图变换器进行预训练，重点学习视图间的交互。

具体而言，MaskedVGGT 通过无标注的掩码自蒸馏对多视图变换器进行预训练：学生根据完整的锚点视图和其余经过高度掩码处理的视图，预测掩码位置处的目标；指数移动平均教师模型则根据未掩码输入提供预测目标。在参数规模最高达 1B 的模型实验中，MaskedVGGT 预训练后再进行微调，其表现优于使用相同训练数据、模型架构和优化更新总次数进行的单阶段监督式 VGGT-Ω 训练。MaskedVGGT 的表现也优于 E-RayZer、CroCo、MuM 和 Muskie，这证明了基于强大二维表征构建多视图预训练的有效性。最后，利用额外的无标签视频扩展预训练能够进一步提升泛化能力，学习到的表征也可迁移至多样化的下游任务。这些结果表明，一种有效的学习分工是：复用二维预训练得到的强大图像表征，并从无标签多视图数据中学习跨视图交互，而不依赖几何标注。

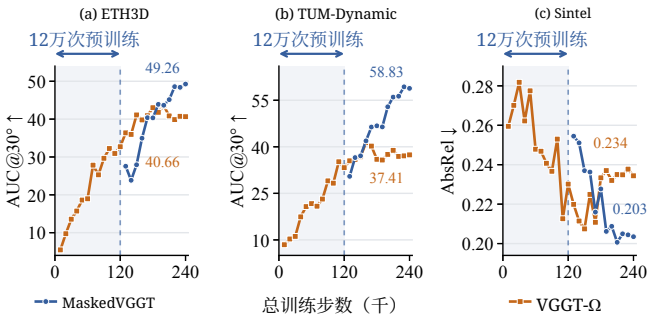

图 1 预训练可提升三维泛化能力。 经过 24 万次总更新后，MaskedVGGT 在 ETH3D 和 TUM-Dynamic 上取得了更高的位姿准确率，在 Sintel 上的深度误差也更低
相比之下，直接监督训练的 VGGT-Ω-1B 的分数在训练后期趋于平稳。两个模型使用相同的架构和训练数据。阴影区域标示了我们进行自监督预训练的前 12 万次更新。评估详情见第 4.3 节。

1 引言

三维重建经历了从基于优化的流程 [106, 120],到学习式成对重建 [84, 156], 再到联合多视图预测 [78,150]的发展。近期的前馈模型随着模型规模增大而性能提升，展现了扩展几何学习的潜力 [151]。然而，几何数据需要成本高昂的标注，因而成为扩展几何训练数据的瓶颈。

大量无标注互联网视频提供了丰富的数据来源，推动了面向多视图三维模型的自监督学习。SelfEvo [64] 和 VGGT-Ω [151]利用无标注视频进一步提升监督训练模型的性能，但这些模型仍须先使用带标注的数据进行训练。正如 VGGT-Ω中所述，“我们原以为自监督重建能够从零开始奏效，但我们最终成功的方法仍然需要一个（监督）预训练模型。”鉴于预训练在语言 [36, 111] 和二维视觉 [22, 56, 57]领域取得的成功，探索无监督三维预训练十分值得。CroCo [164]和 MuM [100]等现有的自监督三维预训练方法，使用跨视图重建解码器对图像编码器进行预训练；而 Muskie [86] 和 E-RayZer [69, 191]则对原生多视图骨干网络进行预训练，这类网络通常会联合学习图像表征和跨视图交互。我们认为，三维预训练应建立在二维视觉的进展之上，而非从头重新学习图像表征，从而能够直接使用快速发展的二维基础模型（例如 DINO [22], 、CLIP [112], 和前沿视觉语言模型）。

为此，我们提出 MaskedVGGT，这是一种面向多视图三维模型的自监督预训练框架。该框架冻结预训练二维图像编码器，并

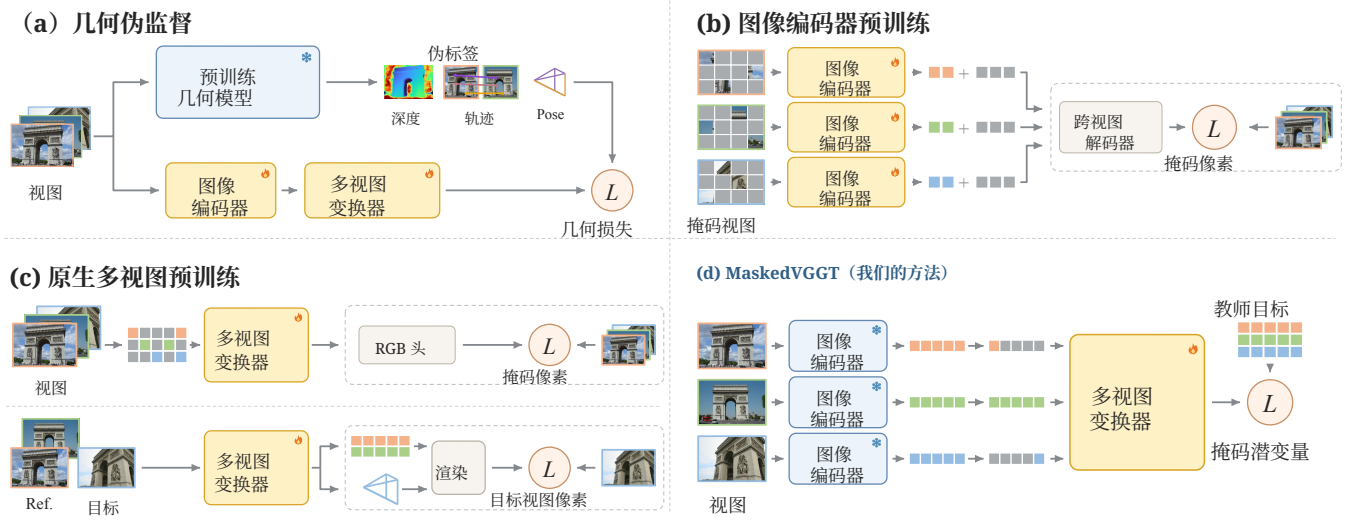

图 2 多视图几何的预训练。 (a) 几何伪监督 [30, 71] 使用预训练几何模型的预测结果作为标签。(b) 图像编码器预训练 [100, 164] 通过利用其他视图的信息补全一个视图，从头开始学习图像特征。(c) 原生多视图预训练通过掩码重建 [86] (上) 或新视图合成 [69, 191] (下)，从头开始联合学习图像表征和跨视图交互。(d) 我们提出的 MaskedVGGT 重用冻结的预训练二维编码器提取的特征，并且仅训练随机初始化的多视图变换器，以学习独立编码图像之间的跨视图交互。

训练新添加且随机初始化的多视图变换器。具体而言，我们通过移动平均教师进行掩码自蒸馏来训练学生。对于学生，我们在二维图像编码器之后对特征进行掩码，保留部分视图完全可见作为锚点，并对其余视图施加大量掩码。教师在不进行掩码的情况下处理所有视图，以生成目标。为构建上下文化监督，我们结合来自多个层的归一化教师特征。按照 CAPI [34],，我们将这些特征转换为在线聚类分配，而非直接对其进行回归；此外，我们在每个图像块位置分别平衡分配，以防止学生利用位置捷径。

在实验中，对于最大规模达 1B 参数的模型，所提出的两阶段预训练-微调方法优于使用相同架构、数据和优化更新总步数的单阶段监督 VGGT-Ω训练。图 1 展示了在 240k 更新步数内两种 L 规模的训练轨迹。与 E-RayZer、CroCo、MuM 和 Muskie 等基线方法的比较表明，基于预训练二维模型构建方法是有效的。此外，使用未标注视频扩展预训练可提升泛化能力，而学习到的表征也有利于多种下游任务。

总而言之，我们做出了三项贡献：

- 我们提出了 MaskedVGGT，这是一种多视图三维预训练框架，采用冻结的预训练二维编码器，并将自监督学习的重点放在随机初始化的多视图变换器上。
- 我们提出了一种掩码跨视图自蒸馏方法，将非对称特征掩码与上下文化、均衡的潜在目标相结合，无需几何标注或监督教师。
- 我们展示了在相同更新步数下，该方法在不同规模和架构上的重建效果均有所提升，并且能够从额外的无标注视频中获益，并迁移到多种下游视觉任务。

2 相关工作

三维重建。 跨视图对应关系是经典运动恢复结构和多视图立体视觉的基础 [1, 55, 106, 120, 121, 128, 137]。通过局部特征 [35, 38, 138], 图像匹配 [23, 39, 40, 90, 119, 130], 可微几何估计 [15, 132–134, 139, 148, 149, 162], 以及学习型立体视觉 [53, 65, 175, 190], 学习逐步将这种推理内化 [84, 156] 。DUSt3R 及相关点图模型 [18, 78, 150, 159, 172, 189] 转向直接的成对预测；联合多视图骨干网络 [74, 124, 144], 将前馈推理扩展到图像集合。后续工作致力于高效聚合 [82, 145, 153, 167, 196], 面向流式输入的持久表示 [63, 73, 88, 185, 186]以及动态场景Ω 。这些方向共同推动了能够整合跨视图观测的架构不断发展。VGGT [70, 89, 151] 及相关扩展研究进一步展现了模型容量和训练数据的价值，同时保留几何监督。我们在这一架构进展的基础上，研究无标注预训练如何改进前馈三维重建模型后续的重建训练。

自监督视觉预训练。 早期工作探索了手工设计的视觉前置任务 [37, 48, 102, 188]。其他方法类别包括对比学习 [25, 27, 56, 58, 104,168], 、聚类 [3, 19, 21], 以及蒸馏或特征正则化 [11, 22, 26, 52, 182]。在预测学习领域，MAE 及相关自编码方法 [10, 24, 57, 107,

142, 170] 围绕缺失的图像内容组织学习。后续设计在掩码方式 [75, 85, 146] 和监督信号上各有不同，涵盖从手工设计或离散目标到预训练特征等形式 [43, 44, 62, 108, 109, 160, 161, 171]。iBOT 及相关的基于教师的方法 [4, 8, 9, 29, 92, 193]则在预训练过程中学习预测目标；CAPI 及相关工作致力于改善潜在目标的质量 [34, 141, 163], , 而寄存器令牌 [33, 72] 和局部目标则有助于提升稠密表示 [12, 59, 103, 166,180]。与这些算法发展并行，大规模自监督学习产生了强大的单图像先验 [20, 42, 50, 51, 105, 125, 135, 143], , 视觉语言学习和模型蒸馏则对其加以补充 [14, 31,68, 112, 113, 183, 184]。预测学习也扩展到了视频 [5, 13, 45, 54, 67, 118, 136, 152, 154, 158] 和统一多模态模型 [6, 7, 99]。这些方向为我们的方法奠定了两项基础：成熟的掩码预测和自蒸馏目标，以及预训练二维图像编码器；在学习用于三维的多视图表示时，可保留其视觉知识。

多视图几何的预训练。 CroCo 在图像对上引入了跨视图补全；后续工作进一步发展了双目和多视图补全、对应关系以及共可见性目标 [2, 86, 94,95, 100, 164, 165]。它们的架构将跨视图推理置于独立图像编码器之上的解码器中（CroCo、MuM），或直接置于联合编码器中（Muskie）。这种区别涉及交互发生的位置：基于解码器的方法同样可以在下游任务中保留其交互模块。互补性工作研究了几何感知图像表示及其迁移 [41, 61, 81, 123]。另一条路线通过渲染进行学习，使用神经场或显式场景表示 [79, 97, 177]。其监督范围从带位姿视图或预训练几何模型 [60, 71, 76, 127, 179, 187] , 到基于学习相机的合成 [28, 47, 66, 69, 98, 110, 117, 147, 155, 191]。因此，图像先验的使用方式与几何监督的来源是相互独立的设计选择。几何伪标签为利用RGB图像集合进行训练提供了另一条途径 [30, 173, 174, 192], , 而自蒸馏和无标注后训练则进一步优化已经训练好的重建器 [64, 151, 195]。我们的方法保留一个冻结的图像编码器，并在无位姿RGB视图上对多视图聚合器进行预训练，以实现跨视图交互。

