

DINOv3

奥里安·西梅奥尼^{*}于伊·V·沃^{*}马克西米利安·塞策^{*}费德里科·巴尔达萨雷^{*}马克西姆·奥夸布^{*}
西约·何塞·瓦西尔·哈利多夫 马克·萨夫拉涅茨 岚根·李·迈克尔·拉马蒙吉索亚 弗朗西斯科·马萨 丹尼尔·哈
齐扎 卢卡·韦斯特德特 王建元
蒂莫西·达塞特 西奥·穆塔卡尼 莱昂内尔·森塔纳 克莱尔·罗伯茨 安德烈亚·韦达利 杰米·托兰 约翰·布兰特¹卡米
尔·库普里 朱利安·梅拉尔²赫尔维·杰古 帕特里克·拉巴图 皮奥特·博亚诺夫斯基

元人工智能研究 ¹WRI ²国立信息与自动化研究院 (Inria)

* 通讯作者: {osimeoni, huyvvo, seitzer, baldassarre, qas}@meta.com

摘要

自监督学习有望彻底消除人工标注数据的需求，使模型能够轻松扩展至海量数据集和大型架构。这种训练范式无需针对特定任务或领域定制，仅需单一算法即可从自然图像到航拍图像等多元数据源中学习视觉表征。本技术报告介绍DINOv3，这是通过运用简单而有效的策略实现该愿景的重要里程碑。首先，我们通过精心的数据准备、设计和优化，实现了数据集和模型规模的双重扩展优势。其次，我们提出名为“格拉姆锚定”的新方法，有效解决了长期训练中密集特征图质量下降这一已知但未解决的问题。最后，我们应用后处理策略进一步提升模型在分辨率、模型规模及文本对齐方面的灵活性。最终，我们推出了一款多功能视觉基础模型，无需微调即可在多种场景中超越现有专业级模型。DINOv3生成的高质量密集特征在各类视觉任务中表现卓越，其性能显著超越以往的自监督和弱监督基础模型。我们还提供DINOv3视觉模型套件，旨在通过为不同资源限制和部署场景提供可扩展解决方案，推动广泛任务和数据领域的前沿技术发展。

1 介绍

基础模型已成为现代计算机视觉的核心构建模块，通过单一可复用模型实现跨任务和领域的广泛泛化。自监督学习（SSL）是训练此类模型的强有力方法，它直接从原始像素数据中学习，并利用图像中模式的自然共现。与需要高质量元数据配对图像的弱监督和完全监督预训练方法不同（Radford等人，2021；Dehghani等人，2023；Bolya等人，2025），SSL解锁了对海量原始图像集合的训练。由于几乎无限的训练数据可用，这对于训练大规模视觉编码器尤其有效。DINOv2（Oquab等人，2024）体现了这些优势，在图像理解任务中取得令人印象深刻的结果（Wang等人，2025），并实现了对组织病理学等复杂领域的预训练（Chen等人，2024）。基于自监督学习（SSL）训练的模型展现出诸多优势：它们能有效应对输入分布变化，提供强大的全局与局部特征，并生成有助于理解物理场景的丰富嵌入向量。由于SSL模型无需针对特定下游任务进行训练，因此能生成通用性强且稳健的特征。例如DINOv2模型无需任务特定微调即可在不同领域和任务中表现优异，使得单一冻结主干网络可实现多场景应用。值得注意的是，自监督学习特别适合利用海量观测数据进行训练。

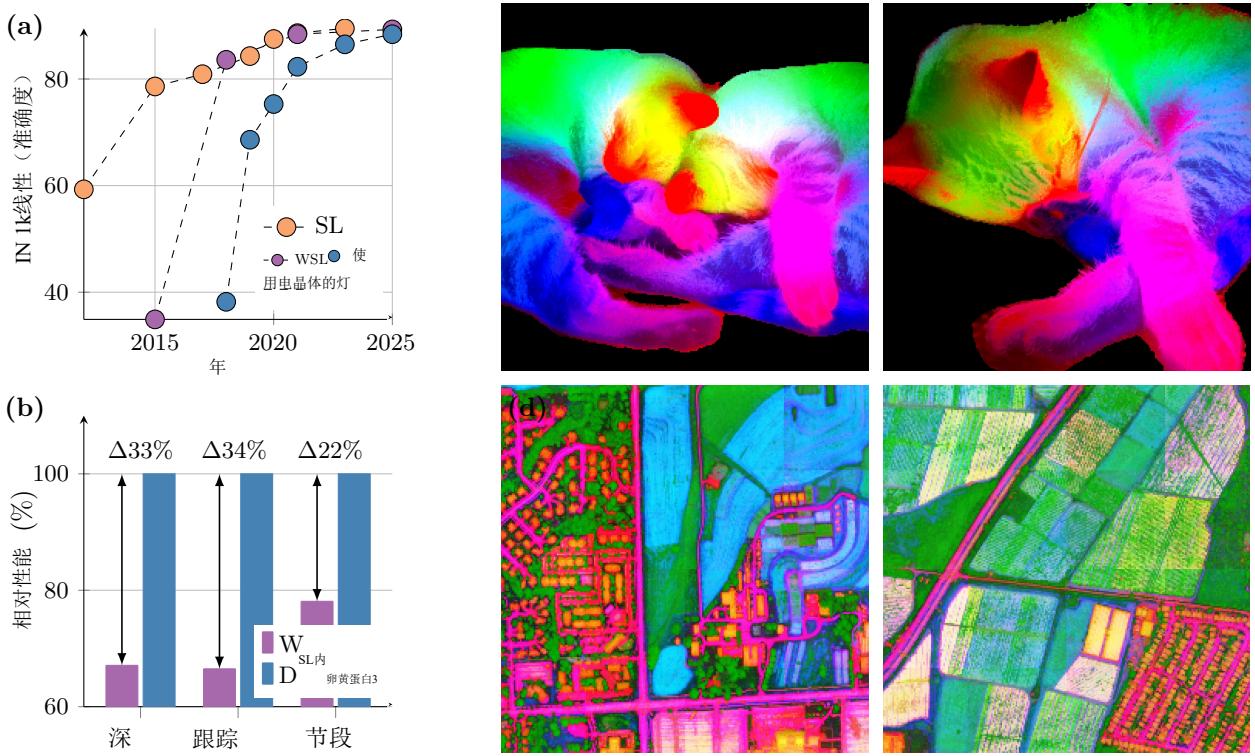


图1: (a)ImageNet1k (IN1k) 数据集上线性探测结果的演变趋势，对比了全监督学习 (SL)、弱监督学习 (WSL) 与自监督学习 (SSL) 方法。尽管SSL出现较晚，但其发展迅速，目前已达到近年来ImageNet的准确率平台期。另一方面，我们证明SSL具有生成高质量密集特征的独特优势。通过DINOv3模型，我们在密集任务上显著优于弱监督模型，这一点可通过最佳WSL模型与DINOv3的相对性能对比(b)得到验证。我们还生成了基于自然图像(c)和航拍图像(d)训练的DINOv3模型对高分辨率图像特征的主成分分析图。

诸如组织病理学（Vorontsov 等人, 2024）、生物学（Kim 等人, 2025）、医学影像（Perez-Garcia 等人, 2025）、遥感（Cong 等人, 2022; Tolan 等人, 2024）、天文学（Parker 等人, 2024）或高能粒子物理学（Dillon 等人, 2022）等领域。这些领域通常缺乏元数据，且已证明能从DINOv2等基础模型中获益。最后，无需人工干预的SSL非常适合在日益增长的网络数据量中实现终身学习。

在实际应用中，SSL技术的核心优势——即通过海量无约束数据生成任意规模的强大模型——在大规模部署时仍面临挑战。虽然Oquab 等人（2024）提出的启发式方法有效缓解了模型不稳定性与崩溃问题，但进一步扩展仍存在三大瓶颈：首先，如何从无标注数据集中获取有效数据仍无定论；其次，常规训练实践中采用余弦调度需要预先确定优化时间范围，这在处理大规模图像语料库时尤为困难；第三，特征性能在初期训练后逐渐下降，通过视觉检查斑块相似度图谱即可验证。当模型规模超过ViT-Large（3亿参数）时，这种现象在更长的训练周期中愈发明显，使得DINOv2的扩展应用价值大打折扣。

为解决上述问题，我们开发了DINOv3这一突破性工作，实现了大规模SSL训练的突破。研究表明，仅需一个冻结的SSL主干网络即可作为通用视觉编码器，在具有挑战性的下游任务中达到顶尖性能，其表现优于监督学习和依赖元数据的预训练策略。本研究以三大目标为导向：(1)构建跨任务、跨领域的通用基础模型；(2)改进现有SSL模型在密集特征处理上的不足；(3)推广可直接使用的模型家族。下文将详细阐述这三个目标。

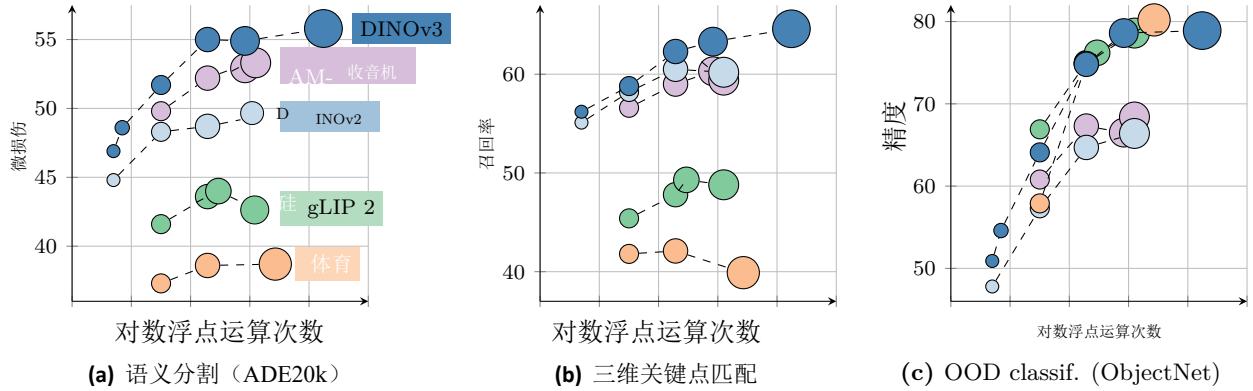


图2：DINOv3模型家族与其他自监督或弱监督模型家族在不同基准上的性能对比。DINOv3在密集基准上显著超越其他模型，包括利用掩码标注先验的模型，如AM-radio（Heinrich 等人，2025）。

强大且多功能的基础模型DINOv3致力于在两个维度上实现高度灵活性，这得益于模型规模与训练数据的可扩展性。首先，自监督学习（SSL）模型的核心优势在于保持模型冻结状态时仍能保持卓越性能，理想情况下可达到与专业模型相当的前沿水平。此时，单次前向传播即可在多个任务中输出尖端成果，从而大幅节省计算资源——这对实际应用（尤其是边缘设备）具有关键优势。我们将在**第6节**展示DINOv3成功应用于的广泛任务范围。其次，不依赖元数据的可扩展SSL训练流程为众多科学应用打开了新天地。通过在多样化图像集（无论是网络图片还是观测数据）上进行预训练，SSL模型能在大量领域和任务中实现泛化。如**图1(d)**所示，从高分辨率航拍图像中提取的DINOv3特征主成分分析（PCA）清晰区分了道路、房屋和绿化区域，充分彰显了该模型的特征质量。

通过Gram锚定实现的优质特征图DINOv3的另一大亮点在于其密集特征图的显著提升。DINOv3的SSL训练策略旨在打造既能处理高级语义任务，又能生成优质特征图的模型——这些特征图可直接用于深度估计、三维匹配等几何任务。具体而言，模型需要生成无需大量后处理即可直接使用的密集特征。在海量图像训练中，密集特征与全局表征之间的平衡尤为难以优化，因为高级理解目标可能与密集特征图质量产生冲突。这种矛盾目标导致大模型和长训练周期下密集特征容易坍塌。我们创新的Gram锚定策略有效缓解了这一问题（详见**第4节**）。最终，DINOv3获得的密集特征图质量远超DINOv2，即便在高分辨率下仍保持清晰（见**图3**）。

DINOv3模型家族通过格拉姆锚定技术解决密集特征图的退化问题，成功解锁了模型扩展的潜力。因此，使用自监督学习训练更大规模模型能显著提升性能。本研究成功训练出参数量达70亿的DINO模型。由于这类大型模型运行需要大量资源，我们采用蒸馏技术将其知识压缩为更小的变体。最终推出DINOv3视觉模型家族，这套全面的模型套件旨在应对各类计算机视觉挑战。该系列模型致力于通过提供可扩展的解决方案，突破现有技术瓶颈，适应不同资源限制和部署场景。蒸馏过程生成的模型变体涵盖多个规模：包括Vision Transformer（ViT）小、基、大三种版本，以及基于ConvNeXt架构的模型。值得注意的是，高效且广泛应用的ViT-L模型在多种任务中表现接近原始70亿参数的教师模型。总体而言，DINOv3家族在广泛的基准测试中表现出强大的性能，在全局任务上与竞争模型的准确率相匹配或超过，而在密集预测任务上显著优于它们，如**图2**所示。



图3：高分辨率密集特征。我们可视化了DINOv3输出特征在红色十字标记区域与其他所有区域之间的余弦相似度图。输入图像尺寸为 4096×4096 。请放大查看，您是否认同DINOv3？

贡献概述在本研究中，我们提出多项贡献以应对将SSL扩展至大型前沿模型的挑战。我们基于自动数据整理的最新进展（Vo等，2024），构建了一个大型“背景”训练数据集，并将其与少量专用数据（ImageNet-1k）进行精心混合。这使得能够利用大量无约束数据来提升模型性能。关于数据扩展的贡献（i）将在第3.1节中详细阐述。

我们通过定制ViT架构的变体，将主模型参数量提升至70亿。该架构不仅采用现代位置嵌入（轴向RoPE），还开发了防止位置伪影的正则化技术。与DINOv2采用的多余弦调度方案不同，我们采用恒定超参数调度进行100万次迭代训练，从而获得性能更优的模型。关于模型架构与训练的这一改进（ii），将在第3.2节中详细阐述。

通过上述技术，我们能够大规模地按照DINOv2算法训练模型。但正如前文所述，规模效应会导致密集特征的退化。为解决这一问题，我们提出在流程中引入格拉姆锚定训练阶段作为核心改进。该方法能有效清除特征图中的噪声，生成出色的相似性图，并显著提升参数化与非参数化密集任务的性能表现。关于格拉姆训练的这项贡献（iii），将在第4节中详细阐述。

遵循既往实践，我们流程的最后阶段包括高分辨率后训练阶段及蒸馏为一系列不同规模的高性能模型。针对后者，我们开发了一种新颖的

高效的单教师多学生蒸馏流程。本贡献 (iv) 将我们7B前沿模型的强大性能转移至一系列更小的实用模型，这些模型适用于常见场景，具体描述见第5.2节。

通过我们全面的基准测试结果，第6节显示我们的方法在密集任务中确立了新标准，在全局任务上与CLIP衍生方法表现相当。特别值得一提的是，通过冻结视觉主干网络，我们在物体检测（COCO检测，mAP 66.1）和图像分割（ADE20k，mIoU 63.0）等长期存在的计算机视觉难题上取得了业界领先水平，超越了专门的微调流程。此外，我们通过将DINOv3算法应用于卫星影像（详见第8节），证明了该方法在不同领域的通用性，其表现超越了所有先前方法。

2 相关工作

自监督学习无需标注的学习需要通过人工学习任务提供监督来替代训练。自监督学习的艺术与挑战在于精心设计这些所谓的预文本任务，以便为下游任务学习强大的表征。语言领域因其离散性提供了设置此类任务的直接方法，这导致了许多针对文本数据的成功无监督预训练方法。示例包括词嵌入（Mikolov 等人，2013；Bojanowski 等人，2017）、句子表征（Devlin 等人，2018；Liu 等人，2019）以及纯语言模型（Mikolov 等人，2010；Zaremba 等人，2014）。相比之下，计算机视觉由于信号的连续性带来了更大挑战。早期尝试模仿语言方法从图像部分提取监督信号以预测其他部分，例如通过预测相对补丁位置（Doersch 等人，2015），图像块重排序（Noroozi 和 Favaro，2016；Misra 和 Maaten，2020），或图像修复（Pathak 等人，2016）。其他任务涉及图像重着色（Zhang 等人，2016）或预测图像变换（Gidaris 等人，2018）。

在这些任务中，基于图像修复的方法因其基于补丁的ViT架构的灵活性而受到广泛关注（He 等人，2021；Bao 等人，2021；El-Nouby 等人，2021）。其目标是重建图像的受损区域，这可视为一种去噪自动编码形式，并在概念上与BERT预训练中的掩码标记预测任务相关（Devlin 等人，2018）。值得注意的是，He 等人（2021）证明基于像素的掩码自动编码器（MAE）可作为下游任务微调的强初始化。随后，Baeviski 等人（2022；2023）；Assran 等人（2023）表明，预测学习到的潜在空间而非像素空间能产生更强大、更高层次的特征——这种学习范式被称为 JEPA：“联合嵌入预测架构”（LeCun，2022）。最近，JEPAs也被扩展到视频训练领域（Bardes 等人，2024；Assran 等人，2025）。

第二类研究方向与我们更为接近，其利用图像间的判别信号来学习视觉表征。这类方法可追溯至早期深度学习研究（Hadsell 等人，2006），但随着实例分类技术的引入而流行起来（Dosovitskiy 等人，2016；Bojanowski 和 Joulin，2017；Wu 等人，2018）。后续进展引入了对比目标和信息论准则（Henaf 等人，2019；He 等人，2020；Chen 和 He，2020；Chen 等人，2020a；Grill 等人，2020；Bardes 等人，2021），以及基于自聚类的策略（Caron 等人，2018；Asano 等人，2020；Caron 等人，2020；2021）。较新的方法，如iBOT（Zhou 等人，2021），将这些判别损失与掩码重建目标相结合。所有这些方法都显示出学习强特征的能力，并在ImageNet等标准基准上取得高性能（Russakovsky 等人，2015）。然而，大多数方法在扩展到更大模型规模时面临挑战（Chen 等人，2021）。

视觉基础模型深度学习革命始于AlexNet的突破（Krizhevsky 等人，2012），这是一种在ImageNet挑战中超越所有先前方法的深度卷积神经网络（Deng 等人，2009；Russakovsky 等人，2015）。早在初期，人们就发现端到端学习的特征在大规模人工标注的ImageNet数据集上对多种迁移学习任务具有高效性（Oquab 等人，2014）。随后关于视觉基础模型的早期研究聚焦于架构开发，包括 VGG（Simonyan 和 Zisserman，2015）、GoogleNet（Szegedy 等人，2015）以及ResNets（He 等人，2016）。

