

DINOv3

Oriane Siméoni* Huy V. Vo* Maximilian Seitzer* Federico Baldassarre* Maxime Oquab*
Cijo Jose Vasil Khalidov Marc Szafranec Seungeun Yi Michaël Ramamonjisoa
Francisco Massa Daniel Haziza Luca Wehrstedt Jianyuan Wang
Timothée Darcet Théo Moutakanni Leonel Sentana Claire Roberts
Andrea Vedaldi Jamie Tolan John Brandt¹ Camille Couprie
Julien Mairal² Hervé Jégou Patrick Labatut Piotr Bojanowski

Meta AI Research¹ WRI² Inria

*corresponding authors: { osimeoni, huyvvo, seitzer, baldassarre, gas } @meta.com

Abstract

自监督学习有望消除对人工数据标注的需求，使模型能够轻松扩展到庞大的数据集和更大规模的架构。由于不针对特定任务或领域，这种训练范式有潜力从多样化的来源学习视觉表示，从自然图像到航空图像，使用单个算法。本文技术报告介绍了 DINOv3，一个通过采用简单而有效的策略实现这一愿景的重要里程碑。首先，我们通过仔细的数据准备、设计和优化，利用数据集和模型规模扩展的好处。其次，我们引入了一种称为 Gram anchoring 的新方法，有效解决了在长时间训练中密集特征图退化的已知未解问题。最后，我们应用事后策略，进一步增强模型在分辨率、模型规模和文本对齐方面的灵活性。结果，我们展示了一个通用的视觉基础模型，在广泛的设置中，无需微调就能超过专业化的现有技术水平。DINOv3 产生的高质量密集特征在各种视觉任务中的表现卓越，远远超过之前的自监督和弱监督基础模型。我们还共享了 DINOv3 视觉模型套件，旨在通过为各种资源限制和部署场景提供可扩展的解决方案，以推动各类任务和数据上的技术进步。

1 引言

基础模型已成为现代计算机视觉中的核心构建模块，通过单一的、可重用的模型在不同任务和领域之间实现广泛的泛化。自监督学习 (SSL) 是一种强大的训练此类模型的方法，通过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性。与需要与高质量元数据配对的图像的弱监督和完全监督预训练方法 (Radford et al., 2021; Dehghani et al., 2023; Bolya et al., 2025) 不同，SSL 解锁了对大量原始图像集合的训练。这对于训练大规模视觉编码器特别有效，因为训练数据几乎是无限可用的。DINOv2 (Oquab et al., 2024) 体现了这些优势，在图像理解任务中取得了令人印象深刻的成果 (Wang et al., 2025) 并且为复杂领域如组织病理学的预训练提供了可能 (Chen et al., 2024)。经过 SSL 训练的模型展现了额外的理想特性：它们对输入分布的变化具有鲁棒性，提供强大的全局和局部特征，并生成丰富的嵌入以促进物理场景理解。由于 SSL 模型不是为任何特定下游任务训练的，它们产生多功能且稳健的通用特征。例如，DINOv2 模型在不同任务和领域中提供了强大的表现，无需任务特定的微调，允许单一冻结的主干为多种用途服务。重要的是，自监督学习特别适合于在组织病理学 (Vorontsov et al., 2024)、生物学 (Kim et al., 2025)、医学成像 (Pérez-García et al., 2025)、遥感 (Cong et al., 2022; Tolan et al., 2024)、天文学 (Parker et al., 2024) 或高能粒子物理 (Dillon et al., 2022) 等领域中利用大量现有的观察数据进行训练。这些领域通常缺乏元数据，并已证明能够从诸如 DINOv2 之类的基础模型中获益。最后，SSL 无需人工干预，非常适合在日益增长的网络数据中进行终身学习。

在实践中，SSL 的承诺，即通过利用大量非约束数据来生成任意大且强大的模型，在规模上仍然具有挑战性。虽然模型的不稳定性和崩溃被 Oquab et al. (2024) 提出的启发式方法缓解了，但随着进一步的扩展，更多问题出现了。首先，如何从未标记的集合中收集有用的数据尚不明确。其次，在通常的训练实践中，采用余弦调度意味着需要预先知道优化的时间范围，这在处理大型图像集时是困难的。第三，经过前期训练后，特征的性能逐渐下降，这通过对补丁相似性图的视觉检查得到了验证。这种现象出现在长时间训练并且模型尺寸在 ViT-Large 以上 (300M 参数) 的情况下，降低了扩展 DINOv2 的实用性。

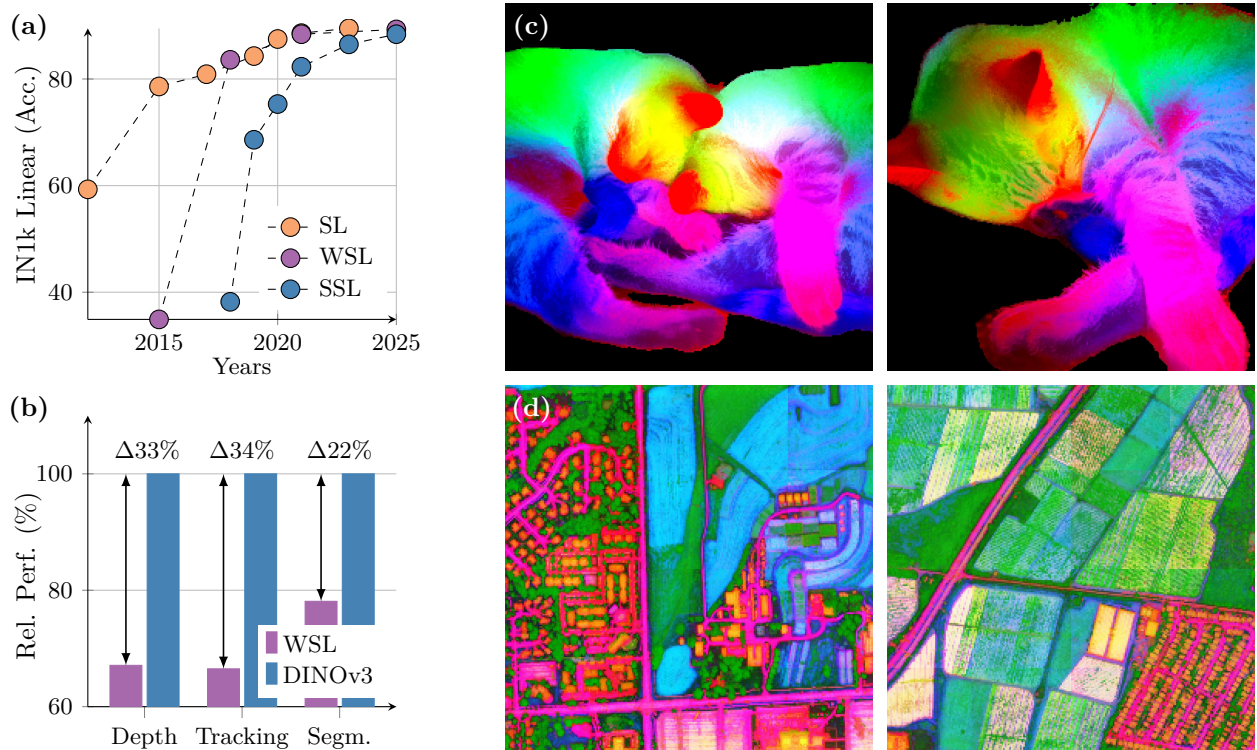


Figure 1: (a) 随着时间的推移，ImageNet1k (IN1k) 的线性探测结果的演变，比较了全监督学习 (SL)、弱监督学习 (WSL) 和自监督学习 (SSL) 方法。虽然起步较晚，但 SSL 迅速发展，现在已经达到了近年来 Imagenet 的准确性水平。另一方面，我们证明了 SSL 提供了高质量密集特征的独特潜力。借助 DINOv3，我们显著提高了弱监督模型在密集任务上的表现，如最佳弱监督模型与 DINOv3 在相对性能上的比较 (b)。我们还生成了通过 DINOv3 在自然 (c) 和航拍图像 (d) 上训练得到的高分辨率图像特征的 PCA 地图。

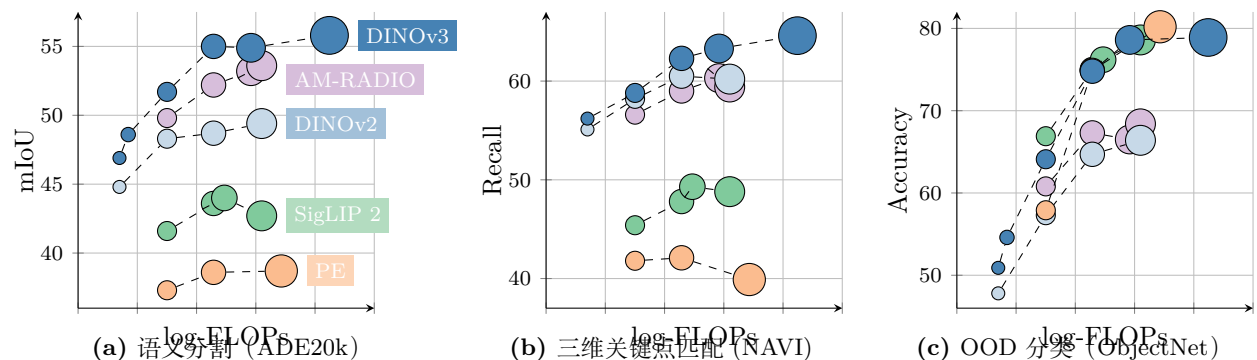


Figure 2: DINOv3 模型家族在不同基准上的表现，相较于其他自监督或弱监督模型家族。DINOv3 在密集基准上显著超越其他模型，包括利用掩码标注先验的模型，如 AM-RADIO (Heinrich et al., 2025)。

解决上述问题导致了这项工作，DINOv3，它在大规模下推进了自监督学习 (SSL) 训练。我们证明，一个单一冻结的 SSL 骨干网络可以作为一个通用的视觉编码器，在具有挑战性的下游任务中实现最先进的性能，优于监督和依赖元数据的预训练策略。我们的研究以以下目标为指导：(1) 训练一个在任务和领域方面都具有通用性的基础模型，(2) 改善现有 SSL 模型在密集特征上的不足，(3) 传播一系列可以直接使用的模型。在下面我们讨论这三个目标。

强大且多功能的基础模型 DINOv3 旨在沿两个方向提供高水平的多功能性，这得益于模型大小和训练数据的扩展。首先，一个自监督学习 (SSL) 模型的重要属性是能够在冻结状态下实现卓越的性能，理想情况下

达到与专门化模型相似的最新结果。在这种情况下，一个前向传递就可以在多个任务中提供前沿的结果，从而带来显著的计算节省——对于实际应用特别是在边缘设备上的应用，此优势至关重要。我们在 [Sec. 6](#) 中展示了 DINOv3 可以成功应用的广泛任务。其次，一个不依赖于元数据的可扩展 SSL 训练流程解锁了众多科学应用。通过对多样化的图像集进行预训练，无论是网页图像还是观测数据，SSL 模型能够跨越大量领域和任务进行泛化。如在 [Fig. 1](#) (d) 中所示，从高分辨率航拍图像中提取的 DINOv3 特征的主成分分析 (PCA) 清晰地分离了道路、房屋和绿地，突出了模型的特征质量。

通过 Gram Anchoring 获得优越的特征图 DINOv3 的另一个重要特性是显著改进了其密集特征图。DINOv3 的自监督学习训练策略旨在生成在高层次语义任务上表现出色的模型，同时生成能够很好地解决几何任务（如深度估计或 3D 匹配）的特征图。特别是，模型应生成密集的特征，这些特征可以直接使用或仅需少量后处理。当使用大量图像进行训练时，密集和全局表示之间的平衡尤其难以优化，因为高层次理解的目标可能与密集特征图的质量产生冲突。这些矛盾的目标会导致密集特征在大模型和长时间训练中崩溃。我们的新 Gram anchoring 策略有效缓解了这种崩溃（见 [Sec. 4](#)）。因此，DINOv3 获得了比 DINOv2 显著更好的密集特征图，即使在高分辨率下也保持清晰（见 [Fig. 3](#)）。

解决密集特征图的退化问题

DINOv3 模型系列 通过使用 Gram anchoring 释放了缩放的力量。因此，使用 SSL 训练更大模型可以显著提升性能。在这项工作中，我们成功训练了一个拥有 7B 参数的 DINO 模型。由于如此庞大的模型需要大量资源运行，我们应用蒸馏技术将其知识压缩至较小的变体。因此，我们呈现了 DINOv3 家族的视觉模型，这是一套全面的系统，旨在解决各种计算机视觉挑战。这个模型家族旨在通过提供可扩展的解决方案来推进技术的进步，以适应各种资源约束和部署场景。蒸馏过程在多个规模上产生了模型变体，包括 Vision Transformer (ViT) Small、Base 和 Large，以及基于 ConvNeXt 的架构。值得注意的是，高效且广泛采用的 ViT-L 模型在各种任务中表现接近原始 7B 教师模型。总体而言，DINOv3 家族在广泛的基准测试上表现出强劲的性能，在全球任务中匹配或超过竞争模型的精度，同时在密集预测任务中显著超越它们，如 [Fig. 2](#) 中所示。

在这项工作中，我们引入了多个贡献来解决将 SSL 扩展到大型前沿模型的挑战。我们构建在最近的自动数据策划进展上 ([Vo et al., 2024](#))，以获得一个大型“背景”训练数据集，我们通过仔细混合一点专业数据 (ImageNet-1k) 来实现。这使得利用大量不受约束的数据来提高模型性能成为可能。这个关于数据扩展的贡献 (i) 将在 [Sec. 3.1](#) 中描述。

我们通过定义 ViT 架构的一个自定义变体，将主模型的参数规模增加到 7B。我们引入了现代位置嵌入（轴向 RoPE）并开发了一种正则化技术，以避免位置伪影。不同于 DINOv2 中的多个余弦调度，我们在 1M 次迭代中使用恒定的超参数调度进行训练。这使得模型性能更强。关于模型架构和训练的贡献 (ii) 将在 ?? 中描述。

通过上述技术，我们能够按照 DINOv2 算法在大规模上训练模型。然而，如前所述，规模会导致密集特征的退化。为了解决这个问题，我们提出了一种具有 Gram anchoring 训练阶段的核心改进。这清理了特征图中的噪声，产生了令人印象深刻的相似性图，并显著提高了参数化和非参数化密集任务的性能。这项关于 Gram 训练的贡献 (iii) 将在 [Sec. 4](#) 中进行描述。

遵循先前的实践，我们的管道的最后步骤包括一个高分辨率的训练后阶段，并将其蒸馏成一系列不同大小的高性能模型。对于后者，我们开发了一种新颖且高效的单教师多学生蒸馏过程。这个贡献 (iv) 将我们 7B 前沿模型的能力转移到一系列较小的实用模型上，以便常规使用，我们在 [Sec. 5.2](#) 中进行了描述。

根据我们的详尽基准测试，[Sec. 6](#) 中的结果表明，我们的方法在密集任务中定义了一个新的标准，在全局任务上表现与 CLIP 衍生方法相近。特别是，使用冻结的视觉骨干网，我们在长期以来的计算机视觉问题上取得了最新的性能，如目标检测 (COCO 检测，mAP 66.1) 和图像分割 (ADE20k，mIoU 63.0)，超过了专门调优的管道。此外，我们通过将 [Sec. 7](#) 中将 DINOv3 算法应用于卫星图像，提供了我们方法在跨领域的普适性的证据，超越了所有先前的方法。

2 相关工作

自监督学习 学习在没有标注的情况下，需要一个为训练提供监督的人造学习任务。SSL 的艺术和挑战在于精心设计这些所谓的前文本任务，以便为后续任务学习强大的表示。由于语言领域的离散性质，提供了设置此类任务的简单方式，这导致了許多成功的文本数据无监督预训练方法。例子包括词嵌入 ([Mikolov et al.,](#)



Figure 3: 高分辨率密集特征。我们可视化了用 DINOv3 输出特征获得的余弦相似性图，这些图是在标记有红叉的补丁和其他所有补丁之间的方法。输入图像大小为 4096x4096。请放大查看，您同意 DINOv3 吗？

2013; Bojanowski et al., 2017)、句子表示 (Devlin et al., 2018; Liu et al., 2019) 和普通语言模型 (Mikolov et al., 2010; Zaremba et al., 2014)。相比之下，由于信号的连续性，计算机视觉提出了更大的挑战。早期模仿语言方法的尝试，通过从图像的部分提取监督信号来预测其他部分，如预测相对补丁位置 (Doersch et al., 2015)、补丁重新排序 (Noroozi and Favaro, 2016; Misra and Maaten, 2020) 或图像修复 (Pathak et al., 2016)。其他任务包括重新上色图像 (Zhang et al., 2016) 或预测图像变换 (Gidaris et al., 2018)。

在这些任务中，由于基于补丁的 ViT 架构的灵活性，基于修补的方式引起了广泛的兴趣 (He et al., 2021; Bao et al., 2021; El-Nouby et al., 2021)。其目标是重建图像的损坏区域，这可以被视为一种去噪自编码，与 BERT 预训练中的屏蔽标记预测任务在概念上相关 (Devlin et al., 2018)。值得注意的是，He et al. (2021) 证明了像素级的掩码自编码器 (MAE) 可以用作下游任务微调的强大初始化。在下文中，Baeviski et al. (2022; 2023); Assran et al. (2023) 表明预测学习的潜在空间而不是像素空间会引导出更强大、更高级的特征——一种称为 JEPA：“联合嵌入预测架构”的学习范式 (LeCun, 2022)。最近，JEPA 也被扩展到视频训练中 (Bardes et al., 2024; Assran et al., 2025)。

第二个与我们更接近的研究方向利用图像之间的判别信号来学习视觉表征。这类方法最早可追溯到深度学习早期研究阶段 (Hadsell et al., 2006)，但随着实例分类技术的引入而流行 (Dosovitskiy et al., 2016; Bojanowski and Joulin, 2017; Wu et al., 2018)。随后的进展引入了对比目标和信息论标准 (Hénaff et al., 2019; He et al., 2020; Chen and He, 2020; Chen et al., 2020a; Grill et al., 2020; Bardes et al., 2021)，以及基于自聚类的策略 (Caron et al., 2018; Asano et al., 2020; Caron et al., 2020; 2021)。更近期的方法，如 iBOT (Zhou et al., 2021)，将这些判别性损失与掩蔽重建目标结合起来。所有这些方法都显示出学习强特

征的能力，并在 ImageNet (Russakovsky et al., 2015) 等标准基准上取得了高性能。然而，大多数方法在扩展到更大模型尺寸时面临挑战 (Chen et al., 2021)。

视觉基础模型 深度学习革命始于 AlexNet 的突破 (Krizhevsky et al., 2012)，这是一种深度卷积神经网络，在 ImageNet 挑战中优于之前的所有方法 (Deng et al., 2009; Russakovsky et al., 2015)。早期时，发现从大型人工标记的 ImageNet 数据集上端到端学习的特征对于广泛的迁移学习任务非常有效 (Oquab et al., 2014)。早期的视觉基础模型研究集中于架构开发，包括 VGG (Simonyan and Zisserman, 2015)、GoogleNet (Szegedy et al., 2015) 和 ResNets (He et al., 2016)。

鉴于扩展的有效性，后续研究探索了在大数据集上训练更大型的模型。Sun et al. (2017) 使用包含 3 亿标记图像的专有 JFT 数据集扩展了监督训练数据，展示了惊人的结果。JFT 也使得 Kolesnikov et al. (2020) 的性能显著提升。同时，使用监督和非监督数据的组合来进行扩展也被探讨。例如，一个 ImageNet 监督模型可以用于为非监督数据生成伪标签，然后用于训练更大的网络 (Yalniz et al., 2019)。随后，像 JFT 这样的大型监督数据集的可用性也促进了将 Transformer 架构适应于计算机视觉的过程 (Dosovitskiy et al., 2020)。特别是，若没有 JFT，要达到与原始视觉 Transformer (ViT) 相当的性能需要大量的努力 (Touvron et al., 2020; 2022)。由于 ViTs 的学习能力，Zhai et al. (2022a) 进一步扩展了扩展工作，最终打造出非常大型的 ViT-22B 编码器 (Dehghani et al., 2023)。

鉴于手动标注大型数据集的复杂性，弱监督训练——其中的标注是从与图像相关的元数据中得出的——提供了对监督训练的有效替代。早期，Joulin et al. (2016) 证明了可以通过简单地预测图像标题中的所有词作为目标来对网络进行预训练。这一初步方法通过利用句子结构 (Li et al., 2017)，结合其他类型的元数据并涉及策展 (Mahajan et al., 2018)，以及规模化 (Singh et al., 2022)，得到了进一步的改进。然而，弱监督算法在引入对比损失和标题表示的联合训练后才充分发挥其潜力，这在 Align (Jia et al., 2021) 和 CLIP (Radford et al., 2021) 中得到了体现。

这种非常成功的方法激发了众多开源重现和扩展努力。OpenCLIP (Cherti et al., 2023) 是首个通过在 LAION 数据集 (Schuhmann et al., 2021) 上训练来复制 CLIP 的开源尝试；后续的工作利用预训练的骨干网络，通过以 CLIP 风格的方式对其进行微调 (Sun et al., 2023; 2024)。意识到数据收集是 CLIP 训练成功的关键因素，MetaCLIP (Xu et al., 2024) 精确遵循原始 CLIP 程序来重现其结果，而 Fang et al. (2024a) 使用监督数据集来策划预训练数据。其他工作则专注于改进训练损失，如在 SigLIP (Zhai et al., 2023) 中使用 Sigmoid 损失 *e.g.*，或利用预训练的图像编码器 (Zhai et al., 2022b)。然而，最终获得尖端基础模型的最关键因素是充足的高质量数据和大量计算资源。在这一方面，SigLIP 2 (Tschannen et al., 2025) 和 Perception Encoder (PE) (Bolya et al., 2025) 在超过 40 亿图文对上训练后获得了令人印象深刻的结果。最大的 PE 模型是在 86 B 亿样本上以全球批大小 131 K 进行训练的。最后，已经提出了一系列更复杂且本地多模态的方法；这些包括对比字幕生成 (Yu et al., 2022)，潜在空间的掩码建模 (Bao et al., 2021; Wang et al., 2022b; Fang et al., 2023; Wang et al., 2023a)，以及自回归训练 (Fini et al., 2024)。

相比之下，关于扩展无监督图像预训练的工作相对较少。早期的努力包括 Caron et al. (2019) 和 Goyal et al. (2019) 利用 YFCC 数据集 (Thomee et al., 2016)。进一步的进展是通过关注更大的数据集和模型 (Goyal et al., 2021; 2022a)，以及初步尝试为 SSL 进行数据策划 (Tian et al., 2021)。细致调整训练算法、更大的架构和更广泛的训练数据导致了 DINOv2 的令人印象深刻的结果 (Oquab et al., 2024)；首次，SSL 模型在多个任务上匹配或超过开源 CLIP 变种。最近，这一方向进一步被推动 Fan et al. (2025)，通过扩展到不用数据策划的更大模型，或者通过 Venkataramanan et al. (2025) 使用公开数据集和改进的训练方法。

密集 Transformer 特征 现代视觉应用中，广泛使用预训练 transformer 的密集特征，包括多模态模型 (Liu et al., 2023; Beyer et al., 2024)、生成模型 (Yu et al., 2025; Yao et al., 2025)、3D 理解 (Wang et al., 2025)、视频理解 (Lin et al., 2023a; Wang et al., 2024b) 和机器人 (Driess et al., 2023; Kim et al., 2024)。此外，传统的视觉任务如检测、分割或深度估计需要精确的局部描述符。为了提高 SSL 训练的局部描述符质量，大量研究工作集中于发展局部 SSL 损失。例子包括利用视频中的时空一致性，*e.g.* 使用点跟踪循环作为训练信号 (Jabri et al., 2020)，利用同一图像不同裁剪之间的空间对齐 (Pinheiro et al., 2020; Bardes et al., 2022)，或在相邻补丁之间施加一致性 (Yun et al., 2022)。Darcet et al. (2025) 显示，预测聚类的局部补丁能提升密集表示。DetCon (Hénaff et al., 2021) 和 ORL (Xie et al., 2021) 在区域提议上进行对比学习，但假设这些提议先验存在；ODIN (Hénaff et al., 2022) 和 SlotCon (Wen et al., 2022) 等方法放宽了这一假设。在不改变训练目标的情况下，Darcet et al. (2024) 表明在输入序列中添加注册 token 可以极大地改善密集特征图，而近期的研究发现这可以在不进行模型训练的情况下实现 (Jiang et al., 2025; Chen et al., 2025)。最近一种趋势是基于蒸馏的“聚集”方法，它结合了来自多个图像编码器的信息，这些编码

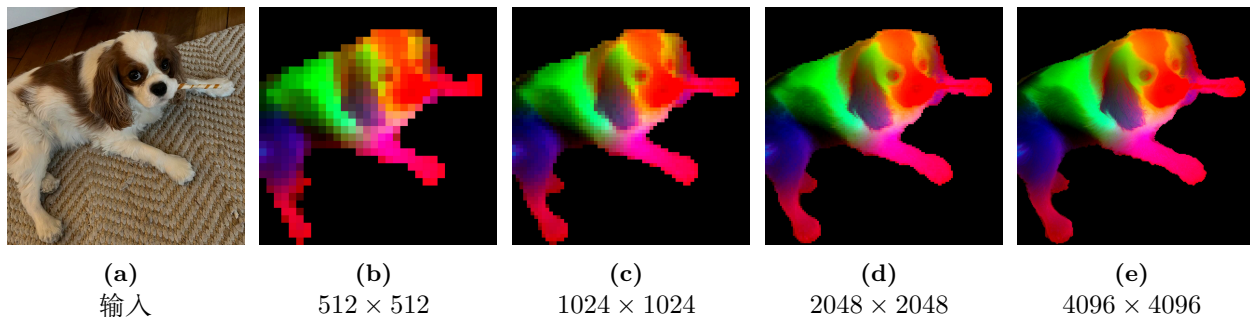


Figure 4: 非常高分辨率下的 DINOv3。我们通过将特征空间上 PCA 计算的前三个分量映射到 RGB 来可视化 DINOv3 的密集特征。为了让 PCA 聚焦于主体，我们通过背景减法对特征图进行遮罩。随着分辨率的增加，DINOv3 产生清晰且语义上有意义的特征。我们在 ?? 中可视化了更多的 PCA。

器在全局和局部特征质量上各不相同，并且使用不同级别的监督进行训练。例如，(Ranzinger et al., 2024; Bolya et al., 2025)：AM-RADIO (Ranzinger et al., 2024) 结合了完全监督的 SAM (Kirillov et al., 2023)、弱监督的 CLIP 和自监督的 DINOv2 的优势，形成一个统一的骨干网络。感知编码器 (Bolya et al., 2025) 类似地将 SAM(v2) 蒸馏成一个称为 PEspatial 的专用密集变体。他们使用的目标是强制学生和教师之间的补丁的余弦相似度较高，其中教师是通过掩膜注释进行训练的。在风格迁移的背景下，类似的损失被证明是有效的，通过减少特征维度的 Gram 矩阵之间的不一致性。在这项工作中，我们采用 Gram 目标来正则化学生和教师之间补丁的余弦相似度，使它们更加接近。在我们的案例中，我们使用更早期的 SSL 模型的迭代作为教师，展示早期阶段的 SSL 模型可以有效指导全局和密集任务的 SSL 训练。