3 方法

3.1 总体框架

图3 总结了我们的预训练框架。在本文中，MaskedVGGT 主要采用 VGGT-Ω 神经网络架构，而只要存在多视图变换器模块，预训练方法便具有通用性。形式化地说，给定来自场景或视频片段的图像 $\mathrm{I} = \{I_i\}_{i=1}^{N}$，预训练二维图像编码器（即 DI- NOv3） E_ϕ 会独立地将每幅图像 I_i 编码为 P d 维图像块令牌 $\mathcal{X} = \{E_\phi(I_i) \in \mathrm{R}^{P\times d}\}_{i=1}^{N}$。然后，多视图 Transformer 聚合器 A_θ 将额外的相机令牌和寄存器令牌与 $\mathcal{X}$ 拼接起来，并交替使用逐帧注意力与全局注意力/寄存器注意力。最后，聚合器输出多视图特征 $(Z_1, \ldots, Z_N) = A_\theta(\mathcal{X})$。在提出的预训练阶段，我们冻结预训练二维图像编码器，并训练随机初始化的变换器聚合器，以学习跨视图交互。在微调阶段，我们添加下游任务头，并使用标注数据对整个模型进行联合训练。

随机初始化的聚合器通过掩码教师-学生自蒸馏进行预训练，该方法维护聚合器的两个副本。教师模型 $A_{\theta_{\text{teacher}}}$ 是学生模型 $A_{\theta_{\text{student}}}$的指数移动平均（EMA）。学生模型选择部分视图作为锚点视图，并将其他视图中的大多数图像块令牌替换为学习得到的掩码令牌，从而得到 $\mathcal{X}_{\text{mask}}$。学生聚合器 $A_{\theta_{\text{student}}}$ 处理$\mathcal{X}_{\text{mask}}$，而教师聚合器 $A_{\theta_{\text{teacher}}}$ 接收未掩码的集合 $\mathcal{X}$。它们分别生成多视图表示 $\mathcal{Z}^{\text{s}}$ 和 $\mathcal{Z}^{\text{t}}$:

$$\mathcal{Z}^{\text{s}} = A_{\theta_{\text{student}}}(\mathcal{X}_{\text{mask}}) \qquad \mathcal{Z}^{\text{t}} = A_{\theta_{\text{teacher}}}(\mathcal{X}) \tag{1}$$

使用根据教师表示构造的目标，我们在掩码位置通过掩码预测损失 L_{pred} 优化学生聚合器，而教师则遵循指数移动平均（EMA）更新规则:

$$\begin{aligned} &\text{Student:} \quad \min_{\theta_{\text{student}}} \mathcal{L}_{\text{pred}}, \\ &\text{Teacher:} \quad \theta_{\text{teacher}} \leftarrow \mu_t \theta_{\text{teacher}} + (1-\mu_t)\theta_{\text{student}}. \end{aligned} \tag{2}$$

在下文第 3.2 节中，我们将详细介绍掩码、目标构造和损失。

3.2 掩码多视图自蒸馏

预训练将非对称特征掩码与上下文化预测目标相结合。

非对称特征掩码。在每次迭代中，冻结的二维图像编码器处理完每张完整图像 I_i 后，我们对 $X_i = E_\phi(I_i)$ 应用掩码。每个视图都以 p_a (= 0.2 在我们的实验中取得了最佳平衡）的概率独立地被选为完全可见锚点。我们确保每个场景至少有一个锚点。这些锚点视图为预测其他视图中的掩码特征提供上下文。在每个非锚点视图中，我们以 ρ (=0.95 可取得最佳性能）的比例随机掩码图像块特征，保留 $\lfloor(1-\rho)P\rfloor$ 个未掩码图像块，从而构造一项具有挑战性的特征预测任务。设 x_{ip} 为视图 i 中图像块 p 的特征， $m_{ip} = 1$ 表示一个被掩码的位置。学生接收 $x_{ip,\ \text{mask}} = (1-m_{ip})x_{ip} + m_{ip}e_m$，其中 e_m 是在所有输入视图间共享的可学习掩码令牌。

上下文化预测目标。 为掩码位置提供监督，教师聚合器 $A_{\theta_{\text{teacher}}}$ 会处理完整的特征集合 $\mathcal{X}$。令 $g_{ip}^{\ell} \in \mathrm{R}^d$ 表示视图 i 中图像块 p 在第 l 个块末尾的特征。为防止高幅值层

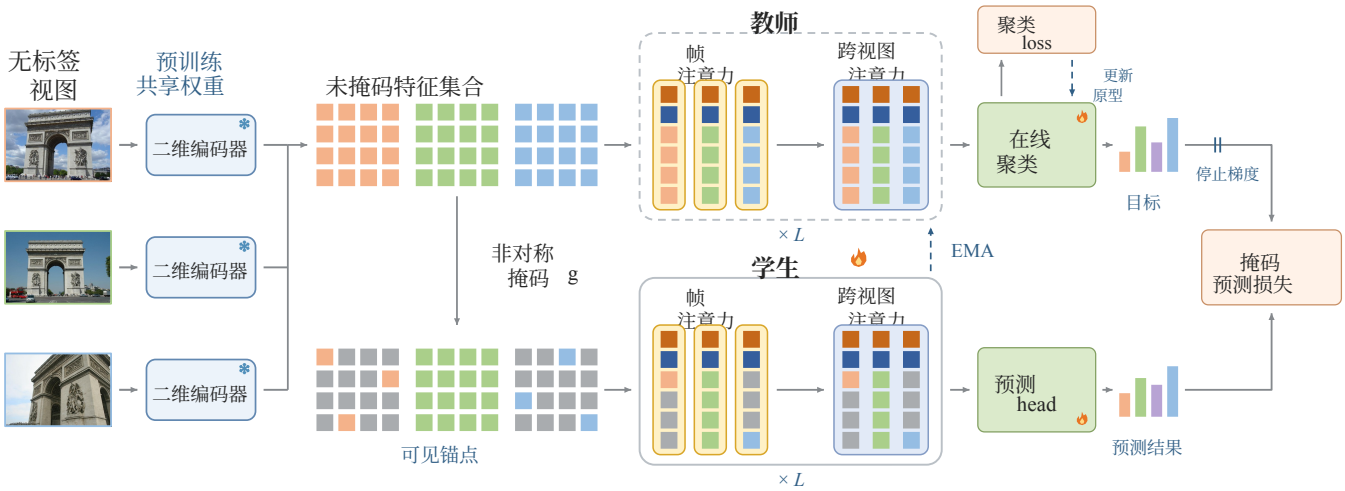

图3 MaskedVGGT 预训练。共享的冻结预训练二维编码器从每个视图中提取特征。教师聚合所有视图的未掩码特征，而学生聚合完整的锚视图特征以及其余视图中大量掩码后的特征。在掩码位置，学生预测根据教师特征得出的平衡目标聚类分配。教师权重是学生权重的指数移动平均。图4 详细展示了目标构造和学生预测。聚类损失仅更新原型 C，而掩码预测损失在目标保持不变的情况下更新学生。

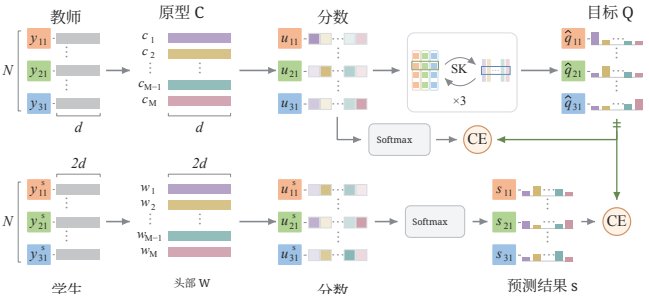

图 4 目标构造与预测。对教师未掩码特征进行在线聚类，得到平衡目标 Q（上），供学生在掩码位置进行预测（下）。上方的交叉熵仅更新原型 C；下方的交叉熵则更新学生及其预测头W 颜色区分不同视图和簇；颜色越深、柱条越高，表示数值越大。

主导目标，遵循 data2vec [8]，，我们对由 L 个块组成的聚合器中最后 K 个块的归一化特征取平均值（= 8 的性能最佳）

$$\text{tor } y_{ip} = \frac{1}{K}\sum_{\ell=L-K+1}^{L} \mathrm{LN}\left(g_{ip}^{\ell}\right).$$

遵循 CAPI [34], 的方法，我们对教师的未掩码特征进行在线聚类，并训练学生在掩码图像块位置预测由此得到的聚类分配（图4）。

我们学习 M 簇原型，并将其存储为 $C \in \mathrm{R}^{M\times d}$ 的行。每个上下文教师特征 y_{ip}都会被分离、进行 L2 归一化，并与这些原型进行比对，得到 u_{ip}。为避免结果塌缩到少数几个簇，我们使用 Sinkhorn–Knopp (SK) [126] 进行三次迭代，为每个图像块生成涵盖 M个簇的目标分布 $\widehat{q}_{ip}$：

$$\begin{aligned} u_{ip} &= C\frac{y_{ip}}{\|y_{ip}\|_2} \in \mathbb{R}^M \\ \{\widehat{q}_{ip}\} &= \mathrm{SK}(\{u_{ip}\}) \quad \widehat{q}_{ip} \in \mathbb{R}^M \end{aligned} \tag{3}$$

对于每个空间图像块位置，SK 会交替对局部小批次中的视图以及簇原型进行归一化。这可以避免学生仅根据输入图像块的空间坐标来预测簇分配。

我们独立于学生训练原型矩阵 C。对于所有教师图像块，我们最小化平衡目标 $\widehat{q}_{ip}$ 与温度为 τ_t 时的簇概率 $q_{ip} = \mathrm{softmax}(u_{ip}/\tau_t)$ 之间的交叉熵：

$$\mathcal{L}_{\text{cluster}} = -\frac{1}{NP}\sum_{i,p} \widehat{q}_{ip}^{\mathsf{T}} \log q_{ip}. \tag{4}$$

该损失产生的梯度仅更新 C，教师特征和平衡目标 $\widehat{q}_{ip}$ 均停止梯度。

学生在掩码位置预测停止梯度的目标 $\widehat{q}_{ip}$。VGGT-Ω 读出模块将最终的逐帧特征与跨视图图像块特征拼接为 $y_{ip}^{s} \in \mathrm{R}^{2d}$。线性头 W 将其映射为 M 簇分数 u_{ip}^{s}，再通过 softmax 得到概率 s_{ip}。我们对小批次中掩码图像块位置集合 M 上的交叉熵取平均：

$$\mathcal{L}_{\text{pred}} = -\frac{1}{\max(1,|\mathcal{M}|)}\sum_{(i,p)\in\mathcal{M}} \widehat{q}_{ip}^{\mathsf{T}} \log s_{ip}. \tag{5}$$

预测损失会更新学生聚合器、掩码嵌入和预测头。教师权重按照学生权重的指数移动平均进行更新（公式2）。

4 实验

我们比较三维重建训练范式，以评估预训练策略的有效性，并考察预训练设计、编码器选择和学习到的特征（第 4.2 节）。随后，我们研究扩展到更大模型、在不同重建架构之间进行迁移，以及

表 1 三维重建的训练范式。 所有方法均在相同的 DL3DV 数据上，使用规模相当的 S 规模骨干网络重新训练：直接监督/伪监督训练进行 120k 次更新，或先进行 60k 次预训练，再进行 60k 次几何微调。