考虑到缩放的有效性，后续研究探索了在大数据集上训练更大模型。Sun 等人（2017）通过包含3亿条数据的专有 JFT 数据集扩展了监督训练数据。

标注图像展示了令人印象深刻的结果。JFT 还为 Kolesnikov 等人（2020）带来了显著的性能提升。与此同时，研究者们探索了结合监督与非监督数据的扩展方法。例如，可以使用 ImageNet 监督模型为非监督数据生成伪标签，进而训练更大的网络（Yalniz 等人，2019）。随后，像 JFT 这样的大型监督数据集的出现，也促进了 Transformer 架构在计算机视觉中的应用（Dosovitskiy 等人，2020）。特别值得注意的是，若不使用 JFT，要达到与原始视觉 Transformer (ViT) 相当的性能需要付出巨大努力（Touvron 等人，2020；2022）。得益于 ViTs 的学习能力，Zhai 等人（2022a）进一步扩展了扩展工作，最终推出了超大规模的 ViT-22B 编码器（Dehghani 等人，2023）。

鉴于手动标注大型数据集的复杂性，弱监督训练——即通过图像关联的元数据生成标注——为监督训练提供了有效替代方案。早期，Joulin 等人（2016）证明网络可通过简单预测图像标题中所有词汇作为目标进行预训练。这一初始方法通过利用句子结构（Li 等人，2017）、整合其他类型元数据并涉及数据整理（Mahajan 等人，2018）以及扩展性（Singh 等人，2022）得到进一步完善。然而，弱监督算法仅在对比损失和标题表示联合训练的引入下才充分发挥潜力，例如 Align（Jia 等人，2021）和 CLIP（Radford 等人，2021）所展示的。

这种方法的成功激发了大量开源复现与扩展努力。OpenCLIP（Cherti 等人，2023）是首个通过 LAION 数据集训练复现 CLIP 的开源项目（Schuhmann 等人，2021）；后续研究采用 CLIP 风格的微调方式利用预训练骨干网络（Sun 等人，2023；2024）。鉴于数据收集是 CLIP 训练成功的关键因素，MetaCLIP（Xu 等人，2024）严格遵循原始 CLIP 流程复现其结果，而 Fang 等人（2024a）则使用监督数据集来整理预训练数据。其他研究聚焦于改进训练损失，例如 SigLIP 采用 Sigmoid 损失（Zhai 等人，2023），或利用预训练图像编码器（Zhai 等人，2022b）。然而最终，获取前沿基础模型最关键的组件是丰富的高质量数据和充足的计算资源。在这方面，SigLIP 2（Tschanne 等人，2025）和感知编码器（PE）（Bolya 等人，2025）在训练超过 400 亿图像-文本对后取得了令人印象深刻的结果。最大的 PE 模型在 860 亿样本上训练，全局批量大小为 131K。最后，一系列更复杂且原生多模态的方法已被提出；这些包括对比式标题生成（Yu 等人，2022）、潜在空间中的掩码建模（Bao 等人，2021；Wang 等人，2022b；Fang 等人，2023；Wang 等人，2023a）以及自回归训练（Fini 等人，2024）。

相比之下，关于无监督图像预训练的扩展研究相对较少。早期尝试包括 Caron 等人（2019）和 Goyal 等人（2019）利用 YFCC 数据集（Thomee 等人，2016）。通过聚焦更大规模数据集和模型（Goyal 等人，2021；2022a），以及 SSL 数据整理的初步尝试（Tian 等人，2021），研究取得了进一步进展。训练算法的精细调优、更大规模架构和更丰富的训练数据带来了 DINOv2 的卓越成果（Oquab 等人，2024）——这是首次有 SSL 模型在多项任务上达到或超越开源 CLIP 变体。该方向近期由 Fan 等人（2025）通过不进行数据整理直接扩展至更大模型，或 Venkataraman 等人（2025）采用开放数据集和改进训练方案所推动。

密集 Transformer 特征 现代视觉应用广泛使用预训练 Transformer 的密集特征，包括多模态模型（Liu 等人，2023；Beyer 等人，2024）、生成模型（Yu 等人，2025；Yao 等人，2025）、三维理解（Wang 等人，2025）、视频理解（Lin 等人，2023a；Wang 等人，2024b）以及机器人技术（Driess 等人，2023；Kim 等人，2024）。此外，诸如检测、分割或深度估计等传统视觉任务需要精确的局部描述符。为提升自监督学习训练的局部描述符质量，大量研究致力于开发局部自监督损失。例如利用视频中的时空一致性，如使用点轨迹环作为训练信号（Jabri 等人，2020），利用在相同图像的不同作物之间进行空间对齐（Pinheiro 等人，2020；Bardes 等人，2022），或强制相邻补丁之间保持一致性（Yun 等人，2022）。Darisetty 等人（2025）表明，预测聚类局部补丁可提升密集表征质量。DetCon（Henaf 等人，2021）和 ORL（Xie 等人，2021）在

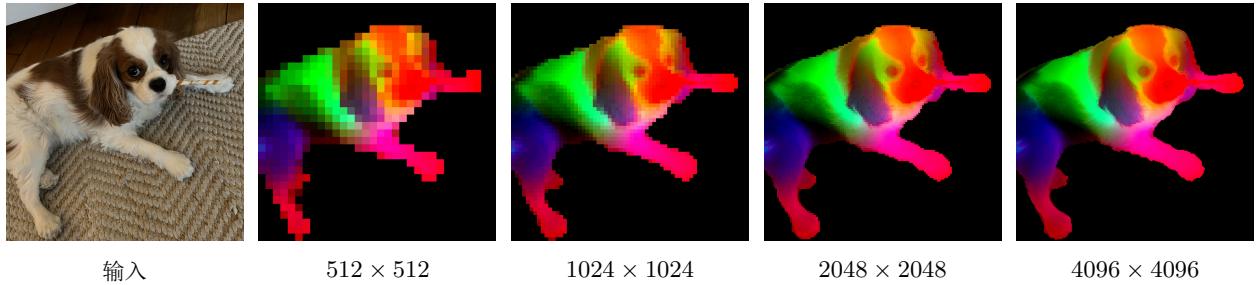


图4：极高分辨率下的DINOv3。我们通过将特征空间上计算的主成分分析（PCA）前三个分量映射到RGB空间，可视化了DINOv3的密集特征。为聚焦PCA主体，我们通过背景减除对特征图进行掩膜处理。随着分辨率提升，DINOv3生成的特征既保持清晰又保持语义意义。更多PCA分析结果详见第6.1.1节。

区域提议但假设此类提议先验存在；这一假设通过ODIN (Henaf 等人, 2022) 和 SlotCon (Wen 等人, 2022) 等方法得到放宽。在不改变训练目标的情况下，Dariset 等人 (2024) 表明在输入序列中添加寄存器标记能显著提升密集特征图质量，近期研究发现无需模型训练即可实现 (Jiang 等人, 2025; Chen 等人, 2025)。

近期趋势是基于蒸馏的“聚类”方法，这些方法结合了来自多个图像编码器的信息，这些编码器具有不同的全局和局部特征质量，并通过不同级别的监督进行训练 (Ranzinger 等人, 2024; Bolya 等人, 2025)：AM-radio (Ranzinger 等人, 2024) 将完全监督的SAM (Kirillov 等人, 2023)、弱监督的CLIP以及自监督的DINOv2的优势整合到一个统一的主干网络中。Perception Encoder (Bolya 等人, 2025) 类似地将SAM (v2) 蒸馏为一种称为PEspatial的专用密集变体。他们使用一个目标函数来强制学生和教师补丁之间的余弦相似度保持较高，其中教师是通过掩码注释进行训练的。类似的损失函数在风格迁移的背景下被证明是有效的，通过减少特征维度的Gram矩阵之间的不一致性 (Gatys 等人, 2016; Johnson 等人, 2016; Yoo 等人, 2024)。本研究采用Gram目标函数对师生图像块的余弦相似度进行正则化处理，以促进其相互接近。具体而言，我们选用自监督学习 (SSL) 模型的早期迭代结果作为教师模型，这表明早期阶段的SSL模型能有效指导全局任务与密集任务的训练。

其他研究聚焦于对SSL训练模型局部特征的事后改进。例如，Ziegler 和 Asano (2022) 通过密集聚类目标对预训练模型进行微调；类似地，Salehi 等人 (2023) 通过时间对齐补丁特征进行微调，两种方法均提升了局部特征质量。更贴近我们研究方向的是，Pariza 等人 (2025) 提出基于补丁排序的目标函数，鼓励学生和教师生成具有一致邻域排序的特征。在不进行微调的情况下，STEGO (Hamilton 等人, 2022) 在冻结的SSL特征基础上学习非线性投影，以形成紧凑的聚类并放大相关模式。另一种方法是，Simoncini 等人 (2024) 通过将不同自监督目标的梯度与冻结的SSL特征拼接来增强自监督特征。最近，Wysoczanska 等人 (2024) 表明，通过加权平均补丁可显著改善噪声特征图。

与SSL相关但非其专属的近期研究，通过ViT特征图生成高分辨率特征图 (Fu 等人, 2024)，但因图像的块化处理常导致分辨率较低。与这些研究不同，我们的模型能原生输出高质量的密集特征图，且在不同分辨率下保持稳定一致，如图4所示。

3 无监督大规模训练

DINOv3是新一代视觉模型，通过突破自监督学习的边界，致力于打造迄今为止最稳健且灵活的视觉表征。我们从大型语言模型 (LLMs) 的成功经验中汲取灵感——这类模型通过扩展容量就能获得卓越的新特性。借助规模大一个数量级的模型和训练数据集，我们旨在充分释放自监督学习的潜力，推动计算机视觉领域实现类似的范式革新，摆脱现有技术的限制。

表1：通过下游任务性能展示训练数据对特征质量的影响。我们比较了采用聚类（Vo 等人，2024）和检索（Oquab 等人，2024）方法整理的数据集与原始数据及我们的数据混合集。该消融研究在较短的20万次迭代周期内完成。

数据集	IN1k k近邻	IN1k线性	对象网络	i自然学家2021	巴黎检索
生的	80.1	84.8	70.3	70.1	63.3
聚类	79.4	85.4	72.3	81.3	85.2
检索	84.0	86.7	70.7	86.0	82.7
LVD -1689M（我方）	84.6	87.2	72.8	87.0	85.9

这与传统监督学习或任务特定方法的本质区别在于，自监督学习（SSL）能生成丰富且高质量的视觉特征，这些特征不会偏向任何特定的监督方式或任务类型，从而为各类下游应用提供了通用基础。尽管此前尝试扩展SSL模型时常受不稳定因素制约，本节将阐述如何通过精心的数据准备、设计和优化来实现扩展优势。我们首先介绍数据集构建流程（第3.1节），随后展示DINOv3模型首阶段训练采用的自监督SSL方案（第3.2节），包括架构选择、损失函数和优化技术。第二阶段训练将聚焦密集特征，具体细节将在第4节中详细说明。

3.1 数据准备

数据规模是大型基础模型成功背后的关键驱动因素之一（Touvron 等人，2023；Radford 等人，2021；Xu 等人，2024；Oquab 等人，2024）。然而，单纯增加训练数据规模并不必然转化为更高的模型质量及下游基准测试中的更好表现（Goyal 等人，2021；Oquab 等人，2024；Vo 等人，2024）：成功的数据规模优化通常涉及精心设计的数据管理流程。这些算法可能具有不同目标：要么聚焦于提升数据多样性与平衡性，要么关注数据实用性——即其与常见实际应用的相关性。在开发DINOv3时，我们结合两种互补方法来提升模型的泛化能力和性能，从而在这两个目标间取得平衡。

数据收集与整理我们通过利用从Instagram公开帖子中收集的海量网络图像数据池来构建大规模预训练数据集。这些图像已通过平台级内容审核以防止有害内容，我们获得约170亿张图像的初始数据池。基于此原始数据池，我们创建了三个数据集部分。第一部分通过应用基于 k -均值的自动整理方法构建，该方法源自Vo 等人（2024）。我们采用DINOv2作为图像嵌入，并使用5级聚类，聚类数量从最低到最高分别为2亿、800万、80万、10万和2.5万。构建聚类层级后，我们应用Vo 等人（2024）提出的平衡采样算法，最终得到16.89亿张图像的整理子集（命名为LVD -1689M），确保对网络上所有视觉概念的均衡覆盖。第二部分采用与Vo 等人（2024）提出的流程类似的检索式整理系统。Oquab 等人（2024）。我们从数据池中检索与选定种子数据集相似的图像，创建一个涵盖下游任务相关视觉概念的数据集。在第三部分中，我们使用公开可用的原始计算机视觉数据集，包括ImageNet1k（Deng 等人，2009）、ImageNet22k（Russakovsky 等人，2015）以及Mapillary街景序列（Warburg 等人，2020）。这一最终部分使我们能够按照Oquab 等人（2024）的方法优化模型性能。

数据采样在预训练阶段，我们使用采样器将不同数据部分混合。对于上述数据组件的混合有多种选择方案：一种是使用来自单个随机选择组件的同质数据批次进行训练；另一种则是通过按特定比例组合所有组件数据来优化异质数据批次。受Charton 和 Kempe（2024）的启发——他们发现由小规模数据集中的高质量数据组成的同质批次更有益——我们在每次迭代中进行随机采样。

表2: DINov2与DINov3模型中教师网络架构的对比。我们保留了模型

深度为40个块，并将嵌入维度增加至4096。重要的是，我们采用16像素的补丁尺寸，从而改变给定分辨率下的有效序列长度。

教师模型	DINov2	DINov3
脊梁骨	ViT-巨	ViT-7B
#参数	1.1B	6.7B
#区块	40	40
补丁大小	14	16
位置嵌入	可学习的	罗佩
寄存器	4	4
嵌入。	1536	4096
FFN 类型	SwiGLU	SwiGLU
FFN 隐形。	4096	8192
收件人: 负责人	24	32
注意: 头部尺寸	64	128
DINO头 MLP	4096-4096-256	8192-8192-512
DINO原型	128k	256k
iBOT 头部 MLP	4096-4096-256	8192-8192-384
iBOT原型机	128k	96k

要么仅使用ImageNet1k的同质批次数据，要么混合使用其他所有组件的数据。在我们的训练中，ImageNet1k的同质批次数据占训练集的10%。

数据消融实验为评估数据清洗技术的效果，我们通过消融实验将数据混合方案与仅采用聚类或检索方法清洗的数据集、以及原始数据池进行对比。具体操作中，我们在每个数据集上训练模型，并通过标准下游任务评估其性能表现。为提升效率，我们采用20万次迭代的精简方案替代了100万次迭代。如表1所示，没有任何单一清洗技术能在所有基准测试中表现最优，而我们的完整清洗流程恰好实现了两全其美。

3.2 大规模自监督训练

虽然基于SSL训练的模型已展现出有趣特性（Chen 等人，2020b；Caron 等人，2021），但大多数SSL算法尚未扩展到更大规模模型。这要么是由于训练稳定性问题（Darcet 等人，2025），要么是过于简化的解决方案未能捕捉视觉世界的全部复杂性。当进行大规模训练时（Goyal 等人，2022a），基于SSL训练的模型未必能表现出令人印象深刻的性能。一个显著例外是DINov2——这个拥有11亿参数的模型在精选数据上训练，其性能与弱监督模型如CLIP相当（Radford 等人，2021）。近期将DINov2扩展至70亿参数的尝试（Fan 等人，2025）在全局任务上展现出良好结果，但在密集预测上表现令人失望。本文旨在扩大模型和数据规模，获得兼具改进全局与局部特性的更强大视觉表征。

学习目标我们采用一种判别式自监督策略训练模型，该策略融合了多个自监督目标，同时包含全局和局部损失项。基于DINov2（Oquab 等人，2024），我们使用图像级目标（Caron 等人，2021） $\mathcal{L}_{\text{DINO}}$ ，并用补丁级潜在重构目标（Zhou 等人，2021） $\mathcal{L}_{\text{iBOT}}$ 进行平衡。同时，我们在两个目标中将DINO的中心化替换为SwAV（Caron 等人，2020）的Sinkhorn-Knopp。每个目标均通过主干网络顶部专用头的输出计算，允许在损失计算前对特征进行一定程度的特化。此外，我们对局部和全局裁剪的主干网络输出应用了专用层归一化。实验发现，这种改变在训练后期稳定了ImageNet kNN分类（+0.2准确率）并提升了密集性能（例如ADE20k分割+1 mIoU，-0.02 RMSE在NYUV2深度估计中）。此外，添加了Koleo正则化器 $\mathcal{L}_{\text{Koleo}}$ 以鼓励批次内的特征在空间中均匀分布（Sablayrolles 等人，2018）。我们使用

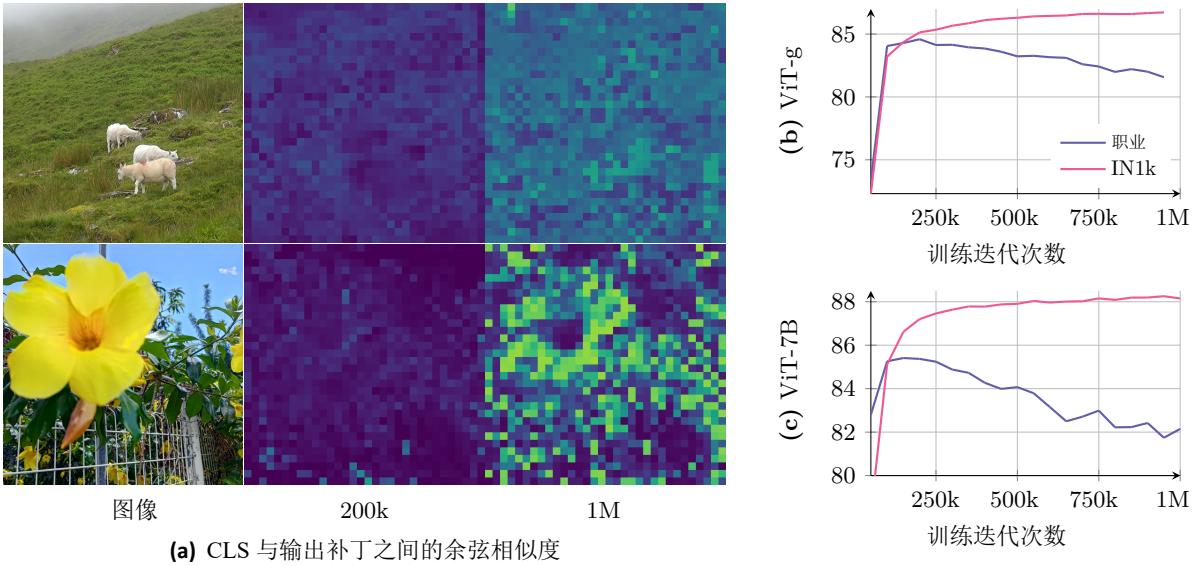


图5：ViT-g(a)和ViT-7B (b、c) 在ImageNet1k线性分类 (IN1k) 和VOC分割任务上的余弦相似度与准确率变化曲线。研究发现，当图像块标记与类别标记的余弦相似度较低时，分割性能达到最佳。随着训练进程，相似度逐渐升高，导致密集任务的性能下降。