其他工作侧重于对经过自监督学习 (SSL) 训练模型的局部特征进行后处理改进。例如，Ziegler and Asano (2022) 用密集聚类目标对一个预训练模型进行微调；类似地，Salehi et al. (2023) 通过在时间上对齐互补特征来进行微调，在这两种情况下都提高了局部特征的质量。与我们的工作更为接近的是，Pariza et al. (2025) 提出了一个基于修补排序的目标，以鼓励学生和教师生成具有一致邻居顺序的特征。在不微调的情况下，STEGO (Hamilton et al., 2022) 在冻结的 SSL 特征之上学习一个非线性映射以形成紧凑的簇并放大相关模式。或者，Simoncini et al. (2024) 通过将来自不同自监督目标的梯度连接到冻结的 SSL 特征上来增强自监督特征。最近，Wysoczańska et al. (2024) 显示，噪声特征图通过修补的加权平均显著提高。

尽管与 SSL 相关，但并不特别针对 SSL，最近一些研究工作从 ViT 特征图 (Fu et al., 2024) 生成高分辨率的特征图，由于图像分块化，这些特征图通常是低分辨率的。与这一系列工作相比，我们的模型本身能够生成高质量的密集特征图，并且在分辨率变化时保持稳定和一致，如 Fig. 4 所示。

3 无监督大规模训练

DINOv3 是一种下一代模型，旨在通过推动自监督学习的界限，创造迄今为止最为强大和灵活的视觉表达。我们从大型语言模型 (LLMs) 的成功中汲取灵感，对于这些模型而言，扩大模型容量带来了显著的新特性。通过利用比以往更大规模的模型和训练数据集，我们希望释放 SSL 的全部潜力，并为计算机视觉带来类似的范式转变，而不受传统监督或任务特定方法的固有限制所束缚。特别是，SSL 能够产生丰富且高质量的视觉特征，这些特征不会偏向于特定的监督或任务，从而为广泛的下游应用提供了一个多用途的基础。虽然之前在扩大 SSL 模型规模的尝试中受到稳定性问题的阻碍，本节描述了我们如何通过仔细的数据准备、设计和优化来利用扩大规模的益处。我们首先描述数据集的创建过程 (Sec. 3.1)，然后介绍用于 DINOv3 第一训练阶段的自监督 SSL 配方 (??)，其中包括架构选择、损失函数和优化技术。第二个训练阶段将在 Sec. 4 中描述，该阶段专注于密集特征。

3.1 数据准备

数据扩展是大型基础模型成功的驱动因素之一 (Touvron et al., 2023; Radford et al., 2021; Xu et al., 2024; Oquab et al., 2024)。然而，简单地增加训练数据的规模并不一定能转化为更高的模型质量和在下游基准测试中的更好表现 (Goyal et al., 2021; Oquab et al., 2024; Vo et al., 2024)：成功的数据扩展工作通常涉及仔细的数据策展流程。这些算法可能有不同的目标：要么关注于提高数据的多样性和平衡性，要么强调数据的

Table 1: 通过在下游任务上的表现展示训练数据对特征质量的影响。我们将通过聚类 (Vo et al., 2024) 和检索 (Oquab et al., 2024) 整理的数据集与原始数据和我们的数据混合进行比较。该消融研究在 200k 次迭代的较短时间表下进行。

Dataset	IN1k k-NN	IN1k Linear	ObjectNet	iNaturalist 2021	Paris Retrieval
Raw	80.1	84.8	70.3	70.1	63.3
Clustering	79.4	85.4	72.3	81.3	85.2
Retrieval	84.0	86.7	70.7	86.0	82.7
LVD-1689M (ours)	84.6	87.2	72.8	87.0	85.9

实用性——即其与常见实际应用的相关性。为开发 DINOv3，我们结合了两种互补的方法，以提高模型的泛化能力和性能，在这两个目标之间取得平衡。

我们通过利用从 Instagram 公共帖子中收集的网络图像的大型数据池来构建我们的大型预训练数据集。这些图像已经通过平台级别的内容审核以帮助防止有害内容，我们获得了大约 170 亿张图像的初始数据池。使用这个原始数据池，我们创建三个数据集部分。我们通过应用基于层次化 k -means 的自动策展方法来构建第一部分。我们使用 DINOv2 作为图像嵌入，并使用 5 个层级的聚类，聚类的数量从最低层到最高层分别为 200 M、8 M、800 k、100 k 和 25 k。在建立聚类层次结构后，我们应用了 Vo et al. (2024) 中提出的平衡采样算法。结果是一个经过策划的 1689 万张图像的子集（命名为 LVD-1689M），确保了对网络上出现的所有视觉概念的平衡覆盖。对于第二部分，我们采用类似于 Oquab et al. (2024) 提出的程序的基于检索的策展系统。从数据池中检索与所选种子数据集相似的图像，创建一个涵盖下游任务相关的视觉概念的数据集。对于第三部分，我们使用原始的公开可用的计算机视觉数据集，包括 ImageNet1k (Deng et al., 2009)、ImageNet22k (Russakovsky et al., 2015) 和 Mapillary Street-level Sequences (Warburg et al., 2020)。最后这部分允许我们优化模型的性能，按照 Oquab et al. (2024) 的方法。

在预训练过程中，我们使用一个采样器将不同的数据部分混合在一起。混合上述数据组件有几种不同的选项。一个选项是在每次迭代中从单个随机选择的组件中训练同质的数据批次。或者，我们可以在由所有组件的数据按某些比例选定而组装的异质批次上优化模型。受到 Charton and Kempe (2024) 的启发，他们观察到从一个小数据集中选出的非常高质量数据组成的同质批次是有益的，我们在每次迭代中随机抽样，只从 ImageNet1k 提取一个同质批次，或混合来自所有其他组件的数据形成异质批次。在我们的训练中，来自 ImageNet1k 的同质批次占训练的 10 %。

为了评估我们数据整理技术的影响，我们进行了一项消融研究，将我们的数据组合与仅使用聚类或基于检索的方法整理的数据集以及原始数据池进行比较。为此，我们在每个数据集上训练一个模型，并比较它们在标准下游任务中的表现。为了提高效率，我们使用了 20 万次迭代的较短训练计划，而不是 100 万次迭代。在 Tab. 1 中可以看到，没有一种整理技术在所有基准测试中表现最好，而我们完整的流程使我们能够获得两者的最佳表现。

尽管使用自监督学习训练的模型展示了有趣的特性，大多数自监督学习算法尚未扩展到更大规模的模型。这要么是由于训练稳定性的问题，要么是因为过于简单的解决方案未能捕捉视觉世界的全部复杂性。当在大规模进行训练时，使用自监督学习训练的模型并不一定表现出令人印象深刻的性能。一个显著的例外是 DINOv2，这个模型拥有 11 亿个参数，经过精选数据训练，匹配了诸如 CLIP 等弱监督模型的性能。最近的一项努力将 DINOv2 扩展到 70 亿个参数，展示了在全局任务上有希望的结果，但在密集预测任务上结果令人失望。在这里，我们的目标是扩大模型和数据的规模，并获得更强大的视觉表征，同时改善全局和局部属性。

我们使用一种判别性的自监督策略来训练模型，该策略混合了多种自监督目标，包含全局和局部损失项。遵循 DINOv2，我们使用图像级目标，并与补丁级潜在重构目标进行平衡。我们还在两个目标中用 SwAV 的 Sinkhorn-Knopp 替换了 DINO 中的中心化。每个目标是通过主干网络顶部的专用头部输出计算的，从而在损失计算之前允许特征的某些专业化。此外，我们使用专用的层归一化应用于局部和全局裁剪的主干输出。实证上，我们发现这种改变在训练后期可以稳定 ImageNet 的 k 近邻分类 (+0.2 准确率) 并提高密集性能 (在 ADE20k 分割上 +1 mIoU，在 NYUv2 深度估计上 -0.02 RMSE)。此外，添加了一个 Kileo 正则化器，以鼓励批次内的特征在空间中均匀分布。我们使用 Kileo 的分布式实现，其中损失应用于可能跨 GPU 的小批量样本。我们的初始训练阶段通过优化以下损失来进行：

Table 2: 比较 DINOv2 和 DINOv3 模型中使用的教师架构。我们保持模型 40 块的深度，并将嵌入维度增加到 4096。重要的是，我们使用了 16 像素的补丁大小，改变了给定分辨率的有效序列长度。

Teacher model	DINOv2	DINOv3
Backbone	ViT-giant	ViT-7B
# Params	1.1B	6.7B
# Blocks	40	40
Patch Size	14	16
Pos. Embeddings	Learnable	RoPE
Registers	4	4
Embed. Dim.	1536	4096
FFN Type	SwiGLU	SwiGLU
FFN Hidden Dim.	4096	8192
Attn. Heads	24	32
Attn. Heads Dim.	64	128
DINO Head MLP	4096-4096-256	8192-8192-512
DINO Prototypes	128k	256k
iBOT Head MLP	4096-4096-256	8192-8192-384
iBOT Prototypes	128k	96k

在这项工作的模型扩展方面，我们将模型的参数量增加到 70 亿，并在中提供与 DINOv2 工作中训练的 11 亿参数模型的相应超参数的比较。我们还采用了一种自定义的 RoPE 变体：我们的基本实现为每个补丁分配一个标准化 $[-1, 1]$ 框中的坐标，然后根据两个补丁的相对位置在多头注意力操作中应用偏差。为了提高模型对分辨率、比例和长宽比的鲁棒性，我们采用了 RoPE 框抖动。坐标框 $[-1, 1]$ 随机缩放到 $[-s, s]$ ，其中 $s \in [0.5, 2]$ 。这些变化使得 DINOv3 能够更好地学习详细和稳健的视觉特征，提高了其性能和可扩展性。

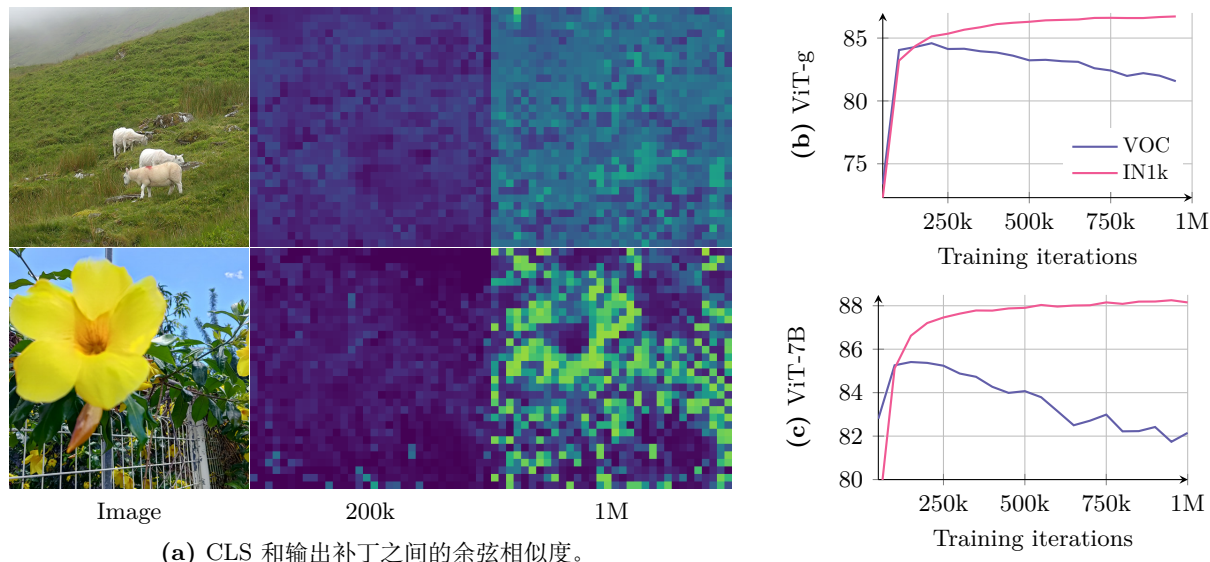
优化 在非常大的数据集上训练大型模型代表了一种复杂的实验流程。由于模型容量与训练数据复杂性之间的相互作用很难事先评估，无法预估正确的优化范围。为了克服这一点，我们取消了所有参数调度，并使用恒定的学习率、权重衰减和教师 EMA 动量进行训练。这有两个主要好处。首先，只要下游性能继续改善，我们就可以继续训练。其次，优化超参数的数量减少，使得更容易正确选择它们。为了使训练正常开始，我们还是对学习率和教师温度进行线性预热。根据常见做法，我们使用 AdamW (Loshchilov and Hutter, 2017)，并将总批次大小设置为跨越 256 个 GPU 的 4096 张图像。我们采用多裁剪策略 (Caron et al., 2020) 来训练我们的模型，每张图像进行 2 次全局裁剪和 8 次局部裁剪。我们使用边长为 256 / 112 像素的方形图像进行全局/局部裁剪，这与改变补丁大小一起，使每张图像具有与 DINOv2 相同的有效序列长度，并且每个批次的总序列长度为 3.7 M 个标记。其他超参数可以在 App. C 中以及代码发布中找到。

4 Gram Anchoring：密集特征的正则化

为了充分利用大规模训练的优势，我们计划对 7B 模型进行长时间训练，并设想该模型可以无限期地训练。预期中，长时间的训练会在全球基准测试中取得进步。然而，随着训练的进行，模型在密集任务 (Figs. 5b and 5c) 上的表现反而下降。这一现象是由于特征表示中出现了补丁级不一致性，这削弱了延长训练的意义。在本节中，我们首先分析了补丁级一致性的丧失，然后提出一个新的目标来缓解这种情况，称为 Gram anchoring。最后，我们讨论了我们的方法对训练稳定性和模型性能的影响。

4.1 训练过程中损失的补丁级一致性

在扩展训练过程中，我们观察到全局指标的一致性提升，但在密集预测任务上的性能显著下降。之前，在训练 DINOv2 时观察到了类似的行为，但程度较轻，并且在 Fan et al. (2025) 的扩展工作中也有讨论。然而，据我们所知，这一问题至今仍未解决。我们在 Figs. 5b and 5c 中展示了这一现象，该文展示了模型在图像分类和分割任务中随迭代的性能表现。对于分类任务，我们使用 CLS 标记在 ImageNet-1k 上训练一个线性分类器，并报告 top-1 准确率。对于分割任务，我们在从 Pascal VOC 提取的 patch 特征上训练一个线性层，并报告平均交并比 (mIoU)。我们观察到，无论是对于 ViT-g 还是 ViT-7B，分类准确率在整个训练过程中单调提升。然而，在大约 200k 次迭代后，分割性能在两种情况下都下降，特别是在 ViT-7B 的情况下，低于早期水平。



(a) CLS 和输出补丁之间的余弦相似度。

Figure 5: 余弦相似度 (a) 以及在 ImageNet1k 线性 (IN1k) 和 VOC 上进行 segmentation 的准确性随 ViT-g (b) 和 ViT-7B (c) 的演变。我们观察到，当 patch 标记和类别标记之间的余弦相似度较低时，分割性能能达到最大。随着训练的进行，这些相似度增加，在密集任务上的性能下降。

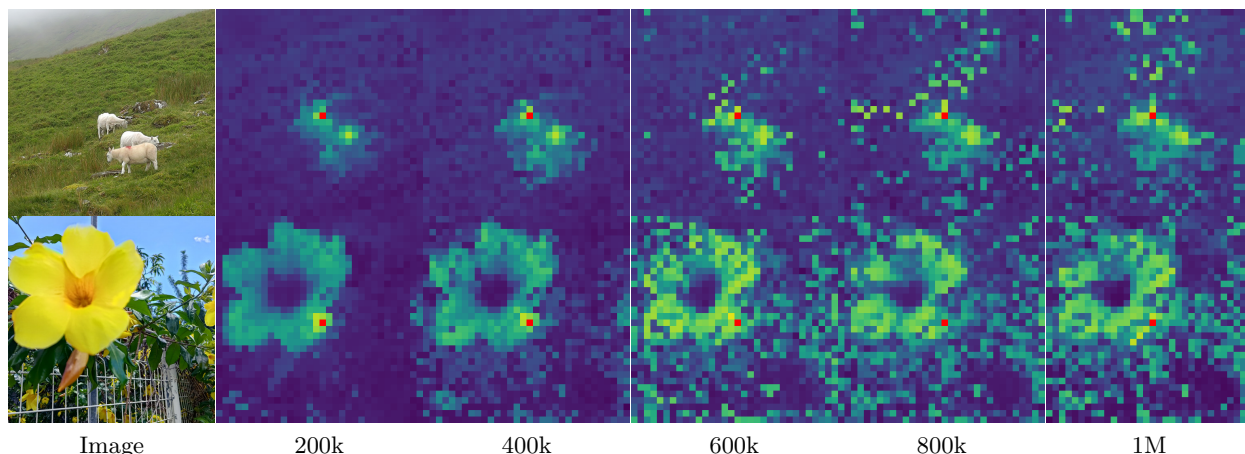


Figure 6: 红色标记区域与所有其他区域的余弦相似度的演变。随着训练的进展，模型产生的特征变得不那么局部化，相似性图也变得更杂乱。

为了更好地理解这种退化，我们通过可视化各个特征块之间的余弦相似性来分析特征块的质量。Fig. 6 显示了主干网络输出的特征块与一个参考特征块（以红色突出显示）之间的余弦相似性图。在 20 万次迭代时，相似性图是平滑且定位良好的，表明特征块级别的表示是一致的。然而，到 60 万次迭代及更长时间后，图像明显退化，出现了更多与参考特征块高度相似但无关的特征块。这种特征块一致性的丧失与密集任务性能的下降有关。

这些 patch 级别的不规则性与 Darcet et al. (2024) 中描述的高范数 patch 异常值不同。具体来说，随着寄存器标记的整合，patch 范数在整个训练过程中保持稳定。然而，我们注意到，在训练过程中 CLS 标记与 patch 输出之间的余弦相似度逐渐增加。这是预期中的，但这意味着 patch 特征的局部性减弱。我们在 Fig. 5a 中可视化了这一现象，其中展示了在 20 万和 100 万次迭代时的余弦图。在密集任务上，为了减轻下降的趋势，我们提出了一个新的目标，专门用于对 patch 特征进行正则化并确保良好的 patch 级别一致性，同时保持高水平的整体性能。

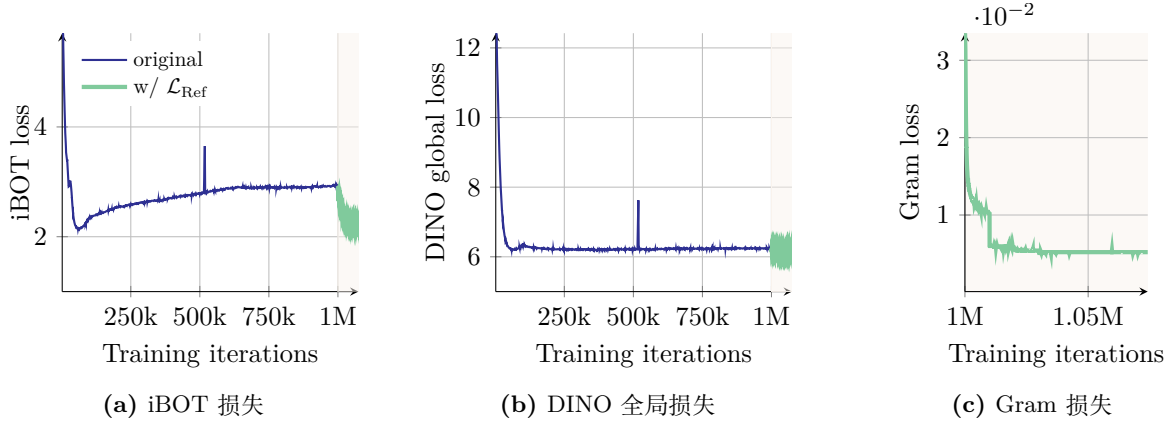


Figure 7: 随着训练迭代的进行，patch-level iBOT 损失、应用于全局裁剪的全局 DINO 损失以及新引入的 Gram 损失的演变。我们突出显示使用 Gram 目标的细化步骤 $\mathcal{L}_{\text{Ref}}$ 的迭代。

4.2 Gram Anchoring 目标

在我们的实验中，我们发现学习强判别特征与保持局部一致性之间具有相对独立性，这可以从全局和密集性能之间缺乏相关性中观察到。尽管将全局 DINO 损失与局部 iBOT 损失结合在一起已经开始解决这个问题，但我们观察到平衡是不稳定的，随着训练的进行，全局表示占据主导地位。基于这一见解，我们提出了一种新颖的解决方案，明确利用这种独立性。

我们引入了一个新的目标，通过强化分块级一致性的质量来缓解分块级一致性退化，而不影响特征本身。这个新的损失函数作用于 Gram 矩阵：图像中所有分块特征的成对点积的矩阵。我们希望将学生的 Gram 矩阵推向一个更早期模型的 Gram 矩阵，这个早期模型称为 Gram 教师。我们通过选择教师网络的早期迭代来选择 Gram 教师，该迭代具有出色的密集属性。通过作用于 Gram 矩阵而不是特征本身，只要相似性的结构保持不变，局部特征可以自由移动。假设我们有一个由 P 块组成的图像以及一个在 d 维度上操作的网络。我们记 $\mathbf{X}_S$ （分别为 $\mathbf{X}_G$ ）为学生（分别为 Gram 教师）的 $P \times d$ 矩阵中进行 L_2 规范化的局部特征。我们将损失 $\mathcal{L}_{\text{Gram}}$ 定义如下：

$$\mathcal{L}_{\text{Gram}} = \left\| \mathbf{X}_S \cdot \mathbf{X}_S^\top - \mathbf{X}_G \cdot \mathbf{X}_G^\top \right\|_F^2. \quad (1)$$

我们仅在全局剪裁上计算此损失。尽管它可以在训练早期应用，但为提高效率，我们仅在 1 次迭代后开始。有趣的是，我们观察到 $\mathcal{L}_{\text{Gram}}$ 的后期应用仍然能够“修复”非常退化的局部特征。为了进一步提高性能，我们每 10k 次迭代更新 Gram 教师，此时 Gram 教师变得与主要的 EMA 教师相同。我们称训练的第二步为精化步骤，它优化目标 $\mathcal{L}_{\text{Ref}}$ ，伴随

$$\mathcal{L}_{\text{Ref}} = w_D \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{\text{DK}} \mathcal{L}_{\text{DKoleo}} + w_{\text{Gram}} \mathcal{L}_{\text{Gram}}. \quad (2)$$

。我们在 Fig. 7 中可视化了不同损失的演变，并观察到应用 Gram 目标显著影响了 iBOT 损失，导致其更快地减少。这表明稳定的 Gram 教师引入的稳定性积极影响了 iBOT 目标。相比之下，Gram 目标对 DINO 损失没有显著影响。这一观察意味着 Gram 和 iBOT 目标以类似的方式影响特征，而 DINO 损失则以不同的方式影响特征。

关于性能，我们观察到新损失的影响几乎是立竿见影的。如 Fig. 8 所示，结合 Gram anchoring 在前 10k 次迭代内对密集任务带来了显著的改进。在 Gram 教师更新后，我们还在 ADE20k 基准测试中看到显著的提升。此外，更长时间的训练进一步改善了 ObjectNet 基准测试的性能，其他全球基准测试显示新损失的影响较小。

4.3 利用更高分辨率的特征

最近的工作表明，通过对补丁特征进行加权平均，能够通过平滑异常补丁和增强补丁级一致性来获得更强的局部表示 (Wysoczańska et al., 2024)。另一方面，将更高分辨率的图像输入主干，能够产生更精细和详细

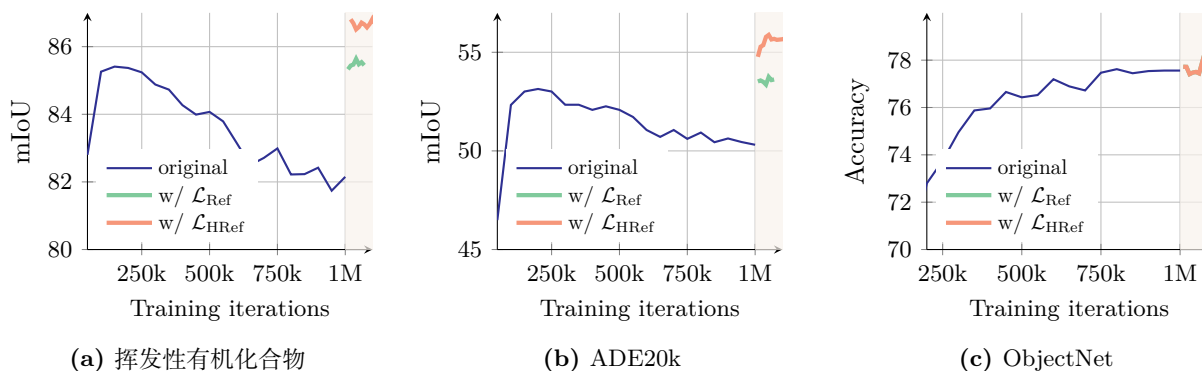


Figure 8: 应用我们提出的 Gram anchoring 方法后在不同基准上的结果演变。我们可视化了在使用我们的细化步骤继续原始训练时的结果，记为 ' $\mathcal{L}_{\text{Ref}}$ '。我们还绘制了使用更高分辨率特征用于 Gram 目标时获得的结果，如后续 Sec. 4.3 中所介绍，并记为 ' $\mathcal{L}_{\text{HRef}}$ '。我们突出使用 Gram 目标的迭代。

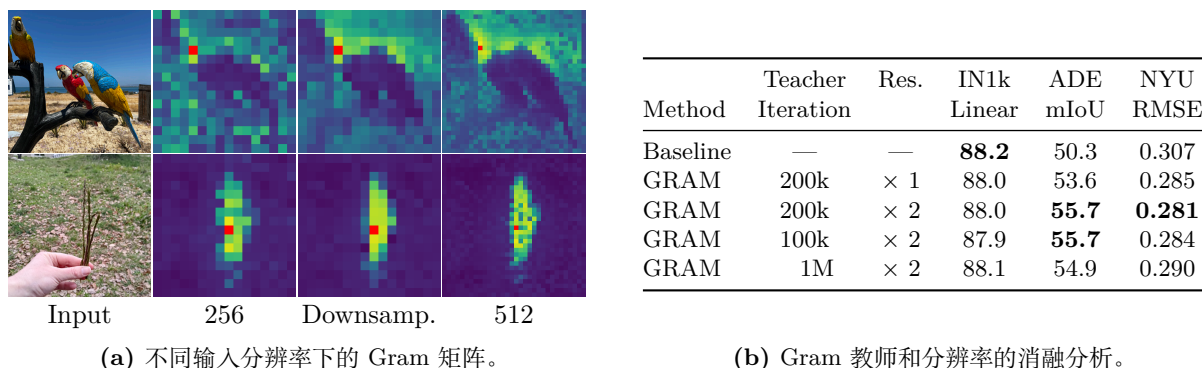


Figure 9: 高分辨率 Gram 影响的定量和定性研究。我们展示了 (a) 将高分辨率地图下采样到较小地图后的改进余弦图，以及 (b) 通过改变训练迭代次数和 Gram 教师的分辨率带来的定量改进。

的特征图。我们利用这两种观察的优点来计算 Gram 老师的高质量特征。具体来说，我们首先将图像以正常分辨率的两倍输入到 Gram 老师，然后用双三次插值对生成的特征图进行下采样，以达到与学生输出大小匹配的平滑特征图。Fig. 9a 可视化了在 256 和 512 分辨率下获取的补丁特征的 Gram 矩阵，以及那些在 512 分辨率下经过 $2\times$ 下采样后获得的特征（标记为“downsamp.”）。我们观察到，较高分辨率特征中优越的补丁级一致性在下采样过程中得以保留，从而形成更平滑和一致的补丁级表示。作为补充说明，由于采用了由 Su et al. (2024) 引入的旋转位置嵌入 (RoPE)，我们的模型能够无缝处理不同分辨率的图像，而不需要进行适应。