方法	DL3DV		ETH3D		Sintel	
	曲线下面积@30◦ ↑	绝对相对误差↓	曲线下面积@30◦ ↑	绝对相对误差↓	曲线下面积@30◦ ↑	绝对相对误差↓
直接几何监督						
VGGT-Ω (从头开始) [151]	18.52	0.2203	1.16	0.2034	7.22	0.4934
VGGT-Ω (DINOv3) [151]	35.89	0.1347	4.15	0.1000	5.98	0.3805
几何伪监督						
Reliev3R [30]	11.33	0.2028	1.41	0.1481	10.77	0.4461
AnySplat [71]	22.68	0.2283	1.30	0.2075	6.33	0.4835
单图像编码器预训练						
CroCo v2 [165]	24.26	0.1818	4.10	0.1431	8.79	0.4727
MuM [100]	25.42	0.1817	3.65	0.1466	7.91	0.4567
Gekko [95]	25.15	0.1801	2.50	0.1436	9.60	0.4656
原生多视图骨干网络预训练						
Muskie [86]	42.68	0.1813	3.24	0.1672	9.47	0.4525
E-RayZer [191]	46.87	0.1965	5.77	0.1879	10.62	0.4674
NAS3R [66]	28.66	0.3509	1.84	0.2902	8.29	0.5452
RayDer [110]	42.95	0.1998	3.52	0.1795	12.04	0.4597
在冻结的二维编码器上进行多视图预训练						
MaskedVGGT	**54.23**	**0.1226**	**9.28**	**0.0997**	**12.36**	**0.3277**

无标注视频、更长时间的预训练和训练资源分配的影响（第 4.3 节）。最后，我们评估学习到的表征向多种下游任务的迁移效果（第 4.4 节），并通过消融实验考察预训练的预测目标和掩码设计选择（第 4.5 节）。

4.1 实验设置

实现细节。 在大多数实验中，Masked- VGGT 使用 S 规模的 VGGT-Ω 架构 [151]，，配备 DINOv3 ViT-S/16 图像编码器和一个具有 12 个注意力组的 384 维多视图聚合器，以便快速验证想法。实验使用两块 A800 进行，每个场景采样 2–24 个视图，每个图形处理器每批最多处理 24 张图像。我们仅使用 RGB 预训练 60k 个更新步数，然后按照 VGGT-Ω，使用相机和深度监督微调 60k 个更新步数。在扩展实验中，我们使用 16 块 A800，在 20 天内将规模扩展至 10 亿个参数。

数据集。 我们的大多数实验使用 DL3DV [91]。为了扩展规模，我们遵循 VGGT-Ω，混合使用真实场景和合成场景。B 尺度训练使用十个数据源：DL3DV、TartanAir [157], ScanNet [32], Dynamic Replica [77],CO3D [115], BlendedMVS [176], MegaDepth [87],Waymo [131], ScanNet++ [178]和 Virtual KITTI 2 [17]。L 尺度训练另外使用十个数据集：MVS-Synth [65], Hy- persim [116], Mapillary Metropolis [96], Mid-Air [46],Taskonomy [181], ParallelDomain-4D [140], Replica [129], WildRGB-D [169], MegaSynth [70],和 EDEN [83]。

4.2 三维重建预训练

训练范式比较。我们比较了用于三维重建的监督、伪监督和自监督训练范式，并使用相同的训练数据和相同的总更新步数（表 1）。

表 2 预训练策略与编码器选择。 在 DL3DV 上训练的 S 尺度模型，预训练 60k 步，随后进行最多 60k 步的微调更新，并采用几何监督。

设置	曲线下面积@30◦ ↑	绝对相对误差↓
预训练策略		
图像编码器预训练	25.22	0.1897
原生多视图预训练	28.23	0.1945
MaskedVGGT（我们的方法）	**54.23**	**0.1226**
预训练期间的编码器更新		
编码器更新	28.29	0.2000
编码器固定	**54.23**	**0.1226**
冻结图像编码器的选择		
PE-Spatial-S	22.26	0.1996
MetaCLIP 2-S	25.32	0.2014
MoCo v3-S	52.32	0.1647
DINOv3-S	**54.23**	**0.1226**

遵循 VGGT-Ω [151]，我们在 DL3DV、ETH3D [122]，和 Sintel [16] 上评估三维重建，每个场景使用十个视图。我们报告位姿曲线下面积@30°（相对旋转和平移方向）和深度 AbsRel（逐帧缩放与平移对齐后的平均绝对相对误差）。MaskedVGGT 在报告的全部六项指标上表现最佳。

预训练设计与学习到的表征。 我们考察预训练策略和图像编码器的选择如何影响下游三维重建，并采用相同的 Ω-S 聚合器和训练计划（表 2）。

为进行更公平的比较，我们将图像编码器和原生多视图预训练适配到我们的框架中。我们采用与所提方法相同的掩码策略、EMA 目标构造和自蒸馏目标。预训练

表3 不同模型规模下的重建。VGGT-Ω 和 MaskedVGGT 在每种规模下均采用匹配的架构、数据和总更新次数进行重新训练。模型总参数量包括图像编码器、聚合器、相机头和深度头。

规模	方法	参数	相机姿态：曲线下面积@30。(%) ↑			深度：绝对相对误差 ↓		
			DL3DV	ETH3D	TUM-Dynamic	DL3DV	Sintel	DyCheck
S	VGGT-Ω	121.6M	35.89	4.15	7.78	0.1347	0.3805	0.1887
	MaskedVGGT	121.6M	**54.23**	**9.28**	**11.03**	**0.1226**	**0.3277**	**0.1591**
B	VGGT-Ω	401.3M	36.33	4.58	5.69	0.1354	0.3947	0.1838
	MaskedVGGT	401.3M	**66.45**	**16.81**	**23.23**	**0.1196**	**0.3140**	**0.1536**
L	VGGT-Ω	1.144B	91.32	40.66	37.41	0.0587	0.2344	0.1302
	MaskedVGGT	1.144B	**92.56**	**49.26**	**58.83**	**0.0552**	**0.2035**	**0.0980**

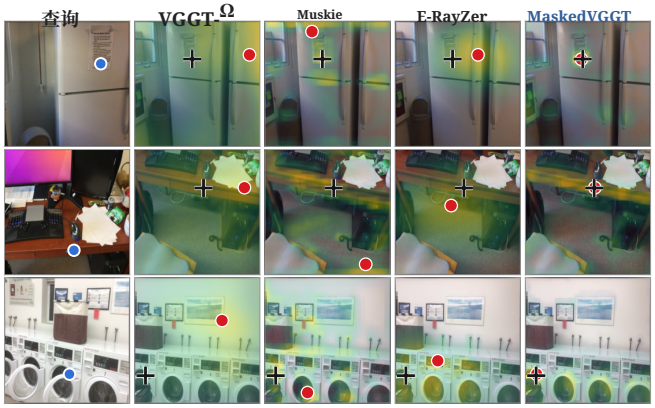

图5 跨视图注意力。 在每个示例中，所有模型接收相同的两张图像。热力图展示每个蓝色查询点对另一视图中图像块的全局注意力；在统一的色标下，颜色越暖表示注意力越高。十字标记真实对应关系，红点标记跨视图注意力峰值。

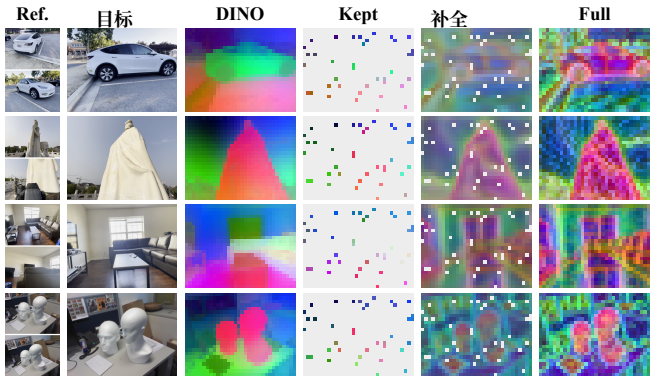

图6 掩码特征补全。 每个示例使用十个视图。参考视图展示两个未掩码的参考视图，目标视图展示八个视图中的一个，其中 95% 的 DINO令牌被掩码以进行补全。图中对 DINO 特征、保留令牌（保留）以及输入为掩码（补全）或完整输入（完整输入）时的聚合器输出进行 PCA可视化。完整输入会保留全部十个未掩码视图。灰色表示可视化中的掩码输入或未显示的特征输出。

基于冻结的二维特征进行的预训练优于图像编码器和原生多视图预训练。在我们的预训练策略中，解冻 DINOv3 会显著降低位姿和深度准确率。重建准确率

表4 跨重建架构的预训练。 所有 S- 规模模型均在 DL3DV 上重新训练，采用 120k 次直接监督更新，或 60k 次预训练和 + 60k 次微调。

架构	直接监督		MaskedVGGT	
	曲线下面积@30。↑	绝对相对误差↓	曲线下面积@30。↑	绝对相对误差↓
VGGT-Ω	35.89	0.1347	**54.23**	**0.1226**
π^3	55.76	0.1241	**74.61**	**0.1086**
MapAnything	27.18	0.1183	**36.53**	**0.1083**
VGGT	45.11	0.1117	**56.79**	**0.1095**

表5 使用无标注视频扩展预训练。仅使用 DL3DV 的基线采用 6 万次预训练更新。加入视频的实验总共进行 12 万次预训练更新，其中包括预计进行的 6 万次 DL3DV 更新。随后，所有模型都在 DL3DV 上进行 6 万次几何微调更新。我们报告相较于仅使用 DL3DV 的基线，相机姿态曲线下面积的提升，单位为百分点。

预训练数据	相机姿态：曲线下面积@30° (%) ↑			
	DL3DV	ETH3D	Sintel	TUM-Dynamic
仅 DL3DV	54.23	9.28	12.36	11.03
+ RE10K	+7.94	+3.02	+0.89	+7.75
+ SA-V	+4.43	+2.26	+0.54	+5.03

不同图像编码器之间差异显著。DINOv3 在这两项指标上表现最佳，而 MoCo v3 的位姿准确率几乎与其持平。

为了考察预训练多视图模块学到了什么，我们将其跨视图注意力和特征补全可视化。注意力可视化结果表明，在冻结的 DINO 特征上进行第一阶段预训练期间，即使没有几何监督，多视图模块也学会了关注不同视图中的对应位置。在所示的 S 尺度示例中，与 VGGT-Ω、Muskie 和 E-RayZer 相比，其全局注意力更集中于真实对应关系周围（图 5）。在特征补全方面，该模块利用这些跨视图关联来恢复掩码特征。借助同一视图中保留的特征以及来自其他视图的上下文信息，即使在 95% 掩码率下，该模块也能恢复出可辨认的物体结构和场景细节（图 6）。输入完整时，这些结构会更加清晰，展现出比 DINO 特征更精细的细节。

表6 使用冻结编码器进行下游评估。我们将经过120k更新步数预训练的MaskedVGGT-L与公开发布的基线方法进行比较（在可用时采用L规模变体；对于E-RayZer，采用多数据集预训练）。对于每项任务，我们都按照相同的协议评估所有模型。在需要训练时，所有模型均使用相同的特定任务模块和更新步数。

方法	三维几何			特征匹配		匹配（RoMa）		分割（SAM）	
	相机 曲线下面积@30°↑	深度 绝对相对误差↓	点 L2↓	直接 EPE↓	线性探针 EPE↓	MegaDepth EPE↓	ScanNet-1500 关键点正确率@3↑	实例分割 O-IoU↑	 T-mIoU↑
单视图图像编码器预训练									
MAE	8.27	0.0942	0.4889	73.08	33.18	1.7706	31.92	57.89	53.00
DINOv3	4.04	0.0745	0.5181	47.16	27.23	1.3575	33.89	71.83	71.77
跨视图图像编码器预训练									
CroCo v2	9.07	0.0947	0.4774	61.57	38.67	1.6748	28.13	66.18	61.65
MuM	13.92	0.0751	0.4182	48.02	27.74	1.4181	31.88	70.13	63.76
原生多视图预训练									
Muskie	28.49	0.0796	0.3520	33.65	18.29	1.3521	36.26	68.68	69.92
E-RayZer	3.72	0.1495	0.5841	124.20	50.75	3.1225	32.01	51.51	45.62
在冻结的二维编码器上进行多视图预训练									
MaskedVGGT	**35.80**	**0.0669**	**0.3046**	**32.85**	**16.50**	**1.3368**	**37.69**	**75.53**	**74.06**