Koleo的分布式实现中，损失函数以16个样本的小批量形式应用——可能跨多个GPU进行。我们的初始训练阶段通过优化以下损失函数来完成：

$$\mathcal{L}_{\text{Pre}} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + 0.1 * \mathcal{L}_{\text{DKoleo}}. \quad (1)$$

更新的模型架构在本研究的模型扩展方面，我们将模型规模扩大至70亿参数，并在表2中提供了与DINOv2研究中训练的11亿参数模型对应的超参数对比。我们还采用了一种定制化的RoPE变体：基础实现为每个图像块分配归一化的[-1,1]坐标框，然后根据两个图像块的相对位置在多头注意力操作中施加偏置。为增强模型对分辨率、尺度和纵横比的鲁棒性，我们采用了*RoPE-box抖动*技术。坐标框[-1,1]被随机缩放至[-s, s]，其中s∈[0.5,2]。这些改进使DINOv3能够更好地学习精细且鲁棒的视觉特征，从而提升其性能和扩展性。

优化在超大规模数据集上训练大型模型是一个复杂的实验流程。由于模型容量与训练数据复杂度之间的相互作用难以先验评估，因此无法预估正确的优化时序。为解决这一问题，我们摒弃了所有参数调度，采用恒定学习率、权重衰减和教师EMA动量进行训练。这具有两大优势：首先，只要下游性能持续提升，我们就能继续训练；其次，优化超参数数量减少，便于合理选择。为确保训练顺利启动，我们仍采用线性热身策略来调整学习率和教师温度。遵循常规做法，我们使用AdamW (Loshchilov 和 Hutter, 2017)，并将总批次大小设置为4096张图像，分配至256个GPU。我们采用多裁剪策略 (Caron 等人, 2020)，每张图像取2个全局裁剪和8个局部裁剪。我们采用边长为256/112像素的方形图像进行全局/局部裁剪，结合补丁尺寸的变化，使得每张图像的有效序列长度与DINOv2保持一致，每批次总序列长度为370万标记。更多超参数详见附录C及代码发布。

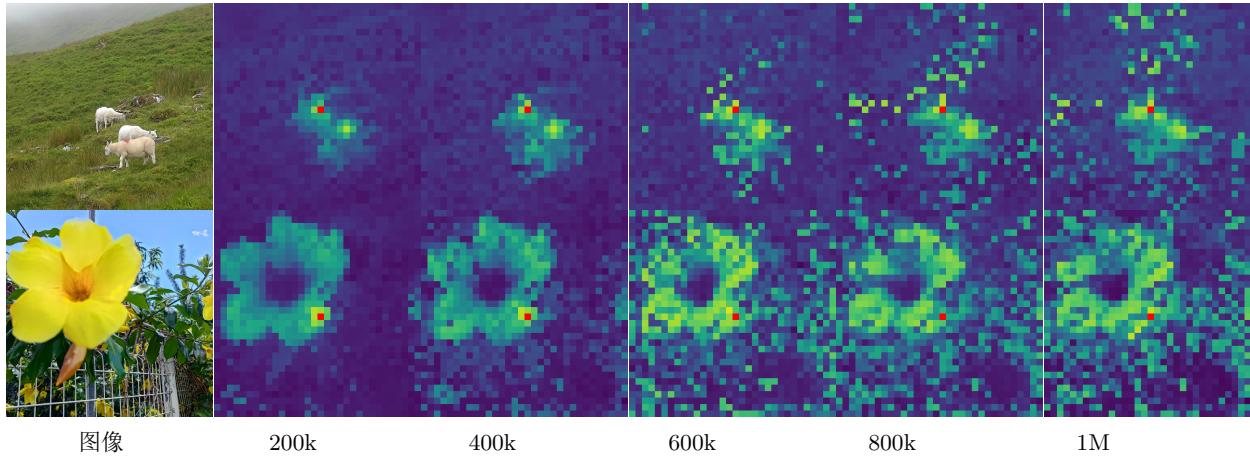


图6：红色标注补丁与其他所有补丁间余弦相似度的演变。随着训练进程，模型生成的特征逐渐失去局部性，相似度图谱的噪声水平随之升高。

4 Gram锚定:一种稠密特征的正则化方法

为充分挖掘大规模训练的优势，我们计划对7B模型进行长期训练，其理念是该模型可能实现无限期训练。正如预期，长时间训练确实提升了全局基准的性能。然而随着训练进程的推进，模型在密集任务上的表现却逐渐下降（[图5b](#)和[图5c](#)）。这种现象源于特征表征中出现的块级不一致性，这削弱了长期训练的初衷。¹在本节中，我们首先分析了块级一致性损失，随后提出了一种名为格拉姆锚定的新目标来缓解该问题。最后，我们将讨论该方法对训练稳定性和模型性能的影响。

4.1 训练过程中补丁级一致性缺失

在长期训练过程中，我们观察到全局指标持续提升，但在密集预测任务上的表现却明显下降。这种现象在DINOv2训练中曾被发现（程度较轻），并在[Fan 等人](#)的扩展研究中有所讨论（2025）。但据我们所知，该问题至今仍未解决。我们在[图5b](#)和[图5c](#)中展示了该现象，这些图表呈现了模型在图像分类和分割任务上的迭代性能表现。分类任务中，我们使用CLS标记在ImageNet-1k数据集上训练线性分类器，并报告前1%准确率；分割任务中，我们基于Pascal VOC提取的图像块特征训练线性层，并报告平均交并比（mIoU）。我们发现，无论是ViT-g还是ViT-7B，分类准确率在训练过程中都呈现单调提升趋势。然而分割性能在约20万次迭代后均出现下滑，其中ViT-7B的性能甚至回落至早期水平以下。

为深入理解这种性能退化现象，我们通过可视化分析斑块特征间的余弦相似度来评估其质量。[图6](#)展示了主干网络输出斑块特征与参考斑块（红色高亮标注）的余弦相似度分布图。在20万次迭代时，相似度分布图呈现平滑且局部集中的特征，表明斑块级表征保持稳定。但当迭代次数达到60万次后，分布图质量显著下降，出现大量与参考斑块高度相似的无关斑块。这种斑块级一致性缺失与密集任务性能的下降呈现相关性。

这些补丁级别的不规则性与[Dariset 等人](#)（2024）描述的高范数补丁异常值不同。具体而言，随着注册标记的整合，补丁范数在整个训练过程中保持稳定。然而我们注意到，CLS标记与补丁输出之间的余弦相似度在训练期间逐渐增加。这虽在意料之中，但意味着补丁特征的局部性有所减弱。我们在[图5a](#)中可视化了这一现象，该图展示了20万次和100万次迭代时的余弦图。为了缓解

¹我们还观察到随着训练的持续进行，出现了不同类型的数据异常值；相关讨论详见[附录A](#)。

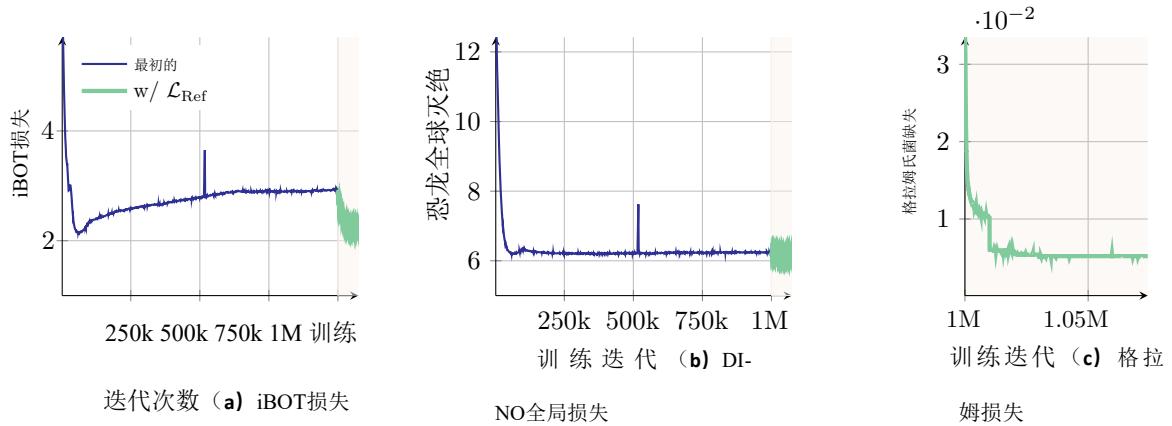


图7：通过训练迭代展示的补丁级iBOT损失、全局损失DINO（应用于全局裁剪）以及新引入的Gram损失的演变过程。我们特别标注了使用Gram目标的精炼步骤 \mathcal{L}_{Ref} 的迭代过程。

针对密集任务，我们提出了一种新目标函数，该函数专门用于正则化补丁特征并确保良好的补丁级一致性，同时保持高全局性能。

4.2 牙槽基台

在实验过程中，我们发现学习强判别特征与保持局部一致性之间存在相对独立性，这体现在全局性能与密集性能之间缺乏相关性。虽然将全局DINO损失与局部iBOT损失相结合已开始解决这一问题，但我们观察到这种平衡并不稳定，随着训练进程的推进，全局表征逐渐占据主导地位。基于这一发现，我们提出了一种创新解决方案，该方案明确利用了这种独立性。

我们引入了一个新目标，通过强制执行补丁级一致质量来缓解补丁级一致性退化，同时不影响特征本身。这个新的损失函数作用于格拉姆矩阵：即图像中所有补丁特征两两点积的矩阵。我们希望将学生模型的格拉姆矩阵推向早期模型（称为格拉姆教师）的格拉姆矩阵。我们通过选取教师网络的早期迭代来选择格拉姆教师，该迭代表现出更优的稠密特性。由于作用于格拉姆矩阵而非特征本身，只要相似性结构保持不变，局部特征即可自由移动。假设我们有一个由 P 个补丁组成的图像，以及一个在 d 维空间中运行的网络。用 \mathbf{X}_S （分别用 $\mathbf{X}\mathbf{G}$ 表示）表示学生（分别表示格拉姆教师）的 $P \times d$ 维 L_2 归一化局部特征矩阵。我们定义损失 $\mathcal{L}_{\text{Gram}}$ 如下：

$$\mathcal{L}_{\text{Gram}} = \|\mathbf{X}_S \cdot \mathbf{X}_S^\top - \mathbf{X}_G \cdot \mathbf{X}_G^\top\|_{\text{F}}^2. \quad (2)$$

我们仅在全局裁剪数据上计算该损失函数。虽然该方法可在训练初期应用，但为提升效率，我们选择在完成100万次迭代后才开始使用。有趣的是，我们发现即使在后期应用 $\mathcal{L}_{\text{Gram}}$ ，仍能有效“修复”严重退化的局部特征。为进一步提升性能，我们每1万次迭代更新一次Gram教师，直至其与主EMA教师完全一致。我们将此训练的第二阶段称为优化阶段，该阶段通过优化目标函数 \mathcal{L}_{Ref} 来实现...

$$\mathcal{L}_{\text{Ref}} = w_D \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{DK} \mathcal{L}_{\text{DKoleo}} + w_{\text{Gram}} \mathcal{L}_{\text{Gram}}. \quad (3)$$

我们在图7中可视化了不同损失函数的演变过程，发现应用Gram目标函数会显著影响iBOT损失函数，使其下降速度明显加快。这表明稳定Gram教师引入的稳定性对iBOT目标函数产生了积极影响。相比之下，Gram目标函数对DINO损失函数的影响并不显著。这一观察结果表明，Gram目标函数和iBOT目标函数对特征的影响方式相似，而DINO损失函数则以不同方式影响特征。

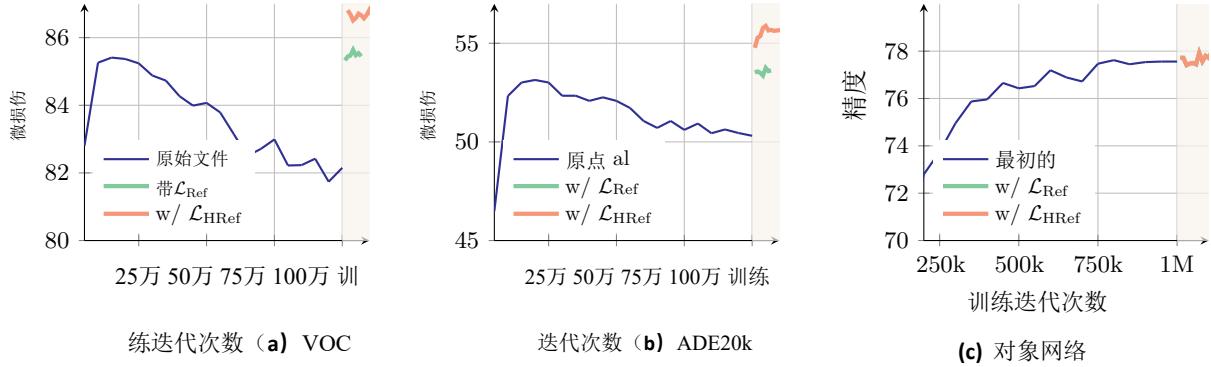


图8：应用我们提出的格拉姆锚定方法后不同基准测试结果的演变。我们可视化了在原始训练基础上添加改进步骤（标记为‘ \mathcal{L}_{Ref} ’）后的结果。同时绘制了使用更高分辨率特征进行格拉姆目标优化时的结果（如第4.3节所述，标记为‘ $\mathcal{L}_{\text{HRef}}$ ’）。我们特别标注了采用格拉姆目标的迭代过程。

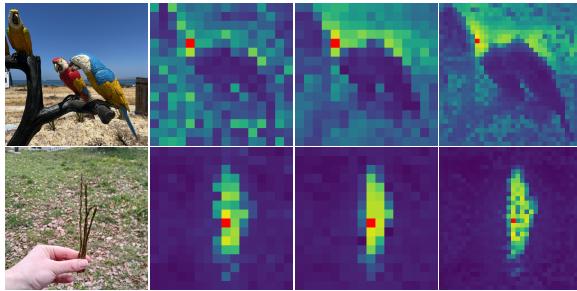
在性能表现方面，我们观察到新损失函数的影响几乎是立竿见影的。如图8所示，在前10k次迭代中，引入Gram锚定技术显著提升了密集任务的性能表现。在Gram教师更新后，ADE20k基准测试也取得了明显提升。此外，延长训练周期进一步优化了ObjectNet基准测试的性能表现，而其他全局基准测试也显示出新损失函数带来的温和影响。

4.3 利用高分辨率特征

最新研究表明，通过加权平均补丁特征可以平滑异常补丁并增强补丁级一致性，从而产生更强的局部表征（Wysoczanska等人，2024）。另一方面，将更高分辨率的图像输入主干网络可生成更精细、更详细的特征图。我们利用这两项优势为Gram教师计算高质量特征。具体而言，我们首先以两倍正常分辨率的图像输入Gram教师，然后使用双三次插值对生成的特征图进行 $2 \times$ 下采样，以获得与学生输出尺寸匹配的平滑特征图。图9展示了使用256和512分辨率图像获得的补丁特征Gram矩阵，以及从512分辨率特征进行 $2 \times$ 下采样后获得的特征（标记为‘downsamp.’）。我们观察到，高分辨率特征中更优的补丁级一致性通过下采样得以保留，从而产生更平滑、更连贯的补丁级表征。值得一提的是，得益于Su等人（2024）提出的旋转位置嵌入（RoPE）技术，我们的模型能够无缝处理不同分辨率的图像而无需调整。

我们计算下采样特征的Gram矩阵，并用它来替换目标函数 $\mathcal{L}_{\text{Gram}}$ 中的 \mathbf{X}_G 。我们将由此产生的新优化目标标记为 $\mathcal{L}_{\text{HRef}}$ 。这种方法使得Gram目标能够有效地将平滑高分辨率特征的改进块一致性提炼到学生模型中。如图8和图9b所示，这种提炼转化为在密集任务上的更好预测，在 \mathcal{L}_{Ref} 带来的优势基础上又获得了额外收益（ADE20k数据集上mIoU提升2）。我们还在图9b中消除了Gram教师的选择。有趣的是，从10万或20万次迭代中选择Gram教师对结果影响不大，但使用更晚迭代次数的Gram教师（100万次迭代）则有害，因为这种教师的块级一致性较差。

最后，我们在图10中定性展示了Gram锚定对斑块级一致性的影响，该图可视化了初始训练和高分辨率Gram锚定优化后获得的Gram矩阵斑块特征。我们观察到，高分辨率优化过程显著提升了特征相关性。



输入 256 下采样。512

(a) 不同输入分辨率下的格拉姆矩阵

(b) 格拉姆氏菌教师与决议的消融。

图9: 高分辨率Gram图像影响的定量与定性研究。我们展示了：(a) 将高分辨率图像降采样为更小尺寸后改进的余弦图；(b) 通过调整训练迭代次数和Gram教师分辨率带来的定量改进。

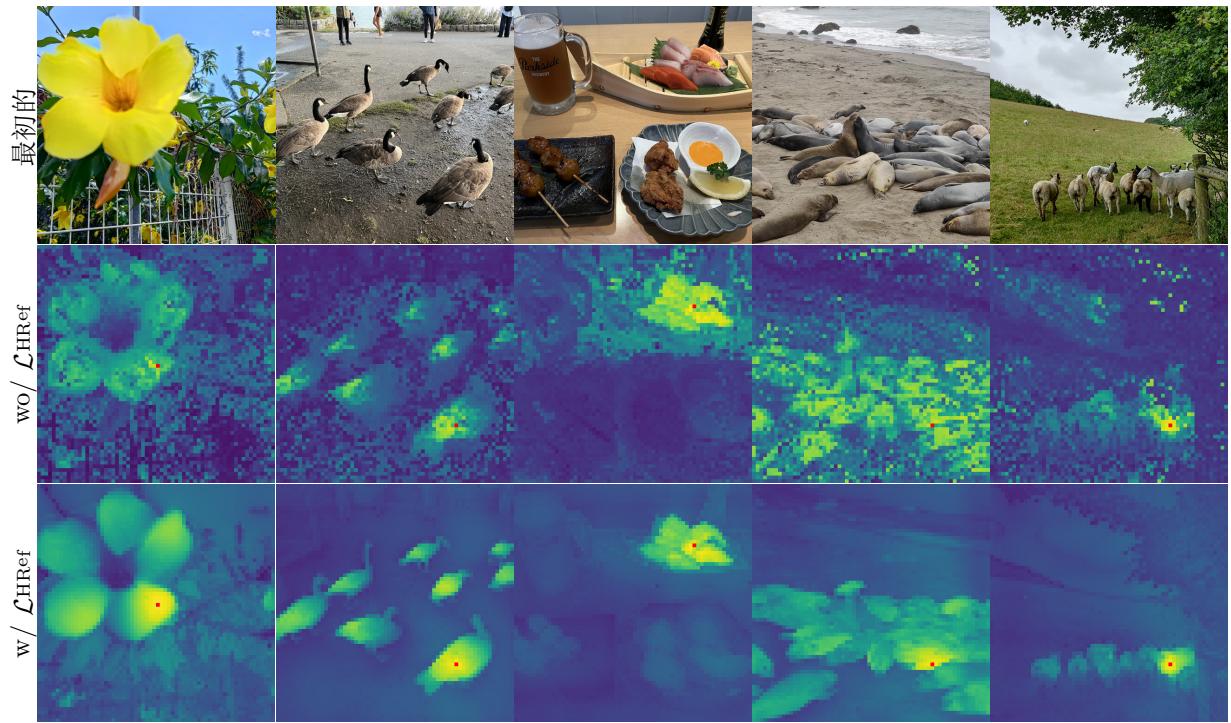


图10: Gram锚定的定性效果。我们可视化了使用精修目标 $\mathcal{L}_{\text{HRef}}$ 前后的余弦图。图像的输入分辨率为 1024×1024 像素。