我们计算下采样特征的 Gram 矩阵，并使用它来替换目标 $\mathbf{X}_G$ 中的 $\mathcal{L}_{\text{Gram}}$ 。我们将新的优化目标记为 $\mathcal{L}_{\text{HRef}}$ 。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的补丁一致性提炼到学生模型中。如 Fig. 8 和 Fig. 9b 所示，这种提炼转化为在密集任务上的更好预测，带来了在 $\mathcal{L}_{\text{Ref}}$ 带来的收益之上的额外增益（在 ADE20k 上增加 2 mIoU）。我们还在 Fig. 9b 中对 Gram 教师的选择进行消融实验。有趣的是，从 100k 或 200k 选择 Gram 教师对结果没有显著影响，但使用更晚期的 Gram 教师（100 万次迭代）是有害的，因为这种教师的补丁级一致性较差。

最后，我们定性地说明了 Fig. 10 中的 Gram 锚定对补丁级一致性的影响，其中可视化了通过初始训练和高分辨率 Gram anchoring 精细化获得的 Gram 矩阵补丁特征。我们观察到高分辨率精细化过程在特征关联性方面带来了巨大的改进。

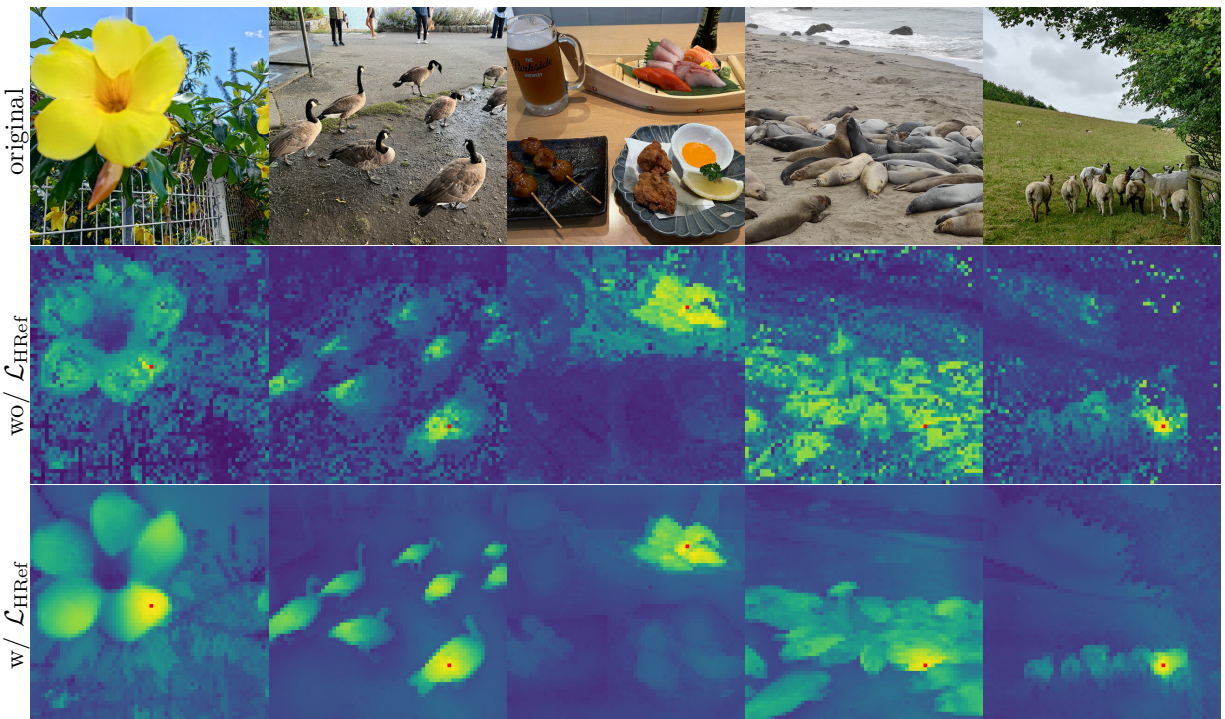


Figure 10: Gram anchoring 的定性效果。我们在使用细化目标 $\mathcal{L}_{HRef}$ 前后可视化余弦图。图像的输入分辨率为 1024×1024 像素。

5 训练后

本节介绍训练后的阶段。这包括高分辨率适应阶段，使得在不同输入分辨率（Sec. 5.1）下能够进行有效的推理，模型蒸馏产生高质量和高效的小尺寸模型（Sec. 5.2），以及文本对齐为 DINOv3（Sec. 5.3）增加零样本能力。

5.1 分辨率缩放

我们在相对较小的分辨率 256 下训练模型，这为我们在速度和效果之间提供了良好的折衷。对于补丁大小为 16 的情况，该设置与使用分辨率 224 和补丁大小为 14 训练的 DINOv2 具有相同的输入序列长度。然而，许多当代计算机视觉应用需要以显著更高的分辨率（通常为 512×512 像素或更大）来处理图像，以捕获复杂的空间信息。推理图像的分辨率在实际中也不是固定的，会根据特定的使用情况而有所变化。为了解决这个问题，我们通过一个高分辨率适配步骤（Touvron et al., 2019）来扩展我们的训练方案。为了确保在不同分辨率范围内的高性能，我们利用混合分辨率，在每个小批量中采样不同比例的全局和局部裁剪。具体来说，我们考虑来自 $\{512, 768\}$ 的全局裁剪大小和来自 $\{112, 168, 224, 336\}$ 的局部裁剪大小，并额外训练模型 1 万次迭代。

类似于主要训练，这一高分辨率适应阶段的一个关键组成部分是添加 Gram anchoring，使用 7B 教师作为 Gram 教师。我们发现这一组成部分是必不可少的：没有它，模型在密集预测任务上的性能会显著下降。Gram anchoring 鼓励模型在空间位置上保持一致且稳健的特征相关性，这在处理高分辨率输入的增加复杂性时至关重要。

根据经验，我们观察到这个相对简短但针对性高的高分辨率步骤显著提升了整体模型的质量，并使其能够在各种输入尺寸上进行泛化，正如在 Fig. 4 中视觉上展示的那样。在 Fig. 11 中，我们比较了我们的 7B 模型在适应前后的表现。我们发现，分辨率缩放在 ImageNet 分类 (a) 上带来了小幅提升，性能相对于分辨率来说相对稳定。然而，在 ObjectNet OOD 转移 (b) 中，我们观察到低分辨率时性能略有下降，而高分辨率时则有所改善。这在很大程度上被通过高分辨率下本地特征质量的提高所弥补了，正如在 ADE20k 上的分割 (c) 和在 DAVIS 上的跟踪 (d) 中的正向趋势显示的那样。适应后，本地特征随着图像大小的增加而改善，

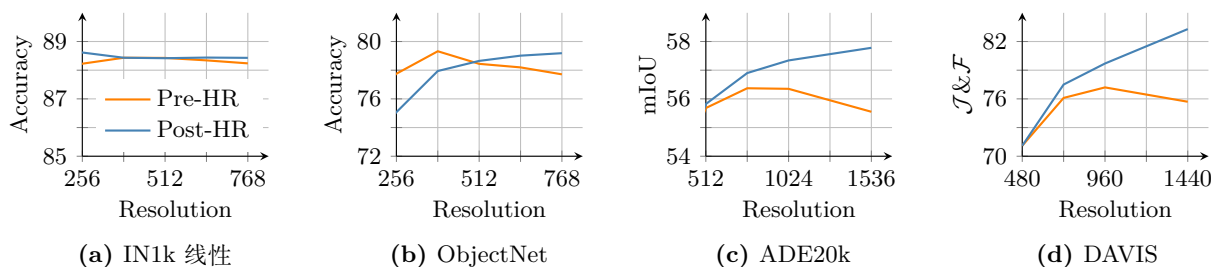


Figure 11: 高分辨率适应的效果。在不同评估分辨率下, (a) ImageNet 上的线性分类, (b) 应用于 ObjectNet 上的 OOD, (c) ADE20k 上的线性语义分割, 以及 (d) DAVIS 上的分割跟踪的分辨率缩放 (Sec. 5.1) 前 (“Pre-HR”) 和后 (“Post-HR”) 的结果。

利用在更高分辨率下可获得的更丰富的空间信息, 有效实现高分辨率推理。有趣的是, 适应后的模型支持远超出最大训练分辨率 768 的更高分辨率——我们在高于 4k 的分辨率下视觉上观察到稳定的特征图 (c.f. Fig. 4)。

5.2 模型蒸馏

一个适用于多种用例的模型家族 我们将 ViT-7B 模型的知识蒸馏到更小的视觉 Transformer 变体 (ViT-S、ViT-B 和 ViT-L), 这些变体因其改进的可管理性和效率而受到社区的高度重视。我们的蒸馏方法使用与第一训练阶段相同的训练目标, 确保学习信号的一致性。然而, 我们直接使用 7B 模型作为教师来指导较小的学生模型, 而不是依赖于模型权重的指数移动平均数 (EMA)。在这种情况下, 教师模型是固定的。我们没有观察到补丁级别的一致性, 因此没有应用 Gram anchoring 技术。该策略使得蒸馏模型能够继承大型教师模型的丰富表示能力, 同时在部署和实验中更具实用性。

我们的 ViT-7B 模型被蒸馏成一系列 ViT 模型, 其规模覆盖了广泛的计算预算, 并允许与同时期的模型进行适当比较。它们包括标准的 ViT-S (21M 参数)、B (86M)、L (0.3B), 还有定制的 ViT-S+ (29M) 和定制的 ViT-H+ (0.8B) 模型, 以缩小与自蒸馏 7B 教师模型的性能差距。实际上, 我们在 DINOv2 中观察到, 较小的学生模型在蒸馏过程中可以达到与其教师相当的性能。因此, 蒸馏模型在部分推理计算消耗中表现出前沿水平的性能, 如我们在 Tab. 14 中所见。我们训练模型 1M 次迭代, 然后在应用上述不含 Gram anchoring 的高分辨率阶段之前, 根据余弦计划进行 250k 次学习率降温。

由于大型教师模型的推理成本可能远高于学生模型 (见 Fig. 16a), 我们设计了一个并行蒸馏流水线, 该流水线可以同时训练多个学生, 并在所有参与训练的节点之间共享教师模型的推理 (见 Fig. 12 的图示)。设 C_T 和 C_S 分别为单样本上的教师推理成本和学生训练成本, 在单教师/单学生蒸馏中, 批量大小为 B , 其中每个 N 个 GPU 处理数据的 B/N 部分, 教师推理成本为每个 GPU 的 $B/N \times C_T$, 学生训练成本为每个 GPU 的 $B/N \times C_S$ 。在多学生蒸馏中, 我们执行如下步骤。每个学生 S_i 被分配一组 N_{Si} 个 GPU 进行训练, 所有 $N_T = \sum N_{Si}$ 个 GPU 都属于全局推理组。在每次迭代中, 我们首先在全局组上运行教师推理, 每个 GPU 的计算成本为 $B/N_T \times C_T$ 。然后, 我们运行一个 all-gather 合集操作, 以与所有计算节点共享输入数据和推理结果。最后, 每个学生组分别进行学生训练, 成本为 $B/N_{Si} \times C_{Si}$ 。

上述计算表明, 向蒸馏管道中添加额外的学生将会 (1) 减少每次迭代时每个 GPU 的计算量, 从而整体上提高蒸馏速度, (2) 总计算量仅增加新学生的训练成本, 因为总教师推理成本现在已固定。实现这一点只需仔细设置 GPU 进程组, 调整数据加载器和教师推理, 确保输入和输出在组间使用 NCCL 集体操作保持同步。由于各组在每次迭代中都同步, 我们调整每个学生的 GPU 数量使其迭代时间大致相同, 以最大化速度。通过这一过程, 我们可以无缝地训练多个学生, 并从我们的旗舰 7B 模型中产生一整系列的蒸馏模型。

5.3 将 DINOv3 与文本对齐

开放词汇图像-文本对齐因其促进灵活可扩展的多模态理解的潜力, 而受到了研究界的极大兴趣和热情。大量研究工作集中在提高 CLIP (Radford et al., 2021) 的质量, CLIP 最初仅学习了图像和文本表示之间的全局对齐。尽管 CLIP 展现了令人印象深刻的零样本能力, 但其对全局特征的关注限制了其捕捉精细化和局部对应关系的能力。最近的研究 (Zhai et al., 2022b) 表明, 通过预训练的自监督视觉骨干网可以实现有效的图

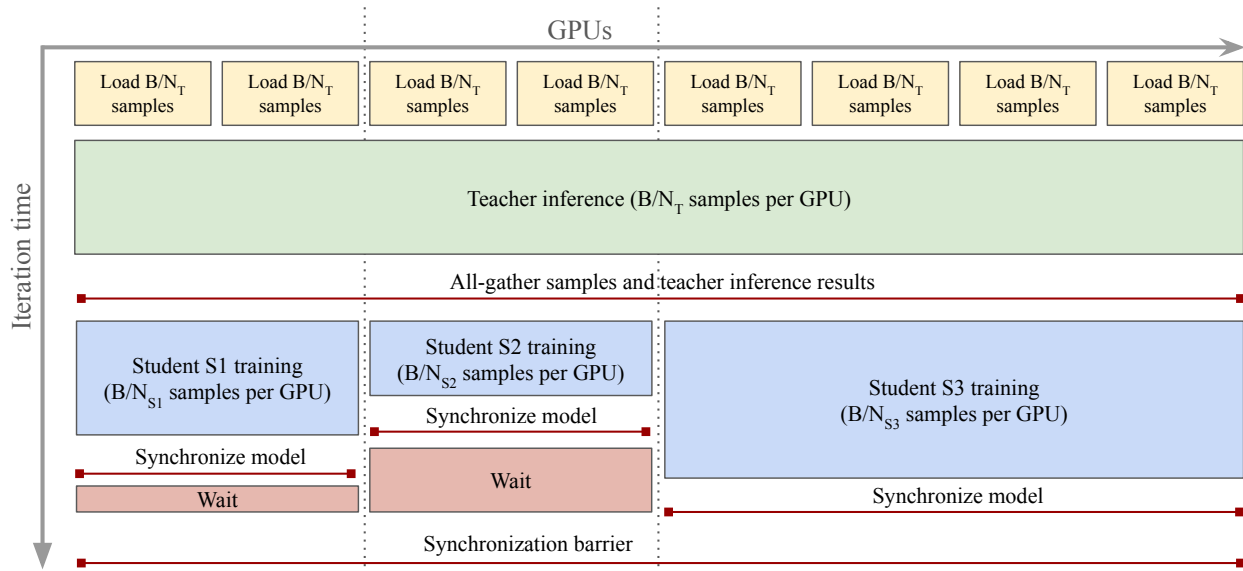


Figure 12: 多学生蒸馏过程。在此图中，我们并行蒸馏 3 个学生：我们首先通过所有 T 节点共享教师推理以节省计算，并在所有 GPU 上收集输入和结果。然后，较小的组执行学生训练。我们调整这些组的大小，使训练步骤在所有学生 S_i 上的持续时间相同，从而最大限度地减少在同步屏障处等待的空闲时间。

像-文本对齐。这使得将这些强大的模型应用于多模态环境中成为可能，促进了更丰富和更精确的文本到图像的关联，超越了全局语义，同时也降低了计算成本，因为视觉编码已经被学习。

我们通过采用先前在 [Jose et al. \(2025\)](#) 中提出的训练策略来将文本编码器与我们的 DINOv3 模型对齐。这种方法遵循 LiT 训练范式 ([Zhai et al., 2022b](#))，从头开始训练一个文本表示，以对比目标将图像与其标题匹配，同时保持视觉编码器冻结。为了在视觉方面提供一些灵活性，在冻结的视觉主干上引入了两个 transformer 层。该方法的一个关键增强是将均值池化的补丁嵌入与输出的 CLS 标记在匹配文本嵌入之前进行连接。这使得将全局和局部视觉特征与文本对齐成为可能，从而在不需要额外的启发式或技巧的情况下提高密集预测任务的性能。此外，我们使用与 [Jose et al. \(2025\)](#) 中建立的相同的数据整理协议以确保一致性和可比性。

6 结果

在本节中，我们在多种计算机视觉任务上评估了我们的旗舰模型 DINOv3 7B。在整个实验中，除非另有说明，我们始终保持 DINOv3 冻结，仅使用其表示。我们证明，使用 DINOv3，无需微调即能获得强大的性能。本节组织如下。我们首先使用轻量级评估协议来探测 DINOv3 的密集 ([Sec. 6.1](#)) 和全局 ([Sec. 6.2](#)) 图像表示的质量，并将其与最强的现有视觉编码器进行比较。我们展示了 DINOv3 学习到的密集特征异常优秀，同时提供强大且多功能的全球图像表示。然后，我们考虑将 DINOv3 作为开发更复杂计算机视觉系统的基础 ([Sec. 6.3](#))。我们展示了在 DINOv3 之上只需付出很少的努力，就能在物体检测、语义分割、3D 视图估计或相对单目深度估计等多样化任务中取得与当前最先进技术相媲美或更优的结果。

6.1 DINOv3 提供卓越的密集特征

我们首先使用一组多样的轻量级评估来研究 DINOv3 密集表示的原始质量。在所有情况下，我们利用最后一层的冻结补丁特征，并用以下方式进行评估：(1) 质量可视化 ([??](#))，(2) 密集线性探测 ([Sec. 6.1.1](#)：分割、深度估计)，(3) 无参数方法 ([Sec. 6.1.2](#)：3D 对应性估计，[Sec. 6.1.3](#)：对象发现，[Sec. 6.1.4](#)：跟踪)，以及 (4) 轻量级注意力探测 ([Sec. 6.1.5](#)：视频分类)。

我们比较了 DINOv3 的密集特征与最强的公开可用图像编码器的密集特征，包括弱监督和自监督的编码器。我们考虑弱监督编码器 Perception Encoder (PE) Core ([Bolya et al., 2025](#)) 和 SigLIP 2 ([Tschannen et al., 2025](#))，它们使用 CLIP 风格的图像文本对比学习。我们还比较了最强的自监督方法：DINOv3 的

Table 3: 利用冻结的骨干网络进行语义分割和单目深度估计的密集线性探测结果。我们报告了分割基准 ADE20k、Cityscapes 和 VOC 的数据平均交并比 (mIoU) 指标。我们为深度基准 NYUv2 和 KITTI 报告均方根误差 (RMSE) 指标。对于分割, 所有模型都在输入分辨率适应到 1024 个 patch 标记的情况下评估 (对于 patch 尺寸为 14 的情况下为 *i.e.* 448×448 , 对于 patch 尺寸为 16 的情况下为 512×512)。

Method	ViT	Segmentation			Depth	
		ADE20k	Citysc.	VOC	NYUv2 ↓	KITTI ↓
Agglomerative backbones						
AM-RADIOv2.5	g/14	53.0	78.4	85.4	0.340	2.918
PEspatial	G/14	49.3	73.2	82.7	0.362	3.082
Weakly-supervised backbones						
SigLIP 2	g/16	42.7	64.8	72.7	0.494	3.273
PEcore	G/14	38.9	61.1	69.2	0.590	4.119
Self-supervised backbones						
Franca	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
Web-DINO	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

前身 DINOv2 (Oquab et al., 2024) 带注册 (Darcet et al., 2024)、Web-DINO (Fan et al., 2025), 这是 DINO 最近的扩展尝试, 以及 Franca (Venkataramanan et al., 2025), 作为最佳开放数据 SSL 模型。最后, 我们比较了聚合模型 AM-RADIOv2.5 (Heinrich et al., 2025), 这是从 DINOv2 提炼出来的, 以及 CLIP (Radford et al., 2021)、DFN (Fang et al., 2024a)、和 Segment Anything (SAM) (Kirillov et al., 2023), 以及 PEspatial, 将 SAM 2 (Ravi et al., 2025) 提炼到 PEcore 中。对于每个基线, 我们报告了最强模型可用的性能, 并在表格中指定了架构。

我们首先对 DINOv3 的密集特征图进行定性分析。为此, 我们使用主成分分析 (PCA) 将密集特征空间投影到 3 维, 并将结果 3D 空间映射到 RGB。由于 PCA 中的符号不确定性 (八种变体) 和主成分与颜色之间的任意映射 (六种变体), 我们探索所有组合并报告视觉上最引人注目的一个。生成的可视化如 Fig. 13 所示。与其他视觉骨干网相比, 可以看出 DINOv3 的特征更为清晰, 噪音明显减少, 并显示出更优的语义一致性。

6.1.1 密集线性探测

我们在稠密特征之上执行线性探测, 用于两个任务: 语义分割和单目深度估计。在这两种情况下, 我们在 DINOv3 冻结的 patch 输出之上训练一个线性变换。对于语义分割, 我们在 ADE20k (Zhou et al., 2017)、Cityscapes (Cordts et al., 2016) 和 PASCAL VOC 2012 (Everingham et al., 2012) 数据集上进行评估并报告平均交并比 (mIoU) 指标。对于深度估计, 我们使用 NYUv2 (Silberman et al., 2012) 和 KITTI (Geiger et al., 2013) 数据集, 并报告均方根误差 (RMSE)。

结果 (Tab. 3) 分割结果展示了我们密集特征的卓越质量。在通用的 ADE20k 数据集上, DINOv3 比自监督基线高出超过 6 个 mIoU 点, 并比弱监督基线高出超过 13 个点。此外, DINOv3 比 PEspatial 高出超过 6 个点, 并比 AM-RADIOv2.5 高出接近 3 个点。这些结果是显著的, 因为两者都是从高度监督的分割模型 SAM (Kirillov et al., 2023) 蒸馏出来的强基线。在自动驾驶基准 Cityscapes 上也观察到类似的结果, DINOv3 实现了最佳的 mIoU 81.1, 超过 AM-RADIOv2.5 2.5 个点, 并且至少超过所有其他骨干 5.5 个点。

在单目深度估计方面, DINOv3 再次以显著的差距超越了所有其他模型: 弱监督模型 PEcore 和 SigLIP 2 仍然落后, 而 DINOv2 和从 SAM 派生的更先进的模型是最接近的竞争对手。有趣的是, 尽管 PEspatial 和 AM-RADIO 在 NYU 上表现出强劲的性能, 但它们在 KITTI 上的表现低于 DINOv2。即使如此, DINOv3 在 RMSE 上也比其前身 DINOv2 高出 0.278。

两组评估均显示出 DINOv3 的密集特征的卓越表示能力, 并反映了来自 Fig. 13 的视觉结果。仅用一个线性预测器, DINOv3 就能稳健地预测对象类别和掩码, 以及场景的物理测量, 如相对深度。这些结果表明, 这些特征不仅在视觉上清晰且定位准确, 还以线性可分的方式表示了基础观测中的许多重要属性。最后, 在

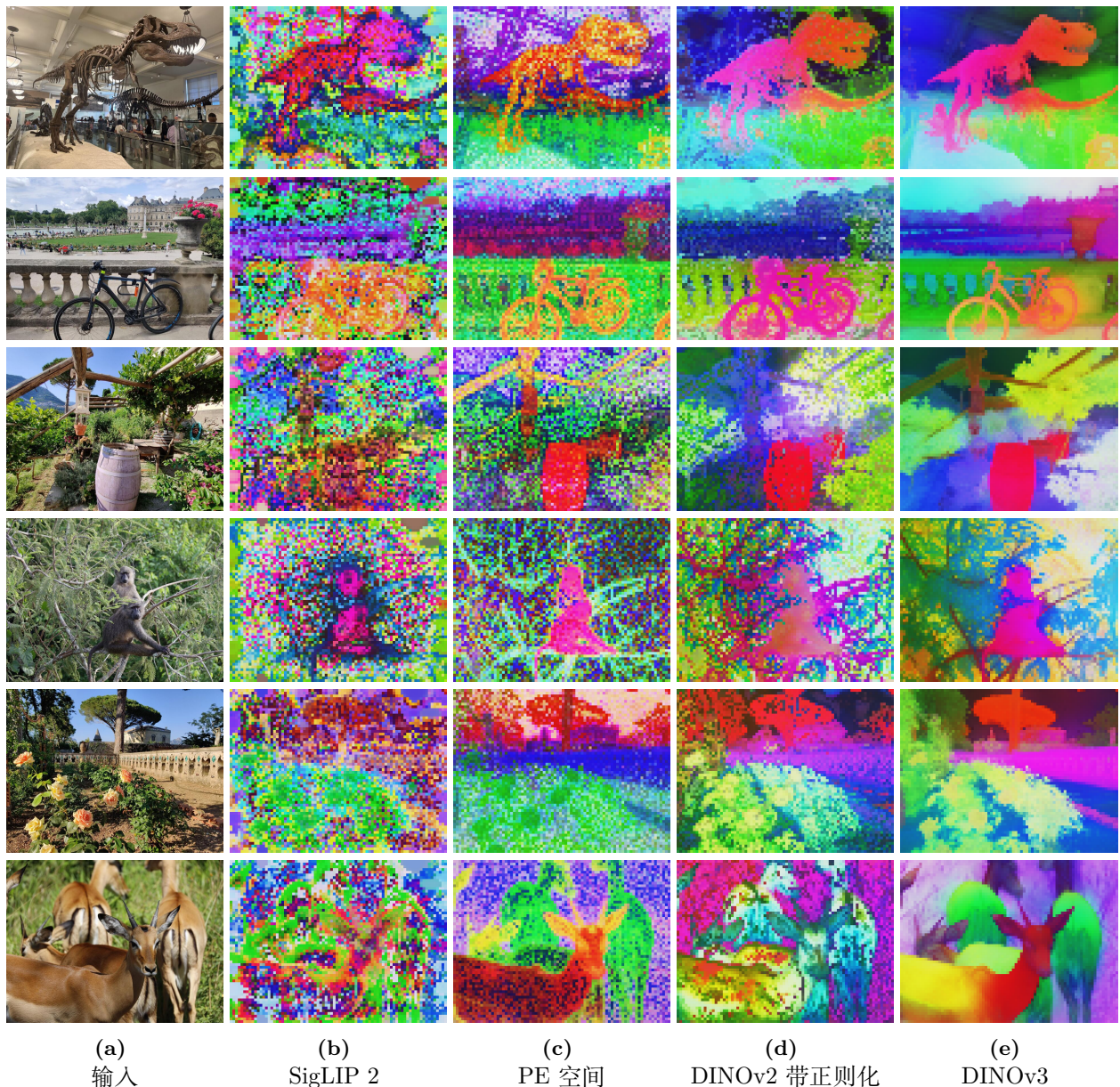


Figure 13: 密集特征比较。我们通过使用 PCA 对它们的密集输出进行投影，并将其映射到 RGB，来比较几种视觉骨干网络。从左到右：SigLIP 2 ViT-g/16，PEspatial ViT-G/14，带寄存器的 DINOv2 ViT-g/14，DINOv3 ViT-7B/16。对于使用补丁 16 的模型，图像以分辨率 1280×960 传递；对于补丁 14，分辨率为 1120×840 。所有特征图的尺寸均为 80×60 。

ADE20k 上使用线性分类器获得的绝对性能 (55.9 mIoU) 本身就令人印象深刻，因为它与该数据集上的绝对最先进水平 (63.0 mIoU) 相差不远。

6.1.2 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。最近，图像基础模型通过提供对三维有感知特征，推动了三维理解的研究。在本节中，我们根据 Probe3D 中定义的协议评估 DINOv3 的多视图一致性，即一个物体在不同视图中同一关键点的特征是否相似。我们区分几何对应估计和语义对应估计。前者指匹配同一物体实例的关键点，而后者指匹配同一物体类别不同实例的关键点。我们在 NAVI 数据集上评估几何对应，在 SPair 数据集上评估语义对应，并在两种情况下用对应召回率评估性能。更多实验详情请参见 App. D.3。

