DINOv3

Oriane Siméoni* **Huy V. Vo*** **Maximilian Seitzer*** **Federico Baldassarre*** **Maxime Oquab***
Cijo Jose **Vasil Khalidov** **Marc Szafraniec** **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 Brandt**[1] **Camille Couprie**
Julien Mairal[2] **Hervé Jégou** **Patrick Labatut** **Piotr Bojanowski**

Meta AI Research [1] *WRI* [2] *Inria*

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

Abstract

Self-supervised learning holds the promise of eliminating the need for manual data annotation, enabling models to scale effortlessly to massive datasets and larger architectures. By not being tailored to specific tasks or domains, this training paradigm has the potential to learn visual representations from diverse sources, ranging from natural to aerial images—using a single algorithm. This technical report introduces DINOv3, a major milestone toward realizing this vision by leveraging simple yet effective strategies. First, we leverage the benefit of scaling both dataset and model size by careful data preparation, design, and optimization. Second, we introduce a new method called Gram anchoring, which effectively addresses the known yet unsolved issue of dense feature maps degrading during long training schedules. Finally, we apply post-hoc strategies that further enhance our models' flexibility with respect to resolution, model size, and alignment with text. As a result, we present a versatile vision foundation model that outperforms the specialized state of the art across a broad range of settings, without fine-tuning. DINOv3 produces high-quality dense features that achieve outstanding performance on various vision tasks, significantly surpassing previous self- and weakly-supervised foundation models. We also share the DINOv3 suite of vision models, designed to advance the state of the art on a wide spectrum of tasks and data by providing scalable solutions for diverse resource constraints and deployment scenarios.

1 Introduction

Foundation models have become a central building block in modern computer vision, enabling broad generalization across tasks and domains through a single, reusable model. Self-supervised learning (SSL) is a powerful approach for training such models, by learning directly from raw pixel data and leveraging the natural co-occurrences of patterns in images. Unlike weakly and fully supervised pretraining methods (Radford et al., 2021; Dehghani et al., 2023; Bolya et al., 2025) which require images paired with high-quality metadata, SSL unlocks training on massive, raw image collections. This is particularly effective for training large-scale visual encoders thanks to the availability of virtually unlimited training data. DINOv2 (Oquab et al., 2024) exemplifies these strengths, achieving impressive results in image understanding tasks (Wang et al., 2025) and enabling pre-training for complex domains such as histopathology (Chen et al., 2024). Models trained with SSL exhibit additional desirable properties: they are robust to input distribution shifts, provide strong global and local features, and generate rich embeddings that facilitate physical scene understanding. Since SSL models are not trained for any specific downstream task, they produce versatile and robust generalist features. For instance, DINOv2 models deliver strong performance across diverse tasks and domains without requiring task-specific finetuning, allowing a single frozen backbone to serve multiple purposes. Importantly, self-supervised learning is especially suitable to train on the vast amount of available observational data in

DINOv3

奥里安·西梅奥尼* **胡·V·Vo*** **马克西米利安·赛策尔*** **费德里科·巴尔达萨雷*** **马克西姆·奥卡布*** **西约·何塞·瓦西尔·哈里多夫** **马克·扎法尼亚茨** **徐胜恩** **米歇尔·拉马蒙吉索亚** **弗朗西斯科·马萨** **丹尼尔·哈齐扎** **卢卡·韦尔斯泰特** **王建元** **蒂莫泰·达塞特** **西奥·穆塔卡尼** **莱昂内尔·森塔纳** **克莱尔·罗伯茨** **安德烈亚·韦达利** **杰米·托兰约翰·布兰德特**[1] **卡米尔·库普里** **朱利安·梅拉尔**[2] **埃尔韦·杰古** **帕特里克·拉巴图** **皮奥特·博亚诺夫斯基**

Meta AI Research [1]瓦[2]*Inria*

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

摘要

自监督学习有望消除手动数据标注的需求，使模型能够轻松扩展到大规模数据集和更大的架构。由于不受特定任务或领域的限制，这种训练范式有潜力从各种来源学习视觉表示，从自然图像到航空图像——使用单一算法。本技术报告介绍了DINOv3，这是实现这一愿景的重要里程碑，通过利用简单而有效的策略。首先，我们通过仔细的数据准备、设计和优化，利用了扩展数据集和模型大小的优势。其次，我们引入了一种名为Gram锚定的新方法，该方法有效解决了密集特征图在长时间训练计划中退化的问题。最后，我们应用了事后策略，进一步增强了我们的模型在分辨率、模型大小以及与文本的alignment方面的灵活性。结果，我们提出了一种通用的视觉基础模型，它在广泛的设置中优于专门的当前最佳技术，而无需微调。DINOv3产生高质量的密集特征，在各种视觉任务上取得了卓越的性能，显著超越了之前的自监督和弱监督基础模型。我们还分享了DINOv3视觉模型套件，旨在通过为不同的资源限制和部署场景提供可扩展的解决方案，在广泛的任务和数据上推进当前最佳技术。