5 培训后

本节介绍后训练阶段，包括高分辨率适配阶段（支持不同输入分辨率下的有效推理，[第5.1节](#)）、模型蒸馏（生成高质量且高效的精简模型，[第5.2节](#)）以及文本对齐（为DINOv3添加零样本能力，[第5.3节](#)）。

5.1 分辨率缩放

我们采用256的相对较低分辨率训练模型，这种设置在速度与效率之间取得了良好平衡。当采用16的图像块尺寸时，该配置产生的输入序列长度与DINOv2相同——后者是基于224分辨率和14图像块尺寸训练的。不过，许多当代计算机视觉

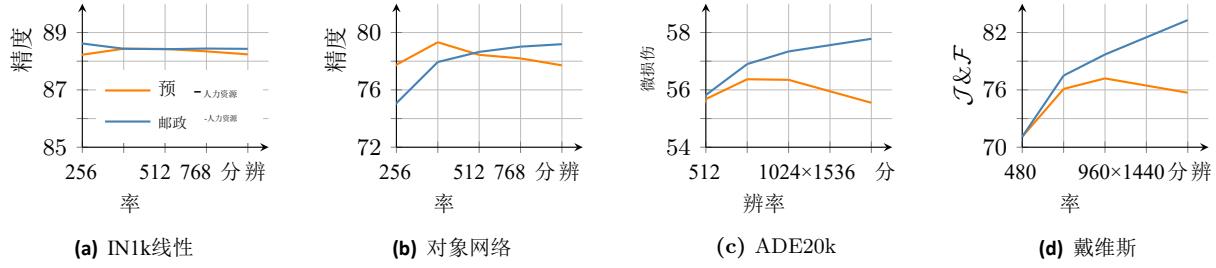


图11: 高分辨率自适应效果。分辨率缩放前（‘Pre-HR’）与缩放后（‘Post-HR’）的结果（第5.1节）对比，分别对应：(a)ImageNet上的线性分类，(b)ObjectNet上的应用 OOD，(c)ADE20k上的线性语义分割，以及(d)不同评估分辨率下davis上的分割跟踪。

应用程序需要处理分辨率显著更高的图像，通常为 512×512 像素或更高，以捕捉复杂的空间信息。推理图像分辨率在实际应用中也不是固定的，而是根据具体使用场景而变化。为解决这一问题，我们在训练机制中引入了高分辨率自适应步骤（Touvron 等人，2019）。为确保在不同分辨率下保持高性能，我们采用混合分辨率策略，即在每个小批量中采样不同尺寸的全局和局部裁剪图像对。具体而言，我们考虑全局裁剪尺寸为{512,768}，局部裁剪尺寸为{112,168,224,336}，并额外训练模型10k次迭代。

与主训练阶段类似，这一高分辨率适应阶段的关键组件是引入格拉姆锚定技术，即使用7B教师作为格拉姆教师。我们发现该组件至关重要：若缺失该组件，模型在密集预测任务上的表现将显著下降。格拉姆锚定技术促使模型在不同空间位置间保持一致且稳健的特征相关性，这对于处理高分辨率输入日益增加的复杂性至关重要。

通过实证研究我们发现，这个相对简短但精准的高分辨率步骤能显著提升模型整体质量，并使其具备跨多种输入尺寸的泛化能力，如图4所示。在图11中，我们对比了7B模型在适应前后的表现。研究发现分辨率缩放对ImageNet分类(a)仅带来小幅提升，且性能相对稳定；但在ObjectNet OOD迁移(b)中，低分辨率时性能略有下降，高分辨率时则有所改善。这种差异主要被高分辨率下局部特征质量的提升所抵消——ADE20k分割(c)和Davis跟踪(d)的正向趋势充分证明了这一点。模型适应后产生的局部特征会随图像尺寸增大而优化，这得益于高分辨率下更丰富的空间信息，从而有效支持高分辨率推理。值得注意的是，适应后的模型支持的分辨率远超最大训练分辨率768——我们直观观察到在4k分辨率以上仍能保持特征图的稳定性（参见图4）。

5.2 模型蒸馏

多场景应用模型家族我们将ViT-7B模型的知识蒸馏为更小型的视觉Transformer变体（ViT-S、ViT-B和ViT-L），这些变体因其更易管理且效率更高而备受学界推崇。我们的蒸馏方法沿用了第一阶段训练的相同目标函数，确保学习信号的一致性。但与传统采用指数移动平均（EMA）对模型权重进行优化不同，我们直接以7B模型作为教师模型来指导小型学生模型。在此方案中，教师模型保持固定，因此无需考虑图像块级一致性问题，也无需应用格拉姆锚定技术。这种策略使蒸馏模型既能继承大型教师模型强大的表征能力，又在实际部署和实验中更具可操作性。

我们的ViT-7B模型经过精简，衍生出多个ViT模型，其规模覆盖了广泛的计算预算范围，便于与现有模型进行有效对比。这些模型包括标准版ViT-S（2100万参数）、B版（8600万）、L版（3亿），以及定制版ViT-S+（2900万）和ViT-H+（8亿），旨在缩小与自蒸馏70亿参数教师模型的性能差距。事实上，在DINOv2中我们观察到，较小规模的模型

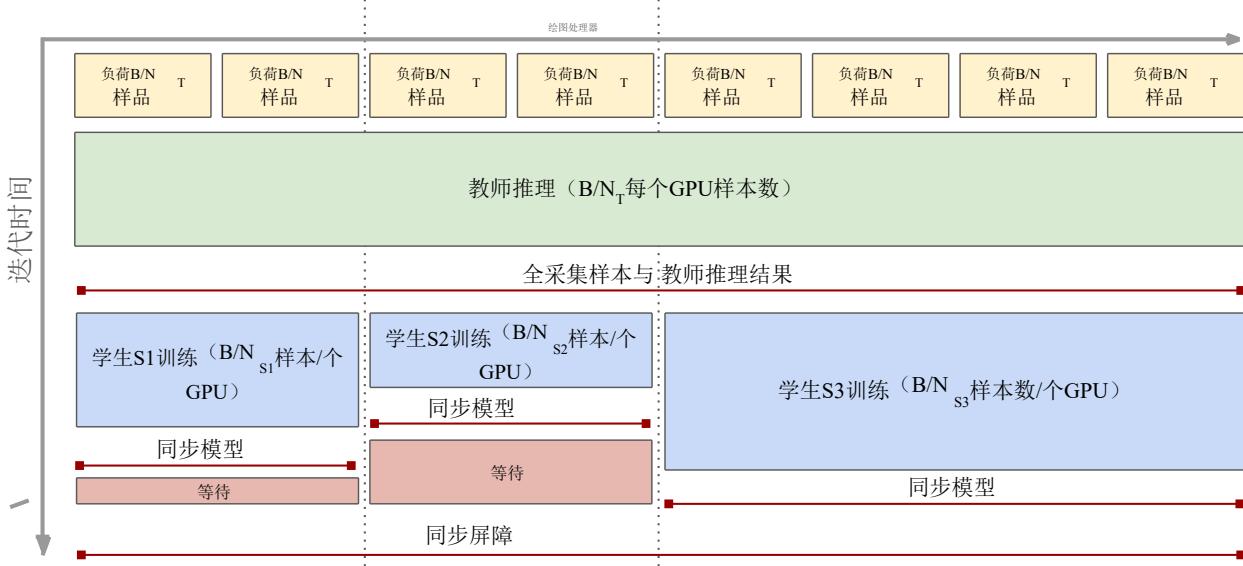


图12：多学生蒸馏流程。该图展示了三名学生并行蒸馏的流程：首先通过共享所有T节点的教师推理来节省计算资源，并在所有GPU上集中收集输入和输出结果。随后，较小规模的小组执行学生训练。我们通过调整小组规模，确保所有学生 S_i 的训练步骤持续时间一致，从而最大限度减少在同步屏障处等待的空闲时间。

学生模型通过蒸馏技术能达到与教师模型相当的性能。如表14所示，蒸馏模型仅需极小的推理计算量即可实现前沿水平的性能表现。我们先对模型进行100万次迭代训练，随后按照余弦调度方案执行25万次学习率冷却迭代，最后在不使用格拉姆锚定的情况下应用第5.1节所述的高分辨率阶段。

高效多学生蒸馏由于大型教师的推理成本可能比学生高出数个数量级（参见图16a），我们设计了一种并行蒸馏流水线，允许同时训练多个学生并将教师推理共享至所有参与训练的节点（示意图见图12）。设 C_T 和 C_S 分别表示在单教师/单学生蒸馏中运行教师推理和学生训练的成本，其中批量大小为 B ，每个 N 个GPU处理 B/N 个数据切片，教师推理成本为 $B/N \times C_T$ ，学生训练成本为 $B/N \times C_S$ （每个GPU）。在多学生蒸馏中，我们按以下方式操作：每个学生 S_i 被分配一组 N_{Si} 个GPU进行训练，所有 $N_T = \epsilon N Si$ 个GPU均属于全局推理组。在每次迭代中，我们首先对全局组进行教师推理，计算每个GPU的 $B/N_T \times C_T$ 计算成本。然后我们运行一个*all-gather*集体操作，将输入数据和推理结果共享给所有计算节点。最后，每个学生组分别执行学生训练，计算 $B/N_{Si} \times C_{Si}$ 成本。

上述计算表明，在蒸馏流程中增加一个学生模型将带来双重效果：(1) 每次迭代时每个GPU的计算量减少，从而整体提升蒸馏速度；(2) 仅需增加新学生模型的训练成本，因为教师模型的总推理成本已固定。具体实现只需仔细配置GPU进程组，调整数据加载器和教师模型推理，确保各组通过NCCL集体同步输入输出。由于各组在每次迭代时保持同步，为实现速度最大化，我们调整每个学生模型的GPU数量使其迭代时间大致相同。通过这种机制，我们能够无缝训练多个学生模型，并从旗舰级70亿参数模型中生成完整的蒸馏模型家族。

5.3 DINOv3与文本的对齐

开放词汇图像-文本对齐因其能够实现灵活且可扩展的多模态理解的潜力，已引起研究界的广泛关注与热情。大量研究

已有研究致力于提升CLIP的质量（Radford 等人，2021），该模型最初仅学习图像与文本表示之间的全局对齐。尽管CLIP展现出令人印象深刻的零样本能力，但其对全局特征的聚焦限制了其捕捉精细、局部对应关系的能力。近期研究（Zhai 等人，2022b）表明，通过预训练的自监督视觉骨干网络可实现有效的图像-文本对齐。这使得这些强大模型能够在多模态场景中发挥作用，促进超越全局语义的更丰富、更精确的文本-图像关联，同时由于视觉编码已预先学习，还能降低计算成本。

我们通过采用Jose 等人（2025）先前提出的训练策略，将文本编码器与我们的DINOv3模型对齐。该方法遵循LiT训练范式（Zhai 等人，2022b），通过对比目标从头训练文本表示以使图像与其标题匹配，同时保持视觉编码器冻结。为在视觉端提供一定灵活性，在冻结的视觉骨干网络上引入了两个Transformer层。该方法的关键改进在于：在匹配文本嵌入前，将均值池化补丁嵌入与输出CLS 标记进行拼接。这使得全局和局部视觉特征都能与文本对齐，从而在无需额外启发式方法或技巧的情况下，提升密集预测任务的性能。此外，我们采用Jose 等人（2025）建立的相同数据整理协议，以确保一致性和可比性。

6 结果

在本节中，我们将在多种计算机视觉任务上评估旗舰级DINOv3 7B模型。除非另有说明，实验过程中我们保持DINOv3模型冻结并仅使用其表征。我们证明了DINOv3无需微调即可获得优异性能。本节结构安排如下：首先通过轻量级评估协议探究DINOv3的密集表征（第6.1节）和全局表征（第6.2节）质量，并与现有最强视觉编码器进行对比。我们发现DINOv3在学习密集特征的同时，还能提供稳健且多样的全局图像表征。随后，我们将DINOv3作为开发更复杂计算机视觉系统的基础（第6.3节）。通过在DINOv3上进行少量优化，我们证明在目标检测、语义分割、三维视图估计或相对单目深度估计等多样化任务中，DINOv3能够达到与当前最先进水平相当甚至超越的性能。

6.1 DINOv3提供卓越的密集特征

我们首先通过多种轻量级评估方法来考察DINOv3密集表征的原始质量。在所有情况下，我们均采用最后一层的冻结补丁特征，并通过以下方式评估：(1) 定性可视化（第6.1.1节）、(2) 密集线性探测（第6.1.2节：分割、深度估计）、(3) 非参数方法（第6.1.3节：3D对应估计、第6.1.4节：目标发现、第6.1.5节：跟踪）、(4) 轻量级注意力探测（第6.1.6节：视频分类）。

基线我们将DINOv3的密集特征与当前最强的公开图像编码器（包括弱监督和自监督类型）进行对比。我们考察了采用CLIP式图像文本对比学习的弱监督编码器Perception Encoder (PE) Core (Bolya等人，2025) 和 SigLIP 2 (Tschannen等人，2025)，同时对比了最强的自监督方法：DINOv3的前身DINOv2 (Oquab等人，2024) 及其注册表 (Darcet等人，2024)、Web-DINO (Fan等人，2025)、DINO近期的扩展尝试，以及最佳开源SSL模型Franca (Venkataraman等，2025)。最后，我们还对比了从DINOv2衍生的AM-RADIOv 2.5 (Heinrich等人，2025)、CLIP (Radford等人，2021)、DFN (Fang等人，2024a) 和 Segment Anything (SAM) (Kirillov等人，2023) 等聚类模型，以及PEspatial，将SAM 2 (Ravi 等人，2025) 蒸馏至PEcore。针对每个基线，我们报告现有最强模型的性能，并在表格中明确架构。

6.1.1 定性分析

我们首先对DINOv3的密集特征图进行定性分析。为此，我们采用主成分分析 (PCA) 将密集特征空间投影至三维空间，并将所得三维空间映射为RGB颜色空间。由于PCA存在符号歧义（八种变体）以及主成分间任意映射关系

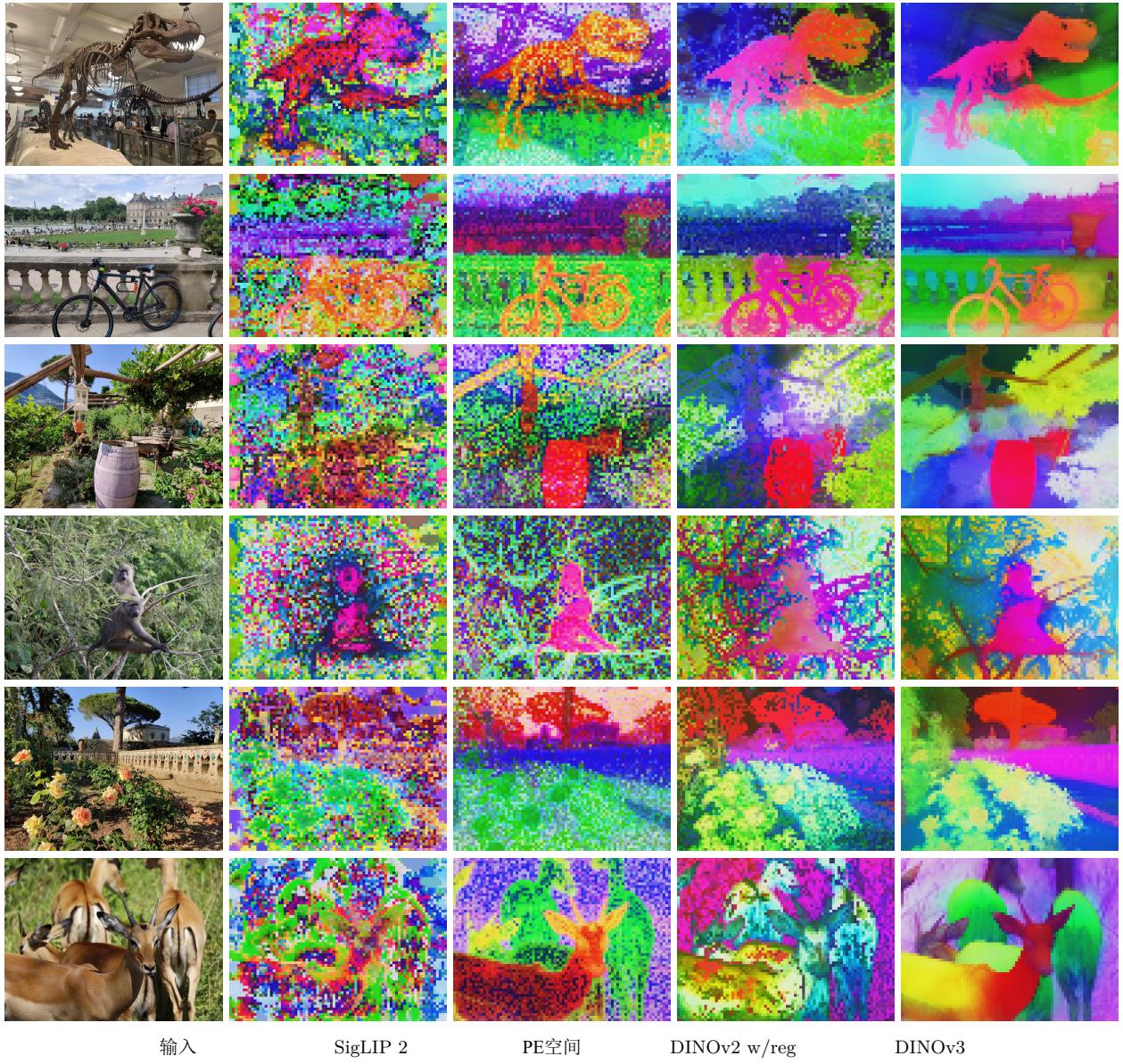


图13：密集特征对比。我们通过主成分分析（PCA）将视觉骨干网络的密集输出投影并映射到RGB空间，对多个视觉骨干网络进行对比。从左至右依次为：SigLIP 2 ViT-g/16、PEspatial ViT-G/14、带配准器的DINOv2 ViT-g/14、DINOv3 ViT-7B/16。使用16号图像块的模型图像以 1280×960 分辨率输入，使用14号图像块的模型图像以 1120×840 分辨率输入，即所有特征图尺寸均为 80×60 。