Table 4: 稠密表示的三维一致性评估。我们遵循 Probe3D (Banani et al., 2024) 的评估协议来估计跨视图的三维关键点对应关系。为了衡量性能，我们报告了对应关系的召回率 *i.e.*，即落入指定距离内的对应关系的百分比。

		Geometric	Semantic
Method	ViT	NAVI	SPair
Agglomerative backbones			
AM-RADIOv2.5	g/14	59.4	56.8
PEspatial	G/14	53.8	49.6
Weakly-supervised backbones			
SigLIP 2	g/16	49.4	42.6
PEcore	G/14	39.9	23.1
Self-supervised backbones			
Franca	g/14	54.6	51.0
DINOv2	g/14	60.1	56.1
Web-DINO	7B/14	55.0	32.2
DINOv3	7B/16	64.4	58.7

对于几何对应，DINOv3 优于所有其他模型，并比第二好的模型（DINOv2）提高了 4.3 % 查准率。其他自监督学习扩展努力（Franca 和 WebSSL）落后于 DINOv2，这表明 DINOv2 仍然是一个强有力的基线。弱监督模型（PEcore 和 SigLIP 2）在此任务中表现不佳，表明对 3D 感知的缺乏。对于具有 SAM 蒸馏的模型，AM-RADIO 几乎达到 DINOv2 的性能，但 PEspatial 仍然落后于它（-11.6 % 查准率），甚至落后于 Franca（-0.8 % 查准率）。这表明自监督学习是此任务中取得强大性能的关键组件。对于语义对应，同样的结论适用。DINOv3 表现最佳，超过了其前任（+2.6 % 查准率）和 AM-RADIO（+1.9 % 查准率）。总体而言，这些在关键点匹配上的出色性能是 DINOv3 在其他 3D 密集型应用中下游使用的非常有希望信号。

6.1.3 无监督目标发现

强大的自监督特征可以在不需要任何标注的情况下促进图像中的对象实例发现 (Vo et al., 2021; Siméoni et al., 2021; Seitzer et al., 2023; Wang et al., 2023c; Siméoni et al., 2025)。我们通过无监督对象发现任务测试不同视觉编码器的这种能力，该任务要求图像中对象的无类别分割 (Russell et al., 2006; Tuytelaars et al., 2010; Cho et al., 2015; Vo et al., 2019)。特别地，我们使用基于图的非参数 TokenCut 算法 (Wang et al., 2023c)，该算法在多种基础模型上表现出了强大的性能。我们在三个广泛使用的数据集上运行了它：VOC 2007, VOC 2012 (Everingham et al., 2015)，和 COCO-20k (Lin et al., 2014; Vo et al., 2020)。我们遵循由 Siméoni et al. (2021) 定义的评估协议，并报告 CorLoc 指标。为了正确比较具有不同特征分布的基础模型，我们对主要的 TokenCut 超参数进行搜索，即在构建用于划分的补丁图时应用的余弦相似度阈值。最初，最好的对象发现结果是使用最后一个注意力层的键 DINO (Caron et al., 2021) 获得的。然而，这种人为选择并不能一致地推广到其他基础模型。为了简化，我们始终对所有模型采用输出特征。

结果 (Fig. 14) 初始版的 DINO 为这一任务设立了一个非常高的标准。有趣的是，虽然 DINOv2 在像素级密集任务中表现非常强劲，但在物体发现方面却失败了。这部分可以归因于密集特征中存在的伪影 (*c.f.* Fig. 13)。DINOv3 拥有干净而精准的输出特征图，其表现优于其前辈，在 VOC 2007 上 CorLoc 提升了 5.9，并且优于其他所有的骨干网络，无论是自监督、弱监督还是聚合方法。这一评估证明了 DINOv3 的密集特征在语义上是强大的并且定位良好。我们相信，这将为更多无类别限制的目标检测方法铺平道路，特别是在标注代价高昂或不可用、相关类别集不被限制于预定义子集的情境下。

6.1.4 视频分割跟踪

除了静态图像，视觉表示的一个重要属性是其时间一致性，*i.e.* 即特征随时间稳定演变的能力。为了测试这一特性，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 et al., 2017)、YouTube-VOS (Xu et al., 2018) 和 MOSE (Ding et al., 2023) 数据集。我们使用标准的 $\mathcal{J}$ & $\mathcal{F}$ -mean 评价指标来评估性能，该指标结合了区域相似度 ($\mathcal{J}$) 和轮廓准确度 ($\mathcal{F}$) (Perazzi et al., 2016)。根据 Jabri et al. (2020)，我们使用一种非参数

Method	ViT	VOC07	VOC12	COCO
Agglomerative backbones				
AM-RADIOv2.5	g/14	55.0	59.7	45.9
PEspatial	G/14	51.2	56.0	43.9
Weakly-supervised backbones				
SigLIPv2	g/16	20.5	24.7	18.6
PEcore	G/14	14.2	18.2	13.5
Self-supervised backbones				
DINO	S/16	61.1	66.0	48.7
DINO	B/16	60.1	64.4	50.5
DINOv2	g/14	55.6	60.4	45.4
Web-DINO	7B/14	26.1	29.7	20.9
DINOv3	7B/16	66.1	69.5	55.1



Figure 14: 无监督物体发现。我们在不同主干网络的输出补丁特征上应用 TokenCut (Wang et al., 2022c) 并报告 CorLoc 指标。我们还展示了通过 DINOv3 获得的预测掩码 (在分辨率 1024 的输入图像上以红色叠加显示), 这些掩码是在没有任何标注和后处理的情况下获得的。

Table 5: 视频分割跟踪评估。我们在不同分辨率下报告了 DAVIS、YouTube-VOS 和 MOSE 上的 $\mathcal{J}\&\mathcal{F}$ - 平均值。对于补丁大小为 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。

Method	ViT	DAVIS			YouTube-VOS			MOSE		
		S	M	L	S	M	L	S	M	L
Agglomerative backbones										
AM-RADIOv2.5	g/14	66.5	77.3	81.4	70.1	78.1	79.2	44.0	52.6	54.3
PEspatial	G/14	68.4	74.5	70.5	68.5	67.5	55.6	39.3	40.2	34.0
Weakly-supervised backbones										
SigLIP 2	g/16	56.1	62.3	62.9	52.0	57.3	55.1	28.0	30.3	29.2
PEcore	G/14	48.2	53.1	49.8	34.7	33.0	25.3	17.8	19.0	15.4
Self-supervised backbones										
Franca	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
Web-DINO	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

的标签传播算法, 该算法考虑跨帧的特征块相似性。我们在三种输入分辨率下进行评估, 使用短边长度为 420/480 (S)、840/960 (M) 和 1260/1440 (L) 像素的模型, 块大小为 14/16 (匹配块标记的数量)。 $\mathcal{J}\&\mathcal{F}$ 分数始终在视频的原始分辨率下计算。有关更详细的实验设置, 请参见 App. D.5。

与之前所有结果一致, 弱监督骨干网络没有提供令人信服的表现。PEspatial 从视频模型 SAMv2 中提取, 表现令人满意, 在较小的分辨率上优于 DINOv2, 但在较大分辨率上则不及。跨分辨率来看, DINOv3 优于所有竞争者, 在 DAVIS-L 上表现出惊人的 83.3 $\mathcal{J}\&\mathcal{F}$, 比 DINOv2 高出 6.7 点。此外, 性能作为分辨率的函数呈现良好的趋势, 证实了我们的模型能够利用更多的输入像素输出精确的高分辨率特征图 (c.f. Figs. 3 and 4)。相对而言, 对于较高分辨率, SigLIP 2 和 PEcore 的性能几乎保持不变, 而 PEspatial 的性能则有所下降。有趣的是, 我们的图像模型无需任何视频上的微调, 能够在时间上正确跟踪对象 (见 Fig. 15)。这使得它成为嵌入视频的优秀候选者, 能够在此基础上构建强大的视频模型。

6.1.5 视频分类

之前的结果已表明 DINOv3 的表示在低级别时序一致性上良好, 可以准确地跟踪时间中的对象。更进一步, 我们在本节中评估其密集特征用于高级视频分类的适用性。与 V-JEPA 2 (Assran et al., 2025) 的设置类

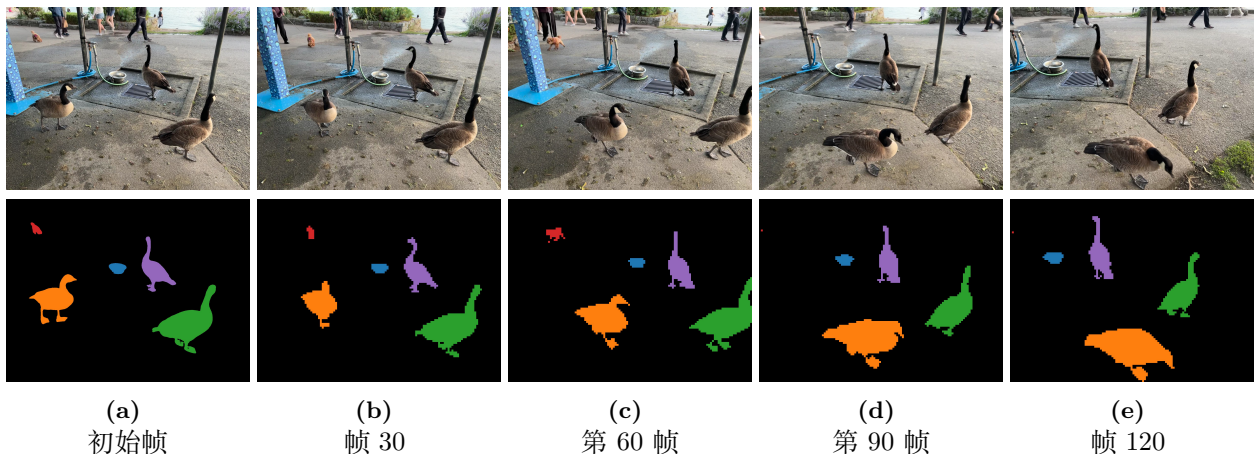


Figure 15: 分割跟踪示例。给定初始帧的真实实例分割掩码，我们根据 DINOv3 特征空间中的图块相似性将实例标签传播到后续帧。输入分辨率为 2048×1536 像素，生成 128×96 个图块。

Table 6: 使用注意力探针进行视频分类评估。我们报告了在 UCF101、Something-Something V2 (SSv2) 和 Kinetics-400 (K400) 上的 top-1 准确率。对于每个模型，我们报告了评估每个视频的单个片段的性能，或通过对多个片段的预测概率取平均来应用测试时增强 (TTA)。

		UCF101		SSv2		K400	
Method	ViT	Single	TTA	Single	TTA	Single	TTA
Agglomerative backbones							
AM-RADIOv2.5	g/14	92.8	92.5	69.1	70.0	84.8	85.2
PEspatial	G/14	92.7	92.8	66.4	68.4	83.5	84.8
Weakly-supervised backbones							
SigLIP 2	g/16	93.6	94.2	68.8	70.2	86.9	87.7
PEcore	G/14	93.1	93.3	69.0	70.4	87.9	88.8
Self-supervised backbones							
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
Web-DINO	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

似，我们在从每帧提取的补丁特征之上训练一个注意力探测器——一个浅层的 4 层基于变压器的分类器。这使得在时空维度上进行推理成为可能，因为这些特征是独立于每一帧提取的。在评估期间，我们要么取每个视频的单个片段，要么通过对每个视频的 3 个空间和 2 个时间裁剪的预测进行平均来使用测试时增强 (TTA)。有关实验细节，请参见 App. D.6。我们在三个数据集上运行此评估：UCF101 (Soomro et al., 2012)，Something-Something V2 (Goyal et al., 2017)，和 Kinetics-400 (Kay et al., 2017)，并报告 top-1 准确率。作为额外的基线，我们报告了 V-JEPA v2 的性能，这是用于视频理解的最先进的自监督学习 (SSL) 模型。

符合前一个实验的结论，我们发现 DINOv3 可以用于提取强大的视频特征。由于这次评估涉及多层自注意力的训练，不同模型之间的差异不那么显著。然而，DINOv3 与 PEcore 和 SigLIP 2 处于同一水平，并且在各个数据集上明显优于其他模型 (DINOv2, AM-RADIO)。UCF101 和 K400 以外观为主，强大的类别级对象理解是性能的主要来源。而 SSv2 则需要对运动有更好的理解——专门的视频模型 V-JEPA v2 在这个数据集上表现出色。有趣的是，在这个数据集上，DINOv3 与弱监督模型之间的差距稍大了一些。这再次证明了 DINOv3 适用于视频任务。

Table 7: 在 ImageNet1k 上用冻结骨干网络训练的线性探测器的分类准确性。弱监督和自监督模型在图像分辨率适配为 1024 个 patch tokens 时进行评估（对于 patch 大小为 14 使用 *i.e.* 448×448 ，对于 patch 大小为 16 使用 512×512 ）。为了参考，我们还列出了使用不同评估协议的 Dehghani et al. (2023) 的结果（用 * 标记）。

Method	ViT	ImageNet			Rendition		Hard		
		Val	V2	ReaL	R	S	A	C ↓	Obj.
Supervised backbones									
Zhai et al. (2022a) *	G/14	89.0	81.3	90.6	91.7	—	78.8	—	69.6
Chen et al. (2023) *	e/14	89.3	82.5	90.7	94.3	—	81.6	—	71.5
Dehghani et al. (2023) *	22B/14	89.5	83.2	90.9	94.3	—	83.8	—	74.3
Agglomerative backbones									
AM-RADIOv2.5	g/14	88.0	80.2	90.3	83.8	67.1	81.3	27.1	68.4
Weakly-supervised backbones									
PEcore	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
Self-supervised backbones									
Web-DINO	7B/14	85.9	77.1	88.6	75.6	64.0	71.6	31.2	69.7
Franca	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

6.2 DINOv3 拥有稳健且多功能的全局图像描述符

在本节中，我们评估 DINOv3 捕捉全局图像统计量的能力。为此，我们考虑使用线性探测器（Sec. 6.2.1）和实例检索基准（Sec. 6.2.2）的经典分类基准。我们同样与最强的公开可用图像编码器进行比较。除了上一节中的模型外，我们还评估了两个弱监督模型 AIMv2（Fini et al., 2024），该模型使用联合自回归像素和文本预测进行训练，以及庞大的 EVA-CLIP-18B（Sun et al., 2024）。

6.2.1 线性探测图像分类

我们在 DINOv3 的输出 CLS token 之上训练一个线性分类器，以此在分类基准上评估模型。我们考虑 ImageNet1k（Deng et al., 2009）数据集及其变体来评估超出分布的鲁棒性，并使用来自不同领域的一系列数据集来理解 DINOv3 区分细粒度类别的能力。有关评估的详细信息，请参见 App. D.7。

在这个实验中，我们在 ImageNet-train 上训练，使用 ImageNet-val 作为验证集来选择超参数，并将找到的最佳分类器转移到不同的测试数据集：ImageNet-V2 和 ReaL 是 ImageNet 的替代图像和标签集，用于检测在 ImageNet 验证集上的过拟合；Rendition 和 Sketch 展示了 ImageNet 类别的风格化和人工版本；Adversarial 和 ObjectNet 包含经过精心选择的困难示例；Corruptions 测量对常见图像损坏的鲁棒性。为了参考，我们还列出了通过 Dehghani et al. (2023) 线性探测的结果，该探测用于在庞大的 JFT 数据集（30 亿到 40 亿张图像）上使用监督分类训练 ViT。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 显著超越了所有以前的自监督模型，在 ImageNet-R 上提高了 +10%，在-Sketch 上提高了 +6%，在 ObjectNet 上超过了之前最强的自监督学习模型 DINOv2，提升了 +13%。我们注意到，最强的弱监督模型 SigLIP 2 和 PE 现在在像 ImageNet-A 和 ObjectNet 这样困难的 OOD 任务中比最强的监督模型（ViT-22B）表现更好。DINOv3 在 ImageNet-R 和-Sketch 上取得了可比的结果，而在 ImageNet-A 和 ObjectNet 这样困难的任務中紧跟在 PE 之后，超越了 SigLIPv2。在 ImageNet 上，虽然验证分数落后于 SigLIPv2 和 PE 0.7–0.9 分，但在“更干净”的测试集-V2 和-ReaL 上的表现几乎相同。值得注意的是，DINOv3 在抗腐败方面（ImageNet-C）达到最佳的鲁棒性。总的来说，这是自监督模型首次在图像分类这一过去由（弱监督）监督训练方法占优的领域达到与弱监督和监督模型可比的结果。考虑到像 ViT-22B、SigLIP 2 和 PE 这样的模型是利用大量人工标注的数据集进行训练的，这是一个显著的成果。相比之下，DINOv3 纯粹从图像中学习，这使得在未来进一步扩大和改进该方法成为可能。

Table 8: 细粒度分类基准。Fine-S 在 12 个数据集上的平均结果，完整结果请参见 Tab. 22。

Method	ViT	Fine-S	Places	iNat18	iNat21
Agglomerative backbones					
AM-RADIOv2.5	g/14	93.9	70.2	79.0	83.7
Weakly-supervised backbones					
SigLIP 2	g/16	93.7	70.5	80.7	82.7
PEcore	G/14	94.5	71.3	86.6	87.0
AIMv2	3B/14	92.9	70.7	80.8	83.2
EVA CLIP	18B/14	92.9	71.1	80.7	83.5
Self-supervised backbones					
Franca	g/14	87.7	64.6	61.4	70.6
DINOv2	g/14	92.6	68.2	80.7	86.1
Web-DINO	7B/14	90.2	69.6	65.3	74.1
DINOv3	7B/16	93.0	70.0	85.6	89.8

Table 9: 实例识别基准。有关其他指标，请参见 Tab. 23。

Oxford-H	Paris-H	Met (GAP)	AmsterTime
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

细粒度分类 (Tab. 8) 我们还在多个数据集上训练线性探测器进行细粒度分类时测量了 DINOv3 的性能。特别是，我们报告了三个大型数据集上的准确性，即用于场景识别的 Places205 (Zhou et al., 2014)，以及用于详细植物和动物物种识别的 iNaturalist 2018 (Van Horn et al., 2018) 和 iNaturalist 2021 (Van Horn et al., 2021)，以及 12 个覆盖场景、物体和纹理的小型数据集的平均值（如在 Oquab et al. (2024) 中，这里称为 Fine-S）。有关这些数据集的个别结果，请参见 Tab. 22。

我们发现，DINOv3 再次超越了所有之前的 SSL 方法。相比于弱监督方法，它也显示出具有竞争力的结果，表明其在各种细粒度分类任务中的鲁棒性和泛化能力。值得注意的是，DINOv3 在困难的 iNaturalist21 数据集上达到了最高的准确率 89.8%，甚至超过了最好的弱监督模型 PEcore 的 87.0%。

6.2.2 实例识别

为了评估我们模型的实例级识别能力，我们采用了一种非参数检索方法。在这里，数据库图像根据与给定查询图像的余弦相似度进行排序，使用输出的 CLS 标记。我们在多个数据集上进行了性能基准测试：用于地标识别的 Oxford 和 Paris 数据集 (Radenović et al., 2018)，包含来自大都会博物馆艺术品的 Met 数据集 (Ypsilantis et al., 2021)，以及由与阿姆斯特丹历史档案图像匹配的现代街景图像构成的 AmsterTime 数据集 (Yildiz et al., 2022)。检索效果通过 Oxford、Paris 和 AmsterTime 的平均平均精度以及 Met 的全局平均精度来量化。更多评估细节请参见 App. D.8。

在所有评估的基准中，DINOv3 以较大优势取得了最强的表现，*e.g.*，比第二好的模型 DINOv2 在 Met 上提升了 +10.8 分，在 AmsterTime 上提升了 +7.6 分。在这个基准上，弱监督模型远远落后于 DINOv3，AM-RADIO 除外，因为它是从 DINOv2 特征中提取出来的。这些发现突显了 DINOv3 在实例级检索任务中的鲁棒性和多功能性，既适用于传统地标数据集，也适用于更具挑战性的领域，如艺术和历史图像检索。

6.3 DINOv3 是复杂计算机视觉系统的基础

前两节已经为 DINOv3 在密集和全局任务中的质量提供了可靠信号。然而，这些结果是在“模型实验”实验协议下获得的，通过使用轻量级线性适配器甚至非参数算法来评估特征的质量。虽然这些简单的评估能够消除涉及实验协议中的混杂因素，但它们不足以评估 DINOv3 作为更大的计算机视觉系统中基础组件的全部潜力。因此，在本节中，我们不再使用轻量级协议，而是训练更复杂的下游解码器，并考虑更强的任务特定基线。特别地，我们使用 DINOv3 作为 (1) 使用 Plain-DETR 进行目标检测 (Sec. 6.3.1)，(2) 使用 Mask2Former 进行语义分割 (Sec. 6.3.2)，(3) 使用 Depth Anything 进行单目深度估计 (Sec. 6.3.3)，以及 (4) 使用视觉几何基础变换器进行 3D 理解 (Sec. 6.3.4) 的基础。这些任务只是对 DINOv3 可能实现的探索。即便如此，我们发现基于 DINOv3 的构建以较少的努力解锁了有竞争力甚至是最先进的结果。

Table 10: 与最先进的系统在目标检测上的比较。我们在一个固定的 DINOv3 骨干网络上训练了一个检测适配器。我们展示了 COCO 和 COCO-O 数据集的验证集上的结果，并报告了跨越 IoU 阈值的 mAP，以及有效鲁棒性 (ER)。我们基于 DINOv3 的检测系统建立了新的技术标准。由于 InternImage-G 检测模型尚未发布，我们无法重现他们的结果或计算 COCO-O 得分。

Model	Detector	FT	Parameters			COCO		COCO-O	
			Encoder	Decoder	Trainable	Simple	TTA	mAP	ER
EVA-02	Cascade	🔥	300M	—	300M	64.1	—	63.6	34.7
InternImage-G	DINO	🔥	6B	—	6B	65.1	65.3	—	—
EVA-02	Co-DETR	🔥	300M	—	300M	65.4	65.9	63.7	34.3
PEspatial	DETA	🔥	1.9B	50M	2B	65.3	66.0	64.0	34.7
DINOv3	Plain-DETR	❄️	7B	100M	100M	65.6	66.1	66.4	36.8

6.3.1 目标检测

作为第一个任务，我们解决了长久以来的计算机视觉问题，即物体检测。给定一幅图像，目标是为所有预定义类别的物体实例提供边界框。这个任务需要精确的定位和良好的识别，因为边框需要匹配物体的边界并与正确的类别对应。虽然在像 COCO (Lin et al., 2014) 这样的标准基准上的表现大多趋于饱和，我们提出用一个冻结的骨干网络来解决这个任务，只在其上训练一个小的解码器。

我们在 COCO 数据集上评估 DINOv3 的目标检测能力，报告的是 COCO-VAL2017 划分的结果。此外，我们还在 COCO-O 评估数据集上评估分布外的性能。该数据集包含相同的类别，但在六种分布移位设置下提供输入图像。对于这两个数据集，我们报告了在 $[0.5 : 0.05 : 0.95]$ 中的 IoU 阈值下的平均平均精度 (mAP)。对于 COCO-O，我们还报告了有效鲁棒性 (ER)。由于 COCO 是一个小数据集，仅包含 11.8 万张训练图像，我们利用更大的 Objects365 数据集来对解码器进行预训练，这是常见做法。

我们基于 Plain-DETR (Lin et al., 2023b)，但进行了以下修改：我们没有将 transformer 编码器融合到骨干网络中，而是将其作为一个单独的模块，类似于原始 DETR (Carion et al., 2020)，这使我们能够在训练和推理中保持 DINOv3 的骨干网络完全冻结。据我们所知，这使它成为第一个使用冻结骨干网络的有竞争力的检测模型。我们在 Objects365 上以分辨率 1536 训练 Plain-DETR 检测器 22 个时期，然后以分辨率 2048 训练 1 个时期，接着在 COCO 上以分辨率 2048 训练 12 个时期。在推理时，我们以分辨率 2048 运行。我们还可以选择通过在多个分辨率（从 1536 到 2880）下传递图像来应用测试时增强 (TTA)。完整的实验细节见 App. D.9。

我们将我们的系统与四个模型进行比较：使用级联检测器的 EVA-02 (Fang et al., 2024b)，使用 Co-DETR 的 EVA-02 (Zong et al., 2023)，使用 DINO 的 InternImage-G (Wang et al., 2023b)，以及使用 DETA 的 PEspatial (Bolya et al., 2025)。我们发现，我们基于冻结的 DINOv3 骨干网络训练的轻量级检测器 (100M 参数) 成功地达到了最新的性能指标。对于 COCO-O，差距十分明显，表明检测模型可以有效利用 DINOv3 的稳健性。有趣的是，我们的模型在训练参数更少的情况下，优于所有先前的模型，最小的对比点仍然使用超过 300M 的可训练参数。我们认为，在不专门化骨干网络的情况下实现如此强劲的性能，能够支持多种实际应用：一个单一的骨干网络前向过程可以提供支持多任务的特征，从而减少计算需求。

6.3.2 语义分割

在之前的实验之后，我们现在评估语义分割，这是另一个长期以来的计算机视觉问题。这个任务同样需要强大且定位良好的表示，并期望每个像素的密集预测。然而，与目标检测相反，模型不需要区分同一物体的不同实例。与检测类似，我们在冻结的 DINOv3 模型之上训练一个解码器。

我们将评估重点放在 ADE20k 数据集上 (Zhou et al., 2017)，该数据集包含 150 个语义类别，涵盖 20k 张训练图像和 2k 张验证图像。我们使用平均交并比 (mIoU) 来衡量性能。为了训练分割模型，我们还使用了 COCO-Stuff (Caesar et al., 2018) 和 Hypersim (Roberts et al., 2021) 数据集。前者包含 164k 张图像，含 171 个语义类别，后者包含 77k 张图像，有 40 个类别。

为了构建一个将 DINOv3 特征映射到语义类别的解码器，我们结合了 ViT-Adapter (Chen et al., 2022) 和 Mask2Former (Cheng et al., 2022)，类似于之前的工作 (Wang et al., 2022b; 2023b;a)。然而，在我们的

Table 11: 与 ADE20k 上的最新系统进行语义分割比较。我们在单尺度或多尺度设置下评估模型（分别为 Simple 和 TTA）。根据常规做法，我们在分辨率 896 下运行此评估，并报告 mIoU 分数。BEIT3、ONE-PEACE 和 DINOv3 使用带有 ViT-Adapter 架构的 Mask2Former，并且解码器参数同时考虑到两者。我们将在 Tab. 24 上报告其他数据集结果。

Model	FT	Parameters			mIoU	
		Encoder	Decoder	Trainable	Simple	TTA
BEIT3	🔥	1.0B	550M	1.6B	62.0	62.8
InternImage-H	🔥	1.1B	230M	1.3B	62.5	62.9
ONE-PEACE	🔥	1.5B	710M	2.2B	62.0	63.0
DINOv3	❄️	7B	927M	927M	62.6	63.0

情况下，DINOv3 骨干在训练过程中保持冻结状态。为了避免改变骨干特征，我们进一步修改了原始的 ViT-Adapter 结构，移除了注入组件。与基准相比，我们还将嵌入维度从 1024 增加到 2048，以支持处理 DINOv3 骨干输出的 4096 维。我们首先在 COCO-Stuff 上预训练分割解码器 80 千次迭代，然后在 Hypersim (Roberts et al., 2021) 上 10 千次迭代。最后，我们在 ADE20k 的训练集上训练 20k 次迭代，并在验证集上报告结果。所有训练均以 896 的输入分辨率完成。在推理时，我们考虑两种设置：单尺度，*i.e.* 以训练分辨率传递图像，或多尺度，*i.e.* 我们对在 $\times 0.9$ 和 1.1 之间的多个图像比例下的预测进行平均，采用原始训练分辨率。有关更多实验细节，请参阅 App. D.10。