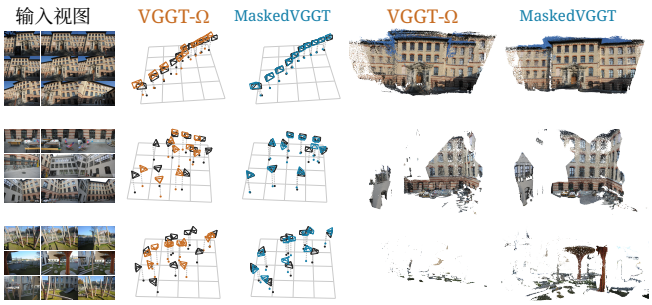

图7 ETH3D 上 L尺度重建的比较。我们展示了每个场景的输入视图（左）、相机位姿（中）和重建点云（右）。黑色相机表示真实值。与 VGGT-Ω 相比，MaskedVGGT 能够恢复更准确的相机位姿和更完整的场景结构。

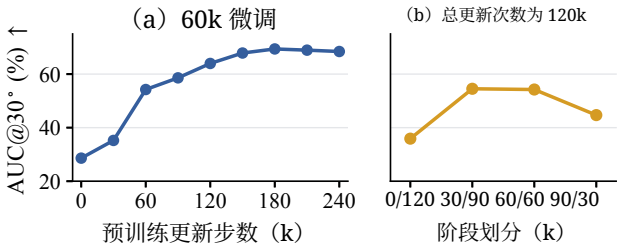

图 8 训练计划分析。 DL3DV 上的 S-scale 结果。（a）预训练更新步数为 0–240k，每次增加 30k，并进行 60k 次微调更新。（b）自监督预训练/有监督微调更新步数。0/120 的划分采用直接监督。

4.3 预训练的扩展性与通用性

跨规模的有效性。 我们将模型规模、总更新次数（120k/180k/240k）、图形处理器数量（2/4/16）和数据集数量（1/10/20）从 S 增至 B 再到 L。我们对全部五个评估数据集均采用 VGGT-Ω评估协议。在所有三个规模下，MaskedVGGT 在全部六项指标上均优于直接监督（表 3）。在最大规模下，两者在 DL3DV 上的差距缩小，但在未见数据集上仍有显著提升，且在 TUM-Dynamic 上的位姿优势进一步扩大。图 7 比较了 L 尺度模型在 ETH3D 上恢复的相机和场景结构。图1 以每 10k 次更新为间隔，跟踪 L 尺度模型在三个未见数据集上的表现。

跨架构预训练。我们测试了该策略能否扩展到 VGGT-Ω 之外的三种具有代表性的前馈重建架构： π^3 [159]、MapAnything [78]，和 VGGT [150]。预训练能够改善这四种架构的相机姿态和深度估计

采用相同训练更新步数的架构（表 4）。因此，我们的策略有利于将预训练图像编码器与不同多视图模块配对的重建模型。

使用额外无标注视频进行预训练。 我们测试没有相机或深度标签的视频能否提升重建效果和跨数据集泛化能力。我们分别将 RE10K [194]，（包含以静态场景为主的相机运动）和 SA-V [114]，（包含明显的前景物体运动）加入 DL3DV。两者都提升了在 DL3DV 和未见数据集上的位姿准确率（表5）。

预训练时长与阶段划分。 我们测试了延长预训练或重新分配固定训练预算时，重建效果如何变化。在微调固定为 60k 次更新的情况下，相机姿态准确率的提升在预训练更新达到约 180k 次时开始趋于饱和（图 8a）。在总预算固定为 120k 次更新时，30k/90k 和 60k/60k 的预训练/微调划分表现相近，且优于直接监督。仅进行 30k 次微调更新时，位姿准确率会下降，但仍高于直接监督（图 8b）。

4.4 迁移至多样化下游任务

我们在冻结编码器的情况下比较下游任务上的预训练表征。MaskedVGGT 在全部九项几何、匹配和分割指标上均优于其他预训练方法（表 6）。

三维几何估计。 我们为每个模型分别训练 VGGT-Ω相机和深度头，进行 60k 次更新，从 DL3DV、ScanNet++ [178]、Blended- 中等量采样，

表 7 掩码预训练的消融研究。我们对 S 尺度模型进行6 万步预训练，再进行另外6 万步微调。两个阶段每个图形处理器最多使用 48 张图像。我们按照与表 1相同的协议，在 DL3DV 上评估相机姿态和深度。

目标	曲线下面积@30° ↑	绝对相对误差↓
RGB 像素	58.09	0.1067
DINO 特征	71.65	0.0964
EMA 特征	69.96	0.1024
聚类标签	**77.33**	**0.0879**

(a) 预测目标。

策略	曲线下面积@30° ↑	绝对相对误差↓
随机	71.63	0.0973
Tube	63.59	0.1040
DINO引导	74.78	0.0929
非对称	**77.33**	**0.0879**

(b) 掩码策略。

比例	曲线下面积@30° ↑	绝对相对误差↓
25%	73.37	0.0941
50%	73.57	0.0939
75%	74.00	0.0926
95%	**77.33**	**0.0879**

(c) 掩码比例。

概率	曲线下面积@30° ↑	绝对相对误差↓
5%	76.85	0.0881
10%	74.90	0.0921
20%	**77.33**	**0.0879**
30%	76.85	0.0908

(d) 锚点概率。

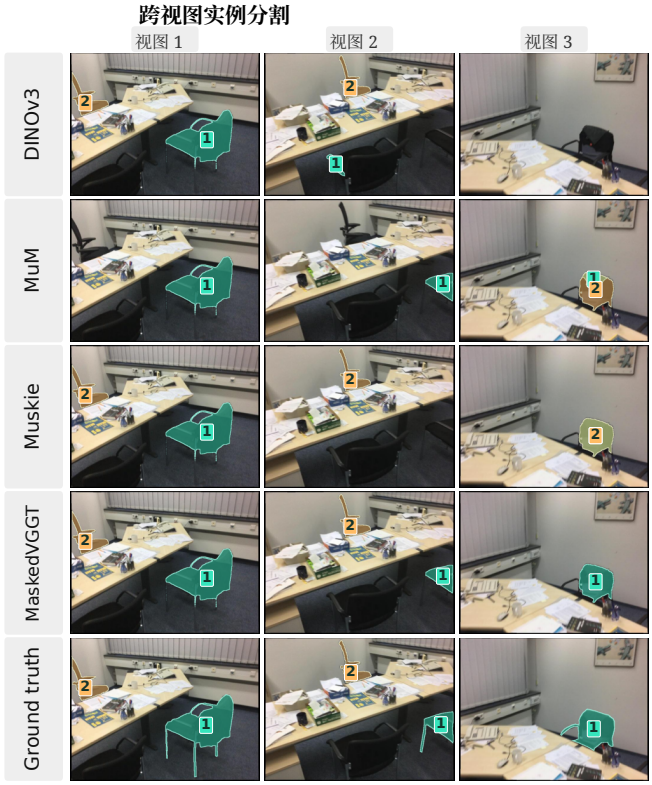

图 9 跨视角实例分割。列展示同一场景的三个视图。各行展示不同方法预测的掩码以及真实值（底行），并叠加在输入图像上。相同的颜色和 ID 表示三个展示视图中的同一实例。

MVS [176], 和 Hypersim [116]。MaskedVGGT 在平均姿态、深度和三维点分数（DL3DV、ScanNet++、7Scenes、DTU）方面表现领先。

跨视角特征匹配。直接余弦匹配使用各编码器扫描所得最佳层的冻结图像块特征。线性探针使用固定层和 20 万个 ScanNet训练对。在这两种设置下，MaskedVGGT 的 ScanNet-1500 端点误差（EPE）均最低。

使用 RoMa 进行稠密匹配。我们将冻结的编码器用作 RoMa-Ω 主干网络 [101]，在所评估的编码器中取得最低的 MegaDepth EPE 和最高的 ScanNet-1500 PCK@3（匹配点相距不超过三个像素）。匹配器初始化、训练对和训练计划均相同（新匹配器进行 50k 次粗匹配更新和 +25k 次细化更新）。

使用 SAM 进行多视图实例分割。 使用相同的 SAM掩码 [80], 源点，并进行投票

规则，我们通过特征匹配在各视图之间传播实例 [93]。MaskedVGGT 在来自 ScanNet 和 ScanNet 的 19 个场景中取得了最高的 O-IoU 和 T-mIoU++ [49]，且无需训练。这些指标分别表示在各视图之间汇总，或在可见视角上取平均后的掩码交并比。图9显示，MaskedVGGT 能够在不同视图中更加一致地分割同一物体，而其他方法会将其与相似物体混淆，或完全遗漏该物体。

4.5 消融研究

预测目标。在掩码位置，我们比较了重建 RGB 像素、预测冻结的 DINOv3 特征、直接预测 EMA 教师特征，以及预测根据 EMA 教师特征计算的软聚类分配（我们的方法）。特征预测优于像素重建，而我们的方法总体表现最佳（表 7a）。

掩码策略。 我们将本文策略与以下方法进行比较：为每个视图独立生成的随机掩码、跨视图共享的管状掩码，以及由 DINO 特征匹配引导的掩码。在总体掩码比例相近的情况下，我们的非对称掩码使锚点视图完全可见，并对其他视图进行大比例掩码处理，从而取得最佳的位姿和深度结果（表 7b）。

掩码比例。 在锚点概率固定为 20% 的情况下，我们比较了非锚点视图中 25%、50%、75% 和 95% 的掩码比例，同时保持锚点完全可见。随着掩码比例增加，位姿和深度准确率均有所提升，其中 95% 的设置表现最佳（表7c）。

锚点概率。 在非锚点掩码比例固定为 95% 的情况下，我们比较了 5%、10%、20% 和 30% 的锚点概率。尽管 20 % 的设置表现最佳，但不同设置下的位姿和深度结果差异很小，表明该设计选择具有较强的鲁棒性（表 7d）。

5 结论

我们提出了 MaskedVGGT，通过掩码自蒸馏，在冻结的二维特征上从头学习多视图变换器。我们的方法优于以往的三维预训练方法。在总更新次数相同的情况下，我们在参数量不超过 10 亿时也超越了直接监督。随着架构变化和无标注视频的增加，性能提升持续扩大。我们的预训练编码器也能迁移到下游任务。这些结果证明了基于强大二维表征进行三维预训练的有效性。