针对六种颜色的组件组合，我们穷尽所有可能的排列组合，最终呈现视觉效果最出众的方案。如图13所示，与现有视觉模型相比，DINOv3的特征呈现更为锐利，噪声显著降低，语义连贯性也更胜一筹。

6.1.2 密集线性探测

我们在密集特征基础上对两个任务进行线性探测：语义分割和单目深度估计。在这两种情况下，我们都在DINOv3的冻结块输出基础上训练线性变换。对于语义分割，我们在ADE20k (Zhou 等人, 2017)、Cityscapes (Cordts 等人, 2016) 和Pascal VOC 2012 (Everingham 等人, 2012) 数据集上进行评估，并报告平均交并比

表3: 冻结骨干网络在语义分割和单目深度估计上的密集线性探测结果。我们报告了分割基准ADE20k、Cityscapes和VOC的平均交并比（mIoU）指标，以及深度基准NYUv2和KITTI的均方根误差（RMSE）指标。对于分割任务，所有模型均采用适配1024个补丁标记的输入分辨率进行评估（即补丁尺寸14为 448×448 ，补丁尺寸16为 512×512 ）。

方法	维生素T	分割			深	
		ADE20k	城市科学。	职业	NYUv2 ↓	KITTI ↓
聚集性骨干						
AM-RADIOv2.5	g/14	53.0	78.4	85.4	0.340	2.918
PE空间	G/14	49.3	73.2	82.7	0.362	3.082
弱监督骨干网络						
SigLIP 2	g/16	42.7	64.8	72.7	0.494	3.273
PE核心	G/14	38.9	61.1	69.2	0.590	4.119
自监督骨干网络						
弗兰卡	g/14	46.3	68.7	82.9	0.445	3.140
DINOv2	g/14	49.5	75.6	83.1	0.372	2.624
网络动态信息网络	7B/14	42.7	68.3	76.1	0.466	3.158
DINOv3	7B/16	55.9	81.1	86.6	0.309	2.346

(mIoU) 指标。对于深度估计，我们使用NYUv2 (Silberman 等人, 2012) 和 KITTI (Geiger 等人, 2013) 数据集，并报告均方根误差 (RMSE)。

结果 (表3) 分割结果表明我们的密集特征具有更优质量。在通用ADE20k数据集上，DINOv3比自监督基线高出6个以上mIoU点，比弱监督基线高出13个以上。此外，DINOv3比PEspatial高出6个以上，比AM-RADIOv2.5高出近3个点。这些结果尤为显著，因为两者都是从强监督分割模型SAM (Kirillov 等人, 2023) 中提炼出的强基线。在自动驾驶基准Cityscapes上也观察到类似结果，DINOv3达到最佳mIoU值81.1，比AM-RADIOv2.5高出2.5个点，比所有其他骨干网络至少高出5.5个点。

在单目深度估计任务中，DINOv3再次以显著优势超越所有其他模型：弱监督模型PEcore和SigLIP2仍处于落后位置，而DINOv2及基于SAM的更先进模型是其最接近的竞争对手。值得注意的是，尽管PEspatial和AM-radio在纽约大学数据集 (NYU) 上表现强劲，但其在 KITTI 上的表现仍逊于DINOv2。即便如此，DINOv3仍以0.278 RMSE 的优势超越其前身DINOv2。

两组评估结果均表明DINOv3的密集特征具有卓越的表征能力，其视觉效果如图13所示。仅使用线性预测器，DINOv3就能实现对物体类别和掩码的稳健预测，同时还能准确测量场景中的相对深度等物理参数。这些结果表明，该模型的特征不仅视觉清晰且定位精准，还能以线性可分的方式呈现底层观测数据的诸多重要属性。最后，DINOv3在ADE20k数据集上使用线性分类器取得的绝对性能 (55.9 mIoU) 同样令人瞩目，其表现与该数据集当前最先进的绝对性能 (63.0 mIoU) 仅差咫尺之遥。

6.1.3 三维对应估计

理解三维世界一直是计算机视觉的重要目标。图像基础模型最近通过提供三维感知特征推动了三维理解的研究。在本节中，我们按照Probe3D (Banani 等人, 2024) 定义的协议评估DINOv3的多视图一致性——即同一关键点在物体不同视图中的补丁特征是否相似。我们区分几何与语义对应估计：前者指同一物体实例的关键点匹配，后者指同一物体类别不同实例的关键点匹配。我们在NAVI数据集 (Jampani 等人, 2023) 上评估几何对应关系，以及语义

表4: 密集表示的3D一致性评估。我们按照Probe3D的评估方案 (Banani 等人, 2024) 估算跨视图的3D关键点对应关系。为衡量性能, 我们报告对应召回率, 即落在指定距离内的对应点百分比。

方法	维生T	几何的		语义学
		导航	SPair	
聚类骨架 AM-RADIOv2.5				
g/14 PE空间	G/14	59.4 53.8	56.8 49.6	
弱监督骨干网络 SigLIP 2				
16 PE核心	g/ G/14	49.4 39.9	42.6 23.1	
自监督骨干网络				
弗兰卡·迪诺v2	g/14 g/14 7 B/14	54.6 60.1 55.0	51.0 56.1 32.2	
DINOv3	7B/16	64.4	58.7	

在SPair数据集上的对应关系 (Min 等人, 2019), 并使用对应召回率在两种情况下衡量性能。更多实验细节请参阅附录D.3。

结果 (表4) 在几何对应任务中, DINOv3表现优于所有其他模型, 其召回率较次优模型DINOv2提升了4.3%。其他SSL扩展研究 (Franca和WebSSL) 则落后于DINOv2, 表明其仍是强劲的基准模型。弱监督模型 (PEcore和SigLIP 2) 在此任务中表现欠佳, 反映出其缺乏三维感知能力。对于采用SAM蒸馏的模型, AM-radio几乎达到DINOv2的性能水平, 但PEspatial仍落后于前者 (召回率-11.6%), 甚至不及Franca (召回率-0.8%)。这表明自监督学习是该任务取得优异表现的关键要素。在语义对应任务中, 结论同样适用: DINOv3表现最佳, 其召回率较前代模型提升2.6%, 较AM-radio提升1.9%。总体而言, 这些在关键点匹配任务中取得的优异表现, 为DINOv3在其他三维密集型应用中的下游使用提供了极具前景的信号。

6.1.4 无监督对象发现

强大的自监督特征有助于在图像中发现对象实例, 而无需任何标注 (Vo 等人, 2021; Simeoni 等人, 2021; Seitzer 等人, 2023; Wang 等人, 2023c; Simeoni 等人, 2025)。我们通过无监督对象发现任务测试了不同视觉编码器的这一能力, 该任务需要对图像中的对象进行类别无关的分割 (Russell 等人, 2006; Tuytelaars 等人, 2010; Cho 等人, 2015; Vo 等人, 2019)。特别地, 我们使用了非参数化的基于图的TokenCut算法 (Wang等人, 2023c), 该算法在多种骨干网络上表现出强劲性能。我们在三个广泛使用的数据集上运行了它: VOC 2007、VOC 2012 (Everingham 等人, 2015) 和COCO-20k (Lin等人, 2014; Vo等人, 2020)。我们遵循Simeoni 等人 (2021) 定义的评估协议, 并报告CorLoc指标。为正确比较具有不同特征分布的骨干网络, 我们对主要TokenCut超参数进行搜索, 即用于构建分区补丁图时应用的余弦相似度阈值。最初, 使用DINO (Caron 等人, 2021) 时, 通过最后一层注意力机制的键获得了最佳目标发现结果。然而, 这种手工选择并不能始终推广到其他骨干网络。为简化起见, 我们始终采用所有模型的输出特征。

结果 (图14) 原始DINO为该任务设定了极高的标准。有趣的是, 虽然DINOv2在像素级密集任务中表现出色, 但在目标发现方面却表现不佳。这在一定程度上可归因于密集特征中存在的伪影 (参见图13)。DINOv3凭借其清晰精确的输出特征图, 超越了前两代模型, 在VOC 2007数据集上实现了5.9的CorLoc提升, 并优于所有其他骨干网络——无论是自监督、弱监督还是聚类方法。该评估结果证实了

方法	维生素T	VOC07	VOC12	COCO
聚集性主干				
AM-RADIOv2.5	g/14	55.0	59.7	45.9
PE空间	G/14	51.2	56.0	43.9
弱监督骨干网络				
信号脂质G蛋白偶联受体	g/16	20.5	24.7	18.6
PE核心	G/14	14.2	18.2	13.5
自监督骨干网络				
意大利人	S/16	61.1	66.0	48.7
意大利人	B/16	60.1	64.4	50.5
DINOv2	g/14	55.6	60.4	45.4
网络动态信息网络	7B/14	26.1	29.7	20.9
DINOv3	7B/16	66.1	69.5	55.1



图14：无监督目标发现。我们对不同骨干网络的输出补丁特征应用TokenCut (Wang 等人, 2022c)，并报告CorLoc指标。同时可视化DINOv3获得的预测掩码（分辨率1024的输入图像上红色叠加层），该结果未经标注且未进行后处理。

DINOv3的密集特征不仅语义性强，定位也精准。我们相信，这将为更多类无关的物体检测方法铺平道路，特别是在标注成本高昂或无法获取，且相关类别集合不受限于预定义子集的场景中。

6.1.5 视频分割跟踪

除了静态图像之外，视觉表征的一个重要特性是其时间一致性，即特征是否随时间以稳定方式演变。为验证这一特性，我们在视频分割跟踪任务上评估DINOv3：给定视频第一帧的真实实例分割掩码，目标是将这些掩码传播到后续帧。我们使用davis 2017 (Pont-Tuset等人, 2017)、YouTube- VOS (Xu 等人, 2018) 和 MOSE (Ding 等人, 2023) 数据集。我们采用标准J&F均值指标评估性能，该指标结合区域相似性 (J) 和轮廓精度 (F) (Perazzi 等人, 2016)。遵循Jabri 等人 (2020) 的方法，我们采用考虑帧间补丁特征相似性的非参数标签传播算法。我们采用三种输入分辨率进行评估，分别为420/480(S)、840/960(M)和1260/1440(L)像素的短边长，针对14/16像素大小的模型（与补丁标记数量匹配）。J&F分数始终在视频的原生分辨率下计算。更多实验设置详见附录D.5。

结果 (表5) 与所有先前结果一致，弱监督骨干网络未能提供令人信服的性能。从视频模型SAMv2中提取的PEspatial表现令人满意，在较小分辨率上超越DINOv2，但在较大分辨率上表现欠佳。在所有分辨率下，DINOv3均优于所有竞争对手，在davis-L上取得惊人的83.3 J&F分，比DINOv2高出6.7分。此外，性能随分辨率变化呈现健康趋势，证实我们的模型能够利用更多输入像素输出精确的高分辨率特征图（参见图3和图4）。相比之下，SigLIP 2和PEcore在更高分辨率下的性能几乎持平，而PEspatial则出现下降。有趣的是，我们的图像模型无需任何视频调优即可正确追踪时间中的物体（见图15），这使其成为嵌入视频的理想候选，从而可在其基础上构建强大的视频模型。

6.1.6 视频分类

先前的研究结果表明，DINOv3的表征具有低层次的时间一致性，能够实现对物体的精准时间追踪。在此基础上，本节我们将评估其密集特征在高级视频分类中的适用性。与V- JEPA 2的设置类似 (Assran 等人, 2025)，我们基于每帧提取的图像块特征训练了一个注意力探测器——一个基于浅层4层Transformer的分类器。这种设计使得特征能够通过独立提取的方式，在时间和空间维度上进行推理。

表5: 视频分割跟踪评估。我们报告了在davis、YouTube- VOS 和MOSE数据集上不同分辨率下的J&F均值。对于块大小为14/16的模型，小、中、大分辨率分别对应视频短边为420/480、840/960、1260/1140像素。

方法	维生素T	戴维斯			YouTube VOS			摩西		
		S	M	L	S	M	L	S	M	L
聚集性主干										
AM-RADIOv2.5	g/14	66.5	77.3	81.4	70.1	78.1	79.2	44.0	52.6	54.3
PE空间	G/14	68.4	74.5	70.5	68.5	67.5	55.6	39.3	40.2	34.0
弱监督骨干网络										
SigLIP 2	g/16	56.1	62.3	62.9	52.0	57.3	55.1	28.0	30.3	29.2
PE核心	G/14	48.2	53.1	49.8	34.7	33.0	25.3	17.8	19.0	15.4
自监督骨干网络										
弗兰卡	g/14	61.8	66.9	66.5	67.3	70.5	67.9	40.3	42.6	41.9
DINOv2	g/14	63.9	73.6	76.6	65.6	73.5	74.6	40.4	47.6	48.5
网络动态信息网络	7B/14	57.2	65.8	69.5	43.9	49.6	50.9	24.9	29.9	31.1
DINOv3	7B/16	71.1	79.7	83.3	74.1	80.2	80.7	46.0	53.9	55.6

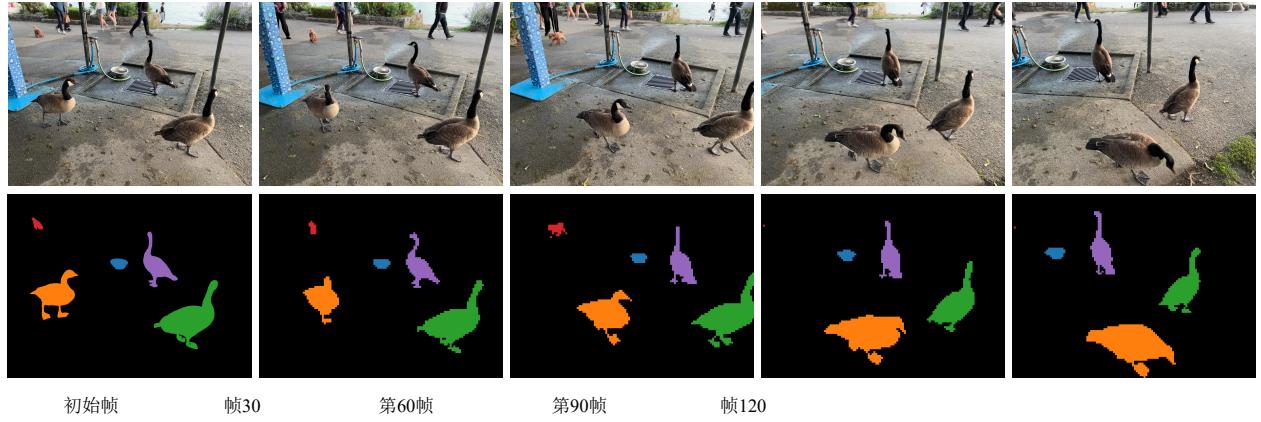


图15: 分割跟踪示例。给定初始帧的实例分割掩码，我们根据DINOv3特征空间中的补丁相似性将实例标签传播到后续帧。输入分辨率为 2048×1536 像素，生成 128×96 个补丁。

帧。在评估过程中，我们要么为每个视频选取单个片段，要么使用测试时增强（TTA），即对每个视频的3个空间和2个时间裁剪的预测结果取平均值。实验细节参见附录D.6。我们在三个数据集上进行评估：UCF101（Soomro 等人，2012）、Something-Something V2（Goyal等人，2017）和Kinetics-400（Kay 等人，2017），并报告top-1准确率。作为额外基线，我们报告了V- JEPA v2的性能，这是视频理解领域的最先进SSL模型。

结果（表6）与先前实验的结论一致，我们发现DINOv3能够成功提取强视频特征。由于本次评估涉及多层自注意力训练，各模型间的差异较为隐蔽。不过DINOv3的表现与PEcore和SigLIP 2处于同一水平区间，并在多个数据集上明显优于其他模型（DINOv2、AM-radio）。UCF101和K400是注重外观特征的数据集，其中对物体的强类别级理解贡献了大部分性能。而SSv2则需要更好的运动理解能力——专用视频模型V- JEPA v2在该数据集上表现突出。有趣的是，DINOv3与弱监督模型之间的差距在该数据集上略大，这再次证实了DINOv3适用于视频任务的特性。

表6: 使用注意力探针进行视频分类评估。我们报告了UCF101、Something-Something V2 (SSv2) 和 Kinetics-400 (K400) 数据集上的top-1准确率。针对每个模型，我们分别报告了单个视频片段评估的性能，以及通过平均多个片段的预测概率来实现测试时增强 (TTA) 的性能。

方法	维生素T	UCF101		SSv2		K400	
		单人	TTA	单人	TTA	单人	TTA
聚集性主干							
AM-RADIOv2.5	g/14	92.8	92.5	69.1	70.0	84.8	85.2
PE空间	G/14	92.7	92.8	66.4	68.4	83.5	84.8
弱监督骨干网络							
SigLIP 2	g/16	93.6	94.2	68.8	70.2	86.9	87.7
PE核心	G/14	93.1	93.3	69.0	70.4	87.9	88.8
自监督骨干网络							
DINOv2	g/14	93.5	93.8	67.4	68.4	84.4	85.6
V-JEPA 2	g/16	94.0	93.8	73.8	75.4	83.3	84.3
网络动态信息网络	7B/14	93.9	94.1	67.3	68.1	86.8	87.2
DINOv3	7B/16	93.5	93.5	70.1	70.8	87.8	88.2

6.2 DINOv3具备稳健且多功能的全局图像描述符

在本节中，我们评估DINOv3捕捉全局图像统计的能力。为此，我们采用经典分类基准（线性探测器第6.2.1节）和实例检索基准（第6.2.2节）进行测试。我们再次将其与最强的公开图像编码器进行对比。除前一节的模型外，我们还评估了两个弱监督模型：AIMv2 (Fini 等人, 2024) ——采用联合自回归像素与文本预测训练，以及大规模EVA-CLIP-18B (Sun 等人, 2024)。

6.2.1 基于线性探测的图像分类

我们基于DINOv3的输出CLS标记训练线性分类器，以评估模型在分类基准上的表现。我们采用ImageNet1k (Deng 等人, 2009) 数据集及其变体来评估模型的分布外鲁棒性，并通过不同领域的数据集组合来验证DINOv3对细粒度类别的区分能力。具体评估细节参见附录D.7。