我们将我们的模型性能与多种最新的基线进行比较，包括 BEIT-3 (Wang et al., 2022b)、InternImage-H (Wang et al., 2023b) 和 ONE-PEACE (Wang et al., 2023a)，并在 Tab. 24 中报告额外数据集的结果。基于冻结的 DINOv3 主干网络的我们的分割模型达到了最新的性能，等同于 ONE-PEACE (63.0 mIoU)。在 COCO-Stuff (Caesar et al., 2018) 和 VOC 2012 (Everingham et al., 2012) 数据集上，它也优于所有先前的模型。由于语义分割需要精确的每像素预测，视觉转换器主干网络构成了一个基本问题。实际上，16 像素宽的输入块使得预测的粒度相对较粗——促使了如 ViT-Adapter 这样的解决方案。另一方面，我们已证明即使在高达 4096 (*c.f.* Figs. 3 and 4) 的非常高的分辨率下，我们也可以获得高质量的特征图；这对应于 512-标记宽的密集特征图。我们希望未来的工作能够利用这些高分辨率特征达到最新的性能，而不必依赖于如 ViT-Adapter with Mask2Former 这样巨大的解码器。

6.3.3 单目深度估计

我们现在考虑构建一个用于单目深度估计的系统。为此，我们采用最新的先进方法 Depth Anything V2 (DAv2) (Yang et al., 2024b) 的设置。DAv2 的关键创新在于使用大量合成生成的图像及其真实深度标注。重要的是，这依赖于 DINOv2 作为特征提取器，能够弥合模拟到真实的差距，这是其他视觉骨干网如 SAM (Kirillov et al., 2023) 所无法展示的能力 (Yang et al., 2024b)。因此，我们在 DAv2 流程中用 DINOv3 替换 DINOv2，以观察我们是否可以获得类似的结果。

和 DAv2 类似，我们使用密集预测 Transformer (DPT) 来预测逐像素的深度场，使用来自 DINOv3 四个等间距层的特征作为输入。我们使用 DAv2 的损失集在 DAv2 的合成数据集上训练模型，提高训练分辨率到 1024×768 ，以利用 DINOv3 的高分辨率能力。与 DAv2 相比，我们保持骨干网络冻结而不是微调它，以测试 DINOv3 的开箱即用能力。我们还发现扩大 DPT 头以获得 DINOv3 7B 的更大特征的全部潜力是有益的。详情请参见 App. D.11。

数据集和指标 我们在五个真实世界的数据集 (NYUv2 (Silberman et al., 2012), KITTI (Geiger et al., 2013), ETH3D (Schöps et al., 2017), ScanNet (来自 Ke et al. (2025)) 和 DIODE (Vasiljevic et al., 2019)) 上评估我们的模型，采用类似于 Ranftl et al. (2020); Ke et al. (2025); Yang et al. (2024b) 的零样本尺度不变深度设置。我们报告标准指标绝对相对误差 (ARel) (越低越好) 和 δ_1 (越高越好)。有关这些指标的说明，请参阅 Yang et al. (2024a)。

我们与目前最新的相对深度估计技术进行比较：MiDaS (Ranftl et al., 2020)、LeReS (Yin et al., 2021)、Omnidata (Eftekhari et al., 2021)、DPT (Ranftl et al., 2021)、集成版本中的 Marigold (Ke et al., 2025) 和 DAv2。我们的深度估计模型在所有数据集上都达到了一个新的最先进水平，只是在 DIODE 的 ARel 上略逊于 DPT。值得注意的是，这在使用冻结的骨干网络的情况下是可能的，而其他所有基准都需要微调骨

Table 12: 与当前最先进系统的相对单目深度估计进行比较。通过将 DINOv3 与 Depth Anything V2 (Yang et al., 2024b) 结合，我们获得了一个用于相对深度估计的最先进模型。

Method	FT	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
		ARel ↓	δ_1 ↑	ARel ↓	δ_1 ↑	ARel ↓	δ_1 ↑	ARel ↓	δ_1 ↑	ARel ↓	δ_1 ↑
MiDaS	🔥	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5
LeReS	🔥	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6
OmniData	🔥	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
Marigold	🔥	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

Table 13: 使用视觉几何基础变压器 (VGGT) 进行 3D 理解 (Wang et al., 2025)。通过简单地将 VGGT 流水线中的图像特征提取器从 DINOv2 替换为 DINOv3 ViT-L，我们能够在各种 3D 几何任务中获得最先进的结果。我们从 Wang et al. (2025) 重现基线结果。我们还报告了使用真实相机信息的方法，并用 * 标记。相机位姿估计结果以 AUC@30 报告。

(a) 相机位姿估计。			(b) 在 DTU 上的多视图估计。				(c) 在 ScanNet-1500 上查看匹配。		
Method	Re10K	CO3Dv2	Method	Acc. ↓	Comp. ↓	Overall ↓	Method	AUC@5	AUC@10
DUST3R	67.7	76.7	Gipuma *	0.283	0.873	0.578	SuperGlue	16.2	33.8
MASt3R	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	MASt3R *	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
FLARE	78.8	83.3	DUST3R	2.677	0.805	1.741	Roma	31.8	53.4
VGGT	85.3	88.2	VGGT	0.389	0.374	0.382	VGGT	33.9	55.2
DINOv3	86.3	89.6	DINOv3	0.375	0.361	0.368	DINOv3	35.2	56.1

干网络以进行深度估计。此外，这验证了 DINOv3 继承了 DINOv2 强大的模拟到现实能力，这一理想的特性为下游任务使用合成生成的训练数据开辟了可能性。

6.3.4 视觉几何基础变压器与 DINOv3

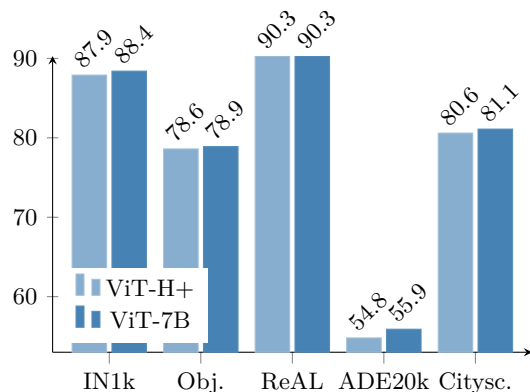
最后，我们考虑使用最近的视觉几何结合变压器 (VGGT) (Wang et al., 2025) 进行 3D 理解。VGGT 在大量 3D 标注的数据上进行训练，学习在一次前馈过程中估计场景的所有重要 3D 属性，例如相机的内在参数和外参、点地图或深度图。使用一个简单、统一的管道，它在许多 3D 任务上达到了最新的成果，同时比专用方法更高效——这构成了 3D 理解的重大进展。

实现 VGGT 使用一个经过 DINOv2 预训练的骨干网络来获取场景不同视角的表示，然后通过一个 transformer 来融合它们。在这里，我们简单地将 DINOv2 骨干替换为 DINOv3，使用我们的 ViT-L 变体 (见 ??)，以匹配原始工作中的 DINOv2 ViT-L/14。我们运行与 VGGT 相同的训练流程，包括对图像骨干网络的微调。我们将图像分辨率从 518×518 切换为 592×592 ，以适应 DINOv3 的 16 号补丁大小，并保持与 VGGT 的结果可比性。我们还采用了一些小的超参数变化，具体详见 App. D.12。

根据 Wang et al. (2025)，我们在 Re10K (Zhou et al., 2018) 和 CO3Dv2 (Reizenstein et al., 2021) 数据集上评估相机位姿估计，在 DTU (Jensen et al., 2014) 上进行稠密多视图估计，以及在 ScanNet-1500 (Dai et al., 2017) 上进行双视图匹配。对于相机位姿估计和双视图匹配，我们报告标准的曲线下面积 (AUC) 指标。对于多视图估计，我们报告预测到真实值之间最小的 L2 距离作为“准确性”，从真实值到预测的最小 L2 距离作为“完整性”，以及它们的平均值作为“总体”结果。关于方法和评估的详细信息，请参考 Wang et al. (2025)。

Model	# Params	Inference GFLOPs	
		Res. 256	Res. 512
CNX-Tiny	29M	5	20
CNX-Small	50M	11	46
CNX-Base	89M	20	81
CNX-Large	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 模型系列。



(b) ViT-H+ 对比 ViT-7B。

Figure 16: (a) 精炼模型特征的展示。CNX 代表 ConvNeXT。我们展示每个模型的参数数量和和在尺寸为 256×256 和 512×512 的图像上估算的 GFLOPs。(b) 我们将 DINOv3 ViT-H+ 与其 7B 大小的教师模型进行比较；尽管参数少了近 $10 \times$ ，ViT-H+ 的性能接近 DINOv3 7B。

结果 (Tab. 13) 我们发现，VGGT 配备 DINOv3 后，在所考虑的所有三项任务上都进一步超越了 VGGT 设置的先前最先进水平——使用 DINOv3 带来了明显且一致的提升。这令人鼓舞，因为我们对 DINOv3 仅进行了最少的调整。这些任务涵盖了不同层次的视觉理解：场景内容的高级抽象（相机位姿估计）、密集几何预测（多视图深度估计）以及细粒度的像素级对应（视图匹配）。结合先前在对应估计 (Sec. 6.1.2) 和深度估计 (Sec. 6.3.3) 上的结果，我们认为这进一步实证证明了 DINOv3 作为 3D 任务基础的强大适用性。此外，我们预计使用更大的 DINOv3 7B 模型还会带来进一步的改进。

在本节中，我们对从我们的 7B 参数模型中提取的模型家族进行定量评估（参见 Sec. 5.2）。该家族包括基于 Vision Transformer (ViT) 和 ConvNeXt (CNX) 架构的变体。我们在 Fig. 16a 中提供了所有模型的详细参数数量和推理 FLOPs。这些模型涵盖了广泛的计算预算，以适应广大用户和部署场景。我们对所有 ViT (??) 和 ConvNeXt 变体进行全面评估，以评估它们在不同任务上的表现。

Figure 2 提供了 DINOv3 家族与其他模型集合的总体对比。DINOv3 家族在密集预测任务上显著优于所有其他模型。这包括从监督骨干网络提取的专用模型，如 AM-RADIO 和 PEspatial。同时，我们的模型在分类任务上取得了类似的结果，使得它们在不同计算预算下成为最佳选择。

在 ?? 中详细介绍了我们的 ViT 模型，并将它们与其他开源替代品进行比较。然后，在 Sec. 6.4 中，我们讨论了 ConvNeXt 模型。最后，继 Sec. 5.3 之后，我们训练了一个文本编码器，使其与我们的 ViT-L 模型的输出对齐。在 Sec. 6.5 中，我们展示了这个模型的多模态对齐结果。

我们 ViT 系列涵盖了从紧凑的 ViT-S 到更大的 840 百万参数 ViT-H+ 模型的架构。前者旨在如笔记本电脑等资源受限的设备上高效运行，而后者则为需要更高性能的应用交付最先进的性能。我们将我们的 ViT 模型与相应大小的最佳开源图像编码器进行比较，即 DINOv2、SigLIP2 和 Perception Encoder。为了公平比较，我们确保输入序列长度在各个模型之间相等。具体来说，对于补丁大小为 16 的模型，我们输入尺寸为 512×512 的图像，而当模型使用补丁大小为 14 时则输入尺寸为 448×448 的图像。

我们的实证研究清楚地表明，DINOv3 模型在密集预测任务中始终优于其竞争对手。尤其是在 ADE20k 基准上，DINOv3 的 ViT-L 模型相比最佳竞争对手 DINOv2 实现了超过 6 个 mIoU 点的提升。ViT-B 版本相比次佳竞争对手显示了大约 3 个 mIoU 点的增益。这些显著的改进突显了 DINOv3 的局部特征在捕捉细粒度空间细节方面的有效性。此外，在深度估计任务上的评估也显示出相较于竞争对手的一贯性能提升。这强调了 DINOv3 系列在不同的密集视觉问题上的多功能性。重要的是，我们的模型在全球识别基准如 ObjectNet 和 ImageNet-1k 上也取得了具有竞争力的结果。这表明增强的密集任务性能并没有以牺牲全局任务准确性为代价。这种平衡证实了 DINOv3 模型提供了一种强大且全面的解决方案，在密集和全局视觉任务中都表现卓越且无妥协。

Table 14: 我们将我们的一系列模型与开源的同等大小替代品进行比较。在一组具有代表性的全局和密集基准上展示了我们的 ViT- $\{S, S+, B, L, H+\}$ 模型：分类 (IN-ReAL, IN-R, ObjectNet)、检索 (Oxford-H)、分割 (ADE20k)、深度 (NYU)、跟踪 (DAVIS at 960px) 和关键点匹配 (NAVI, SPair)。我们匹配补丁令牌的数量，以便在不同补丁大小的模型之间进行公平比较。

Size	Model	Global Tasks				Dense Tasks				
		IN-ReAL	IN-R	Obj.	Ox.-H	ADE20k	NYU ↓	DAVIS	NAVI	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	PEcore	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	PEcore	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
SO400m	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

另一方面，我们还希望验证我们蒸馏得到的最大模型是否能够捕捉教师模型的所有信息。为此，我们对比了我们最大的 ViT-H+ 模型与 7B 教师模型。如 Fig. 16b 所示，最大的学生模型取得了与大 8 倍的 ViT-7B 模型相当的性能。这一结果不仅验证了我们蒸馏过程的有效性，还表明当由高质量的教师模型引导时，较小的模型可以学习达到相当水平的性能。此发现强化了我们相信训练非常大的模型能为更广泛的社区带来益处。较大的模型的优势可以成功地蒸馏到更高效、较小的模型中，而几乎不损失质量。

6.4 用于资源受限环境的高效 ConvNeXt

在本节中，我们评估了从 7B 教师模型蒸馏出来的 ConvNeXt (CNX) 模型的质量。ConvNeXt 模型在 FLOPs 方面非常高效，并且非常适合部署在为卷积计算优化的设备上。此外，变换器模型通常不易于量化，然而卷积网络的量化是一个研究广泛的话题。我们蒸馏出 T、S、B 和 L 大小的 CNX 架构 (参见 Fig. 16a)，并将其与原始 ConvNeXt 模型进行比较。这些基线在 ImageNet-1k 上取得了高性能，因为它们是使用 ImageNet-22k 标签以监督方式训练的，因此代表着一个强大的竞争者。在这个实验中，我们为输入分辨率为 256 和 512 的全球任务，分辨率为 512 的 ADE20k，以及分辨率为 640 的 NYU 提供了结果。

结果 (Tab. 15) 我们发现，在分辨率为 256 的分布内图像分类中，我们的模型略微落后于监督模型 (*e.g.* -0.7 IN-ReAL for CNX-T)。然而，在分辨率为 512 时，趋势发生逆转，监督的 ConvNeXt 表现显著下降，而我们的模型则随着输入分辨率的增加而扩展。在分布外分类 (IN-R, ObjectNet) 中，两种模型系列之间在所有尺寸上都有显著差距——这证明了 DINOv3 CNX 模型的鲁棒性。此外，DINOv3 模型在密集任务上提供了非常大的改进。实际上，对于 CNX-T，我们的模型带来了 +17.9 mIoU 的提升 (42.7 对比 24.8)，而对于 CNX-L，我们的模型获得了 +14.5 mIoU 的提升 (47.8 对比 33.3)。高性能和计算效率的结合使蒸馏的 ConvNeXt 模型在资源受限的情况下尤为适用于现实世界的应用。另外，将 ViT-7B 模型蒸馏成较小的 ConvNeXt 模型尤为令人兴奋，因为它连接了两种根本不同的架构。虽然 ViT-7B 基于有 CLS token 的 transformer 模块，ConvNeXt 依赖于没有 CLS token 的卷积操作，这使得这种知识转移并不简单。这个成就突出了我们蒸馏过程的多样性和有效性。

6.5 基于 DINOv3 的零样本推理 dino.txt

如 Sec. 5.3 中所详述，我们训练了一个文本编码器，以对齐蒸馏后的 DINOv3 ViT-L 模型的 CLS 标记和输出补丁与文本，遵循 dino.txt Jose et al. (2025) 的方法。我们在标准基准上从全局和补丁层面评估对齐质量。我们使用 CLIP 协议 (Radford et al., 2021) 在 ImageNet-1k、ImageNet-Adversarial、ImageNet-Rendition

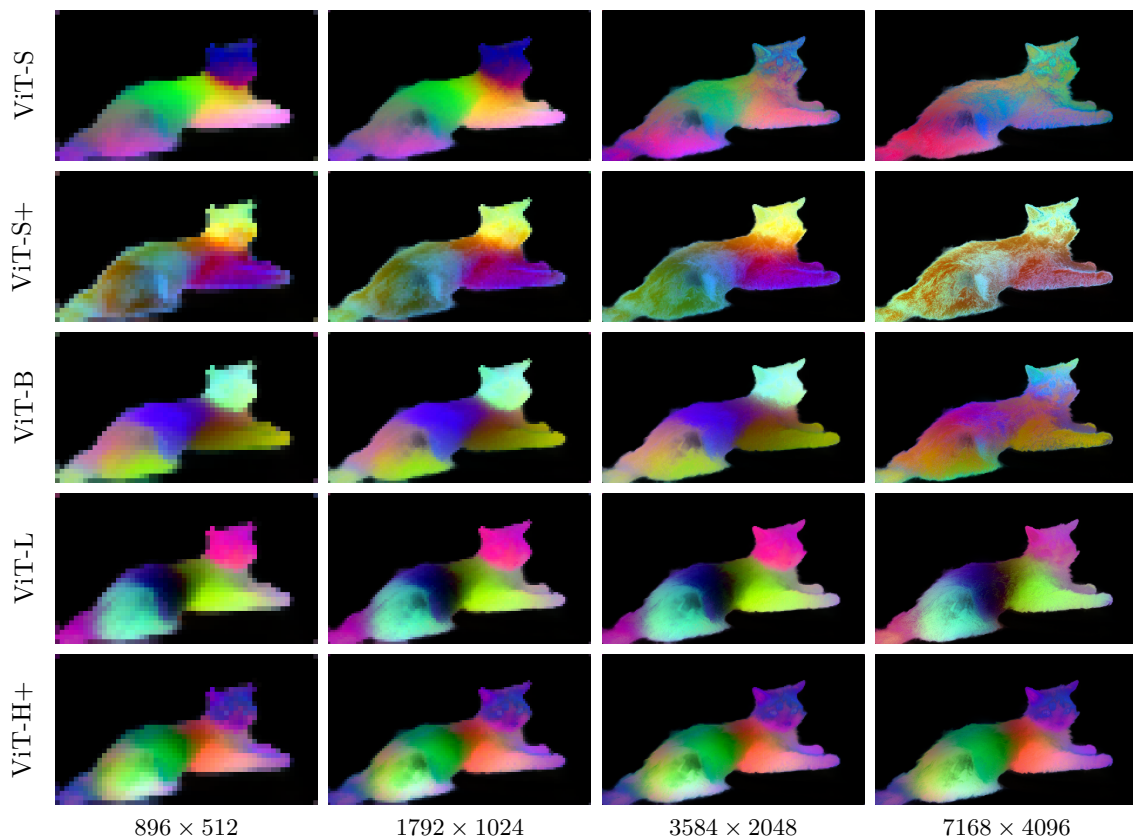


Figure 17: DINOv3 ViT 模型家族在分辨率下特征稳定性。从上到下: ViT-S、S+、B、L、H+。我们对一幅图像在多个分辨率上进行推理, 然后对计算出的 1792×1024 图像 (112×64 图像标记) 特征进行主成分分析。接着, 我们将所有分辨率下的特征投影到为可视化而映射到 RGB 空间的主成分 5-7 上。虽然这些模型在所有分辨率下都能正常工作, 但我们观察到特征在相当大范围的分辨率上保持一致, 然后才开始漂移: 例如, ViT-S+ 特征在 896×512 和 3584×2048 输入之间稳定, 而 ViT-L 在最大分辨率 7168×4096 上几乎开始漂移。ViT-H+ 在整个测试范围内保持稳定。

和 ObjectNet 基准上报告零样本分类准确性。在图像文本检索中, 我们在 COCO2017 数据集 (Tsung-Yi et al., 2017) 上进行评估, 并报告图像到文本 ($I \rightarrow T$) 和文本到图像 ($T \rightarrow I$) 任务的 Recall@1。为了探测补丁层面对齐的质量, 我们在使用常用基准 ADE20k 和 Cityscapes 的开放词汇分割任务上评估我们的模型, 并报告 mIoU 指标。

我们将我们文本对齐的 DINOv3 ViT-L 与相同大小类别的竞争对手进行比较。与将 DINOv2 对齐文本的 Jose et al. (2025) 相比, DINOv3 在所有基准测试中表现显著更好。在全局对齐任务中, 我们在对比原始 CLIP (Radford et al., 2021) 和如 EVA-02-CLIP (Sun et al., 2023) 等强基线时表现良好, 但稍逊于 SigLIP2 (Tschannen et al., 2025) 和感知编码器 (Bolya et al., 2025)。在密集对齐任务中, 我们的文本对齐模型在两个具有挑战性的基准测试 ADE20K 和 Cityscapes 上表现出色, 这得益于 DINOv3 的干净特征图。

7 在地理空间数据上的 DINOv3

我们的自监督学习方法是通用的, 可以应用于任何图像领域。在这一部分, 我们通过为卫星图像构建一个 DINOv3 7B 模型来展示这种通用性, 这些图像与最初开发 DINOv3 所使用的网络图像具有非常不同的特征 (e.g. 物体纹理、传感器噪声和焦距视图)。

Table 15: 我们对蒸馏后的 DINOv3 ConvNeXt 模型进行了评估。我们将自己的模型与在 ImageNet-22k 上进行监督训练的现成 ConvNeXts 进行比较。对于全局任务，我们在输入分辨率 256 和 512 上提供结果，因为我们发现监督模型在 512 分辨率时显著退化。

Size	Model	Global Tasks						Dense Tasks	
		IN-ReAL		IN-R		Obj.		ADE20k	NYU ↓
		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

Table 16: 比较我们文本对齐的 DINOv3 ViT-L 与最新技术。我们的模型在保持全局对齐任务竞争力的同时，实现了优秀的密集对齐性能。所有比较的模型均为 ViT-L 大小，并在相同的 576 序列长度上运行。

Method	Classification				Retrieval		Segmentation	
	IN1k	A	R	Obj.	I → T	T → I	ADE20k	Cityscapes
CLIP	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
dino.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
PE	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

7.1 预训练数据和基准测试

我们的卫星 DINOv3 7B 模型是在 SAT-493M 上进行预训练的，该数据集由 493 百万幅 512×512 图像组成，这些图像是从 Maxar 的 RGB 正射影像中随机采样的，分辨率为 0.6 米。我们使用与 web 版 DINOv3 7B 模型完全相同的超参数集，除了适用于卫星图像的 RGB 均值和标准差归一化，以及训练长度。与 web 模型类似，我们的卫星模型训练流程由 100k 次初始预训练的全局裁剪（ 256×256 ）组成，随后进行 10k 次的 Gram 正则化，并以 8k 次高分辨率微调步骤结束，分辨率为 512。同样与 web 模型类似，我们将我们的 7B 卫星模型蒸馏为一个更易管理的 ViT-Large 模型，以便在低预算条件下使用。

我们在多个地球观测任务中评估了 DINOv3 卫星和网络模型。在全球树冠高度测绘任务中，我们使用了 App. D.13 中描述的 Satlidar 数据集，该数据集由 100 万张 512×512 图像组成，具有 LiDAR 地面真值，并按 8/1/1 的比例分为训练/验证/测试集。分集包括了 Tolan et al. (2024) 使用的 Neon 和 São Paulo 数据集。对于国家尺度的树冠高度测绘，我们在 Open-Canopy (Fogel et al., 2025) 上进行评估，该数据集将 SPOT 6-7 卫星影像与横跨法国 87000 平方公里的航空 LiDAR 数据结合起来。由于此数据集中的图像有 4 个通道，包括额外的红外 (IR) 通道，我们通过将嵌入块模块的权重中取三个通道的平均值并添加为第四个通道来调整我们的骨干。我们在将图像裁剪并调整到 1667 以匹配 Maxar 地面样本分辨率的 512×512 上训练了一个 DPT 解码器。

语义地理空间任务通过 GEO-Bench (Lacoste et al., 2023) 进行评估，该评估包含六个分类任务和六个分割任务，涵盖了各种空间分辨率和光学波段。GEO-Bench 任务多样，包括检测屋顶上的光伏系统、分类地方气候区、测量森林砍伐的驱动因素以及检测树冠。对于高分辨率的语义任务，我们考虑土地覆盖分割数据集 LoveDA (Wang et al., 2022a)，物体分割数据集 iSAID (Zamir et al., 2019)，以及水平检测数据集 DIOR (Li et al., 2020)。

Table 17: 不同骨干网络对高分辨率郁闭高度预测的评估。所有模型均使用 DPT 解码器进行训练。结果展示为使用在 SatLidar 上训练的解码器在 IID 样本 (SatLidar Val) 和 OOD 测试集 (SatLidar Test、Neon 和 São Paulo) 进行评估的实验，或是在 Open-Canopy 数据集上训练和评估的解码器的实验。我们列出了平均绝对误差 (MAE) 和来自 Tolan et al. (2024) 的块 R^2 指标。为了完整性，我们还额外评估了 Tolan et al. (2024) 原始解码器，其在 Neon 数据集上训练 (用 * 表示)。

Method	Arch.	SatLidar								Open Canopy
		SatLidar Val		SatLidar Test		Neon Test		São Paulo		
		MAEâ “	R^2 â ‘	MAEâ “	R^2 â ‘	MAEâ “	R^2 â ‘	MAEâ “	R^2 â ‘	
Tolan et al. (2024) *	ViT-L	2.8	0.86	4.0	0.61	2.7	0.73	5.4	0.42	—
Tolan et al. (2024)	ViT-L	2.4	0.90	3.4	0.81	2.9	0.69	5.4	0.48	2.42
DINOv3 Web	ViT-7B	2.4	0.90	3.6	0.74	2.7	0.75	5.9	0.34	2.17
DINOv3 Sat	ViT-L	2.2	0.91	3.2	0.81	2.4	0.81	5.8	0.42	2.07
DINOv3 Sat	ViT-7B	2.2	0.92	3.2	0.82	2.6	0.74	5.5	0.51	2.02

7.2 树冠高度估计

从卫星图像估计树冠高度是一个具有挑战性的度量任务，需要准确恢复连续的空间结构，尽管在坡度、视角、太阳角度、大气散射和量化伪影等方面存在随机变化。该任务对于全球碳监测以及森林和农业管理至关重要 (Harris et al., 2021)。遵循 Tolan et al. (2024)，即首个利用在卫星图像上训练的 SSL 骨干进行该任务的工作，我们在冻结的 DINOv3 之上训练一个 DPT 头部，使用 SatLidar1M 训练集，然后在 i.i.d. 的 SatLidar1M 验证集样本以及包括 SatLidar1M 测试、Neon 和圣保罗在内的分布外测试集上进行评估。我们还在 Open-Canopy 数据集上进行训练和评估。