参考文献

[1] Sameer Agarwal、Yasutaka Furukawa、Noah Snavely、Ian Simon、Brian Curless、Steven M Seitz 和 Richard Szeliski。《一天建成罗马》。*ACM*通信，54(10):105–112，2011年。[2] Honggyu An、Jin Hyeon Kim、Seonghoon Park、Jaewoo Jung、Jisang Han、Sunghwan Hong 和 Seungryong Kim。跨视图补全模型是零样本对应关系估计器。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第1103–1115页。IEEE，2025年。[3] Yuki Markus Asano、Christian Rupprecht 和 Andrea Vedaldi。通过同步聚类和表征学习进行自标注。*arXiv*预印本 *arXiv:1911.05371*，2019年。[4] Mahmoud Assran、Quentin Duval、Ishan Misra、Piotr Bojanowski、Pascal Vincent、Michael Rabbat、Yann LeCun 和 Nicolas Ballas。利用联合嵌入预测架构从图像中进行自监督学习。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第15619–15629页。IEEE，2023年。[5] Mido Assran、Adrien Bardes、David Fan、Quentin Garrido、Russell Howes、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zholus 等。V-JEPA 2：自监督视频模型助力理解、预测与规划。*arXiv*预印本 *arXiv:2506.09985*，2025年。[6] Roman Bachmann、David Mizrahi、Andrei Atanov 和 Amir Zamir。MultiMAE：多模态多任务掩码自编码器。载于欧洲计算机视觉会议，第348–367页。施普林格，2022年。[7] Roman Bachmann、Oğuzhan F Kar、David Mizrahi、Ali Garjani、Mingfei Gao、David Griffiths、Jiaming Hu、Afshin Dehghan 和 Amir Zamir。4M-21：面向数十种任务和模态的任意到任意视觉模型。神经信息处理系统进展，37:61872–61911，2024年。[8] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli。data2vec：用于语音、视觉和语言自监督学习的通用框架。载于国际机器学习会议，第1298–1312页。PMLR，2022年。[9] Alexei Baevski、Arun Babu、Wei-Ning Hsu 和 Michael Auli。利用上下文化目标表示，高效地对视觉、语音和语言进行自监督学习。载于国际机器学习会议，第1416–1429页。PMLR，2023年。[10] Hangbo Bao、Li Dong、Songhao Piao 和 Furu Wei。BEiT：图像变换器的 BERT 预训练。*arXiv*预印本 *arXiv:2106.08254*，2021年。[11] Adrien Bardes、Jean Ponce 和 Yann LeCun。VICReg：用于自监督学习的方差-不变性-协方差正则化。*arXiv*预印本 *arXiv:2105.04906*，2021年。[12] Adrien Bardes、Jean Ponce 和 Yann LeCun。VICRegL：局部视觉特征的自监督学习。神经信息处理系统进展，35:8799–8810，2022年。[13] Adrien Bardes、Quentin Garrido、Jean Ponce、Xinlei Chen、Michael Rabbat、Yann LeCun、Mahmoud Assran 和 Nicolas Ballas。重新审视特征预测，从视频中学习视觉表征。*arXiv*预印本 *arXiv:2404.08471*，2024年。[14] Daniel Bolya、Po-Yao Huang、Peize Sun、Jang Hyun Cho、Andrea Madotto、Chen Wei、Tengyu Ma、Jiale Zhi、Jathushan Rajasegaran、Hanoona Bangalath等。感知编码器：最佳视觉嵌入并不位于网络的输出端。神经信息处理系统进展，38:60884–60937，2026年。[15] Lucas Brynte、José Pedro Iglesias、Carl Olsson和Fredrik Kahl。利用图注意力网络学习运动恢复结构。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第4808–4817页。IEEE，2024年。[16] Daniel J Butler、Jonas Wulff、Garrett B Stanley和Michael J Black。用于光流评估的自然场景开源电影。载于欧洲计算机视觉会议，第611–625页。施普林格，2012年。[17] Yohann Cabon、Naila Murray和Martin Humenberger。Virtual KITTI 2。*arXiv*预印本 *arXiv:2001.10773*，2020年。[18] Yohann Cabon、Lucas Stoffl、Leonid Antsfeld、Gabriela Csurka、Boris Chidlovskii、Jerome Revaud和Vincent Leroy。MUST3R：用于立体三维重建的多视图网络。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第1050–1060页。IEEE，2025年。[19] Mathilde Caron、Piotr Bojanowski、Armand Joulin和Matthijs Douze。用于无监督视觉特征学习的深度聚类。载于欧洲计算机视觉会议，第139–156页。施普林格，2018年。[20] Mathilde Caron、Piotr Bojanowski、Julien Mairal和Armand Joulin。在未经筛选的数据上对图像特征进行无监督预训练。载于*2019*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第2959–2968页。IEEE，2019年。[21] Mathilde Caron、Ishan Misra、Julien Mairal、Priya Goyal、Piotr Bojanowski和Armand Joulin。通过对比聚类分配进行视觉特征的无监督学习。神经信息处理系统进展，33:9912–9924，2020年。[22] Mathilde Caron、Hugo Touvron、Ishan Misra、Hervé Jégou、Julien Mairal、Piotr Bojanowski和Armand Joulin。自监督视觉变换器中的新兴特性。载于*2021*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第9630–9640页。IEEE，2021年。[23] Hongkai Chen、Zixin Luo、Jiahui Zhang、Lei Zhou、Xuyang Bai、Zeyu Hu、Chiew-Lan Tai和Long Quan。利用种子图匹配网络学习特征匹配。载于*2021*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第6281–6290页。IEEE，2021年。[24] Mark Chen、Alec Radford、Rewon Child、Jeffrey Wu、Heewoo Jun、David Luan和Ilya Sutskever。从像素进行生成式预训练。载于国际机器学习会议，第1691–1703页。PMLR，2020年。[25] Ting Chen、Simon Kornblith、Mohammad Norouzi和Geoffrey Hinton。用于视觉表征对比学习的简单框架。载于国际机器学习会议，第1597–1607页。PMLR，2020年。[26] Xinlei Chen和Kaiming He。探索简单的孪生表征学习。载于*2021*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第15745–15753页。IEEE，2021年。

[27] Xinlei Chen、Saining Xie和Kaiming He。自监督视觉变换器训练的实证研究。载于*2021*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第9620–9629页。IEEE，2021年。[28] Yu Chen和Gim Hee Lee。Dbarf：深度束调整可泛化神经辐射场。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第24–34页。IEEE，2023年。[29] Yabo Chen、Yuchen Liu、Dongsheng Jiang、Xiaopeng Zhang、Wenrui Dai、Hongkai Xiong和Qi Tian。Sdae：自蒸馏掩码自编码器。载于欧洲计算机视觉会议，第108–124页。施普林格，2022年。[30] Youyu Chen、Junjun Jiang、Yueru Luo、Kui Jiang、Xian-ming Liu、Xu Yan和Dave Zhenyu Chen。Reliev3R：摆脱多视图几何标注对前馈重建的依赖。*arXiv*预印本 *arXiv:2604.00548*，2026年。[31] Yung-Sung Chuang、Yang Li、Dong Wang、Ching-Feng Yeh、Kehan Lyu、Ramya Raghavendra、Jim Glass、Lifei Huang、Jason Weston、Luke Zettlemoyer等。Meta Clip 2：全球扩展方案。神经信息处理系统进展，38:48009–48036，2026年。[32] Angela Dai、Angel X Chang、Manolis Savva、Maciej Hal-ber、Thomas Funkhouser和Matthias Nießner。ScanNet：带有丰富标注的室内场景三维重建。载于*2017*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第2432–2443页。IEEE，2017年。[33] Timothée Darcet、Maxime Oquab、Julien Mairal和Piotr Bojanowski。视觉变换器需要寄存器。载于国际学习表征会议，第2632–2652页，2024年。[34] Timothée Darcet、Federico Baldassarre、Maxime Oquab、Julien Mairal和Piotr Bojanowski。聚类并预测潜在图像块，以改进掩码图像建模。*arXiv*预印本 *arXiv:2502.08769*，2025年。[35] Daniel DeTone、Tomasz Malisiewicz和Andrew Rabi-novich。SuperPoint：自监督兴趣点检测与描述。载于*2018*年*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*），第337–33712页。IEEE，2018年。[36] Jacob Devlin、Ming-Wei Chang、Kenton Lee和Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，2019年。[37] Carl Doersch、Abhinav Gupta和Alexei A Efros。通过上下文预测进行无监督视觉表征学习。载于*IEEE*国际计算机视觉会议论文集，第1422–1430页，2015年。[38] Mihai Dusmanu、Ignacio Rocco、Tomas Pajdla、Marc Pollefeys、Josef Sivic、Akihiko Torii和Torsten Sattler。D2-Net：用于局部特征联合描述与检测的可训练卷积神经网络。载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第8084–8093页。IEEE，2019年。[39] Johan Edstedt、Ioannis Athanasiadis、Mårten Wadenbäck和Michael Felsberg。DKM：用于几何估计的稠密核化特征匹配。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第17765–17775页。IEEE，2023年。

[40] Johan Edstedt、Qiyu Sun、Georg Bökman、Mårten Waden- bäck 和 Michael Felsberg。RoMa：鲁棒稠密特征匹配。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第19790–19800页。IEEE，2024年。[41] Mohamed El Banani、Amit Raj、Kevis-Kokitsi Maninis、Abhishek Kar、Yuanzhen Li、Michael Rubinstein、Deqing Sun、Leonidas Guibas、Justin Johnson 和 Varun Jam-pani。探究视觉基础模型的三维感知能力。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第21795–21806页。IEEE，2024年。[42] David Fan、Shengbang Tong、Jiachen Zhu、Koustuv Sinha、Zhuang Liu、Xinlei Chen、Michael Rabbat、Nicolas Ballas、Yann LeCun、Amir Bar 等。扩展无语言视觉表征学习。载于*2025*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第1–13页。IEEE，2025年。[43] Yuxin Fang、Wen Wang、Binhui Xie、Quan Sun、Ledell Wu、Xinggang Wang、Tiejun Huang、Xinlong Wang 和 Yue Cao。EVA：探索规模化掩码视觉表征学习的极限。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第19358–19369页。IEEE，2023年。[44] Yuxin Fang、Quan Sun、Xinggang Wang、Tiejun Huang、Xinlong Wang 和 Yue Cao。EVA-02：面向新世纪福音战士的视觉表征。*Image and Vision Computing*，149：105171，2024年。[45] Christoph Feichtenhofer、Yanghao Li、Kaiming He 等。作为时空学习器的掩码自编码器。神经信息处理系统进展，35：35946–35958，2022年。[46] Michael Fonder 和 Marc Van Droogenbroeck。Mid-air：用于超低空无人机飞行的多模态数据集。载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*），第553–562页。IEEE，2019年。[47] Yang Fu、Ishan Misra 和 Xiaolong Wang。MonoNeRF：从单目视频中学习可泛化神经辐射场，无需相机位姿。载于国际机器学习会议，第10392–10404页。PMLR，2023年。[48] Spyros Gidaris、Praveer Singh 和 Nikos Komodakis。通过预测图像旋转进行无监督表征学习。*arXiv*预印本 *arXiv:1803.07728*，2018年。[49] Jiangshan Gong、Yuqun Wu、Qiqian Fu、Yao Xiao、Chuhang Zou、Shenlong Wang 和 Derek Hoiem。SAM-v：面向多视图实例分割的几何感知万物分割。*arXiv*预印本 *arXiv:2609.25490*，2026年。[50] Priya Goyal、Dhruv Mahajan、Abhinav Gupta 和 Ishan Misra。扩展并基准测试自监督视觉表征学习。载于*2019*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第6390–6399页。IEEE，2019年。[51] Priya Goyal、Mathilde Caron、Benjamin Lefaudeux、Min Xu、Pengchao Wang、Vivek Pai、Mannat Singh、Vitaliy Liptchinsky、Ishan Misra、Armand Joulin 等。在自然环境中对视觉特征进行自监督预训练。*arXiv*预印本 *arXiv:2103.01988*，2021年。[52] Jean-Bastien Grill、Florian Strub、Florent Altché、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Do-ersch、Bernardo Avila Pires、Zhaohan Guo、Mohammad