ImageNet领域泛化 (表7) 在本实验中，我们使用ImageNet-训练集进行训练，采用ImageNet-验证集作为验证集来选择超参数，并将找到的最佳分类器迁移至不同测试数据集：ImageNet-V2 (Recht 等人, 2019) 和 ReaL (Beyer 等人, 2020) 是ImageNet的替代图像与标签集，用于在ImageNet验证集上测试过拟合；Rendition (Hendrycks等人, 2021a) 和 Sketch (Wang 等人, 2019) 展示了ImageNet类别的风格化与人工版本；Adversarial (Hendrycks 等人, 2021b) 和 ObjectNet (Barbu 等人, 2019) 包含刻意选择的困难样本；Corruptions (Hendrycks 与 Dietterich, 2019) 用于测量对常见图像破坏的鲁棒性作为参考，我们还列出了Dehghani 等人 (2023) 在大规模JFT数据集 (3B-4B图像) 上使用监督分类训练的ViTs的线性探测结果。请注意，这些结果遵循略有不同的评估协议，与我们的结果不能直接比较。

DINOv3在所有自监督模型中表现显著领先，相较于此前最强的自监督模型DINOv2，其在ImageNet-R上的增益达到+10%，在-Sketch上提升+6%，在ObjectNet上增长+13%。值得注意的是，在ImageNet-A和ObjectNet等高难度OOD任务中，目前最强的弱监督模型SigLIP 2和PE的表现已超越最强的监督模型ViT-22B。DINOv3在ImageNet-R和-Sketch上表现相当，在高难度任务ImageNet-A和ObjectNet上紧随PE之后，同时超越了SigLIPv2。在ImageNet测试中，虽然验证分数比SigLIPv2和PE低0.7-0.9分，但在“更干净”的测试集-V2和-ReaL上表现几乎持平。特别值得一提的是，DINOv3在ImageNet-C任务中展现出最佳的抗干扰能力。总体而言，这是自监督模型首次达到

表7: 基于ImageNet1k数据集训练且采用冻结主干网络的线性探测器分类准确率。弱监督与自监督模型的评估均采用适配1024个图像块标记的分辨率（即块大小14时为 448×448 ，块大小16时为 512×512 ）。作为参考，我们还列出了Dehghani等人（2023）采用不同评估协议的结果（标有*）。

方法	维生素T	图像网			演奏		困难的		
		瓦尔	V2	瑞乐	R	S	A	C ↓	目标
监督式骨干网络									
Zhai等人（2022a）*	G/14	89.0	81.3	90.6	91.7	—	78.8	—	69.6
Chen等人（2023）*	e/14	89.3	82.5	90.7	94.3	—	81.6	—	71.5
Dehghani等人（2023）*	22B/14	89.5	83.2	90.9	94.3	—	83.8	—	74.3
聚集性主干									
AM-RADIOv2.5	g/14	88.0	80.2	90.3	83.8	67.1	81.3	27.1	68.4
弱监督骨干网络									
PE核心	G/14	89.3	81.6	90.4	92.2	71.9	89.0	22.7	80.2
SigLIP 2	g/16	89.1	81.6	90.5	92.2	71.8	84.6	30.0	78.6
AIMv2	3B/14	87.9	79.5	89.7	82.3	67.1	74.5	29.5	69.0
EVA-CLIP	18B/14	87.9	79.3	89.5	85.2	64.0	81.6	33.0	71.9
自监督骨干网络									
网络动态信息网络	7B/14	85.9	77.1	88.6	75.6	64.0	71.6	31.2	69.7
弗兰卡	g/14	84.8	75.3	89.2	67.6	49.5	56.5	40.0	54.5
DINOv2	g/14	87.3	79.5	89.9	81.1	65.4	81.7	24.1	66.4
DINOv3	7B/16	88.4	81.4	90.4	91.1	71.3	86.9	19.6	79.0

表8: 细粒度分类基准。Fine-S在12个数据集上取平均值，完整结果参见表22。

方法	维生素T	细S	地方	iNat18	iNat21
聚集性主干					
AM-RADIOv2.5	g/14	93.9	70.2	79.0	83.7
弱监督骨干网络					
SigLIP 2	g/16	93.7	70.5	80.7	82.7
PE核心	G/14	94.5	71.3	86.6	87.0
AIMv2	3B/14	92.9	70.7	80.8	83.2
EVA夹	18B/14	92.9	71.1	80.7	83.5
自监督骨干网络					
弗兰卡	g/14	87.7	64.6	61.4	70.6
DINOv2	g/14	92.6	68.2	80.7	86.1
网络动态信息网络	7B/14	90.2	69.6	65.3	74.1
DINOv3	7B/16	93.0	70.0	85.6	89.8

表9: 实例识别基准测试。详见表23为其他指标。

	牛津-H	巴黎-	甲基化(GAP)	阿姆斯特丹时间
	47.5	85.7	30.5	23.1
	25.1	60.9	13.9	15.5
	32.7	68.9	10.6	23.1
	28.8	71.4	29.5	14.6
	27.1	65.6	0.5	18.9
	14.3	51.6	27.2	21.1
	58.2	84.6	44.6	48.9
	31.2	80.3	35.2	30.6
	60.7	87.1	55.4	56.5

在图像分类领域，其表现与弱监督和监督模型不相上下——而该领域曾是（弱）监督训练方法的强项。考虑到ViT-22B、SigLIP 2和PE等模型都是基于海量人工标注数据集训练的，这一结果实属难得。相比之下，DINOv3完全通过图像学习，这为未来进一步扩展和改进该方法提供了可能。

细粒度分类（表8） 我们还测量了DINOv3在多个数据集上训练线性探针进行细粒度分类时的性能。具体而言，我们报告了3个大型数据集的准确率，即用于场景识别的Places205（Zhou等人，2014），以及用于详细动植物物种识别的iNaturalist 2018（Van Horn等人，2018）和iNaturalist 2021（Van Horn等人，2021），以及覆盖场景、物体和纹理的12个较小数据集的平均值（如Oquab等人（2024）中所示，本文称为Fine-S）。有关这些数据集的个别结果，请参见表22。

研究发现，DINOv3再次超越所有先前的SSL方法。与弱监督方法相比，其表现同样具有竞争力，这表明该方法在多种细粒度分类任务中均展现出强大的鲁棒性和泛化能力。值得注意的是，DINOv3在难度较高的iNaturalist21数据集上取得了89.8%的最高准确率，甚至超越了表现最佳的弱监督模型PEcore（87.0%）。

6.2.2 实例识别

为评估模型的实例级识别能力，我们采用了非参数检索方法。该方法通过输出CLS标记，根据数据库图像与给定查询图像的余弦相似度进行排序。我们在多个数据集上进行性能基准测试：用于地标识别的牛津和巴黎数据集（Radenovic等人，2018）、包含大都会博物馆艺术品的Met数据集（Ypsilantis等人，2021），以及AmsterTime数据集（由现代街景图像与阿姆斯特丹历史档案图像匹配组成）（Yildiz等人，2022）。检索效果通过牛津、巴黎和AmsterTime数据集的平均精度均值，以及Met数据集的全局平均精度进行量化。更多评估细节参见附录D.8。

结果（表9和23） 在所有评估基准中，DINOv3以显著优势实现最强性能，例如在Met数据集上比第二佳模型DINOv2提升+10.8分，在AmsterTime数据集上提升+7.6分。在此基准上，弱监督模型远落后于DINOv3，AM-radio除外——该模型基于DINOv2特征提炼而来。这些发现凸显了DINOv3在实例级检索任务中的稳健性和通用性，既涵盖传统地标数据集，也适用于艺术与历史图像检索等更具挑战性的领域。

6.3 DINOv3是复杂计算机视觉系统的基础

前两节已经为DINOv3在密集任务和全局任务中的质量提供了坚实信号。然而，这些结果是通过“模型探测”实验方案获得的，使用轻量级线性适配器甚至非参数算法来评估特征质量。虽然这种简单的评估允许从涉及的实验方案中去除干扰因素，但它们不足以评估DINOv3作为更大计算机视觉系统基础组件的全部潜力。因此，在本节中，我们脱离轻量级方案，转而训练更复杂的下游解码器，并考虑更强、任务特定的基线模型。具体而言，我们以DINOv3为基础进行以下任务：(1) 使用Plain-DETR进行目标检测（第6.3.1节）、(2) 使用Mask2Former进行语义分割（第6.3.2节）、(3) 使用Depth Anything进行单目深度估计（第6.3.3节）以及(4) 使用Visual Geometry Grounded Transformer进行3D理解（第6.3.4节）。这些任务仅作为探索DINOv3可能性的尝试。尽管如此，我们发现基于DINOv3进行构建能够以较小的努力解锁具有竞争力甚至达到最先进水平的结果。

6.3.1 目标检测

作为首要任务，我们着手解决计算机视觉领域长期存在的目标检测难题。给定一张图像，我们的目标是为所有预定义类别的物体实例提供边界框。该任务既需要精准定位，又要求准确识别——边界框必须与物体边界完全吻合，并对应正确的类别。虽然COCO等标准基准测试（Lin等人，2014）的性能已基本饱和，但我们提出采用冻结主干网络的方案，仅在顶层训练小型解码器来攻克这一难题。

数据集与指标 我们使用COCO数据集（Lin等，2014）评估DINOv3的物体检测能力，并报告COCO-VAL2017分割集上的结果。此外，我们在COCO-O评估数据集（Mao等，2023）上评估了分布外性能。该数据集包含相同的类别，但提供六种分布偏移设置下的输入图像。对于两个数据集，我们报告平均精度（mAP）的均值，其IoU阈值为[0.5 : 0.05 : 0.95]。对于COCO-O，我们还报告了有效鲁棒性（ER）。由于COCO是一个小型数据集，仅包含11.8万张训练图像，我们采用更大的Objects365数据集（Shao等，2019）对解码器进行预训练，这是常见做法。

表10：与当前最先进的目标检测系统对比。我们在冻结的DINOv3主干网络上训练检测适配器，展示了COCO和COCO-O数据集验证集上的结果，并报告了不同IoU阈值下的平均精度（mAP）及有效鲁棒性（ER）。基于DINOv3的检测系统创下了新的技术标杆。由于InternImage-G检测模型尚未公开，我们无法复现其结果或计算COCO-O评分。

模型	探测器	参数			COCO		COCO-O	
		FT	编码器	解码器	可训练的	简单	TTA	磁盘映射程序
EVA-02	级联	🔥	300M	—	300M	64.1	—	63.6
InternImage-G	意大利人	🔥	6B	—	6B	65.1	65.3	—
EVA-02	Co-DETR	🔥	300M	—	300M	65.4	65.9	63.7
PE空间	二亚乙基三胺	🔥	1.9B	50M	2B	65.3	66.0	64.0
DINOv3	普通-DETR	✳️	7B	100M	100M	65.6	66.1	66.4
								36.8

实现我们基于Plain-DETR (Lin 等人, 2023b) 进行改进，具体修改如下：我们未将Transformer编码器融合到主干网络中，而是保持其作为独立模块，类似于原始DETR (Carion 等人, 2020)，这使得我们能够在训练和推理过程中完全冻结DINOv3主干网络。据我们所知，这使其成为首个采用冻结主干网络的有竞争力检测模型。我们在Objects365数据集上以1536分辨率训练Plain-DETR检测器22个周期，随后以2048分辨率训练1个周期，接着在COCO数据集上以2048分辨率训练12个周期。推理时我们以2048分辨率运行。可选地，我们还通过在多个分辨率（从1536到2880）转发图像来应用测试增强（TTA）。完整实验细节参见附录D.9。

结果 (表10) 我们将系统与四个模型进行对比：采用Cascade检测器的EVA-02 (Fang等人, 2024b)、采用Co-DETR的EVA-02 (Zong 等人, 2023)、采用DINO的InternImage-G (Wang 等人, 2023b) 以及采用Deta的PEspatial (Bolya 等人, 2025)。我们发现，基于冻结DINOv3主干训练的轻量级检测器（1亿参数）已能达到业界顶尖水平。COCOO，差距显著，表明检测模型能有效利用DINOv3的鲁棒性。值得注意的是，我们的模型在训练参数大幅减少的情况下仍优于所有先前模型，最小的对比样本仍需超过3亿可训练参数。我们认为，无需主干专门化即可实现如此强的性能，这为多种实际应用提供了可能：单一主干前向网络即可提供支持多任务的特征，从而降低计算需求。

6.3.2 语义分割

在完成前期实验后，我们转而评估语义分割这一长期存在的计算机视觉难题。该任务同样需要强效且定位精准的表征，并要求对每个像素进行密集预测。但与目标检测不同，模型无需区分同一对象的不同实例。与检测方法类似，我们在冻结的DINOv3模型基础上训练了一个解码器。

数据集与指标 我们将评估重点放在ADE20k数据集 (Zhou 等人, 2017) 上，该数据集包含150个语义类别，涵盖2万张训练图像和2千张验证图像。我们使用平均交并比 (mIoU) 来衡量性能。为训练分割模型，我们还使用了COCOStuf (Caesar 等人, 2018) 和Hypersim (Roberts 等人, 2021) 数据集。这些数据集包含16.4万张图像。171个语义类别，以及7.7万张图像，分别涵盖40个类别。

实现 为构建将DINOv3特征映射到语义类别的解码器，我们结合了ViTAdapter (Chen 等人, 2022) 和Mask2Former (Cheng 等人, 2022)，类似于先前的研究 (Wang等人, 2022b; 2023b; a)。然而，在我们的案例中，DINOv3主干在训练期间保持冻结。为了避免改变主干特征，我们进一步修改了原始ViT-Adapter架构，移除了注入器组件。与基线模型相比，我们还将嵌入维度从1024增加到2048，以支持处理DINOv3主干的4096维输出。我们首先对

表11：与ADE20k语义分割最先进系统的对比。我们分别在单尺度和多尺度设置（即Simple和TTA）下评估模型。遵循常规做法，我们在896分辨率下进行评估并报告mIoU分数。BEIT3、ONE-peace和DINOv3采用Mask2Former结合ViT-Adapter架构，解码器参数同时考虑了两者。我们在表24中报告了其他数据集的结果。

模型	FT	参数			微损伤	
		编码器	解码器	可训练的	简单	TTA
BEIT3	煜	1.0B	550M	1.6B	62.0	62.8
InternImage-H	煜	1.1B	230M	1.3B	62.5	62.9
一和	煜	1.5B	710M	2.2B	62.0	63.0
DINOv3	✳	7B	927M	927M	62.6	63.0

在COCO-Stuf上进行80k次迭代的分割解码器训练，随后在Hypersim上进行10k次迭代（Roberts等人，2021）。最后，我们在ADE20k的训练集上进行20k次迭代训练，并在验证集上报告结果。所有训练均在896的输入分辨率下完成。推理时我们考虑两种设置：单尺度，即以训练分辨率前向传播图像；多尺度，即在原始训练分辨率的×0.9至1.1之间多个图像比例下取平均预测值。更多实验细节参见附录D.10。

结果（表11） 我们将模型性能与多个前沿基线进行对比，包括BEIT-3（Wang等人，2022b）、InternImage-H（Wang等人，2023b）和ONE-peace（Wang等人，2023a），并在表24中报告了其他数据集的结果。基于冻结DINOv3主干的分割模型达到前沿性能，与ONE-peace持平（63.0 mIoU）。在COCO-Stuf（Caesar等人，2018）和VOC 2012（Everingham等人，2012）数据集上，该模型也优于所有先前模型。由于语义分割需要精确的逐像素预测，视觉变换器主干提出了一个根本性问题。事实上，16像素宽的输入补丁使得预测粒度相对粗糙——这促使了ViT-Adapter等解决方案的出现。另一方面，我们已证明即使在高达4096的分辨率下（参见图3和4），仍能获得高质量的特征图；这一该模型可响应512个标记宽的密集特征图。我们期望未来研究能够利用这些高分辨率特征，在不依赖Mask2Former等大型解码器的情况下，达到最先进的性能水平。

6.3.3 单眼深度估计

我们现在考虑构建一个单目深度估计系统。为此，我们采用Depth Anything V2（DAv2）（Yang等人，2024b）的架构，这是一种近期最先进的方法。DAv2的关键创新在于使用大量带有真实深度标注的合成图像。关键在于，这依赖于DINOv2作为特征提取器，能够弥合模拟与真实的鸿沟，而其他视觉骨干网络如SAM（Kirillov等人，2023）则不具备这一能力（Yang等人，2024b）。因此，我们在DAv2流程中用DINOv3替换DINOv2，以观察是否能获得类似结果。

实现 与DAv2类似，我们采用密集预测变换器（DPT）（Ranftl等人，2021）作为输入，利用DINOv3四个等间距层的特征进行像素级深度场预测。我们使用DAv2在合成数据集上的损失函数集进行训练，并将训练分辨率提升至1024×768，以充分利用DINOv3的高分辨率能力。与DAv2不同，我们保持主干网络冻结而非微调，从而测试DINOv3的开箱即用能力。我们还发现扩展DPT头部有助于充分发挥DINOv3 7B更大特征的潜力。详见附录D.11。

数据集与指标 我们在5个真实世界数据集上评估模型（NYUv2（Silberman等人，2012）、KITTI（Geiger等人，2013）、ETH3D（Schops等人，2017）、ScanNet（来自Ke等人（2025））和二极管（Vasiljevic等人，2019））的零样本尺度不变深度设置中，类似于Ranftl等人（2020）；Ke等人（2025）；

表12: 与最先进的单目深度估计系统的比较。通过将DINOv3与Depth Anything V2 (Yang 等人, 2024b) 结合, 我们获得了一个用于相对深度估计的SotA模型。

方法	FT	纽约大学病毒2			KITTI		ETH3D		扫描网		二极管	
		ARel ↓	型 $\delta_1 \uparrow$	ARel ↓	型 $\delta_1 \uparrow$							
MiDaS	🔥	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	
利瑞普单抗	🔥	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	
奥姆尼达	🔥	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	
DPT	🔥	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	18.2	75.8	
万寿菊	🔥	5.5	96.4	9.9	91.6	6.5	96.0	6.4	95.1	30.8	77.3	
DAv2 (ViT-g)	🔥	4.4	97.9	7.5	94.7	13.1	86.5	—	—	—	—	
DINOv3	✳️	4.3	98.0	7.3	96.7	5.4	97.5	4.4	98.1	25.6	82.2	

Yang 等人 (2024b)。我们报告了标准指标绝对相对误差 (ARel) (数值越小越好) 和 δ_1 (越高越好)。关于这些指标的描述, 可参考 Yang 等人 (2024a) 的研究。

结果 (表12) 我们与相对深度估计领域的最先进方法进行对比: MiDaS (Ranftl 等人, 2020)、LeReS (Yin 等人, 2021)、Omnidata (Eftekhar 等人, 2021)、DPT (Ranftl 等人, 2021)、集成版本的Marigold (Ke 等人, 2025) 以及DAv2。我们的深度估计模型在所有数据集上均达到新的最先进水平, 仅在二极管上的 ARel方面稍逊于DPT。值得注意的是, 这通过使用冻结主干实现, 而其他所有基线模型都需要对主干进行深度估计的微调。此外, 这验证了DINOv3继承了DINOv2的强模拟到真实能力, 这一理想特性为下游任务使用合成生成的训练数据提供了可能性。