我们比较了不同的 SSL 骨干网络，将训练 SAT-493M 数据集的模型称为“DINOv3 Sat”，将训练 LVD-1689M 的模型称为“DINOv3 Web” (见 Sec. 3.1)。可以看出，DINOv3 卫星模型在大多数基准上表现出色。我们的 7B 卫星模型在 SatLidar1M 验证集、SatLidar1M 测试集和 Open-Canopy 上设立了新的技术水平，将 MAE 分别从 2.4 降低到 2.2，从 3.4 降低到 3.2，从 2.42 降低到 2.02。这些结果表明，DINOv3 的训练方法具有通用性，可以有效地直接应用于其他领域。有趣的是，我们的蒸馏 ViT-L 卫星模型与 7B 模型表现相当，在 SatLidar1M 和 Open-Canopy 上取得了相似的结果，而在 Neon 测试集上表现出色，达到最低的 MAE 为 2.4，相比 7B 模型的 2.6 和 Tolan et al. (2024) 的 2.9。我们的 DINOv3 7B Web 模型在基准测试中表现良好，在 SatLidar1M 验证集、Neon 和 Open-Canopy 上超过了 Tolan et al. (2024)，但仍不如卫星模型。这突出了特定领域预训练在物理基础任务如树冠高度估计中的优势，在这些任务中，传感器特定的先验知识和辐射一致性是重要的。

7.3 与地球观测前沿技术的比较

我们在 Tab. 18 和 Tab. 19 中比较了不同方法在地球观测任务中的性能。冻结的 DINOv3 卫星和网络模型在 15 个分类、分割和水平目标检测任务中的 12 个上创下了新的最先进结果。我们的 Geo-Bench 结果超过了之前的模型，包括使用 6 个以上波段的 Prithvi-v2 (Szwarcman et al., 2024) 和 DOFA (Xiong et al., 2024)，以及针对 Sentinel-2 和 Landsat 任务进行的任务特定微调 (Tab. 18)。尽管只用 RGB 输入的冻结骨干网络，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都优于以前的方法。有趣的是，DINOv3 7B 网络模型在这些基准测试中竞争力很强。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中实现了相当或更强的性能。如 Tab. 18 和 Tab. 19 所示，冻结的 DINOv3 网络模型在 Geo-Bench 任务以及 LoveDA 和 DIOR 数据集上的分割和检测任务中确立了新的领先结果。

这些发现对地理空间基础模型的设计具有更广泛的影响。最近强调了启发式技术，例如多时态聚合、多传感器融合或结合卫星特定元数据 (Brown et al., 2025; Feng et al., 2025)。我们的结果表明，对于依赖于精确物体边界 (例如分割或物体检测) 的任务而言，通用 SSL 可以匹配或超过卫星特定的方法。这支持了新兴的证据，发现领域无关的预训练即使在专业的下游领域中也能够提供强大的泛化能力 (Lahrichi et al., 2025)。

总体而言，我们的结果表明领域特定预训练的任务依赖性益处。DINOv3 卫星模型在深度估计等度量任务中表现出色，充分利用了卫星特定先验知识。相比之下，DINOv3 网络模型通过多样化的、通用的表示在语义

Table 18: 我们的 DINOv3 模型在 Geo-Bench 任务中与强有力的基线 DOFA (Xiong et al., 2024)、Prithvi-v2 (Szwarcman et al., 2024) 和 Tolan et al. (2024) 的比较。虽然 Prithvi-v2 和 DOFA 充分利用了所有可利用的光学波段，但我们的模型仅凭 RGB 输入就取得了显著更好的性能。

(a) 分类任务。

Method	Arch.	FT	Bands	m-BEnet	m-brick-kiln	m-eurosat	m-forestnet	m-pv4ger	m-so2sat	Mean
DOFA	ViT-L	👉	all	68.7	98.4	96.6	55.7	98.2	61.6	79.9
Best of Prithvi-v2	ViT-L/H	👉	all	71.2	98.8	96.4	54.1	98.1	59.1	79.6
Tolan et al. (2024)	ViT-L	🌟	RGB	66.0	97.1	95.2	56.3	94.3	58.1	77.8
DINOv3 Sat	ViT-L	🌟	RGB	73.0	96.5	94.1	60.6	96.0	57.4	79.6
DINOv3 Sat	7B	🌟	RGB	74.0	97.2	94.8	62.3	96.1	62.1	81.1
DINOv3 Web	7B	🌟	RGB	74.6	97.7	97.0	57.9	98.3	63.8	81.6

(b) 分割任务。

Method	Arch.	FT	Bands	m-cashew *	m-chesapeake	m-NeonTree	m-nz-cattle	m-pv4ger-seg	m-SA-crop	Mean
DOFA	ViT-L	👉	all	81.2	61.6	58.5	77.4	95.1	35.7	68.3
Best of Prithvi-v2	ViT-L/H	👉	all	90.2	69.4	59.1	81.0	95.3	41.9	72.8
Tolan et al. (2024)	ViT-L	🌟	RGB	92.8	73.7	58.1	83.1	94.7	35.1	72.9
DINOv3 Sat	ViT-L	🌟	RGB	94.2	75.6	61.8	83.7	95.2	36.8	74.5
DINOv3 Sat	7B	🌟	RGB	94.1	76.6	62.6	83.4	95.5	37.6	75.0
DINOv3 Web	7B	🌟	RGB	96.0	76.5	66.4	83.7	95.9	36.8	75.9

* Conversion to 6 classes following Szwarcman et al. (2024).

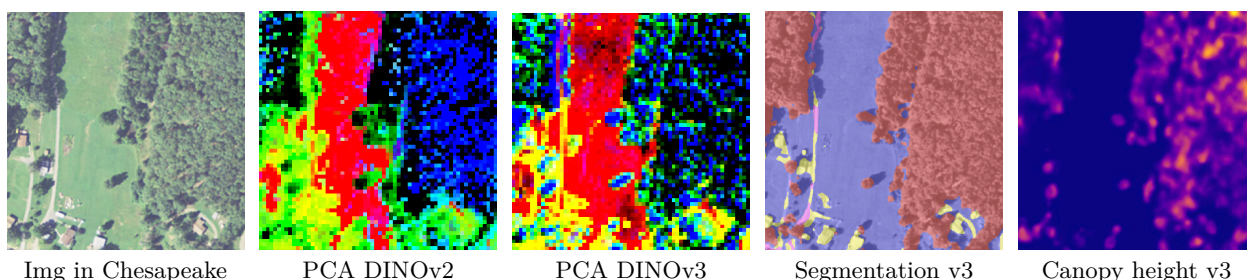


Figure 18: 展示了一个单一的 DINOv3 模型在遥感中的多种应用。DINOv3 特征的 PCA 显示出了比 DINOv2 更细致的细节。分割图是仅使用 GEO-Bench chesapeake 标签进行计算的。冠层高度模型解码器是在 Open-Canopy 数据集上进行训练的，使用了 4 个通道 (RGB + 红外)，而推理则仅在 RGB 通道上进行。

地理空间任务中取得了最先进的成果。这两个模型的互补优势展示了 DINOv3 自监督学习范式的广泛适用性和有效性。

8 环境影响

为了估算我们预训练的碳排放，我们遵循了之前在自然语言处理 (Strubell et al., 2019; Touvron et al., 2023) 和 SSL (Oquab et al., 2024) 研究中使用的方法。我们将所有外生变量的值固定，*i.e.* 能源使用效率 (PUE) 和电网碳强度因素设定为 Touvron et al. (2023) 所用的相同值，*i.e.* 我们假设 PUE 为 1.1，并且美国平均碳强度因素为 0.385 千克 CO₂ 当量/千瓦时。对于 GPU 的功耗，我们采用它们的热设计功耗：A100 GPU 为 400W，H100 GPU 为 700W。我们在 Tab. 20 中报告了 ViT-7B 预训练计算的详细信息。为了对比，我们提供了 DINOv2 和 MetaCLIP 的类似数据。作为另一个比较点，训练一个 DINOv3 模型所需的能量 (47 MWh) 大致相当于使用一辆平均电动车行驶 240,000 公里的能量需求。

为了计算整个项目的碳足迹，我们粗略估计总共需要 900 万 GPU 小时。使用和上面相同的网格参数，我们估计总碳足迹大约为 2600 tCO₂ eq。作为比较，一次完整的巴黎和纽约之间的波音 777 往返航班大约对应 560 tCO₂ eq。假设每天有 12 次这样的航班，我们项目的环境影响就相当于这两座城市之间一天所有航班的一半。该估算仅考虑为 GPU 供电的电力，并忽略了其他排放，例如冷却、制造和处理。

Table 19: 我们将 DINOv3 的性能与最先进的模型 Privthi-v2 (Szwarcman et al., 2024)、BillionFM (Cha et al., 2024) 和 SkySense V2 (Zhang et al., 2025) 用于高分辨率语义地理空间任务进行比较。我们报告分割数据集 LoveDA (1024 ×) 和 iSAID (896 ×) 的 mIoU，以及检测数据集 DIOR (800×) 的 mAP。

Method	Arch.	FT	LoveDA	iSAID	DIOR
Prev. SotA		🔥	BillionFM, ViT-G 54.4	SkySense V2, Swin-G * 71.9	SkySense V2, Swin-G * 79.5
Decoder Arch.			UPerNet	UPerNet	Faster-RCNN
Privthi-v2	ViT-H	🔥	52.2	62.8	—
DINOv3 Sat	ViT-L	❄️	54.4	62.9	72.7
DINOv3 Sat	ViT-7B	❄️	55.3	64.8	76.6
DINOv3 Web	ViT-7B	❄️	56.2	71.4	80.5

* Uses modified DINOv2 SSL with supervised pretraining alignment on OpenStreetMap, reporting +0.8 mIoU on iSAID.

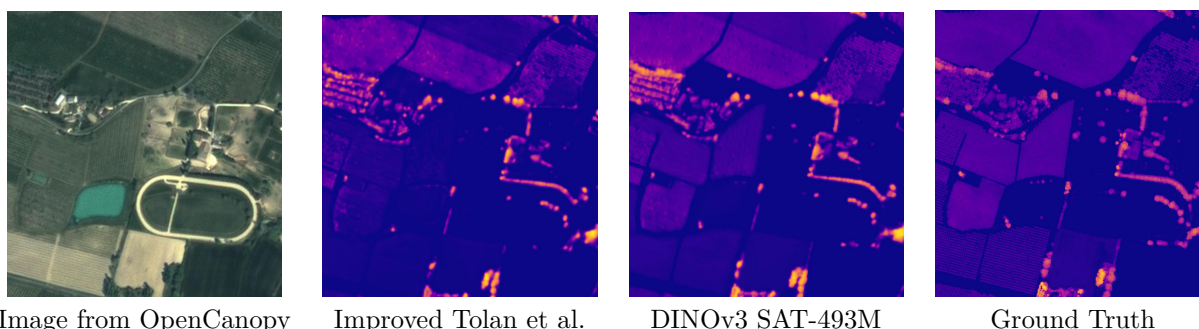


Figure 19: 对于开放树冠数据集，将 DINOv3 7B 卫星模型与 Tolan et al. (2024) 进行定性比较。对于两个模型，解码器都在 448 × 448 的输入图像上进行训练。可以看到 DINOv3 生成的地图更准确，例如对于田野上树木的准确高度。

9 结论

DINOv3 代表了自监督学习领域的一项重大进展，展示了改变各个领域视觉表示学习方式的潜力。通过精细的数据准备、设计和优化，扩大数据集和模型规模，DINOv3 展示了自监督学习的强大力量，能够消除对人工标注的依赖。引入的 Gram anchoring 方法有效地减轻了在长时间训练中密集特征图的退化问题，确保了稳健可靠的性能。

结合后期修饰策略的实施，如高分辨率后期训练和蒸馏，我们在广泛的视觉任务中实现了最先进的性能，而无需对图像编码器进行微调。DINOv3 视觉模型套件不仅设定了新的基准，还提供了在各种资源限制、部署场景和应用使用案例中多功能的解决方案。DINOv3 所取得的进步证明了自监督学习在推进计算机视觉及其他领域的最新技术中的潜力。

Table 20: 模型训练的碳足迹。我们报告了重现完整模型预训练的潜在碳排放量，其计算使用了 1.1 的 PUE 和 0.385kg CO₂ eq/KWh 的碳强度系数。

Model	Arch.	GPU type	Power (W)	Steps	GPU hours	PUE	Total power (MWh)	Emission (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

References

- Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. ICLR , 2025.
- Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In ICLR , 2020.
- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. arXiv preprint arXiv:2301.08243 , 2023.
- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 , 2025.
- Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 , 2022.
- Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In ICML , 2023.
- Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In CVPR , 2024.
- Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 , 2021.
- Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In NeurIPS , 2019.
- Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 , 2021.
- Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. arXiv preprint arXiv:2210.01571 , 2022.
- Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 , 2024.
- Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? CoRR , abs/2006.07159, 2020.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 , 2024.
- Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In ICCV , Oct 2017.
- Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In ICML , 2017.

-
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* , 5:135–146, 2017.
- Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181* , 2025.
- Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Conference on Empirical Methods in Natural Language Processing* , 2021.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV* , 2014.
- Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaeart foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025.
- Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR* , 2018.
- Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Enhancing end-to-end object detection with aligned loss, 2024.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision* , pages 213–229. Springer, 2020.
- Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV* , 2018.
- Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *ICCV* , 2019.
- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS* , 2020.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV* , 2021.
- Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* , page 1–17, 2024. ISSN 2151-1535. doi: 10.1109/jstars.2024.3401772.
- François Charton and Julia Kempe. Emergent properties with repeated examples, 2024.
- Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine* , 2024.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *preprint arXiv:2002.05709* , 2020a.
- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* , 33:22243–22255, 2020b.

-
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In ICLR , 2023.
- Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. preprint arXiv:2011.10566 , 2020.
- Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 , 2021.
- Yinjie Chen, Zipeng Yan, Chong Zhou, Bo Dai, and Andrew F Luo. Vision transformers with self-distilled registers. arXiv preprint arXiv:2505.21501 , 2025.
- Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 , 2022.
- Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In CVPR , 2022.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In CVPR , 2023.
- Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In CVPR , 2015.
- M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In CVPR , 2014.
- Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. NeurIPS , 2022.
- Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition , pages 3213–3223, 2016.
- Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In CVPR , 2017.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In ICLR , 2024.
- Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. TMLR , 2025.
- Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim M. Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Paveti’c, Dustin Tran, Thomas Kipf, Mario Luvci’c, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In ICML , 2023.

-
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR , 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. preprint arXiv:1810.04805 , 2018.
- Barry M Dillon, Gregor Kasieczka, Hans Olschlager, Tilman Plehn, Peter Sorrenson, and Lorenz Vogel. Symmetries, safety, and self-supervision. SciPost Physics , 12(6):188, 2022.
- Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In CVPR , 2023.
- Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In ICCV , 2015.
- Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE TPAMI , 2016.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. preprint arXiv:2010.11929 , 2020.
- Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In ICML , 2023.
- Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In ICCV , 2021.
- Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint arXiv:2112.10740 , 2021.
- Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes challenge 2007 (VOC2007) results, 2007.
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.
- Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. IJCV , 111(1):98–136, 2015.
- David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. arXiv preprint arXiv:2504.01017 , 2025.
- Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In ICLR , 2024a.
- Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In CVPR , 2023.
- Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. Image and Vision Computing , 149:105171, 2024b.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In 2004 conference on computer vision and pattern recognition workshop , pages 178–178. IEEE, 2004.

-
- Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis, 2025.
- Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrise da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402* , 2024.
- Fajwel Fogel, Yohann Perron, Nikola Besic, Laurent Saint-André, Agnès Pellissier-Tanon, Martin Schwartz, Thomas Boudras, Ibrahim Fayad, Alexandre d’Aspremont, Loic Landrieu, et al. Open-canopy: Towards very high resolution forest monitoring. In *CVPR* , 2025.
- Stephanie Fu, Mark Hamilton, Laura E Brandt, Axel Feldmann, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. In *ICLR* , 2024.
- Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 2414–2423, 2016.
- Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* , 32(11):1231–1237, 2013.
- Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.
- Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV* , 2019.
- Priya Goyal, Mathilde Caron, Benjamin Lefaudeaux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *preprint arXiv:2103.01988* , 2021.
- Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncured images without supervision. *arXiv preprint arXiv:2202.08360* , 2022a.
- Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *2022 ACM Conference on Fairness, Accountability, and Transparency* , pages 70–88, 2022b.
- Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The " something something " video database for learning and evaluating visual common sense. In *ICCV* , 2017.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS* , 2020.
- Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR* , 2006.
- Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR* , 2022.
- Nancy Harris, David Gibbs, A. Baccini, Richard Birdsey, Sytze de Bruin, Mary Farina, Lola Fatoyinbo, Matthew Hansen, Martin Herold, Richard Houghton, Peter Potapov, Daniela Requena Suarez, Rosa Maria Roman-Cuesta, Sassan Saatchi, Christy Slay, Svetlana Turubanova, and Alexandra Tyukavina. Global maps of twenty-first century forest carbon fluxes. *Nature Climate Change* , 11:1–7, 03 2021. doi: 10.1038/s41558-020-00976-6.

-
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR , 2016.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR , 2020.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 , 2021.
- Greg Heinrich, Mike Ranzinger, Hongxu, Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. RADIOv2.5: Improved baselines for agglomerative vision foundation models. arXiv preprint arXiv:2412.07679 , 2025.
- Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. preprint arXiv:1905.09272 , 2019.
- Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. arXiv preprint arXiv:2103.10957 , 2021.
- Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In ECCV , 2022.
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. ICLR , 2019.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In ICCV , pages 8340–8349, 2021a.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In CVPR , pages 15262–15271, 2021b.
- Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In European Conference on Computer Vision , pages 289–305. Springer, 2024.
- Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. In NeurIPS , 2020.
- Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuezhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In NeurIPS , 2023.
- Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanaes. Large scale multi-view stereopsis evaluation. In CVPR , June 2014.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning , pages 4904–4916. PMLR, 2021.
- Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don’t need trained registers. arXiv preprint arXiv:2506.08010 , 2025.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV , 2016.
- Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. DINOv2 meets text: A unified framework for image-and pixel-level vision-language alignment. In CVPR , 2025.

-
- Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In ECCV , 2016.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 , 2017.
- Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. arXiv preprint arXiv:2505.09358 , 2025.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 , 2024.
- Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, Marah Halawa, Tim König, David Gnutz, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. Scientific Reports , 15(1):4876, 2025.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In CVPR , 2023.
- Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In ECCV , pages 491–507. Springer, 2020.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13) , Sydney, Australia, 2013.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. NeurIPS , 2012.
- Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. NeurIPS , 2023.
- Saad Lahrichi, Zion Sheng, Shufan Xia, Kyle Bradbury, and Jordan Malof. Is self-supervised pre-training on satellite imagery better than imagenet? a systematic study with sentinel-2, 2025.
- Yann LeCun. A path towards autonomous machine intelligence. openreview , 2022.
- Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In ICCV , 2017.
- Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS journal of photogrammetry and remote sensing , 159: 296–307, 2020.
- Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , 2023a.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft COCO: common objects in context. In ECCV , 2014.

-
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.
- Yutong Lin, Yuhui Yuan, Zheng Zhang, Chen Li, Nanning Zheng, and Han Hu. Detr does not need multi-scale or locality design. In ICCV , 2023b.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS , 2023.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 , 2019.
- Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In CVPR , 2022.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR , 2017.
- Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In ECCV , 2018.
- S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 , 2013.
- Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. COCO-O: A benchmark for object detectors under natural distribution shifts. In ICCV , 2023.
- Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In Interspeech , 2010.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In NeurIPS , 2013.
- Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence, 2019.
- Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In CVPR , 2020.
- Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In Indian Conference on Computer Vision, Graphics and Image Processing , Dec 2008.
- Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision – ECCV 2016 , pages 69–84, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4.
- Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR , 2014.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. TMLR , 2024.
- Valentin Pariza, Mohammadreza Salehi, Gertjan J Burghouts, Francesco Locatello, and Yuki M Asano. Near, far: Patch-ordering enhances vision foundation models’ scene understanding. In ICLR , 2025.
- Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, et al. Astroclip: a cross-modal foundation model for galaxies. Monthly Notices of the Royal Astronomical Society , 531(4):4990–5011, 2024.

-
- Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition* , 2012.
- Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR* , 2016.
- Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 724–732, 2016.
- Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* , pages 1–12, 2025.
- Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmaleck, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. In *NeurIPS* , 2020.
- Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. preprint [arXiv:1704.00675](https://arxiv.org/abs/1704.00675) , 2017.
- Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR* , 2018.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* , pages 8748–8763. PMLR, 2021.
- Jathushan Rajasegaran, Ilija Radosavovic, Rahul Ravishankar, Yossi Gandelsman, Christoph Feichtenhofer, and Jitendra Malik. An empirical study of autoregressive pre-training from videos. [arXiv preprint arXiv:2501.05453](https://arxiv.org/abs/2501.05453) , 2025.
- René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI* , 44(3):1623–1637, 2020.
- René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 12179–12188, 2021.
- Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR* , 2024.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR* , 2025.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML* , pages 5389–5400, 2019.
- Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV* , 2021.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)* , 2015.
- Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019.

-
- Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In ICCV , 2021.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. IJCV , 2015.
- Bryan Russell, William Freeman, Alexei Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In CVPR , 2006.
- Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. arXiv preprint arXiv:1806.03198 , 2018.
- Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, and Yuki M. Asano. Time does tell: Self-supervised time-tuning of dense image representations. In ICCV , 2023.
- Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In CVPR , 2017.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 , 2021.
- Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In ICLR , 2023.
- Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF international conference on computer vision , pages 8430–8439, 2019.
- Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European conference on computer vision , pages 746–760. Springer, 2012.
- Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. arXiv preprint arXiv:2109.14279 , 2021.
- Oriane Siméoni, Éloi Zablocki, Spyros Gidaris, Gilles Puy, and Patrick Pérez. Unsupervised object localization in the era of self-supervised vits: A survey. IJCV , 133(2):781–808, 2025.
- Walter Simoncini, Andrei Bursuc, Spyridon Gidaris, and Yuki Asano. No train, all gain: Self-supervised gradients improve deep frozen representations. NeurIPS , 2024.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR , 2015.
- Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting Weakly Supervised Pre-Training of Visual Perception Models. In CVPR , 2022.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 , 2012.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. arXiv preprint arXiv:1906.02243 , 2019.

-
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* , 568:127063, 2024.
- Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV* , 2017.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* , 2023.
- Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252* , 2024.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR* , 2015.
- Daniela Szwarzman, Sujit Roy, Paolo Fraccaro, Thorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732* , 2024.
- Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *preprint arXiv:1905.11946* , 2019.
- Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM* , 59(2):64–73, 2016.
- Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV* , 2021.
- Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, et al. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment* , 300:113888, 2024.
- Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In *NeurIPS* , 2019.
- Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877* , 2020.
- Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118* , 2022.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* , 2023.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786* , 2025.
- Tsung-Yi, Genevieve Patterson, Matteo R. Ronchi, Yin Cui, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays Georgia, Pietro Perona, Deva Ramanan, Larry Zitnick, and Piotr Dollár. COCO 2017: Common Objects in Context, 2017.

-
- Tinne Tuytelaars, Christoph Lampert, Matthew Blaschko, and Wray Buntine. Unsupervised object discovery: A comparison. *IJCV* , 2010.
- Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR* , 2018.
- Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* , pages 12884–12893, 2021.
- Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *ArXiv 1908.00463* , 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS* , 2017.
- Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Spyros Gidaris, Elias Ramzi, Andrei Bursuc, and Yuki M. Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. *arXiv preprint arXiv:2507.14137* , 2025.
- Huy V. Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *CVPR* , 2019.
- Huy V. Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)* , 2020.
- Huy V. Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *Advances in Neural Information Processing Systems 35 (NeurIPS)* , 2021.
- Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *TMLR* , 2024.
- Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine* , 30(10):2924–2935, 2024.
- Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. Mtp: Advancing remote sensing foundation model via multi-task pretraining, 2024a.
- Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR* , 2025.
- Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation, 2022a.
- Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172* , 2023a.
- Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR* , 2023b.
- Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* , 2022b.

-
- Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In CVPR , 2019.
- Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In CVPR , pages 14543–14553, 2022c.
- Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. IEEE TPAMI , 45(12):15790–15801, 2023c.
- Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. In ECCV , 2024b.
- Frederik Warburg, Søren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In Computer Vision and Pattern Recognition (CVPR) , 2020.
- P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In NeurIPS , 2022.
- Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In CVPR , 2018.
- Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks. ECCV , 2024.
- J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition , pages 3485–3492, June 2010. doi: 10.1109/CVPR.2010.5539970.
- Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding, 2018.
- Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In NeurIPS , 2021.
- Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In CVPR , 2017.
- Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the Earth crossing modalities. arXiv preprint arXiv:2403.15356 , 2024.
- Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In ICLR , 2024.
- Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-VOS: Sequence-to-sequence video object segmentation. In Proceedings of the European conference on computer vision (ECCV) , pages 585–601, 2018.
- I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 , 2019.
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In CVPR , 2024a.

-
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS* , 2024b.
- Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR* , 2025.
- Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. *arXiv preprint arXiv:2203.16291* , 2022.
- Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR* , 2021.
- Yongseon Yoo, Seonggyu Kim, and Jong-Min Lee. Sagagan: Style applied using Gram matrix attribution based on stargan v2. In *BMVC* , 2024.
- Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* , 2021.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* , 2022.
- Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR* , 2025.
- Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *CVPR* , 2022.
- Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images, 2019.
- Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* , 2014.
- Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR* , 2022a.
- Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* , pages 18123–18133, 2022b.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 11975–11986, 2023.
- Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV* , 2016.
- Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. SkySense V2: A Unified Foundation Model for Multi-modal Remote Sensing. *arXiv preprint arXiv:2507.13812* , 2025.
- Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS* , 2014.
- Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 633–641, 2017.
- Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832* , 2021.

-
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In SIGGRAPH , 2018.
- Adrian Ziegler and Yuki M Asano. Self-supervised learning of object parts for semantic segmentation. In CVPR , 2022.
- Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF international conference on computer vision , pages 6748–6758, 2023.
- Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. preprint arXiv:2006.06882 , 2020.