Gheshlaghi Azar 等。自举潜在表示——一种自监督学习的新方法。神经信息处理系统进展，33：21271–21284，2020年。[53] Xiaodong Gu、Zhiwen Fan、Siyu Zhu、Zuozhuo Dai、Feitong Tan 和 Ping Tan。面向高分辨率多视图立体视觉与立体匹配的级联代价体。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第2492–2501页。IEEE，2020年。[54] Agrim Gupta、Jiajun Wu、Jia Deng 和 Fei-Fei Li。孪生掩码自编码器。神经信息处理系统进展，36：40676–40693，2023年。[55] Richard Hartley、Andrew Zisserman 等。计算机视觉中的多视图几何。剑桥大学出版社，剑桥，2003年。[56] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第9726–9735页。IEEE，2020年。[57] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Dollár 和 Ross Girshick。掩码自编码器是可扩展的视觉学习器。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第15979–15988页。IEEE，2022年。[58] Olivier Henaff。通过对比预测编码实现数据高效的图像识别。载于国际机器学习会议，第4182–4192页。PMLR，2020年。[59] Olivier J Hénaff、Skanda Koppula、Jean-Baptiste Alayrac、Aaron Van den Oord、Oriol Vinyals 和 Joao Carreira。通过对比检测实现高效视觉预训练。载于*2021*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第10066–10076页。IEEE，2021年。[60] Sunghwan Hong、Jaewoo Jung、Heeseong Shin、Jisang Han、Jiaolong Yang、Chong Luo 和 Seungryong Kim。Pf3plat：用于新视角合成的无位姿前馈式3D高斯泼溅。载于第四十二届国际机器学习会议，2025年。[61] Ji Hou、Saining Xie、Benjamin Graham、Angela Dai 和 Matthias Nießner。Pri3d：3D先验能否助力2D表征学习？载于*2021*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第5673–5682页。IEEE，2021年。[62] Zejiang Hou 和 Sun-Yuan Kung。语言辅助表征上的掩码图像预训练。载于*ICASSP 2025-2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第1–5页。IEEE，2025年。[63] Jia hui Huang、Qunjie Zhou、Hesam Rabeti、Aleksandr Korovko、Huan Ling、Xuanchi Ren、Tianchang Shen、Jun Gao、Dmitry Slepichev、Chen-Hsuan Lin 等。Vipe：用于3D几何感知的视频位姿引擎。*arXiv*预印本 *arXiv:2508.10934*，2025年。[64] Nan Huang、Pengcheng Yu、Weijia Zeng、James M Rehg、Angjoo Kanazawa、Haiwen Feng 和 Qianqian Wang。通过自蒸馏实现自改进4D感知。*arXiv*预印本 *arXiv:2604.08532*，2026年。[65] Po-Han Huang、Kevin Matzen、Johannes Kopf、Narendra Ahuja 和 Jia-Bin Huang。DeepMVS：学习多视图立体重建。载于*2018*年*IEEE/CVF*计算机视觉与模式识别会议，第2821–2830页。IEEE，2018年。

[66] Ranran Huang、Weixun Luo、Ye Mao 和 Krystian Mikolajczyk。从无到有：通过新视角合成进行自监督三维重建。*arXiv* 预印本 *arXiv:2603.27455*，2026年。[67] Allan Jabri、Andrew Owens 和 Alexei Efros。时空对应关系作为对比随机游走。神经信息处理系统进展，33：19545–19560，2020年。[68] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。通过噪声文本监督扩展视觉与视觉语言表征学习。载于国际机器学习会议，第4904–4916页。PMLR，2021年。[69] Hanwen Jiang、Hao Tan、Peng Wang、Haian Jin、Yue Zhao、Sai Bi、Kai Zhang、Fujun Luan、Kalyan Sunkavalli、Qixing Huang 等。Rayzer：一种自监督大规模视图合成模型。载于*2025*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第4918–4929页。IEEE，2025年。[70] Hanwen Jiang、Zexiang Xu、Desai Xie、Ziwen Chen、Haian Jin、Fujun Luan、Zhixin Shu、Kai Zhang、Sai Bi、Xin Sun 等。MegaSynth：利用合成数据扩展三维场景重建。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第16441–16452页。IEEE，2025年。[71] Lihan Jiang、Yucheng Mao、Linning Xu、Tao Lu、Kerui Ren、Yichen Jin、Xudong Xu、Mulin Yu、Jiangmiao Pang、Feng Zhao 等。AnySplat：从无约束视图进行前馈式三维高斯泼溅。*ACM* 图形学汇刊（*TOG*），44(6)：1–16，2025年。[72] Nicholas Jiang、Amil Dravid、Alexei Efros 和 Yossi Gandelsman。视觉变换器不需要训练的寄存器。神经信息处理系统进展，38：56557–56595，2026年。[73] Zeren Jiang、Chuanxia Zheng、Iro Laina、Diane Larlus 和 Andrea Vedaldi。Geo4d：利用视频生成器进行几何四维场景重建。载于*2025*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第01–14页。IEEE，2025年。[74] Haian Jin、Rundi Wu、Tianyuan Zhang、Ruiqi Gao、Jonathan T Barron、Noah Snavely 和 Aleksander Hołyński。Zipmap：通过测试时训练实现线性时间的有状态三维重建。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第21748–21759页，2026年。[75] Ioannis Kakogeorgiou、Spyros Gidaris、Bill Psomas、Yannis Avrithis、Andrei Bursuc、Konstantinos Karantzalos 和 Nikos Komodakis。该向学生隐藏什么：注意力引导的掩码图像建模。载于欧洲计算机视觉会议，第300–318页。施普林格，2022年。[76] Gyeongjin Kang、Jisang Yoo、Jihyeon Park、Seungtae Nam、Hyeonsoo Im、Sangheon Shin、Sangpil Kim 和 Eunbyung Park。Selfsplat：无位姿、无三维先验的可泛化三维高斯泼溅。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第22012–22022页。IEEE，2025年。[77] Nikita Karaev、Ignacio Rocco、Benjamin Graham、Natalia Neverova、Andrea Vedaldi 和 Christian Rupprecht。Dynamicstereo：从立体视频中获得一致的动态深度。载于*2023*年*IEEE/CVF*计算机视觉会议

与模式识别会议（*CVPR*），第13229–13239页。IEEE，2023年。[78] Nikhil Keetha、Norman Müller、Johannes Schönberger、Lorenzo Porzi、Yuchen Zhang、Tobias Fischer、Arno Knapitsch、Duncan Zauss、Ethan Weber、Nelson Antunes 等。MapAnything：通用前馈度量三维重建；Map-Anything。github.io。载于*2026*年国际三维视觉会议（*3DV*），第499–509页。IEEE，2026年。[79] Bernhard Kerbl、Georgios Kopanas、Thomas Leimkühler、George Drettakis 等。用于实时辐射场渲染的3D 高斯泼溅。*ACM* 图形学汇刊，42(4)：139–1，2023年。[80] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo 等。分割一切。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第3992–4003页。IEEE，2023年。[81] Karim Knaebel、Kadir Yilmaz、Daan De Geus、Alexander Hermans、David Adrian、Timm Linder 和 Bastian Leibe。DINO在室内：利用二维基础模型进行三维分割。载于*2026*年国际三维视觉会议（*3DV*），第1884–1894页。IEEE，2026年。[82] Yushi Lan、Yihang Luo、Fangzhou Hong、Shangchen Zhou、Honghua Chen、Zhaoyang Lyu、Bo Dai、Shuai Yang、Chen Change Loy 和 Xingang Pan。Stream3r：采用因果变换器的可扩展序列三维重建。载于国际学习表征会议，第42746–42768页，2026年。[83] Hoang-An Le、Thomas Mensink、Partha Das、Sezer Karaoglu 和 Theo Gevers。EDEN：封闭式花园场景的多模态合成数据集。载于*2021*年*IEEE* 冬季计算机视觉应用会议（*WACV*），第1578–1588页。IEEE，2021年。[84] Vincent Leroy、Yohann Cabon 和 Jérôme Revaud。利用MASt3R在三维空间中实现图像匹配定位。载于欧洲计算机视觉会议，第71–91页。施普林格，2024年。[85] Gang Li、Heliang Zheng、Daqing Liu、Chaoyue Wang、Bing Su 和 Changwen Zheng。SemMAE：用于学习掩码自编码器的语义引导掩码。神经信息处理系统进展，35：14290–14302，2022年。[86] Wenyu Li、Sidun Liu、Peng Qiao、Yong Dou 和 Tongrui Hu。Muskie：用于三维视觉预训练的多视图掩码图像建模。*arXiv* 预印本 *arXiv:2511.18115*，2025年。[87] Zhengqi Li 和 Noah Snavely。MegaDepth：从互联网照片中学习单视图深度预测。载于*2018*年*IEEE/CVF*计算机视觉与模式识别会议，第2041–2050页。IEEE，2018年。[88] Zhengqi Li、Richard Tucker、Forrester Cole、Qianqian Wang、Linyi Jin、Vickie Ye、Angjoo Kanazawa、Aleksander Hołyński 和 Noah Snavely。MegaSAM：从日常动态视频中快速、准确且稳健地恢复结构与运动。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第10486–10496页。IEEE，2025年。[89] Haotong Lin、Sili Chen、Junhao Liew、Donny Y Chen、Zhenyu Li、Guang Shi、Jiashi Feng 和 Bingyi Kang。Depth Anything 3：从任意视图恢复视觉空间。*arXiv* 预印本 *arXiv:2511.10647*，2025年。

[90] Philipp Lindenberger、Paul-Edouard Sarlin 和 Marc Pollefeys。《Lightglue：以光速进行局部特征匹配》。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第17581–17592页。IEEE，2023年。[91] Lu Ling、Yichen Sheng、Zhi Tu、Wentian Zhao、Cheng Xin、Kun Wan、Lantao Yu、Qianyu Guo、Zixun Yu、Yawen Lu 等。《Dl3dv-10k：用于基于深度学习的三维视觉的大规模场景数据集》。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第22160–22169页。IEEE，2024年。[92] Xingbin Liu、Jinghao Zhou、Tao Kong、Xianming Lin 和 Rongrong Ji。《探索掩码自编码器的目标表示》。*arXiv* 预印本 *arXiv:2209.03917*，2022年。[93] Yang Liu、Muzhi Zhu、Hengtao Li、Hao Chen、Xinlong Wang 和 Chunhua Shen。《Matcher：使用通用特征匹配通过单次处理实现万物分割》。*arXiv*预印本 *arXiv:2305.13310*，2023年。[94] Thibaut Loiseau、Guillaume Bourmaud 和 Vincent Lepetit。《Alligat0r：通过共可见性分割进行预训练，以实现相对相机位姿回归》。神经信息处理系统进展，38：13762–13789，2026年。[95] Thibaut Loiseau、Guillaume Bourmaud 和 Vincent Lepetit。《重新审视跨视图补全：通过重建误差比较进行自监督预训练》。*arXiv*预印本 *arXiv:2609.01530*，2026年。[96] Mapillary。Mapillary Metropolis数据集。数据集及官方SDK，日期不详。访问于2026年10月。[97] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。《NeRF：将场景表示为用于视图合成的神经辐射场》。*ACM*通信，65（1）：99–106，2021年。[98] Thomas Mitchel、Hyunwoo Ryu 和 Vincent Sitzmann。《真正的自监督新视角合成具有可迁移性》。载于国际学习表征会议，第17882–17898页，2026年。[99] David Mizrahi、Roman Bachmann、Oguzhan Kar、Teresa Yeo、Mingfei Gao、Afshin Dehghan 和 Amir Zamir。《4M：大规模多模态掩码建模》。神经信息处理系统进展，36：58363–58408，2023年。[100] David Nordström、Johan Edstedt、Fredrik Kahl 和 Georg Bökman。《MuM：用于三维视觉的多视图掩码图像建模》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第21736–21747页，2026年。[101] David Nordström、Xinyue Zhang、Thibaut Loiseau、Vincent Lepetit 和 Fredrik Kahl。《RoMa-Ω：前馈式三维模型对图像匹配的认知》。*arXiv*预印本 *arXiv:2609.09507*，2026年。[102] Mehdi Noroozi 和 Paolo Favaro。《通过解决拼图游戏进行视觉表征的无监督学习》。载于欧洲计算机视觉会议，第69–84页。施普林格，2016年。[103] Pedro O O Pinheiro、Amjad Almahairi、Ryan Benmalek、Florian Golemo 和 Aaron C Courville。《稠密视觉表征的无监督学习》。神经信息处理系统进展，33：4489–4500，2020年。