6.3.4 基于DINOv3的几何视觉接地变压器

最后, 我们采用近期提出的视觉几何基础转换器 (VGTT) (Wang 等人, 2025) 进行三维理解研究。该模型基于海量三维标注数据训练, 通过单次前向传播即可估算场景中所有关键三维属性, 包括相机内参与外参、点云图及深度图。借助简洁统一的处理流程, VGTT 在多项三维任务中取得业界领先成果, 其效率甚至超越专业方法——这标志着三维理解领域的重要突破。

实现VGTT 采用DINOv2预训练主干网络获取场景不同视角的表征, 随后通过Transformer进行融合。本研究中, 我们直接用DINOv3主干网络替换DINOv2主干网络, 并采用ViT-L变体 (参见第7节) 以匹配原论文中的DINOv2 ViT-L/14模型。我们沿用 VGTT 的完整训练流程, 包括图像主干网络的微调。为适配DINOv3的16像素块尺寸, 我们将图像分辨率从 518×518 调整为 592×592 , 同时保持结果与 VGTT 的可比性。此外, 我们还采用了附录D.12中详述的少量超参数调整。

数据集与指标 根据 Wang 等人 (2025) 的研究, 我们在 Re10K (Zhou 等人, 2018) 和 CO3Dv2 (Reizenstein 等人, 2021) 数据集上评估相机姿态估计, 在 DTU (Jensen 等人, 2014) 上评估密集多视图估计, 在 ScanNet-1500 (Dai 等人, 2017) 上评估双视图匹配。对于相机姿态估计和双视图匹配, 我们报告标准曲线下面积 (AUC) 指标; 对于多视图估计, 我们将预测与真实值之间的最小L2距离定义为“准确性”, 真实值与预测之间的最小L2距离定义为“完整性”, 并取两者的平均值作为“总体”。关于方法和评估的详细信息, 请参阅 Wang 等人 (2025)。

结果 (表13) 我们发现, 配备DINOv3的 VGTT 在所有三个任务上都 VGTT 了先前最先进的水平——使用 DINOv3带来了清晰且持续的性能提升。考虑到我们仅对DINOv3进行了最小程度的调优, 这一结果令人鼓舞。这些任务涵盖了不同层次的视觉理解: 场景内容的高层次抽象 (相机姿态估计)、密集几何预测 (多视角深度估计) 以及精细的像素级对应 (视图匹配)。To-

表13: 基于视觉几何的Transformer（VG GT）在三维理解中的表现（Wang 等人，2025）。通过将 VG GT 流程中的图像特征提取器从DINOv2替换为DINOv3 ViT-L，我们便能在各类三维几何任务中获得最先进成果。我们复现了Wang 等人（2025）的基线结果，并报告了使用标注真实相机信息的方法（标记为*）。相机姿态估计结果以AUC@30指标呈现。

(a) 相机姿态估计		(b) DTU 上的多视图估计			(c) 在ScanNet-1500上查看匹配结果。				
方法	Re10K	CO3Dv2	方法	Acc. \downarrow	比较 \downarrow	总体 \downarrow	方法	5小时曲线下面积	AUC@10
杜斯3R	67.7	76.7	Gipuma*	0.283	0.873	0.578	超级胶	16.2	33.8
MAS3R	76.4	81.8	苹果酒*	0.417	0.437	0.427	LoFTR	22.1	40.8
VG GSFm v2	78.9	83.4	MAS3R*	0.403	0.344	0.374	DKM	29.4	50.7
CUT3R	75.3	82.8	GeoMVSNet*	0.331	0.259	0.295	CasmTR	27.1	47.0
烧旺	78.8	83.3	杜斯3R	2.677	0.805	1.741	罗玛来源于拉丁语	31.8	53.4
VG GT	85.3	88.2	VG GT	0.389	0.374	0.382	VG GT	33.9	55.2
DINOv3	86.3	89.6	DINOv3	0.375	0.361	0.368	DINOv3	35.2	56.1

结合之前关于对应关系估计（第6.1.3节）和深度估计（第6.3.3节）的研究结果，我们进一步将DINOv3作为3D任务基础的强适应性实证依据。此外，我们预计使用更大的DINOv3 7B模型将带来进一步改进。

7 DINOv3模型全系评估

在本节中，我们对基于7B参数模型（参见第5.2节）提炼出的模型家族进行定量评估。该家族包含基于视觉Transformer（ViT）和ConvNeXt（CNX）架构的变体模型。所有模型的详细参数数量及推理浮点运算次数均展示在图16a中。这些模型覆盖了广泛的计算预算范围，以适应不同用户群体和部署场景的需求。我们对所有ViT（第7.1节）和ConvNeXt变体模型进行了全面评估，以检验其在各类任务中的性能表现。

图2展示了DINOv3家族与其他模型集合的对比分析。在密集预测任务中，DINOv3家族的表现显著优于其他所有模型，包括从AM-radio和PEspatial等监督式主干网络提炼出的专用模型。与此同时，我们的模型在分类任务上也取得了相近的性能表现，这使得它们成为不同计算预算下的最优选择。

在第7.1节中，我们详细阐述了ViT模型，并将其与其他开源方案进行对比。随后在第7.2节中，我们讨论了ConvNeXt模型。最后，继第5.3节之后，我们训练了一个与ViT-L模型输出对齐的文本编码器。我们在第7.3节中展示了该模型的多模态对齐结果。

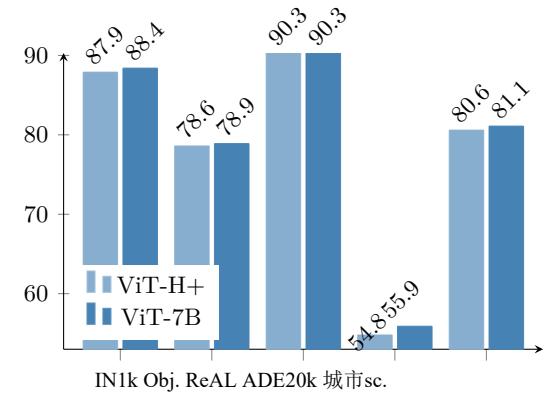
7.1 一种适用于所有用例的视觉变换器

我们的ViT家族涵盖了从紧凑型ViT-S到参数量达8.4亿的ViTH+模型的架构。前者专为笔记本电脑等资源受限设备高效运行而设计，后者则为更严苛的应用场景提供前沿性能。我们将ViT模型与对应规模的最佳开源图像编码器进行对比，包括DINOv2（Oquab 等人，2024）、SigLIP 2（Tschanne 等人，2025）以及Perception Encoder（Bolya 等人，2025）。为确保公平比较，我们保证各模型的输入序列长度一致。具体而言，当模型使用16像素块时，我们输入 512×512 的图像；而使用14像素块时则为 448×448 的图像。

我们的实证研究充分证明，DINOv3模型在密集预测任务中始终表现优于同类方案。特别是在ADE20k基准测试中，DINOv3 ViT-L模型较最佳竞品DINOv2实现了超过6个mIoU点的提升。ViT-B变体相较于次优方案也取得了约3个mIoU点的显著进步。这些亮眼表现充分展现了DINOv3模型在捕捉精细空间细节方面的优势。此外，深度估计任务的评估结果同样显示，DINOv3模型在性能提升方面持续领先于其他竞争方案。

模型	#参数	推理GFLOPs			
		决议256	储	备	金
CNX -微小	29M	5		512	20
CNX 小	50M	11			46
CNX 碱基	89M	20			81
CNX -大	198M	38			152
ViT-S	21M	12			63
ViT-S+	29M	16			79
ViT-B	86M	47			216
ViT-L	300M	163			721
ViT-H+	840M	450			1903
ViT-7B	6716M	3550			14515

(a) DINOV3模型家族



(b) ViT-H+ v.s. ViT-7B.

图16: (a) 经过蒸馏的模型特征展示。CNX 代表ConvNeXT。我们展示了每个模型在 256×256 和 512×512 尺寸图像上的参数数量和GFLOPs估算值。(b) 我们将DINOv3 ViT-H+与其7B规模的教师模型进行比较；尽管参数数量减少了近 $10 \times$ ，ViT-H+的性能与DINOv3 7B非常接近。

表14: 我们模型家族与同等规模开源方案的对比分析。我们展示了ViT-{S、S+、B、L、H+}系列模型在多类代表性基准测试中的表现：分类（IN-ReAL、IN-R、ObjectNet）、检索（Oxford-H）、分割（ADE20k）、深度（NYU）、跟踪（960像素的davis）以及关键点匹配（NAVI、SPair）。通过统一补丁标记数量，确保不同补丁尺寸模型间的公平比较。

尺寸	模型	全局任务				密集任务				
		IN-瑞亚	IN-R	目标	Ox.-H	ADE20k	纽约大学I	戴维斯	导航	SPair
S	DINOv2	87.3	54.0	47.8	39.5	45.5	0.446	73.6	53.4	51.6
S	DINOv3	87.0	60.4	50.9	49.5	47.0	0.403	72.7	56.3	50.4
S+	DINOv3	88.0	68.8	54.6	50.0	48.8	0.399	75.5	57.1	55.2
B	PE核心	87.5	68.4	57.9	20.2	37.4	0.641	44.5	41.8	13.7
B	SigLIP 2	89.3	80.6	66.9	20.2	41.6	0.512	63.2	45.4	32.8
B	DINOv2	89.0	68.4	57.3	51.0	48.4	0.416	72.9	56.9	57.1
B	DINOv3	89.3	76.7	64.1	58.5	51.8	0.373	77.2	58.8	57.2
L	PE核心	90.1	87.7	74.9	25.6	39.7	0.650	48.2	42.1	19.2
L	SigLIP 2	90.1	89.2	75.0	21.4	43.6	0.484	66.3	47.8	41.9
L	DINOv2	89.7	79.1	64.7	55.7	48.8	0.394	73.4	59.9	57.0
L	DINOv3	90.2	88.1	74.8	63.1	54.9	0.352	79.9	62.3	61.2
硫酸盐含量	SigLIP 2	90.3	90.4	76.2	23.0	44.0	0.402	64.8	48.8	38.7
H+	DINOv3	90.3	90.0	78.6	64.5	54.8	0.352	79.3	63.3	56.3

这充分展现了DINOv3家族在各类密集视觉任务中的全能性。值得注意的是，我们的模型在ObjectNet和ImageNet-1k等全局识别基准测试中均取得优异成绩。这表明增强的密集任务性能并未以牺牲全局任务精度为代价。这种平衡性验证了DINOv3模型能提供稳健且全面的解决方案，在密集视觉与全局视觉任务中均表现出色且无妥协。

另一方面，我们还想验证提炼出的最大模型是否能完整捕捉到教师模型的所有信息。为此，我们对最大规模的ViT-H+模型与7B教师模型进行了对比。如图16b所示，最大规模的学生模型表现与体积大8倍的ViT-7B模型相当。

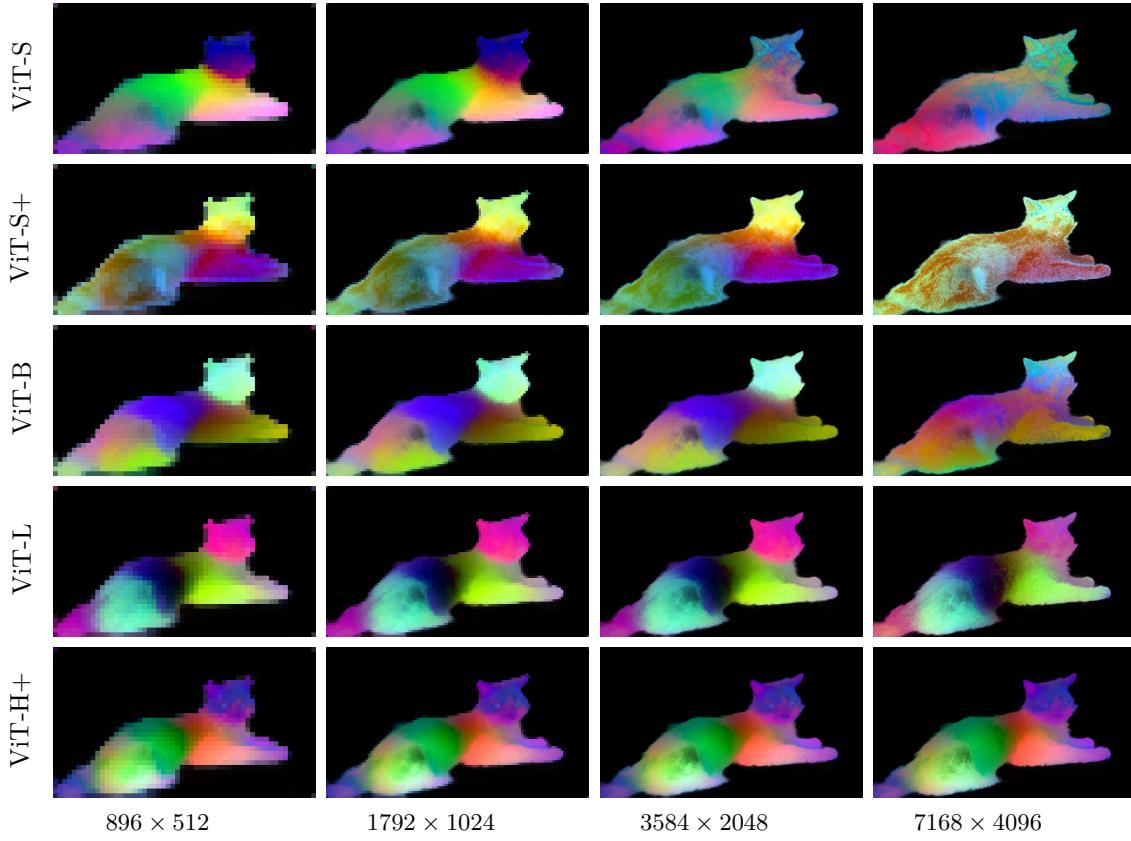


图17：DINOv3 ViT模型家族在多分辨率下的特征稳定性。从上至下依次为：ViT-S、S+、B、L、H+。我们在多分辨率图像上进行推理，随后对 1792×1024 图像（ 112×64 图像标记）计算的特征进行主成分分析。接着将所有分辨率的特征投影到主成分5-7上，并映射至RGB空间进行可视化。虽然模型在所有分辨率下均保持功能，但我们观察到特征在漂移前的较大分辨率范围内保持稳定：例如ViT-S+特征在 896×512 至 3584×2048 输入范围内保持稳定，而ViT-L在最大分辨率 7168×4096 时才开始漂移。ViT-H+在整个测试范围内始终保持稳定。

这一结果不仅验证了我们蒸馏过程的有效性，还表明在优质导师的指导下，小型模型也能学习并达到相当的性能水平。该发现强化了我们的信念：**训练超大规模模型对整个社区都有益**。大型模型的优势能够被成功蒸馏为更高效的小型模型，且几乎不会损失质量。

7.2 资源受限环境下的高效卷积神经网络

在本节中，我们评估了从7B教师模型中提炼出的ConvNeXt (CNX) 模型的质量。ConvNeXt模型在浮点运算效率方面表现优异，非常适合部署在专为卷积计算优化的设备上。此外，Transformer模型通常不适用于量化（Bondarenko 等人, 2021），而卷积网络的量化则是研究较为深入的课题。我们提炼了T、S、B和L四种规模的CNX架构（参见图16a），并与原始ConvNeXt模型（Liu 等人, 2022）进行对比。这些基线模型在ImageNet-1k上表现优异，因为它们是使用ImageNet-22k标签以监督方式训练的，因此具有强大的竞争力。本实验中，我们提供了输入分辨率256和512的全局任务结果、512分辨率的ADE20k任务结果，以及640分辨率的NYU任务结果。

表15：我们对蒸馏DINOv3 ConvNeXt模型的评估。我们将模型与基于ImageNet-22k数据集（Liu 等人，2022）进行监督训练的现成ConvNeXt模型进行对比。针对全局任务，我们在256和512输入分辨率下给出结果，因为发现监督模型在512分辨率下性能显著下降。

尺寸	模型	全局任务						密集任务	
		IN-REAL		IN-R		目标		ADE20k	纽约大学I
		256	512	256	512	256	512		
T	Sup.	87.3	83.0	45.0	33.0	44.5	27.1	24.8	0.666
T	DINOv3	86.6	87.7	73.7	74.1	52.6	58.7	42.7	0.448
S	Sup.	88.9	86.8	52.8	39.1	50.8	40.0	22.6	0.630
S	DINOv3	87.9	88.7	73.7	74.1	52.6	58.7	44.8	0.432
B	Sup.	89.3	87.8	57.3	46.2	53.6	46.5	26.5	0.596
B	DINOv3	88.5	89.2	77.2	78.2	56.2	61.3	46.3	0.420
L	Sup.	89.6	88.1	58.4	46.6	55.0	47.7	33.3	0.567
L	DINOv3	88.9	89.4	81.3	82.4	59.3	65.2	47.8	0.403

结果（表15） 我们发现，在256分辨率的分布内图像分类任务中，我们的模型略逊于监督学习模型（例如CNX-T的-0.7 IN-ReAL）。然而在512分辨率时趋势反转，监督学习的ConvNeXts模型性能显著下降，而我们的模型却能随输入分辨率提升而扩展。对于分布外分类任务（IN-R、ObjectNet），两种模型在所有分辨率下均存在显著差距——这充分证明了DINOv3 CNX 模型的鲁棒性。此外，DINOv3模型在密集任务中展现出巨大改进：对于CNX-T任务，我们的模型实现了+17.9 mIoU的提升（42.7对比24.8）；对于CNX-L任务，提升幅度达+14.5 mIoU（47.8对比33.3）。高性能与计算效率的结合，使得蒸馏后的ConvNeXt模型在资源受限的现实应用中极具潜力。更令人振奋的是，将ViT-7B模型蒸馏为更小的ConvNeXt模型，成功实现了两种根本不同架构的融合。虽然ViT-7B基于带有CLS标记的Transformer模块，但ConvNeXt依赖于不带CLS标记的卷积操作，这使得知识迁移变得非同寻常。这一成果凸显了我们蒸馏过程的多功能性和有效性。

7.3 基于DINOv3的dino.txt进行零样本推理

如第5.3节所述，我们按照dino.txt Jose等人（2025）的方案，训练文本编码器使蒸馏DINOv3 ViT-L模型的CLS标记和输出补丁与文本对齐。我们在标准基准上从全局和补丁层面评估对齐质量。报告使用CLIP协议（Radford 等人，2021）在ImageNet-1k、ImageNetAdversarial、ImageNet-Rendition和ObjectNet基准上的零样本分类准确率。对于图像-文本检索，我们在COCO2017数据集（Tsung-Yi 等人，2017）上评估，并报告图像到文本（I → T）和文本到图像（T → I）任务的召回率@1。为探究补丁级对齐质量，我们使用通用基准ADE20k和Cityscapes在开放词汇分割任务上评估模型，并报告mIoU指标。