附录

A 大规模训练中的伪影和异常值

本节讨论了在大型模型训练中最近观察到的伪影和离群点的出现，这些模型涉及 LLM (An et al., 2025) 和视觉域 (Darcet et al., 2024)。引用 An et al. (2025) 的定义，离群点通常被描述为其值显著偏离其分布平均值的网络激活。在 DINOv3 的训练过程中，我们在不同层次上发现了此类离群点：有些发生在补丁层面，有些发生在特征维度层面。我们在下面讨论了观察到的不同类型的离群点，它们对训练和结果的影响。我们还讨论了它们修复它们的不同尝试和我们的初步结论。

A.1 高范数补丁异常点

Darcet et al. (2024) 发现异常的图像块会对 DINOv2 的性能产生负面影响。这些异常值主要表现为高范数的标记，通常位于图像中信息量较低的背景区域。这些标记在图像块与 CLS 标记之间的内部通信中起到了关键作用。此外，这一现象也影响其他模型，无论是有监督还是无监督训练的，如 CLIP (Radford et al., 2021)。当模型扩展到 7B 时，我们观察到这种高范数图像块的出现，主要位于背景区域。在本节中，我们展示了在 150k 次迭代训练的 7B 模型的结果，虽然有限，但为我们的决策提供了初步的信号。我们在 Fig. 20a 中绘制了输出图像块的范数（在层规范之前），在 ‘ $\emptyset$ ’ 列中，黄色表示的高范数图像块出现在天空和其他信息量较低的区域。

为了减轻此类标记异常值的出现，(Darcet et al., 2024) 提出了一个简单而有效的解决方案：在 ViT 的输入序列中引入称为寄存器的附加标记。它们的作用是接管补丁与 CLS 之间的内部通信。根据结论，我们使用 4 个寄存器，并由于实验成本高而不做进一步消融。Figure 20a 展示了该策略实际应用的例子，我们观察到高范数异常值被消除，进一步确认了相应范数分布的直方图。此外，在 Fig. 20b 中我们定量观察到，在 ImageNet-1k (IN1k) 基准上引入附加寄存器标记的好处。

在注意力机制中整合偏见 最近，An et al. (2025) 的研究调查了不同模型和架构中的 LLM 领域中的异常值现象。作者分析了与注意力机制内在相关的不同类型的异常值。他们提出通过若干解决方案来缓解这一问题，我们选择了两个看起来相关且对注意力机制要求最小改动的有前景的解决方案，具体是显式固定偏置，我们称为“值门控”，以及注意力偏置策略。值门控策略相当于在注意力的输出中加入一个可学习的值偏置 $\mathbf{v}' \in \mathbb{R}^d$ ，具体通过重新定义注意力机制为

$$\text{Attn}(Q, K, V; \mathbf{k}', \mathbf{v}') = \text{softmax}\left(\frac{Q[K^T]}{\sqrt{d}}\right)V + \mathbf{v}', \quad (3)$$

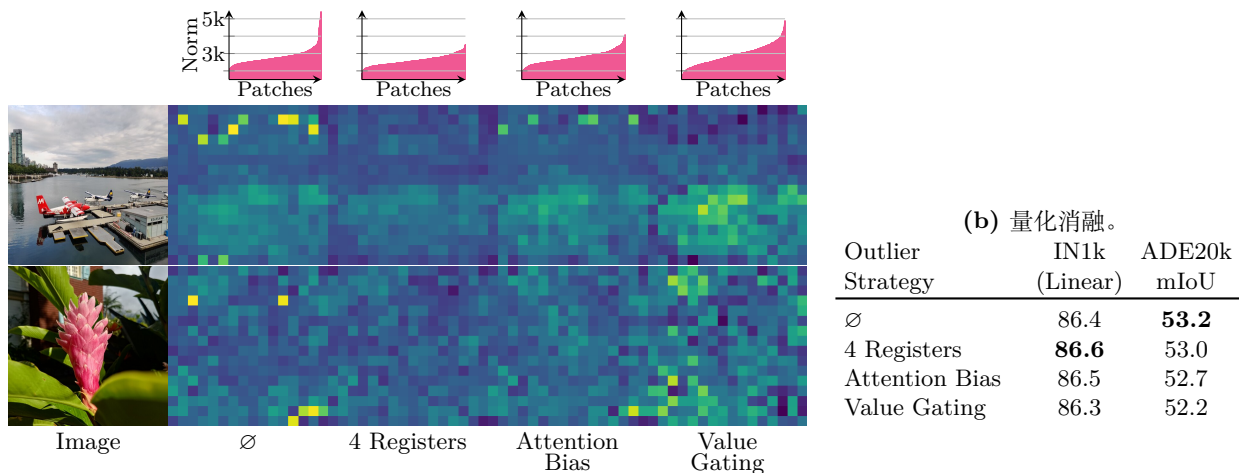
其中 $Q, K, V \in \mathbb{R}^{T \times d}$ 为查询、键和值矩阵， d 为隐藏空间的维度。或者，注意力偏置，定义在 Eq. 4 中，包含在键和值矩阵上分别整合两个可学习的偏置项 $\mathbf{k}', \mathbf{v}' \in \mathbb{R}^d$ 。其定义如下：

$$\text{Attn}(Q, K, V; \mathbf{k}', \mathbf{v}') = \text{softmax}\left(\frac{Q[K^T \mathbf{k}']}{\sqrt{d}}\right) \begin{bmatrix} V \\ \mathbf{v}' \end{bmatrix}. \quad (4)$$

我们在 Fig. 20a 中观察到，值门控策略显著修改了分块范数的分布，导致了通常更高的范数值并消除了明显的异常值。尽管注意力机制减轻了高范数标记的出现，但它并没有完全解决这一问题，因为一些高范数分块仍然存在——当与我们使用寄存器标记的结果比较时，这在顶部行图片中是可见的。值得注意的是，最佳性能是在结合寄存器标记后实现的，这也是为什么我们在论文报告的所有实验中采用这一策略。

A.2 特征维度离群点

将额外寄存器引入模型架构可以有效解决高范数补丁离群值的问题。然而，在训练 7B 模型时，我们观察到一种不同类型的离群值出现，不是在不同补丁之间，而是在学习表示的特征（通道）维度内。具体来说，对转换层和训练迭代过程中的补丁激活分析显示，即使各补丁的范数保持稳定，一小部分特征维度的幅度会变得异常大。有趣的是，这些特征维度的离群值在不同补丁和图像中表现出一致的高值，这与 (An et al., 2025) 中报告的结果形成对比。此外，这些离群维度在给定模型的层中持续存在，其幅度随深度增加并在输出层达到最大值。在整个训练过程中，它们的幅度也逐步增加。



(a) 通过离群点策略可视化补丁规范。底部两行每行共享一个色彩图，从深蓝色（低）到黄色（高）。

Figure 20: 评估不同策略减轻高范数补丁异常值影响的方法，包括 (a) 定性和 (b) 定量评估。我们使用我们的方案训练的一个 7B 模型进行了 150k 次迭代，在没有任何高范数处理策略 ‘ $\emptyset$ ’ 的情况下，使用四个注册标记 (Darcet et al., 2024)，或注意力偏置和值门控策略 (An et al., 2025)。在 (a, 第一行) 中，我们绘制了为三个图像计算的输出补丁范数（按升序排序）的分布。我们还可可视化了每个图像的输出补丁范数（底部两行），使用相同的颜色映射--最小值和最大值在每个图像中的不同异常值策略下计算。

我们进行实验，尝试在训练和推理过程中中和这些维度。我们的研究表明，这些维度在训练过程中扮演重要角色，因为应用 L2 正则化以抑制它们会导致性能下降。然而，在推理阶段去除这些维度并不会导致显著的性能变化，这表明它们主要携带无关紧要或无信息量的信号。此外，我们观察到最终层归一化被训练为显著缩小这些异常维度。因此，我们建议将最终层的特征应用最终层归一化以用于下游任务。或者，应用批归一化也可以抑制这些特征维度的异常值，因为它们在不同的 patch 和图像中保持较高的一致性。

使用来自较早层的特征需要注意。正如上文所讨论的，这些较早层也会受到特征维度异常值的影响，从而导致特征不良。这种情况下，虽然最终层的归一化适用于归一化最终特征的分布，但它的学习参数可能在应用于较早层特征时并不理想。事实上，我们注意到在某些任务中这样做会导致性能下降。在这些情况下，我们发现标准的特征缩放技术（使用批归一化或主成分分析的 *e.g.* 归一化）在处理特征维度异常值方面效果显著。例如，对于我们的语义分割 (Sec. 6.3.2) 和深度估计实验 (Sec. 6.3.3)，在使用中间层的特征时，我们应用了批归一化。

B 附加结果

在 Fig. 1 中，我们提供了近年来最先进性能的大致演变。在此，我们提供了图中所报告的精确参考和性能。请在 Tab. 21 中查阅。

B.1 逐层分析

在本节中，我们评估 DINOv3 7B 模型各个层次特征的质量。具体来说，我们展示了五个代表性任务的结果：分类 (IN-1k 验证、ImageNet-ReAL 和 ObjectNet)、分割 (ADE20k)、深度估计 (NYU)、跟踪 (DAVIS) 和 3D 对应估计 (NAVI)。对于前三个基准测试，在每个主干层的输出上训练一个线性层来评估特征性能，如 Secs. 6.1.1 and 6.2.1 中所述。对于跟踪和对应估计，我们采用 Secs. 6.1.2 and 6.1.4 中的非参数方法。

结果展示在 Fig. 21 中。我们发现对于分类和密集任务，性能随着层数的增加而平稳提高。深度估计、跟踪和 3D 对应估计在第 32 层附近达到峰值，这表明对于几何性占重要角色的任务，通过考虑更早的层次可以提高 DINOv3 的下游性能。另一方面，中间层的性能与最后一层相比仅略有改善，因而是一个不错的默认选择。

Table 21: 关于 Fig. 1 中所使用的数据的出版年份、性能和参考文献的详细信息。对于所有论文，我们报告该算法在 ImageNet 上使用最大模型的 top-1 准确率。对于弱监督和自监督模型，我们提供线性探测性能。关于日期，我们使用首次在 arXiv 上出现的年份。

Year	Supervised		Weakly-Supervised		Self-Supervised	
	Top-1	Reference	Top-1	Reference	Top-1	Reference
2012	59.3	Krizhevsky et al. (2012)				
2013						
2014						
2015	78.6	He et al. (2016)	34.9	Joulin et al. (2016)		
2016						
2017	80.9	Xie et al. (2017)				
2018			83.6	Mahajan et al. (2018)	38.2	Caron et al. (2018)
2019	84.3	Tan and Le (2019)			68.6	He et al. (2020)
2020	87.5	Kolesnikov et al. (2020)			75.3	Caron et al. (2020)
2021	88.6	Dosovitskiy et al. (2020)	88.4	Radford et al. (2021)	82.3	Zhou et al. (2021)
2022						
2023	89.5	Dehghani et al. (2023)			86.5	Oquab et al. (2024)
2024						
2025			89.3	Bolya et al. (2025)	88.4	This work

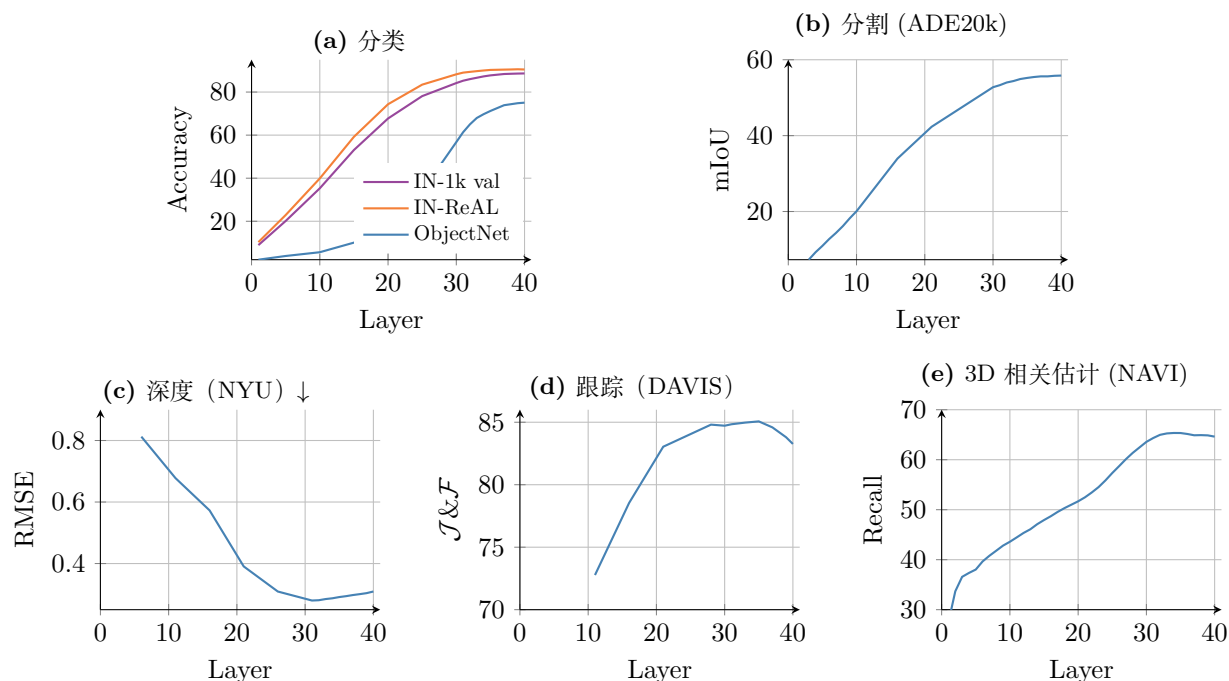


Figure 21: 使用 DINOv3 7B 的中间层特征对五个基准进行了测试。评估 (a-c) 使用线性层（见 Secs. 6.1.1 and 6.2.1），而 (d, e) 使用非参数方法（见 Secs. 6.1.2 and 6.1.4）。

B.2 主要结果部分的附加结果

我们提供了额外的实验结果，以补充 Sec. 6 中的主要结果。在 Tab. 22 中，我们展示了对小数据集进行细粒度分类的逐数据集结果，使用线性探测技术（Fine-S，见 Sec. 6.2.1）。在 Tab. 23 中，我们提供了实例识别评估的完整结果（Sec. 6.2.2），并增加了更多的指标。最后，在 Tab. 24 中，我们提供了对于我们最新的语

Table 22: 对小型数据集进行细粒度分类的每个数据集结果，使用线性探测（Fine-S，见 [Sec. 6.2.1](#)），根据 [Oquab et al. \(2024\)](#)。

Method	ViT	Food	C10	C100	SUN	Cars	Aircr.	VOC	DTD	Pets	Cal101	Flowers	CUB	Avg
Agglomerative backbones														
AM-RADIOv2.5	g/14	96.5	99.5	95.0	82.8	95.4	91.7	90.3	88.6	96.7	98.8	99.7	91.5	93.9
Weakly-supervised backbones														
SigLIP	g/16	97.7	99.3	92.7	85.1	96.5	88.7	91.0	87.7	98.7	90.3	99.7	90.3	93.7
PE-core	G/14	97.8	99.5	95.3	85.2	96.5	92.0	90.5	88.2	98.7	93.3	99.5	93.3	94.5
AIMv2	3B/14	96.6	99.3	93.3	83.4	95.6	84.2	90.5	87.4	96.8	90.7	99.7	90.7	92.9
EVA CLIP	18B/14	96.9	99.5	95.4	85.0	95.4	81.6	90.2	87.1	98.4	90.6	99.6	90.6	92.9
Self-supervised backbones														
Franca	g/14	89.2	98.6	90.4	73.7	89.7	74.1	89.4	80.6	93.2	97.6	97.8	78.4	87.7
DINOv2	g/14	95.6	99.5	94.5	78.9	94.6	88.5	88.4	86.8	96.8	95.9	99.7	91.6	92.6
Web-DINO	7B/14	96.1	99.5	93.4	77.5	95.0	88.8	87.0	79.9	92.9	93.1	99.6	78.9	90.2
DINOv3	7B/16	96.9	99.6	96.0	81.1	95.0	88.2	88.2	87.2	97.0	94.8	99.7	92.4	93.0

义分割模型（[Sec. 6.3.2](#)）的补充结果，这些结果基于 COCO-Stuff（[Caesar et al., 2018](#)）、PASCAL VOC 2012（[Everingham et al., 2012](#)）和 Cityscapes（[Geiger et al., 2013](#)）数据集。

Table 23: 实例识别的完整结果，呈现了 [Sec. 6.2.2](#) 的附加指标。

		Oxford		Paris		Met			AmsterTime
Method	ViT	M	H	M	H	GAP	GAP-	ACC	mAP
Agglomerative backbones									
AM-RADIOv2.5	g/16	72.8	50.7	93.3	85.3	30.5	65.9	69.0	46.7
Weakly-supervised backbones									
SigLIPv2	g/16	49.3	25.1	79.3	60.9	0.0	0.0	0.2	15.5
PE-core	G/14	57.4	32.7	83.6	68.9	10.6	34.8	44.9	23.1
AIMv2	3B/14	55.0	28.8	85.6	71.4	29.5	67.3	69.9	23.1
EvaCLIP	18B/14	55.2	27.1	81.8	65.6	0.5	4.3	11.0	18.9
Self-supervised backbones									
Franca	g/14	44.6	14.3	73.8	51.6	27.2	54.3	57.7	21.1
DINOv2	g/14	78.2	58.2	92.7	84.6	44.6	73.0	75.2	48.9
Web-DINO	7B/14	64.1	31.2	89.8	80.3	35.2	67.3	71.3	30.6
DINOv3	7B/16	81.1	60.7	93.3	87.1	55.4	77.7	80.7	56.5

Table 24: 与最新系统在其他数据集上的语义分割比较，作为 [Tab. 11](#) 中 ADE20k 结果的补充。我们报告了单尺度或多尺度（TTA）设置下评估模型时的 mIoU 分数，并与每个数据集先前发表的最佳结果进行比较：COCO-Stuff 164k 为 [Fang et al. \(2023\)](#)，Cityscapes 为 [Wang et al. \(2023b\)](#)，VOC 2012 为 [Zoph et al. \(2020\)](#)。我们使用的输入分辨率为 COCO-Stuff 为 1280，VOC 2012 为 1024，Cityscapes 为 1280。所有基线都需要对骨干网络进行微调，而我们保持 DINOv3 骨干网络冻结。

Method	FT	COCO-Stuff 164k		Cityscapes		VOC 2012	
		Single	TTA	Single	TTA	Single	TTA
Previous Best	🔥	53.7	53.7	86.3	87.0	—	90.0
DINOv3	❄️	53.8	54.0	86.1	86.7	90.1	90.4

Table 25: 对 OCR 复杂数据集上 DINOv3 分类性能的比较。这些数据集对 SSL 来说是出了名的困难。我们将 DINOv3 与最佳的 DINOv2 模型 (g) 进行比较, 同时也与最佳的弱监督 PE 核心模型 (G) 进行比较。

Model		GTSRB	Logo-2K+	FlickrLogos-32	RP2K	Products-10K	SOPProducts
DINOv2	ViT-g	78.2	52.9	83.6	91.4	70.8	57.6
PE-core	ViT-G	94.8	93.2	99.0	93.1	80.6	80.7
DINOv3-7B	ViT-7B	87.5	86.0	86.3	94.7	74.5	65.2

Table 26: 对不同收入组和地区进行地理公平性和多样性分析, 遵循 Goyal et al. (2022b) 的协议。

Method	Arch.	Income Buckets			Regions			
		low	medium	high	Africa	Asia	Americas	Europe
SEERv2	RG-10B	59.7	78.5	86.6	65.9	76.3	81.1	85.6
DINOv2	ViT-g/14	67.4	83.3	90.5	74.0	81.6	86.2	89.7
DINOv3	ViT-7B	69.6	85.7	90.9	76.7	83.0	88.0	90.7

B.3 在 OCR 重度数据集上的分类

在这个实验中, 我们在需要某种形式字符识别的分类任务中评估 DINOv3。这些任务包括街道标志、标识和产品分类。我们将我们的模型与最佳自监督模型 (DINOv2 g) 和最佳弱监督模型 (PE-core G) 进行比较。我们在分辨率为 512 的图像上运行此评估, 并调整补丁大小以匹配其他模型的序列长度。我们在 Tab. 25 中报告了这个实验的结果。

我们看到我们的新模型 DINOv3 显著优于其前身 DINOv2。然而, 与弱监督模型的差距仍然很大。由于我们的模型在训练期间不利用成对的图像-文本数据, 因此在学习字形关联方面更为困难。最近 Fan et al. (2025) 的工作暗示了训练数据对这类任务性能的影响。由于我们工作的主要重点是改善密集特征, 缩小这一差距将留待未来工作。

B.4 公平性分析

我们评估了 DINOv3 7B 模型在不同收入层级和地区的地理公平性和多样性, 遵循了 Goyal et al. (2022b) 的协议。作为参考, 我们包括了使用 DINOv2 和 SEERv2 获得的结果。结果表明, DINOv3 在各收入类别中表现相对一致, 尽管与最高收入层相比, 在低收入层中性能下降了 23 %。中等和高收入层表现相似。在地区方面, DINOv3 在不同地区取得了相对良好的得分; 然而, 欧洲和非洲之间观察到了超过 14 % 的相对差异, 这比 DINOv2 中观察到的超过 17 % 的相对差异有所改善。

C 实现细节

我们使用多裁剪 (Caron et al., 2020), 学生模型看到的裁剪包括 2 个全局裁剪 (256×256 px) 和 8 个局部裁剪 (112×112 px), 最终总序列长度为 3.7 M 个标记。教师的 EMA (学生的指数移动平均) 仅处理全局裁剪。我们将 $\mathcal{L}_{\text{DINO}}$ 损失应用于学生所有局部裁剪和教师全局裁剪的类标记, 以及两个模型不同全局裁剪对之间。学生看到的一部分全局裁剪的补丁标记以 $[0.1, 0.5]$ 的随机比例被屏蔽, 概率为 50 %, 我们在这些标记和教师 EMA 看到的可见标记之间应用 $\mathcal{L}_{\text{iBOT}}$ 损失。我们对学生看到的第一个全局裁剪的 16 个类标记小批次应用 $\mathcal{L}_{\text{DKoleo}}$ 损失。我们使用 Pytorch 中的全分片数据并行设置进行 1M 次迭代训练, 使用 bfloat16 和 8 位浮点矩阵乘法。我们使用一个常数学习率 0.0004, 经过 100k 次迭代的预热, 权重衰减为 0.04, 分层学习率衰减因子为 0.98, 随机深度 (层丢弃) 值为 0.4, 教师的 EMA 因子为 0.999。其余超参数可以在代码发布的配置文件中找到。

对于 Gram anchoring 步骤, 我们使用 $w_{\text{Gram}} = 2$ 的损失权重, 并每 10k 步更新一次 Gram 教师, 最多更新三次。对于高分辨率适应 (Sec. 5.1), 我们以以下概率从以下全局/局部/Gram 教师裁剪分辨率组合中进行采样: (512, 112, 768) 和 $p = 0.3$, (768, 112, 1152) 和 $p = 0.3$, (768, 168, 1152) 和 $p = 0.3$, (768, 224, 1152) 和 $p = 0.05$, (768, 336, 1152) 和 $p = 0.05$ 。这些值是通过实验获得的。

D 实验细节

在本节中，我们详细描述了本文中所有基准测试使用的数据集和评估指标。

D.1 语义分割：线性探测

数据集和指标 我们通过线性探测，在三个基准数据集 (ADE20k (Zhou et al., 2017), VOC12 (Everingham et al., 2012), 和 Cityscapes (Cordts et al., 2016)) 上评估 DINOv3 的语义分割性能。报告的评价指标是标准化的平均交并比 (mIoU)。

为了评估密集特征的质量，我们在每个基准的训练集上训练一个线性分类器。这个线性层应用于冻结骨干网络的补丁输出特征之上（经过层归一化后），并利用训练好的批归一化层进一步对特征进行归一化。对于所有的骨干网络，我们使用 AdamW 优化器进行超参数搜索，学习率在 $\{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$ 之间变化，权重衰减在 $\{1 \times 10^{-4}, 1 \times 10^{-3}\}$ 之间变化。

D.2 深度估计：线性探测

我们在深度基准 NYUv2 (Silberman et al., 2012) 和 KITTI (Geiger et al., 2013) 数据集上评估 DINOv3 特征在几何任务中的质量。结果使用均方根误差 (RMSE) 指标报告。

为了评估密集特征的质量，我们在每个基准的训练集上训练一个线性分类器。这个线性层应用于冻结主干网络的补丁输出特征（在层归一化之后），特征进一步使用经过训练的批归一化层进行归一化。对于所有主干网络，我们使用 AdamW 优化器进行超参数搜索，学习率变化范围为 $[1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}]$ ，权重衰减变化范围为 $[1 \times 10^{-4}, 1 \times 10^{-3}]$ 。

D.3 3D 关键点匹配

数据集和指标 几何对应应在 NAVI 数据集 (Jampani et al., 2023) 上进行评估，语义对应应在 SPair 数据集 (Min et al., 2019) 上进行评估。对于 NAVI，我们使用调整到边长为 448/512 像素的图像，用于补丁大小为 14/16 的模型。对于 SPair，我们使用调整到边长为 896/1024 像素的图像，用于补丁大小为 14/16 的模型。为了衡量性能，我们报告了对应召回率，即在特定距离内的对应的百分比。*i.e.*

对于 NAVI，我们遵循 Probe3D (Banani et al., 2024) 中定义的协议。具体来说，我们对物体视图进行四分之一的子采样，并且对于每个源视图，选择另一个在最大旋转 120 度范围内的目标视图来创建图像对（源，目标）以进行图块匹配。对于每对图像，每个源（物体内）的图块与目标中的图块进行匹配。保留余弦相似度最高的前 1000 个匹配用于评估，并基于已知的相机姿态和两幅图像的深度图计算每个匹配的 3D 距离误差。这允许计算具有不同阈值的召回错误，我们使用 1 厘米、2 厘米和 5 厘米的阈值。然后，我们计算跨阈值的平均召回率作为对应召回率。

对于每个评估的主干网络，我们使用最后一层的特征，并在应用或不应用最后一层的标准化情况下进行评估。这是因为我们注意到，在应用最后一层标准化时，一些模型的表现不佳。我们报告两种结果中的最大值。

D.4 无监督对象发现

对于此任务，目标是为每张图片生成一个单一的边界框，以突出显示场景中描绘的任何对象。我们遵循 Siméoni et al. (2021) 的无监督对象发现协议，并在检测基准 VOC07 (Everingham et al., 2007)、VOC12 (Everingham et al., 2012) 和 COCO20K (Lin et al., 2014; Vo et al., 2020) 上评估所有骨干网络。COCO20K 是 COCO2014 trainval 数据集 (Lin et al., 2014) 的一个子集，由 (Vo et al., 2020) 中提出，包括 19,817 张随机选择的图像，通常用于该任务。对于每张图像，生成一个边界框。在评估时，我们使用正确定位 (CorLoc) 指标，该指标计算正确定位框的百分比。如果预测框与任何真实边界框的交并比 (IoU) 超过 0.5，则认为该预测框是正确的。

为了评估图像编码器的质量，我们采用了 TokenCut 策略 (Wang et al., 2023c)。该方法将图像块组织成一个完全连接的图，其中边缘代表块对之间的相似度分数，这些分数是使用由 transformer 学习得到的骨干特征计算的。通过应用 Normalized Cut 算法确定显著对象块，这个算法解决一个图切割问题。然后对生成的显著对象掩码进行拟合生成边界框。所有图像均以全分辨率输入编码器，我们使用所有图像编码器的块状输

Table 27: 针对 DAVIS 2017 (Pont-Tuset et al., 2017) 的训练集进行视频分割跟踪评估的超参数列表。选出的最佳超参数适用于所有数据集。

Max context length	Neighborhood mask size	Neighborhood mask shape	Top-K	Temperature
7	12	Square	5	0.2
7	12	Circle	5	0.2
7	5	Square	5	0.2
7	24	Square	5	0.2
7	∞	—	5	0.2
7	12	Square	5	0.01
7	12	Square	5	0.1
7	12	Square	5	0.7
4	12	Square	5	0.2
10	12	Square	5	0.2
15	12	Square	5	0.2
7	12	Square	3	0.2
7	12	Square	10	0.2
7	12	Square	15	0.2
15	12	Circle	10	0.1
15	24	Circle	10	0.1
15	36	Circle	10	0.1
15	∞	—	10	0.1

出特征。为了考虑模型之间特征分布的差异，我们对 TokenCut 的唯一超参数进行扫描：用于图构建的相似度阈值。具体来说，我们将阈值从 0 到 0.4 按 0.05 的增量进行调整。

D.5 视频分割追踪

对于这个任务，我们使用了 DAVIS 2017、YouTube-VOS 和 MOSE 数据集。DAVIS 定义了一个由 60 个视频组成的训练集和一个由 30 个视频组成的验证集，其中所有帧都标注了真实实例分割掩码。对于 YouTube-VOS，只有训练集被标注且公开可用，而验证集则通过评估服务器进行访问。为了模拟 DAVIS 的设置，我们从中随机选取了 2758 个标注过的视频（80%）作为训练集，将剩下的 690 个视频（20%）作为验证集。以类似的方式，我们将 MOSE 数据集分为 1206 个视频用于验证，301 个用于测试。对于所有数据集，我们使用标准的 $\mathcal{J}\&\mathcal{F}$ -均值度量来评估性能，该度量结合了区域相似性 ($\mathcal{J}$) 和轮廓精度 ($\mathcal{F}$) 得分。只有在第一帧中标注的对象会被跟踪和评估，而视频中后出现的对象会被忽略，即使它们的真实掩码被标注了也不例外。