[104] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。《基于对比预测编码的表征学习》。*arXiv*预印本 *arXiv:1807.03748*，2018年。[105] Maxime Oquab、Timothée Darcet、Théo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。《DINOv2：无监督学习鲁棒视觉特征》。*arXiv*预印本 *arXiv:2304.07193*，2023年。[106] Linfei Pan、Dániel Baráth、Marc Pollefeys 和 Johannes L Schönberger。《重新审视全局运动恢复结构》。载于欧洲计算机视觉会议，第58–77页。施普林格，2024年。[107] Deepak Pathak、Philipp Krahenbuhl、Jeff Donahue、Trevor Darrell 和 Alexei A Efros。《上下文编码器：通过图像修复进行特征学习》。载于*IEEE*计算机视觉与模式识别会议论文集，第2536–2544页，2016年。[108] Zhiliang Peng、Li Dong、Hangbo Bao、Qixiang Ye 和 Furu Wei。《BEiT v2：使用向量量化视觉分词器进行掩码图像建模》。*arXiv*预印本 *arXiv:2208.06366*，2022年。[109] Zhiliang Peng、Li Dong、Hangbo Bao、Qixiang Ye 和 Furu Wei。《掩码图像建模的统一视角》。*arXiv*预印本 *arXiv:2210.10615*，2022年。[110] Ulrich Prestel、Stefan Andreas Baumann、Nick Stracke 和 Björn Ommer。《RayDer：从真实世界视频中进行可扩展的自监督新视角合成》。载于欧洲计算机视觉会议，第558–578页。施普林格，2026年。[111] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。《语言模型是无监督的多任务学习器》。*OpenAI*博客，1（8）：9，2019年。[112] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《从自然语言监督中学习可迁移视觉模型》。载于国际机器学习会议，第8748–8763页。PMLR，2021年。[113] Mike Ranzinger、Greg Heinrich、Jan Kautz 和 Pavlo Molchanov。《AM-RADIO：凝聚式视觉基础模型将所有领域归一为一个整体》。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第12490–12500页。IEEE，2024年。[114] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman Rädle、Chloe Rolland、Laura Gustafson 等。《SAM 2：对图像和视频进行万物分割》。载于国际学习表征会议，第28085–28128页，2025年。[115] Jeremy Reizenstein、Roman Shapovalov、Philipp Henzler、Luca Sbordone、Patrick Labatut 和 David Novotny。《三维常见物体：真实世界三维类别重建的大规模学习与评估》。载于*2021*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第10881–10891页。IEEE，2021年。[116] Mike Roberts、Jason Ramapuram、Anurag Ranjan、Atulit Kumar、Miguel Angel Bautista、Nathan Paczan、Russ Webb 和 Joshua M Susskind。《Hypersim：用于整体室内场景的照片级真实感合成数据集》

standing. 在*2021*年*IEEE/CVF*国际计算机视觉会议 *(ICCV)*，第10892–10902页。IEEE，2021年。[117] Mehdi SM Sajjadi、Aravindh Mahendran、Thomas Kipf、Etienne Pot、Daniel Duckworth、Mario Lučić 和 Klaus Greff。Rust：从未摆姿势的图像中学习潜在神经场景表示。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第17297–17306页。IEEE，2023年。[118] Mohammadreza Salehi、Michael Dorkenwald、Fida Mohammad Thoker、Efstratios Gavves、Cees GM Snoek 和 Yuki M Asano。Sigma：Sinkhorn引导的掩码视频建模。载于欧洲计算机视觉会议，第293–312页。施普林格，2024年。[119] Paul-Edouard Sarlin、Daniel DeTone、Tomasz Malisiewicz 和 Andrew Rabinovich。SuperGlue：利用图神经网络学习特征匹配。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第4937–4946页。IEEE，2020年。[120]Johannes L Schonberger 和 Jan-Michael Frahm。重新审视运动恢复结构。载于*IEEE*计算机视觉与模式识别会议论文集，第4104–4113页，2016年。[121] Johannes L Schönberger、Enliang Zheng、Jan-Michael Frahm 和 Marc Pollefeys。面向非结构化多视图立体视觉的逐像素视图选择。载于欧洲计算机视觉会议，第501–518页。施普林格，2016年。[122] Thomas Schops、Johannes L Schonberger、Silvano Galliani、Torsten Sattler、Konrad Schindler、Marc Pollefeys 和 Andreas Geiger。一个包含高分辨率图像和多摄像机视频的多视图立体视觉基准。载于*IEEE*计算机视觉与模式识别会议论文集，第3260–3269页，2017年。[123] David Shavin 和 Sagie Benaim。Splat and Distill：通过前馈式三维重建增强教师，实现三维感知蒸馏。*arXiv*预印本 *arXiv:2602.06032*，2026年。[124] You Shen、Zhipeng Zhang、Yansong Qu、Xiawu Zheng、Jiayi Ji、Shengchuan Zhang 和 Liujuan Cao。FastVGGT：快速视觉几何变换器。载于第十四届国际学习表征会议，2026年。[125] Oriane Siméoni、Huy V Vo、Maximilian Seitzer、Federico Baldassarre、Maxime Oquab、Cijo Jose、Vasil Khalidov、Marc Szafraniec、Seungeun Yi、Michaël Ramamonjisoa 等。DINOv3。*arXiv*预印本 *arXiv:2508.10104*，2025年。[126] Richard Sinkhorn 和 Paul Knopp。论非负矩阵与双随机矩阵。《太平洋数学杂志》，21(2):343–348，1967年。[127] Brandon Smart、Chuanxia Zheng、Iro Laina 和 Victor Adrian Prisacariu。SplatT3R：从未标定图像对进行零样本高斯泼溅。*arXiv*预印本 *arXiv:2408.13912*，2024年。[128] Noah Snavely、Steven M Seitz 和 Richard Szeliski。Photo Tourism：在三维空间中探索照片集合。载于*2006*年*ACM SIGGRAPH*论文集，第835–846页。2006年。[129] Julian Straub、Thomas Whelan、Lingni Ma、Yufan Chen、Erik Wijmans、Simon Green、Jakob J Engel、Raul Mur-Artal、Carl Ren、Shobhit Verma 等。Replica数据集：室内空间的数字副本。*arXiv*预印本 *arXiv:1906.05797*，2019年。[130] Jiaming Sun、Zehong Shen、Yuang Wang、Hujun Bao 和 Xiaowei Zhou。LoFTR：利用变换器进行无检测器局部特征匹配。*arXiv*预印本 *arXiv:2104.00680*，2021年。

[131] Pei Sun、Henrik Kretzschmar、Xerxes Dotiwalla、Aurelien Chouard、Vijaysai Patnaik、Paul Tsui、James Guo、Yin Zhou、Yuning Chai、Benjamin Caine 等。自动驾驶感知的可扩展性：Waymo 开放数据集。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第2443–2451页。IEEE，2020年。[132] Chengzhou Tang 和 Ping Tan。BA-Net：稠密束调整网络。*arXiv*预印本 *arXiv:1806.04807*，2018年。[133] Zachary Teed 和 Jia Deng。DeepV2D：利用可微分运动恢复结构实现视频转深度。*arXiv*预印本 *arXiv:1812.04605*，2018年。[134] Zachary Teed 和 Jia Deng。DROID-SLAM：面向单目、立体和RGB-D相机的深度视觉SLAM。神经信息处理系统进展，34:16558–16569，2021年。[135] Yonglong Tian、Olivier J Henaff 和 Aäron Van den Oord。分而对比：从未经整理的数据中进行自监督学习。载于*2021*年*IEEE/CVF*国际计算机视觉会议 *(ICCV)*，第10043–10054页。IEEE，2021年。[136] Zhan Tong、Yibing Song、Jue Wang 和 Limin Wang。VideoMAE：掩码自编码器是适用于自监督视频预训练的数据高效学习器。神经信息处理系统进展，35:10078–10093，2022年。[137] Bill Triggs、Philip F McLauchlan、Richard I Hartley 和 Andrew W Fitzgibbon。束调整——现代综述。载于国际视觉算法研讨会，第298–372页。施普林格，1999年。[138] Michał Tyszkiewicz、Pascal Fua 和 Eduard Trulls。DISK：利用策略梯度学习局部特征。神经信息处理系统进展，33:14254–14265，2020年。[139] Benjamin Ummenhofer、Huizhong Zhou、Jonas Uhrig、Nikolaus Mayer、Eddy Ilg、Alexey Dosovitskiy 和 Thomas Brox。DeMoN：用于学习单目立体的深度与运动网络。载于*2017*年*IEEE*计算机视觉与模式识别会议 *(CVPR)*，第5622–5631页。IEEE，2017年。[140] Basile Van Hoorick、Rundi Wu、Ege Ozguroglu、Kyle Sargent、Ruoshi Liu、Pavel Tokmakov、Achal Dave、Changxi Zheng 和 Carl Vondrick。生成式摄影机推轨：极端单目动态新视角合成。载于欧洲计算机视觉会议，第313–331页。施普林格，2024年。[141] Shashanka Venkataramanan、Valentinos Pariza、Mohammadreza Salehi、Lukas Knobel、Elias Ramzi、Spyros Gidaris、Andrei Bursuc 和 Yuki M Asano。FRANCA：用于可扩展视觉表征学习的嵌套套娃聚类。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10533–10544页，2026年。[142] Pascal Vincent、Hugo Larochelle、Isabelle Lajoie、Yoshua Bengio、Pierre-Antoine Manzagol 和 Léon Bottou。堆叠去噪自编码器：利用局部去噪准则在深度网络中学习有用表征。机器学习研究期刊，11(12)，2010年。[143] Huy V Vo、Vasil Khalidov、Timothée Darcet、Théo Moutakanni、Nikita Smetanin、Marc Szafraniec、Hugo Touvron、Camille Couprie、Maxime Oquab、Armand Joulin 等人。面向自监督学习的自动数据整理：一种基于聚类的方法。*arXiv* 预印本 *arXiv:2405.15613*，2024年。[144]Chung-Shien Brian Wang、Christian Schmidt、Jens Piekenbrinck 和 Bastian Leibe。用于高效多视图几何变换器的块稀疏全局注意力。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第14546–14555页，2026年。[145] Hengyi Wang 和 Lourdes Agapito。利用空间记忆进行三维重建。载于*2025*年国际三维视觉会议（*3DV*），第78–89页。IEEE，2025年。[146] Haochen Wang、Kaiyou Song、Junsong Fan、Yuxi Wang、Jin Xie 和 Zhaoxiang Zhang。面向掩码图像建模的困难图像块挖掘。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第10375–10385页。IEEE，2023年。[147] Haoru Wang、Kai Ye、Minghan Qin、Yangyan Li、Wen- zheng Chen 和 Baoquan Chen。依赖越少，学得越多：从稀疏、无位姿图像合成新视图，无需任何三维知识。载于国际学习表征会议，第41717–41739页，2026年。[148]Jianyuan Wang、Christian Rupprecht 和 David Novotny。PoseDiffusion：通过扩散辅助束调整解决位姿估计问题。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第9739–9749页。IEEE，2023年。[149] Jianyuan Wang、Nikita Karaev、Christian Rupprecht 和 David Novotny。VGGSfM：以视觉几何为基础的深度运动恢复结构。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第21686–21697页。IEEE，2024年。[150] Jianyuan Wang、Minghao Chen、Nikita Karaev、An- drea Vedaldi、Christian Rupprecht 和 David Novotny。VGGT：以视觉几何为基础的变换器。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第5294–5306页。IEEE，2025年。[151] Jianyuan Wang、Minghao Chen、Shangzhan Zhang、Nikita Karaev、Johannes Schönberger、Patrick Labatut、Piotr Bojanowski、David Novotny、Andrea Vedaldi 和 Christian Rupprecht。VGGT-Ω。载于计算机视觉与模式识别会议 *2026*，2026年。[152] Limin Wang、Bingkun Huang、Zhiyu Zhao、Zhan Tong、Yinan He、Yi Wang、Yali Wang 和 Yu Qiao。VideoMAE v2：利用双重掩码扩展视频掩码自编码器。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第14549–14560页。IEEE，2023年。[153] Qianqian Wang、Yifei Zhang、Aleksander Holynski、Alexei A Efros 和 Angjoo Kanazawa。具有持久状态的连续三维感知模型。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第10510–10522页。IEEE，2025年。[154] Rui Wang、Dongdong Chen、Zuxuan Wu、Yinpeng Chen、Xiyang Dai、Mengchen Liu、Lu Yuan 和 Yu-Gang Jiang。掩码视频蒸馏：重新思考自监督视频表征学习中的掩码特征建模。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6312–6322页，2023年。[155] Ruoyu Wang、Yi Ma 和 Shenghua Gao。Pensieve 中的回忆：通过学习从