结果（表16） 我们将文本对齐的DINOv3 ViT-L与同规模的竞争对手进行对比。相较于Jose等人（2025）将DINOv2对齐到文本的方法，DINOv3在所有基准测试中均展现出显著提升的性能。在全局对齐任务中，我们优于原始CLIP（Radford 等人，2021）和EVA-02-CLIP（Sun 等人，2023）等强基线模型，但稍逊于SigLIP2（Tschanne 等人，2025）和Perception Encoder（Bolya 等人，2025）。在密集对齐任务中，得益于DINOv3的清晰特征图，我们的文本对齐模型在ADE20K和Cityscapes两个具有挑战性的基准测试中表现优异。

表16：我们将文本对齐的DINOv3 ViT-L与当前最先进模型进行对比。该模型在保持全局对齐任务竞争力的同时，展现出优异的密集对齐性能。所有对比模型均采用ViT-L规模，且处理的序列长度统一为576。

方法	分类				检索		分割	
	IN1k	A	R	目标	I → T	T → I	ADE20k	城市景观
夹子	76.6	77.5	89.0	72.3	57.9	37.1	6.0	11.5
EVA-02-CLIP	80.4	82.9	93.2	78.5	64.1	47.9	10.9	14.1
恐龙.txt	81.6	83.2	88.8	74.5	62.5	45.0	19.2	27.4
SigLIP 2	83.1	84.3	95.7	84.4	71.4	55.3	10.8	16.3
体育	83.5	89.0	95.2	84.7	75.9	57.1	17.6	21.4
DINOv3 dino.txt	82.3	85.4	93.0	80.5	63.7	45.6	24.7	36.9

8 地理空间数据上的DINOv3

我们的自监督学习方案具有通用性，可应用于任何图像领域。在本节中，我们通过构建DINOv3 7B模型来展示这种通用性，该模型专门用于处理卫星图像——这类图像的特征（例如物体纹理、传感器噪声和焦距视角）与DINOv3最初开发的网络图像存在显著差异。

8.1 预训练数据与基准

我们的DINOv3 7B卫星模型基于SAT-493M数据集进行预训练，该数据集包含4.93亿张 512×512 分辨率的图像，这些图像随机采样自Maxar公司0.6米分辨率的RGB正射影像。除针对卫星图像调整的RGB均值和标准差归一化参数及训练时长外，我们沿用了与网页版DINOv3 7B模型完全相同的超参数配置。与网页版模型类似，卫星模型的训练流程包含三个阶段：首先进行10万次全局裁剪（ 256×256 ）的初始预训练，接着采用格拉姆正则化进行1万次迭代，最后以 512 分辨率的8000步高精度微调完成最终训练。同样地，我们还将7B卫星模型精简为更易操作的ViT-Large模型，以便在预算有限的场景中灵活使用。

我们在多个地球观测任务中评估DINOv3卫星与网络模型。对于全球树冠高度制图任务，我们使用[附录D.13](#)中描述的Satlidar数据集，该数据集包含一百万张 512×512 图像，其LiDAR真实值按8/1/1比例划分为训练/验证/测试集。这些划分包括[Tolan 等人 \(2024\)](#)使用的Neon和Sao Paulo数据集。对于国家级树冠高度制图任务，我们在Open-Canopy ([Fogel 等人, 2025](#)) 上进行评估，该数据集整合了法国境内87,000平方公里²的SPOT 6-7卫星影像与航空LiDAR数据。由于该数据集图像包含四个通道（含额外的红外IR通道），我们通过取三个通道的平均值作为块嵌入模块权重，并将其作为第四个通道加入权重，从而调整主干网络。我们使用 512×512 图像裁剪（调整至1667分辨率以匹配Maxar地面样本分辨率）训练DPT解码器。

语义地理空间任务通过GEO-Bench ([Lacoste 等人, 2023](#)) 进行评估，该工具包含六个分类任务和六个分割任务，涵盖多种空间分辨率和光学波段。GEO-Bench任务类型多样，包括屋顶光伏系统的检测、局部气候区的分类、森林砍伐驱动因素的测量以及树冠的检测。对于高分辨率语义任务，我们考虑了土地覆盖分割数据集LoveDA ([Wang 等人, 2022a](#))、物体分割数据集iSAID ([Zamir 等人, 2019](#)) 以及水平检测数据集DIOR ([Li 等人, 2020](#))。

8.2 树冠高度估算

从卫星影像估算冠层高度是一项具有挑战性的度量任务，需要在坡度、观测几何、太阳角度、大气散射和量化伪影等随机变化条件下，准确恢复连续的空间结构。该任务对全球碳监测以及森林与农业管理至关重要 ([Harris 等人, 2021](#))。继[Tolan 等人 \(2024\)](#)之后，首个利用SSL

表17: 不同骨干网络在高分辨率冠层高度预测中的评估。所有模型均采用DPT解码器进行训练。结果分别展示在基于SatLidar训练并在IID样本(SatLidar Val)和OOD测试集(SatLidar Test、Neon和Sao Paulo)上评估的实验，或基于Open-Canopy数据集训练和评估的实验。我们列出平均绝对误差(MAE)以及Tolan等人(2024)提出的块 R^2 指标。为完整性，我们额外评估了Tolan等人(2024)在Neon数据集上训练的原始解码器(标记为*)。

方法	存档	卫星激光雷达								开盖 MAE↓	
		SatLidar Val		星载激光雷达 测试		氟	试验	圣保罗			
		MAE↓	$R^2\uparrow$	MAE↓	$R^2\uparrow$	MAE↓	$R^2\uparrow$	MAE↓	$R^2\uparrow$		
Tolan等人(2024)*	ViT-L	2.8	0.86	4.0	0.61	2.7	0.73	5.4	0.42	—	
Tolan等人(2024)	ViT-L	2.4	0.90	3.4	0.81	2.9	0.69	5.4	0.48	2.42	
DINOv3网络	ViT-7B	2.4	0.90	3.6	0.74	2.7	0.75	5.9	0.34	2.17	
DINOv3卫星	ViT-L	2.2	0.91	3.2	0.81	2.4	0.81	5.8	0.42	2.07	
DINOv3卫星	ViT-7B	2.2	0.92	3.2	0.82	2.6	0.74	5.5	0.51	2.02	

基于卫星图像训练的骨干网络，我们在冻结的DINOv3模型上训练DPT头，使用SatLidar1M训练集进行训练，随后在SatLidar1M验证集的独立同分布样本以及包括SatLidar1M测试集、Neon和Sao Paulo在内的分布外测试集上进行评估。此外，我们还在Open-Canopy数据集上进行了训练和评估。

结果(表17) 我们比较了不同的SSL骨干网络，其中“DINOv3 Sat”表示在SAT-493M数据集上训练的模型，“DINOv3 Web”表示在LVD-1689M上训练的模型(参见第3.1节)。可以看出DINOv3卫星模型在大多数基准测试中都取得了最先进的性能。我们的7B卫星模型在SatLidar1M验证集、SatLidar1M测试集和Open-Canopy上均刷新了新纪录，将MAE分别从2.4降至2.2、从3.4降至3.2、从2.42降至2.02。这些结果表明DINOv3训练方案具有通用性，可直接应用于其他领域。值得注意的是，我们蒸馏的ViT-L卫星模型表现与其7B版本相当，在SatLidar1M和Open-Canopy上取得相近结果，而在Neon测试集上表现更出色，最低MAE仅为2.4，远低于7B模型的2.6和Tolan等人(2024)的2.9。我们的DINOv3 7B网络模型在基准测试中表现良好，在SatLidar1M验证集、Neon和Open-Canopy上优于Tolan等人(2024)，但仍落后于卫星模型。这凸显了领域特定预训练在基于物理原理的任务(如树冠高度估算)中的优势，其中传感器特异性先验信息与辐射测量一致性至关重要。

8.3 与地球观测技术现状的比较

我们在表18和表19中比较了不同方法在地球观测任务中的性能。冻结的DINOv3卫星模型和网络模型在15个分类、分割和水平物体检测任务中的12个取得了新的最先进的结果。我们的Geo-Bench结果超越了先前模型，包括Prithvi-v2(Szwarcman等人，2024)和DOFA(Xiong等人，2024)，这些模型使用6+波段进行Sentinel-2和Landsat任务，以及任务特定的微调(表18)。尽管使用仅含RGB输入的冻结主干网络，DINOv3卫星模型在三个非饱和分类任务和六个分割任务中的五个上优于先前方法。有趣的是，DINOv3 7B网络模型在这些基准测试中表现非常具有竞争力。它在许多GEO-Bench任务以及大规模高分辨率遥感分割和检测基准测试中实现了相当或更强的性能。如表18和表19所示，冻结的DINOv3网络模型在Geo-Bench任务中建立了新的领先结果，以及在分割在LoveDA和DIOR数据集上的分类与检测任务。

这些发现对地理空间基础模型的设计具有更广泛的影响。近期研究强调了启发式技术，例如多时相聚合、多传感器融合或整合卫星特定元数据(Brown等人，2025；Feng等人，2025)。我们的研究结果表明，通用SSL在依赖精确对象边界(seg-

表18：我们的DINOv3模型与强基线 DOFA (Xiong 等人, 2024)、Prithiv2 (Szwarcman 等人, 2024) 以及 Tolan等人 (2024) 在Geo-Bench任务中的对比。虽然Privthiv2和DOFA利用了所有可用的光学波段，但我们的模型仅使用RGB输入即可实现显著更好的性能。

(a) 分类任务										
方法	存档	FT	乐队	m-贝内特	m砖窑	m-eurosat	m-forestnet	m-pv4ger	m-二氧化硫卫星	平均
DOFA	ViT-L	🌟	所有	68.7	98.4	96.6	55.7	98.2	61.6	79.9
Prithvi-v2最佳	ViT-L/H	🌟	所有	71.2	98.8	96.4	54.1	98.1	59.1	79.6
Tolan 等人 (2024)	ViT-L	* RGB	RGB	66.0	97.1	95.2	56.3	94.3	58.1	77.8
DINOv3卫星	ViT-L	* RGB	RGB	73.0	96.5	94.1	60.6	96.0	57.4	79.6
DINOv3卫星	7B	* RGB	RGB	74.0	97.2	94.8	62.3	96.1	62.1	81.1
DINOv3网络	7B	* RGB	RGB	74.6	97.7	97.0	57.9	98.3	63.8	81.6

(b) 分割任务											平均
方法	存档	FT	乐队	m-腰果	*	m-切萨皮克	m-NeonTree	m-nz-牛	m-pv4ger-seg	m-短臂	平均
DOFA	ViT-L	🌟	所有	81.2	61.6	58.5	77.4	95.1	35.7	68.3	
Prithvi-v2最佳	ViT-L/H	🌟	所有	90.2	69.4	59.1	81.0	95.3	41.9	72.8	
Tolan 等人 (2024)	ViT-L	* RGB	RGB	92.8	73.7	58.1	83.1	94.7	35.1	72.9	
DINOv3卫星	ViT-L	* RGB	RGB	94.2	75.6	61.8	83.7	95.2	36.8	74.5	
DINOv3卫星	7B	* RGB	RGB	94.1	76.6	62.6	83.4	95.5	37.6	75.0	
DINOv3网络	7B	* RGB	RGB	96.0	76.5	66.4	83.7	95.9	36.8	75.9	

* 根据 Szwarcman 等人 (2024) 的研究，转换为6个类别。

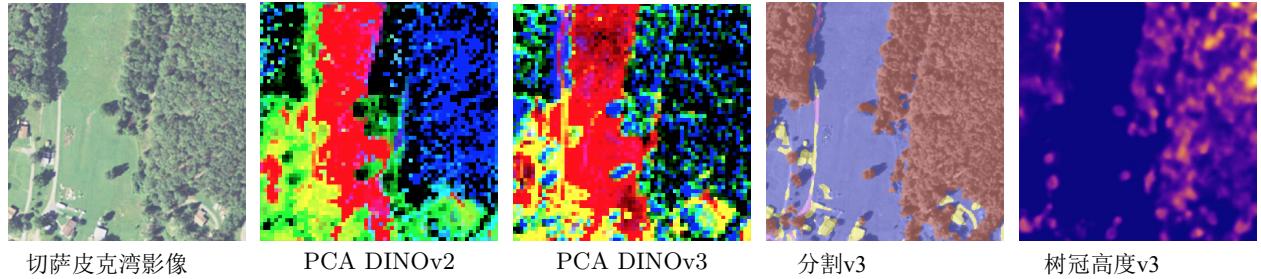


图18：通过单一DINOv3模型实现的遥感多功能应用示意图。DINOv3特征的主成分分析 (PCA) 比DINOv2呈现更精细的细节。分割图仅使用GEO-Bench Chesapeake标签计算。冠层高度模型解码器在Open-Canopy数据集上使用4个通道 (RGB + 红外) 进行训练，而推理仅基于RGB通道完成。

(如记忆或物体检测)。这支持了新兴证据，表明领域无关的预训练即使在专业下游领域也能提供强大的泛化能力 (Lahrichi 等人, 2025)。

综合来看，我们的研究结果表明领域特定预训练具有任务依赖性优势。DINOv3卫星模型凭借卫星特有的先验知识，在深度估计等度量任务中表现突出；而DINOv3网络模型则通过多样化的通用表征，在语义地理空间任务中取得业界领先水平。两种模型的互补优势充分证明了DINOv3 SSL范式的广泛适用性和有效性。

9 环境影响

为了估算我们预训练的碳排放，我们遵循自然语言处理领域先前研究 (Strubell 等人, 2019; Touvron 等人, 2023) 和SSL (Oquab 等人, 2024) 中使用的方法。我们将所有外生变量的值固定为即电网的电力使用效率 (Pue) 和碳强度因子，与Touvron 等人 (2023) 使用的相同，即我们假设Pue为1.1，美国平均碳强度因子为0.385 kgCO₂当量/KWh。对于GPU的功耗，我们取

表19: 我们比较了DINOv3与当前最先进的模型Privthi-v2 (Szwarcman等人, 2024)、BillionFM (Cha 等人, 2024) 和 SkySense V2 (Zhang 等人, 2025) 在高分辨率语义地理空间任务中的性能。我们报告了分割数据集LoveDA (1024 ×) 和iSAID (896 ×) 的mIoU, 以及检测数据集DIOR (800 ×) 的mAP。

方法	存档	FT	洛达	iSAID	迪奥
Prev. SotA		🔥	BillionFM, ViT-G 54.4	SkySense V2, Swin-G* 71.9	SkySense V2, Swin-G* 79.5
解码器			上皮细胞网络	上皮细胞网络	快速 RCNN
Privthi-v2	ViT-H	🔥	52.2	62.8	—
DINOv3卫星	ViT-L	❄️	54.4	62.9	72.7
DINOv3卫星	ViT-7B	❄️	55.3	64.8	76.6
DINOv3网络	ViT-7B	❄️	56.2	71.4	80.5

* 采用基于OpenStreetMap 监督预训练对齐的改进型DINOv2 SSL 模型, 在iSAID 数据集上报告+0.8 mIo U 的性能。

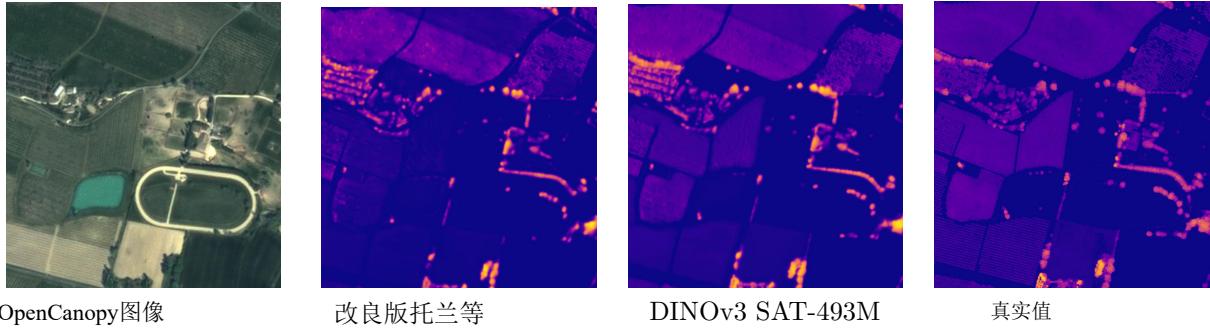


图19: DINOv3 7B卫星模型与Tolan 等人 (2024) 在Open Canopy数据集上的定性比较。对于这两个模型, 解码器均在448 × 448的输入图像上进行训练。可以看出DINOv3生成了更精确的地图, 例如对田野中树木高度的准确测量。

其热设计功耗: A100 GPU为400W, H100 GPU为700W。我们已在**表20**中详细列出了ViT-7B预训练的计算参数。作为参考, 我们同时提供了DINOv2和MetaCLIP的对应数据。另一个对比点是, 训练一个DINOv3模型所需的47兆瓦时 (MWh) 电量, 与普通电动汽车行驶24万公里的能耗大致相当。

项目全生命周期碳足迹 为计算项目全生命周期碳足迹, 我们采用约900万GPU小时的粗略估算。基于前文所述网格参数, 估算总碳足迹约为2600tCO₂。作为对比, 一架波音777客机往返巴黎与纽约的完整航班碳排放量约为560tCO₂。假设每日运营12个此类航班, 本项目环境影响相当于巴黎与纽约间单日航班总量的一半。该估算仅考虑GPU运行用电, 未包含冷却、制造及废弃处理等其他排放源。

表20: 模型训练的碳足迹。我们报告了使用Pue为1.1和碳强度因子为0.385kg CO₂当量/KWh计算的完整模型预训练的潜在碳排放。

模型	存档	GPU类型	权力 (W)	步骤	GPU小时	原因不明发热	总功率 兆瓦时	发出 (tCO ₂ eq)
MetaCLIP	ViT-G	A100-40GB	400W	390k	368,640	1.1	160	62
DINOv2	ViT-g	A100-40GB	400W	625k	22,016	1.1	9.7	3.7
DINOv3	ViT-7B	H100-SXM5	700W	1,000k	61,440	1.1	47	18

10 结论

DINOv3在自监督学习领域实现了重大突破，展现出革新多领域视觉表征学习方式的潜力。通过精心设计数据集和模型规模，并进行细致的数据准备与优化，DINOv3充分展现了自监督学习消除人工标注依赖的强大能力。引入的Gram锚定方法有效缓解了长时间训练导致的密集特征图退化问题，确保了模型性能的稳健可靠。

通过结合后处理优化策略（如高分辨率训练后处理和蒸馏技术），我们在无需对图像编码器进行微调的情况下，实现了跨多种视觉任务的顶尖性能。DINOv3视觉模型套件不仅刷新了行业基准，更在不同资源限制、部署场景和应用场景中提供了灵活多样的解决方案。DINOv3取得的突破性进展，充分证明了自监督学习在推动计算机视觉领域乃至更广泛领域实现技术突破方面的巨大潜力。