类似于 Rajasegaran et al. (2025)，我们实现了一种基于补丁相似性的标签传播非参数协议，该相似性是通过从冻结的 DINOv3 骨干网上提取的特征之间的余弦相似性计算得出的。我们假设视频的第一帧已经使用实例分割掩码进行了标记，我们将其表示为每个补丁的一个独热向量。对于每一帧，我们计算其所有补丁特征与第一帧的所有补丁以及少量过去帧的所有补丁之间的余弦相似性。专注于当前帧中的单个补丁，我们考虑空间邻域内最相似的 k 个补丁，并对它们的标签进行加权平均以获得对当前补丁的预测。在处理完一帧后，我们移动到下一帧，把之前的预测作为软实例分割标签。在通过骨干网络转发单个帧时，我们调整图像大小，使最短边符合某个尺寸，同时保持纵横比直到接近补丁大小的倍数。¹ 补丁相似性和标签传播是在生成的特征分辨率上计算的，然后掩码概率被双线性调整到原始分辨率以进行 $\mathcal{J}\&\mathcal{F}$ 。我们考虑了几种超参数组合：*e.g.* 使用的过去帧的数量、邻居的数量 k 和空间邻域的大小，如 Tab. 27 中总结。我们在 DAVIS 的训练集上进行超参数选择，然后将最佳组合应用于所有数据集的测试分割。

¹例如，DAVIS 视频本身是 480×854 ，我们希望在分辨率为 960 的情况下处理它们。对于补丁大小为 16 的模型，我们将帧调整为 960×1712 ，稍微向水平方向拉伸，得到一个 60×107 特征图。而对于补丁大小为 14 的模型，我们将帧调整为 966×1708 ，稍微向垂直方向拉伸，得到一个 69×122 特征图。

D.6 视频分类

我们在视频分类任务中使用 UCF101、Something-Something V2 和 Kinetics-400 数据集评估 DINOv3。首先，我们从每个视频中提取固定数量的帧，使用一个冻结的骨干网络对其进行编码，将所有分块特征收集成一个扁平序列，然后将其输入到一个浅层的基于转换器的分类器，该分类器在一组标注视频上接受常规监督学习进行训练。在之前的工作中，这个流程被称为注意探测，类似于用于图像分类的线性探测。在接下来的段落中，我们将描述我们对这一流程的具体实施。

在训练时，我们从每个视频中随机选择 16 帧，记录对应的时间戳。我们还会采样一个空间裁剪的参数，其覆盖范围在 40 % 和 100 % 之间——这些参数将在视频的所有帧中共享，以避免抖动。然后我们将每一帧作为独立的 256×256 图像用 DINOv3 处理，提取 16×16 patch 特征并丢弃 CLS token。对于每个 patch，我们记录在 $[0, 1]^2$ 框上定义的空间坐标。所有帧的 patch 特征以线性方式投影到 1024 维度，连接成一个长度为 $16 \times 16 \times 16 = 4096$ 的平坦序列，然后输入四个自注意力块，建模 patch 之间的空间和时间关系。为确保模型能访问位置信息，我们在这些块之前将每个 patch 的时间戳和空间坐标作为加性的正余弦嵌入注入，并在每个注意力头的 3D 因子化 RoPE 中加入随机空间旋转。经过四个块之后，我们应用一个交叉注意力块，用一个没有位置的可学习查询来聚合所有 patch 的信息为一个单一向量，然后线性投影以获得最终的分 logits。自注意力块堆栈、交叉注意力块、位置嵌入及最终投影构成视频分类器，我们使用标准交叉熵损失进行 20 个 epochs 的训练，批量大小为 64。在实践中，我们并行训练一组分类器，每个分类器对应一个学习率和权重衰减的组合。对于每个数据集，我们使用 90 % 的训练集来更新模型参数，10 % 的训练集来选择最佳的学习率和权重衰减组合，最终在验证分割上报告所选模型的性能。

在推理阶段，我们遵循一种确定性策略从视频中抽取单个剪辑：我们取第一个帧、最后一个帧以及中间均匀间隔的帧，总计 16 帧。对于每个帧，我们裁剪最大的中心方块并将其调整为 256×256 像素，这可能会丢失来自长方形视频两侧的信息。然后我们将这些帧输入 DINOv3 和分类器以获得视频的预测。或者，我们遵循 Assran et al. (2025) 并通过选择多个帧序列和多个空间裁剪进行测试时增强 (TTA)，独立处理这些结果，然后平均分类概率以获得最终预测。剪辑采样在 Fig. 22 中举例说明。

对于选择的基线模型，我们使用相同的评估协议、*i.e.* 特征提取、分类器架构、训练过程和推理协议，只有少许差异。对于使用 16 补丁大小的模型，输入分辨率为 256×256 像素，对于 14 补丁大小则为 224×224 像素。通过这种方式，所有骨干网络生成相同数量的标记，因此在分类器中提供相同数量的计算。由于模型是在图像上训练的，因此所有模型都独立地逐帧处理视频。唯一的例外是 V-JEPA 2，我们为其提供整个剪辑以提取时间感知特征。因为 V-JEPA 2 将时间轴缩小了一倍，从而在输入 16 帧的情况下得到 8 个时间步，我们复制每个补丁标记以匹配其他模型的序列长度。

D.7 使用线性探针的图像分类

数据集和指标 我们使用广泛采用的线性探测评估来评估 DINOv3 模型的整体质量。我们在 ImageNet-1k (Deng et al., 2009) 的训练集上训练一个线性变换，并在验证集上评估结果。通过评估对分类测试集的迁移来评估模型的泛化质量：ImageNet-V2 (Recht et al., 2019) 和 ReaL (Beyer et al., 2020)，它们提供了 ImageNet 的替代图像和标签集，旨在测试对原始 ImageNet 验证集的过拟合。此外，我们将考虑 Rendition (Hendrycks et al., 2021a) 和 Sketch (Wang et al., 2019) 数据集，它们展示了 ImageNet 类别的风格化和人工版本；Adversarial (Hendrycks et al., 2021b) 和 ObjectNet (Barbu et al., 2019) 数据集，其中包含故意具有挑战性的示例；以及 Corruptions (Hendrycks and Dietterich, 2019) 数据集，它评估对常见图像损坏的鲁棒性。我们报告所有数据集的 top-1 分类准确率作为评估指标，但 ImageNet-C 除外，对于 ImageNet-C 我们报告平均损坏错误 (mCE，见 (Hendrycks and Dietterich, 2019))。

对于细粒度数据集，我们考虑从 Oquab et al. (2024) 中收集的 12 个数据集，我们在此称为 Fine-S：Food-101 (Bossard et al., 2014)、CIFAR-10 (Krizhevsky et al., 2009)、CIFAR-100 (Krizhevsky et al., 2009)、SUN397 (Xiao et al., 2010)、StanfordCars (Krause et al., 2013)、FGVC-Aircraft (Maji et al., 2013)、VOC 2007 (Everingham et al., 2007)、DTD (Cimpoi et al., 2014)、Oxford Pets (Parkhi et al., 2012)、Caltech101 (Fei-Fei et al., 2004)、Flowers (Nilsback and Zisserman, 2008) 和 CUB200 (Welinder et al., 2010)，以及更大的数据集 Places205 (Zhou et al., 2014)、iNaturalist 2018 (Van Horn et al., 2018) 和 iNaturalist 2021 (Van Horn et al., 2021)。

对于较大的数据集 ImageNet、Places205、iNaturalist 2018 和 iNaturalist 2021，我们使用以下过程。对于每个基准测试，我们使用 ImageNet-1k 训练集在 CLS token 的最终特征（经过层归一化后）上训练一个线性层。具体来说，我们使用动量为 0.9 的 SGD，并使用批大小为 1024 训练 10 个周期。我们对学习率

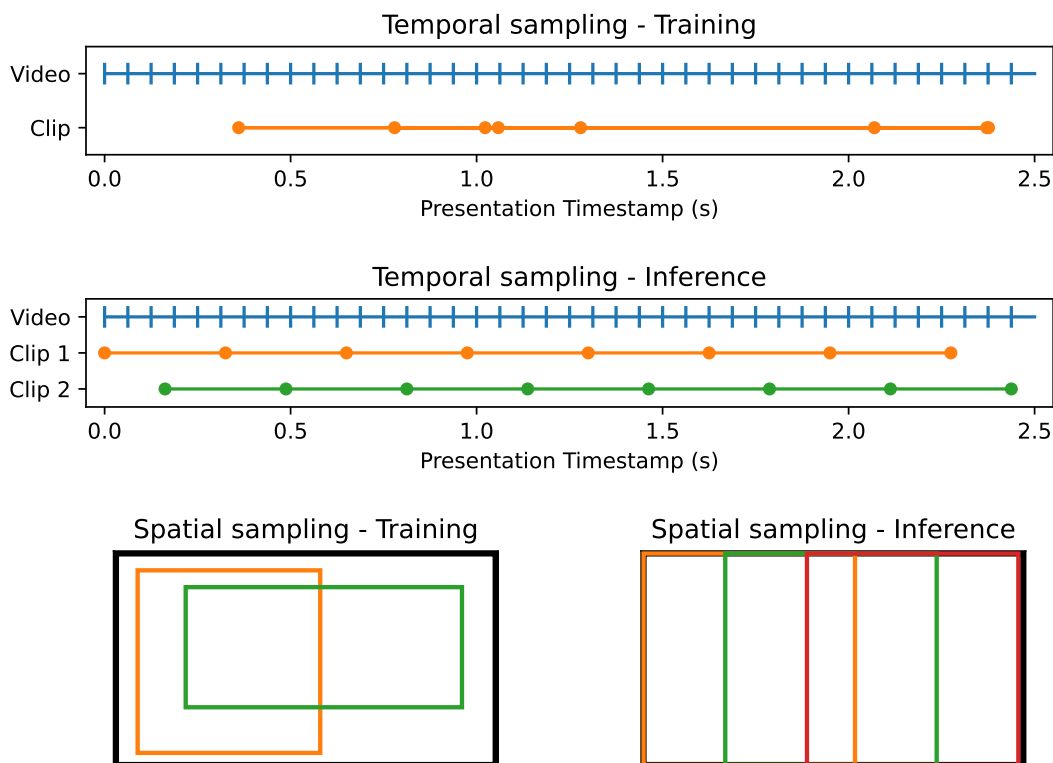


Figure 22: 视频分类的采样片段。为训练或推断选择片段意味着确定空间裁剪的坐标以及要采样的帧/时间戳。在训练时，我们通过从整个视频中随机选择帧并应用覆盖 $\geq 40\%$ 区域的空间裁剪来随机采样片段。在推断时，我们以确定性的方式选择片段。在空间上，我们取三个最大的正方形裁剪，分别对齐于左、中和右。在时间上，我们取两个重叠的帧集，使它们尽可能多地覆盖视频，并且它们的时间戳交错。

$\{1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 2 \times 10^{-1}, 5 \times 10^{-1}, 1 \times 10^0, 2 \times 10^0, 5 \times 10^0\}$ 和权重衰减值 $\{0, 1e-5\}$ 进行调优，并使用 ImageNet-1k 的验证集来选择最佳组合。在训练期间，我们使用带有标准 Inception-crop 参数的随机调整裁剪增强。对于 Fine-S 数据集，按照 [Oquab et al. \(2024\)](#)，我们使用 scitkit-learn 的 LogisticRegression 实现和 L-BFGS 求解器进行较轻量的评估。

在这两种情况下，我们在导致 1024 个 patch token 的分辨率下评估模型，即对于 patch 大小为 14 是 448×448 ，对于 patch 大小为 16 是 512×512 。图像被调整为较短的一边与选择的边长相匹配，然后取中央的正方形部分。

D.8 实例识别

数据集和指标 我们使用牛津和巴黎数据集进行地标识别 ([Radenović et al., 2018](#))，大都会博物馆的艺术作品数据集 Met ([Ypsilantis et al., 2021](#))，以及 AmsterTime，它包含与阿姆斯特丹历史档案图像匹配的现代街景图像 ([Yildiz et al., 2022](#))。在 [Tab. 9](#) 中，我们报告了 Oxford-Hard、Paris-Hard 和 AmsterTime 的平均精准度 (mAP)，以及 Met 的全局精准度 (GAP)。在 [Tab. 23](#) 中，我们额外给出了 Oxford-Medium 和 Paris-Medium 的 mAP，以及附加指标 GAP-和准确性 (见 ([Ypsilantis et al., 2021](#)))。对于牛津和巴黎，我们将所有图像调整为较长边为 224 像素，保持纵横比不变，然后进行中心裁剪，得到分辨率为 224×224 的图像。对于 AmsterTime，我们将所有图像调整为较短边为 256 像素，保持纵横比不变，然后进行大小为

224×224 的中心裁剪。对于 Met，我们在接近原始分辨率的情况下评估所有图像，将两者调整到最接近的补丁大小的倍数（补丁大小为 14/16 时，长边分别为 508/512）。

评估协议 图像相似度是通过计算查询图像和目标图像的 CLS 令牌之间的余弦距离得到的。我们遵循 (Radenović et al., 2018) 对 Oxford 和 Paris 的评估协议、(Yildiz et al., 2022) 对 AmsterTime 的评估协议以及 Ypsilantis et al. (2021) 对 Met 的评估协议。对于 Met，这包括通过网格搜索调整超参数 k 和 τ ，优化 Met 验证集上的 GAP，并使用在 Met 训练集上估计的 PCA 对白化特征。

D.9 目标检测

我们在 COCO (Lin et al., 2014) 和 COCO-O (Mao et al., 2023) 数据集上评估 DINOv3 的物体检测性能。COCO 是一个用于物体检测的标准基准，涵盖 80 个物体类别，包括 118k 张训练图像和 5k 张验证图像。COCO-O 是一个仅用于评估的数据集，类别与 COCO 相同，但在更具挑战性的视觉条件下，例如具有显著遮挡、背景杂乱和光照条件变化的场景。为了训练物体检测模型，我们还利用了 Objects365 (Shao et al., 2019) 数据集，该数据集包含 250 万张图像，涵盖 365 个物体类别，其中一部分可以直接映射到 COCO 类。对于 COCO 和 COCO-O，我们报告在 IoU 阈值为 $[0.5 : 0.05 : 0.95]$ 时计算的平均平均精度 (mAP)。

架构 我们的方法基于 Plain-DETR (Lin et al., 2023b) 的实现，并做了几个修改。我们不会将 transformer 编码器融合到主干中，而是将其作为一个独立的模块，类似于原始 DETR (Carion et al., 2020)。这使我们在训练和推理期间可以完全冻结 DINOv3 主干，从而使其成为第一个做到此的竞争检测模型。从 DINOv3 ViT-7B/16 主干中，我们从四个中间层中提取特征，即 $[10, 20, 30, 40]$ 。对于每个补丁，我们进行通道级别的特征级联，得到的特征维度为 $4 \cdot 4096 = 16384$ ，该维度进一步受到下面描述的窗口策略的增加影响。主干特征输入编码器，该编码器是由嵌入维度为 768 的 6 个自注意力块组成的堆栈。解码器是由相同嵌入维度的 6 个交叉注意力块组成的堆栈，1500 个“一对一”查询和 1500 个“一对多”查询会参与到编码器的补丁令牌中以预测边界框和类别标签。

将训练分为三个阶段，如下所述：一个阶段使用 1536 像素的基础图像分辨率，另外两个阶段使用 2048 像素的基础分辨率。遵循 DETR，我们应用随机水平翻转 ($p = 0.5$)，然后执行以下之一：(i) 随机调整大小，其中最短边在 920 像素（或 1228 像素）和该阶段的基础分辨率（1536 或 2048）之间均匀采样，或 (ii) 随机裁剪，保留原始图像面积的 60% 到 100%，然后按 (i) 中的方式调整大小。在评估时，图像被调整大小，使得最短边为 2048，不进行额外的增强，并将两边都向上取整为补丁尺寸的最接近的倍数。

然后，我们应用一种窗口策略，该策略将图像的全局视图与较小的视图相结合，以使骨干网络能够处理所有尺度的对象。窗口的数量固定为 3×3 个，并根据输入分辨率改变它们的大小。例如，对于大小为 1536×2304 的图像：

1. 图像被分成 3×3 个大小为 512×768 的不重叠窗口。每个窗口通过骨干网络，得到 32×48 个维度为 16384 的补丁标记。所有窗口的特征在空间上重新组装成一个 $(3 \cdot 32) \times (3 \cdot 48)$ 特征图。
2. 整个图像被调整为 512×768 ，并通过骨干网络传递，产生一个维度为 16384 的 32×48 块标记的特征图。然后，这些特征被双线性上采样到 96×144 ，以匹配窗口特征图的大小。
3. 最后，来自步骤 1 和 2 的特征图在通道维度上进行拼接，生成一个维度为 $2 \cdot 16384 = 32768$ 的 96×144 特征图。然后将该特征图展平为一系列 $96 \cdot 144$ 令牌并输入编码器。

训练 我们在三个阶段中遵循一个训练课程，使用 Objects365 数据集 (Shao et al., 2019) 和 COCO 数据集 (Lin et al., 2014) 在逐渐增加的分辨率下进行训练。在整个训练过程中，我们使用 AdamW 优化器 (Loshchilov and Hutter, 2017)，权重衰减为 0.05。按照 DETR 的方法，我们使用 Focal Loss (Lin et al., 2018) 作为分类损失，权重为 2，L1 损失作为边界框损失，权重为 1，并辅以权重为 2 的 GIoU (Rezatofighi et al., 2019) 损失。各阶段如下：

1. 我们在 Objects365 上以基础分辨率 1536 像素开始训练。我们用全局批量大小 32 训练 22 个周期，这个批量大小分布在 32 个 GPU 上。经过最初的 1000 步热身后，学习率被设置为 $5 \cdot 10^{-5}$ ，并在第 20 个周期之后除以 10。
2. 然后我们继续在 Objects365 上以基础分辨率 2048 像素进行训练。我们训练 4 个周期，学习率为 $2.5 \cdot 10^{-5}$ 。

- 我们在基础分辨率 2048 下通过 12 个周期在 COCO 上完成训练。在 2000 次迭代的线性预热后，学习率遵循余弦衰减计划，从 $2.5 \cdot 10^{-5}$ 开始并在第 8 个周期达到 $2.5 \cdot 10^{-6}$ 。在这一部分中，我们使用 IA-BCE 分类损失 (Cai et al., 2024)，而不是 DETR 的简单 Focal Loss。我们观察到这种损失在迁移时能提高模型性能，但在预训练时却没有效果。由于这种损失混合了类别和框的信息，它能在框预测已被良好初始化时发挥其全部潜力。在这一部分中，GIoU 损失权重被设为 4，以鼓励更好的框对齐。

在测试时，我们遵循上述的推理程序，将图像调整为短边为 1536 或 2048。在这些分辨率下，COCO mAP 分别为 65.4 和 65.6。或者，我们可以应用来自 Bolya et al. (2025) 的测试时增强 (TTA) 策略，这包括翻转和将图像调整为多种分辨率，并使用 SoftNMS (Bodla et al., 2017) 合并预测。具体而言，每幅图像在 [1536, 1728, 1920, 2112, 2304, 2496, 2688, 2880] 的分辨率下处理，产生 66.1 的 mAP。

D.10 语义分割

我们在 ADE20k (Zhou et al., 2017)、Cityscapes (Cordts et al., 2016)、COCO-Stuff (Caesar et al., 2018) 和 VOC 2012 (Everingham et al., 2012) 数据集上评估了 DINOv3 在语义分割方面的表现。ADE20k 是一个广泛使用的语义分割基准，涵盖 150 个语义类别，从室外景观到房屋内的人员和物体图像。另外，COCO-Stuff 和 Hypersim (Roberts et al., 2021) 数据集被用于模型的预训练。COCO-Stuff 比 ADE20k 更大的数据集 (11.8 万张训练图像)，包含 80 个事物类和 91 个背景类，而 Hypersim 是一个展示室内场景的逼真合成数据集，具有 40 个语义类别，标注更加清晰准确。超过一半的 Hypersim 图像包含 21 个或更多的对象，使其成为帮助模型学习丰富场景信息的良好候选。所有数据集上报告的评估指标均为 mIoU。

我们调整了其他基线使用的 ViT-Adapter 和 Mask2former 的配置，有几个区别。首先，为了确保我们的主干不被修改并且其激活不被改变，我们移除了 ViT-Adapter 的注入器组件。这使得我们的主干输出特征可以直接在提取模块中使用。其次，将 Mask2former 解码器中的嵌入维度调整为 2048，而不是默认的 1024，以适应我们的主干输出维度 4096，其他基线的主干通常输出维度为 1024 或 1536。作为解码器的输入，我们从 DINOv3 7B/16 主干的四个中间层中提取特征，即层 [10, 20, 30, 40]。我们对所有层的特征应用最终的层归一化，并添加了学习的批归一化。

为了在 COCO-Stuff 上生成结果，我们使用余弦调度器训练模型，进行 6000 次线性暖启动，最大学习率为 $1.5e-5$ 。该模型在分辨率为 1280 像素下训练了 80,000 次迭代。至于在其他数据集上的训练——ADE20k、Cityscapes 和 VOC 2012——我们首先在 COCO-Stuff 上对解码器进行 80,000 次迭代的预训练，进行 6000 次线性暖启动，学习率为 $1.5e-5$ ，并遵循余弦调度器。这有助于模型在比 ADE20k 更大的数据集上学习到多样化的语义类别 (171 个类别)。然后，模型在 Hypersim 上进行 10,000 次迭代训练，学习率为 $2.5e-5$ ，遵循余弦调度器，并进行 1500 次线性暖启动。大约相当于 2 个周期，这一步帮助我们的模型学习高质量的图像到遮罩的对应关系，因为它们具有真实感的合成特性。最后，我们的模型在 ADE20k 上进行了 20,000 次迭代训练，学习率为 $3e-5$ ，同样进行 1500 次线性暖启动和余弦调度器。我们在验证集上报告最终结果。对于 Cityscapes 和 VOC 2012，分别使用 $1.5e-5$ 和 $1e-5$ 的学习率。在所有训练过程中，使用 16 的批量大小和 AdamW 优化器。

对于单尺度评估，使用滑动推理来评估模型——图像首先在训练分辨率下进行调整 (ADE20k 的数据会将具有 400×500 的图像调整为 896×1120 像素大小，因为模型在分辨率 896 下进行训练)。然后，使用滑动窗口方法，在方形裁剪上 (e.g. 为 896×896 像素) 使用步幅 (ADE20k 的步幅为 596 像素)，生成每个裁剪的预测，滑动通过图像。这些结果随后聚合并缩放到原图像大小以生成最终预测。对于测试时增强，ADE20k 和 VOC 2012 的图像被调节到评估分辨率的比例为 [0.9, 0.95, 1.0, 1.05, 1.1]，且每张图像也进行水平翻转，以产生每个样本总共有 10 个预测。在对每张图进行滑动推理后，它们被缩放到原始图像形状并取平均值。COCO-Stuff 164K 的 TTA mIoU 是通过简单地为每个样本使用额外的水平翻转图像获得的，而对于 Cityscapes，使用评估分辨率比例为 [1.0, 1.5, 2.0]。

D.11 单目深度估计

我们的方法与 Depth Anything v2 (DAv2) (Yang et al., 2024b) 的主要区别在于图像分辨率的配置，分辨率设置为 768×1024 pixels，以及网络架构。训练过程中主干网络保持冻结状态，而在 DPT 头部 (Ranftl et al., 2021) 的末尾施加了 0.05 的 dropout 率。作为解码器输入，我们从 DINOv3 7B/16 主干的四个中间层中提取特征，即 [10, 20, 30, 40] 层。我们对所有层的特征应用最终层归一化，并添加了一个学习的批量归一化。深度估计输出被离散化为 256 个均匀分布的箱，覆盖从 0.001m 到 100m 的范围。训练使用基准学习

Table 28: Satlidar 数据集的描述。

Subdataset	Path	Amount of tiles	Purpose
Kalimantan	https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1540	86	train/val/test
OpenDC	https://opendata.dc.gov/datasets/2020-lidar-classified-las/about	68	train/val/test
Brazil	https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1644	37	train/val/test
Mozambique	https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1521	144	train/val/test
Neon	https://data.neonscience.org/data-products/DP3.30015.001	5366	train/val/test
CA20Graup	https://portal.opentopography.org/datasetMetadata?otCollectionID=OT.092021.6339.1	99	train/val/test
CA17Duvall	https://portal.opentopography.org/datasetMetadata?otCollectionID=OT.042020.6339.2	56	train/val/test
Netherlands	https://geotiles.citg.tudelft.nl/	13	train/val/test
Sao Paulo	https://daac.ornl.gov/CMS/guides/LiDAR_Forest_Inventory_Brazil.html	4	test
CA brande	https://doi.org/10.5069/G9C53J18	1	test

率 $1e-3$ ，以 PolyLR 作为调度并设置 3.5 的幂，初始线性预热阶段持续 12k 次迭代。为了增强鲁棒性和泛化性，我们应用了一系列增强：高斯模糊、高斯噪声、自动对比度、自动均衡、颜色抖动、旋转和左右翻转。

我们在 DAv2 的数据集上训练模型，该数据集由来自 IRS、TartanAir、BlendedMVS、Hypersim 和 VKITTI2 的数据集的合成图像组成。我们在五个数据集上进行评估：NYUv2 (Silberman et al., 2012)、KITTI (Geiger et al., 2013)、ETH3D (Schöps et al., 2017)、来自 (Ke et al., 2025) 的 ScanNet 和 DIODE (Vasiljevic et al., 2019)。我们采用零样本尺度不变深度设置，并报告标准指标绝对相对误差和 δ_1 (参见 (Yang et al., 2024a))。

D.12 基于 DINOv3 的视觉几何锚定变换器

实现细节 与原始的 VGGT (Wang et al., 2025) 相比，我们进行了以下更改：(1) 我们将图像尺寸从 518 调整为 592；这是为了匹配 DINOv2 生成的 patch tokens 数量，(2) 采用较小的学习率，具体从 0.0002 调整为 0.0001，(3) 使用 DINOv3 ViT-L 的四个中间层的级联作为下游模块的输入，而不仅仅是最后一层。有趣的是，我们发现使用四个中间层对 DINOv3 有益，而对 DINOv2 则没有带来额外的性能提升。我们还尝试了一个更接近原始 VGGT 设置的版本（图像尺寸为 512，相同的学习率，使用最终层），即使是在没有调优的版本中，也发现在所有测试的基准上均优于原始的 VGGT 工作。

D.13 地理空间

评估细节 在所有评估中，我们保持骨干网络冻结状态，只训练针对任务专门设计的轻量级分类器或解码器。对于 GEO-Bench 分类任务，我们训练一个线性分类器，进行 2400 次迭代，批量大小为 32。我们使用 SGD 优化器、余弦学习率退火，并在 $1e-5$ 到 1 之间选择最佳学习率。除非另有说明，分割评估使用 DPT 解码器 (Ranftl et al., 2021)，学习率基于验证集表现选择，通过对 $[3e-5, 1e-4, 3e-4, 1e-3]$ 中的四个值进行网格搜索。

在 LoveDA 和 iSAID 数据集上，我们训练了 UPerNet 解码器 (Xiao et al., 2018) 进行 80k 次迭代，批量大小为 8，并使用了 1500 次迭代的线性热身，符合 (Wang et al., 2024a) 的要求。所有其他超参数，如裁剪尺寸和权重衰减，与 (Wang et al., 2024a) 中相同。根据之前的工作 (Tolan et al., 2024; Wang et al., 2022a)，我们使用 DPT 头进行冠层高度预测评估，并为物体检测任务训练 Faster RCNN (Ren et al., 2015) 检测器，训练 12 个周期。

Satlidar 数据集 Satlidar 数据集由一百万张 512×512 Maxar 图像及其相应的稠密激光雷达测量组成，这些数据从不同地点收集，如表 Tab. 28 所述。图像是从较大的瓦片中提取的，每个子数据集的瓦片数量在表中有说明。