未标定视频合成新视图。*arXiv* 预印本 *arXiv:2505.13440*，2025年。[156] Shuzhe Wang、Vincent Leroy、Yohann Cabon、Boris Chidlovskii 和 Jerome Revaud。DUSt3R：让几何三维视觉变得简单。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第20697–20709页。IEEE，2024年。[157] Wenshan Wang、Delong Zhu、Xiangwei Wang、Yaoyu Hu、Yuheng Qiu、Chen Wang、Yafei Hu、Ashish Kapoor 和 Sebastian Scherer。TartanAir：一个旨在突破视觉 SLAM 极限的数据集。载于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第4909–4916页。IEEE，2020年。[158] Xiaolong Wang、Allan Jabri 和 Alexei A Efros。通过时间的循环一致性学习对应关系。载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第2561–2571页。IEEE，2019年。[159] Yifan Wang、Jianjun Zhou、Haoyi Zhu、Wenzheng Chang、Yang Zhou、Zizun Li、Junyi Chen、Jiangmiao Pang、Chun- hua Shen 和 Tong He。 π^3：置换等变视觉几何学习。载于国际学习表征会议，第10481–10497页，2026年。[160] Chen Wei、Haoqi Fan、Saining Xie、Chao-Yuan Wu、Alan Yuille 和 Christoph Feichtenhofer。面向自监督视觉预训练的掩码特征预测。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第14648–14658页。IEEE，2022年。[161] Longhui Wei、Lingxi Xie、Wengang Zhou、Houqiang Li 和 Qi Tian。MVP：多模态引导视觉预训练。载于欧洲计算机视觉会议，第337–353页。施普林格，2022年。[162] Xingkui Wei、Yinda Zhang、Zhuwen Li、Yanwei Fu 和 Xiangyang Xue。DeepSfM：通过深度束调整实现运动恢复结构。载于欧洲计算机视觉会议，第230–247页。施普林格，2020年。[163] Yibing Wei、Abhinav Gupta 和 Pedro Morgado。迈向用于自监督视觉表征学习的潜在掩码图像建模。载于欧洲计算机视觉会议，第1–17页。施普林格，2024年。[164] Philippe Weinzaepfel、Vincent Leroy、Thomas Lucas、Romain Brégier、Yohann Cabon、Vaibhav Arora、Leonid Antsfeld、Boris Chidlovskii、Gabriela Csurka 和 Jérôme Revaud。CroCo：通过跨视图补全进行面向三维视觉任务的自监督预训练。神经信息处理系统进展，35：3502–3516，2022年。[165] Philippe Weinzaepfel、Thomas Lucas、Vincent Leroy、Yohann Cabon、Vaibhav Arora、Romain Brégier、Gabriela Csurka、Leonid Antsfeld、Boris Chidlovskii 和 Jérôme Revaud。CroCo v2：改进的跨视图补全预训练，适用于立体匹配和光流。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第17923–17934页。IEEE，2023年。[166] Xin Wen、Bingchen Zhao、Anlin Zheng、Xiangyu Zhang 和 Xiaojuan Qi。通过语义分组进行自监督视觉表征学习。神经信息处理系统进展，35：16423–16438，2022年。[167] Yuqi Wu、Wenzhao Zheng、Jie Zhou 和 Jiwen Lu。Point3R：利用显式空间指针记忆进行流式三维重建。神经信息处理系统进展，38：69675–69699，2026年。[168] Zhirong Wu、Yuanjun Xiong、Stella X Yu 和 Dahua Lin。通过非参数方法进行无监督特征学习

实例判别。载于*IEEE*计算机视觉与模式识别会议论文集，第3733–3742页，2018年。[169] Hongchi Xia、Yang Fu、Sifei Liu 和 Xiaolong Wang。自然环境中的 RGB-D 物体：利用 RGB-D 视频扩展真实世界的三维物体学习。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第22378–22389页。IEEE，2024年。[170] Zhenda Xie、Zheng Zhang、Yue Cao、Yutong Lin、Jianmin Bao、Zhuliang Yao、Qi Dai 和 Han Hu。SimMIM：一种用于掩码图像建模的简单框架。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第9643–9653页。IEEE，2022年。[171] Hongwei Xue、Peng Gao、Hongyang Li、Yu Qiao、Hao Sun、Houqiang Li 和 Jiebo Luo。凝视所见：无需重建的掩码图像建模。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第22732–22741页。IEEE，2023年。[172] Jianing Yang、Alexander Sax、Kevin J Liang、Mikael Henaff、Hao Tang、Ang Cao、Joyce Chai、Franziska Meier 和 Matt Feiszli。Fast3R：迈向单次前向传播实现图像三维重建。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第21924–21935页。IEEE，2025年。[173] Lihe Yang、Bingyi Kang、Zilong Huang、Xiaogang Xu、Jiashi Feng 和 Hengshuang Zhao。Depth Anything：释放大规模无标注数据的潜力。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第10371–10381页。IEEE，2024年。[174] Lihe Yang、Bingyi Kang、Zilong Huang、Zhen Zhao、Xiaogang Xu、Jiashi Feng 和 Hengshuang Zhao。Depth Anything V2。神经信息处理系统进展，37：21875–21911，2024年。[175] Yao Yao、Zixin Luo、Shiwei Li、Tian Fang 和 Long Quan。MVSNet：非结构化多视图立体视觉的深度推断。载于欧洲计算机视觉会议，第785–801页。施普林格，2018年。[176] Yao Yao、Zixin Luo、Shiwei Li、Jingyang Zhang、Yufan Ren、Lei Zhou、Tian Fang 和 Long Quan。BlendedMVS：用于广义多视图立体网络的大规模数据集。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第1787–1796页。IEEE，2020年。[177] Lior Yariv、Yoni Kasten、Dror Moran、Meirav Galun、Matan Atzmon、Ronen Basri 和 Yaron Lipman。通过解耦几何与外观实现多视图神经表面重建。神经信息处理系统进展，33：2492–2502，2020年。[178] Chandan Yeshwanth、Yueh-Cheng Liu、Matthias Nießner 和 Angela Dai。ScanNet++：高保真的三维室内场景数据集。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第12–22页。IEEE，2023年。[179] Alex Yu、Vickie Ye、Matthew Tancik 和 Angjoo Kanazawa。pixelNeRF：从一张或少数几张图像生成神经辐射场。载于*2021*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第4576–4585页。IEEE，2021年。[180] Sukmin Yun、Hankook Lee、Jaehyung Kim 和 Jin-woo Shin。面向自监督视觉变换器的补丁级表征学习。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第8344–8353页。IEEE，2022年。

[181] Amir R Zamir、Alexander Sax、William Shen、Leonidas J Guibas、Jitendra Malik 和 Silvio Savarese。Taskonomy：解耦任务迁移学习。载于*IEEE*计算机视觉与模式识别会议论文集，第3712–3722页，2018年。[182] Jure Zbontar、Li Jing、Ishan Misra、Yann LeCun 和 Stéphane Deny。Barlow Twins：通过冗余消除实现自监督学习。载于国际机器学习会议，第12310–12320页。PMLR，2021年。[183] Xiaohua Zhai、Xiao Wang、Basil Mustafa、Andreas Steiner、Daniel Keysers、Alexander Kolesnikov 和 Lucas Beyer。LiT：通过锁定图像文本调优实现零样本迁移。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第18102–18112页。IEEE，2022年。[184] Xiaohua Zhai、Basil Mustafa、Alexander Kolesnikov 和 Lucas Beyer。用于语言-图像预训练的Sigmoid损失。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第11941–11952页。IEEE，2023年。[185] Chuhan Zhang、Guillaume Le Moing、Skanda Koppula、Ignacio Rocco、Liliane Momeni、Junyu Xie、Shuyang Sun、Rahul Sukthankar、Joëlle K Barral、Raia Hadsell 等。一次处理一个 d4rt，高效重建动态场景。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7382–7392页，2026年。[186] Junyi Zhang、Charles Herrmann、Junhwa Hur、Varun Jampani、Forrester Cole、Deqing Sun、Ming-Hsuan Yang 等。Monst3r：一种在存在运动时估计几何的简单方法。载于国际学习表征会议，第82863–82886页，2025年。[187] Kai Zhang、Sai Bi、Hao Tan、Yuanbo Xiangli、Nanxuan Zhao、Kalyan Sunkavalli 和 Zexiang Xu。GS-LRM：用于3D 高斯泼溅的大规模重建模型。载于欧洲计算机视觉会议，第1–19页。施普林格，2024年。[188] Richard Zhang、Phillip Isola 和 Alexei A Efros。多彩的图像着色。载于欧洲计算机视觉会议，第649–666页。施普林格，2016年。[189] Shangzhan Zhang、Jianyuan Wang、Yinghao Xu、Nan Xue、Christian Rupprecht、Xiaowei Zhou、Yujun Shen、以及 Gordon Wetzstein。Flare：从未标定稀疏视图进行前馈几何、外观和相机估计。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 21936–21947 页。IEEE，2025年。[190] Zhe Zhang、Rui Peng、Yuxi Hu 和 Ronggang Wang。GeoMVSNet：通过几何感知学习多视图立体视觉。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 21508–21518 页。IEEE，2023年。[191] Qitao Zhao、Hao Tan、Qianqian Wang、Sai Bi、Kai Zhang、Kalyan Sunkavalli、Shubham Tulsiani 和 Hanwen Jiang。E-RayZer：将自监督三维重建作为空间视觉预训练。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7525–7535 页，2026 年。[192] Bo Zhou 、 Qiuxia Lai 、Zeren Sun、Xiangbo Shu、Yazhou Yao 和 Wenguan Wang。从无姿态多视图图像中学习用于空间智能的三维表示。*arXiv* 预印本 *arXiv:2604.10573*，2026年。[193] Jinghao Zhou、Chen Wei、Huiyu Wang、Wei Shen、Cihang Xie、Alan Yuille 和 Tao Kong。iBOT：借助在线分词器进行图像 BERT 预训练。*arXiv* 预印本 *arXiv:2111.07832*，2021年。[194] Tinghui Zhou、Richard Tucker、John Flynn、Graham Fyffe 和 Noah Snavely。立体放大：利用多平面图像学习视图合成。*arXiv* 预印本 *arXiv:1805.09817*，2018年。[195] Yaolong Zhu、Peng Li、Jiahao Lu、Chengfeng Zhao、Mengfei Li、Chao Wang、Wenyue Chen、Zekai Gu、Chengwei Ren、Ronggang Wang 等。Pure：通过无标注后训练优化前馈三维重建。*arXiv* 预印本 *arXiv:2610.14698*，2026年。[196] Dong Zhuo、Wenzhao Zheng、Jiahe Guo、Yuqi Wu、Jie Zhou 和 Jiwen Lu。流式视觉几何 Transformer。载于国际学习表征会议，第 88055–88072 页，2026年。