1 引言

基础模型已成为现代计算机视觉的核心构建模块，通过单个可复用的模型实现跨任务和领域的广泛泛化。自监督学习（SSL）是一种强大的方法，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来训练此类模型。与需要图像配对高质量元数据的弱监督和全监督预训练方法（Radford 等人， 2021； Dehghani 等人， 2023； Bolya 等人， 2025）不同，SSL 解锁了在大量原始图像集合上的训练。这得益于几乎无限训练数据的可用性，对于训练大规模视觉编码器特别有效。DINOv2（Oquab 等人， 2024）体现了这些优势，在图像理解任务（王等人， 2025）中取得了令人印象深刻的结果，并能够为病理学等复杂领域进行预训练（陈等人,2024）。使用SSL训练的模型表现出其他额外的理想特性：它们对输入分布偏移具有鲁棒性，提供强大的全局和局部特征，并生成丰富的嵌入，从而促进物理场景理解。由于SSL模型不是为任何特定的下游任务而训练的，因此它们产生通用且鲁棒的一般化特征。例如，DINOv2模型在多样化的任务和领域中提供强大的性能，而无需任务特定的微调，允许单个冻结主干服务于多个目的。重要的是，自监督学习特别适合在病理学（<style id='51'>Vorontsov 等人<style id='53'>， <style id='55'>2024<style id='57'>）、生物学（<style id='59'>金等人<style id='61'>， <style id='63'>2025<style id='65'>）、医学影像（<style id='67'>Pérez- García 等人<style id='69'>， <style id='71'>2025<style id='73'>）、遥感（<style id='75'>Cong 等人<style id='77'>， <style id='79'>2022<style id='81'>； <style id='83'>Tolan 等人<style id='85'>， <style id='87'>2024<style id='89'>）、天文学（<style id='91'>Parker 等人<style id='93'>， <style id='95'>2024<style id='97'>）或高能粒子物理（<style id='99'>Dillon 等人<style id='101'>， <style id='103'>2022<style id='105'>）等缺乏元数据的领域进行训练，并且已经证明这些领域可以从DINOv2等基础模型中受益。最后，SSL无需人工干预，非常适合在网络数据不断增长的环境中进行终身学习。

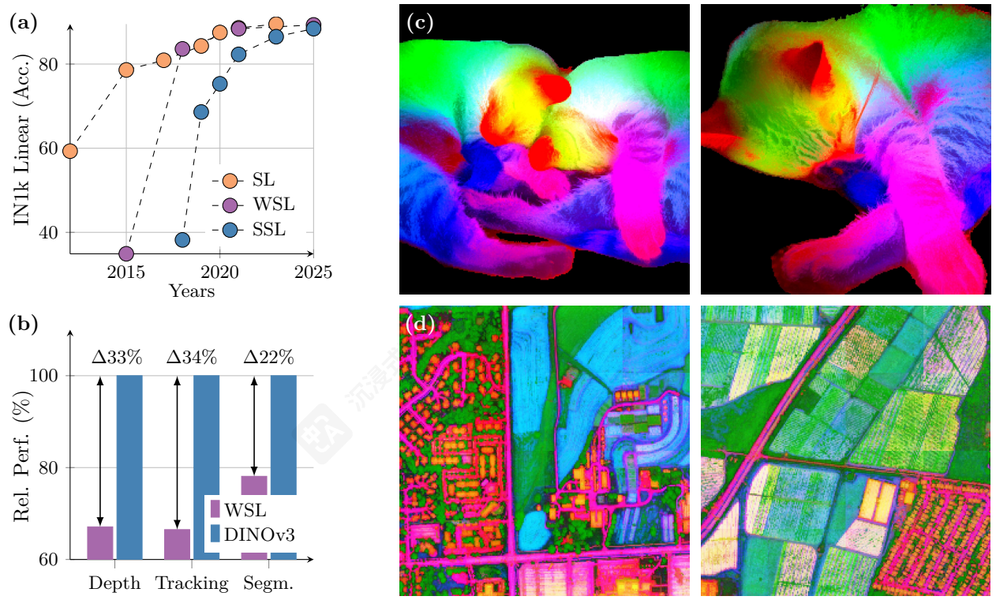


Figure 1: (a) Evolution of linear probing results on ImageNet1k (IN1k) over the years, comparing fully- (SL), weakly- (WSL) and self-supervised learning (SSL) methods. Despite coming into the picture later, SSL has quickly progressed and now reached the Imagenet accuracy plateau of recent years. On the other hand, we demonstrate that SSL offers the unique promise of high-quality dense features. With DINOv3, we markedly improve over weakly-supervised models on dense tasks, as shown by the relative performance of the best-in-class WSL models to DINOv3 (b). We also produce PCA maps of features obtained from high resolution images with DINOv3 trained on natural (c) and aerial images (d).

domains like histopathology (Vorontsov et al., 2024), biology (Kim et al., 2025), medical imaging (Pérez-García et al., 2025), remote sensing (Cong et al., 2022; Tolan et al., 2024), astronomy (Parker et al., 2024), or high-energy particle physics (Dillon et al., 2022). These domain often lack metadata and have already been shown to benefit from foundation models like DINOv2. Finally, SSL, requiring no human intervention, is well-suited for lifelong learning amid the growing volume of web data.

In practice, the promise of SSL, namely producing arbitrarily large and powerful models by leveraging large amounts of unconstrained data, remains challenging at scale. While model instabilities and collapse are mitigated by the heuristics proposed by Oquab et al. (2024), more problems emerge from scaling further. First, it is unclear how to collect useful data from unlabeled collections. Second, in usual training practice, employing cosine schedules implies knowing the optimization horizon a priori, which is difficult when training on large image corpora. Third, the performance of the features gradually decreases after early training, confirmed by visual inspection of the patch similarity maps. This phenomenon appears in longer training runs with models above ViT-Large size (300M parameters), reducing the usefulness of scaling DINOv2.

Addressing the problems above leads to this work, *DINOv3*, which advances SSL training at scale. We demonstrate that a *single frozen SSL backbone* can serve as a universal visual encoder that achieves state-of-the-art performance on challenging downstream tasks, outperforming supervised and metadata-reliant pre-training strategies. Our research is guided by the following objectives: (1) training a foundational model versatile across tasks and domains, (2) improving the shortcomings of existing SSL models on dense features, (3) disseminating a family of models that can be used off-the-shelf. We discuss the three aims in the following.

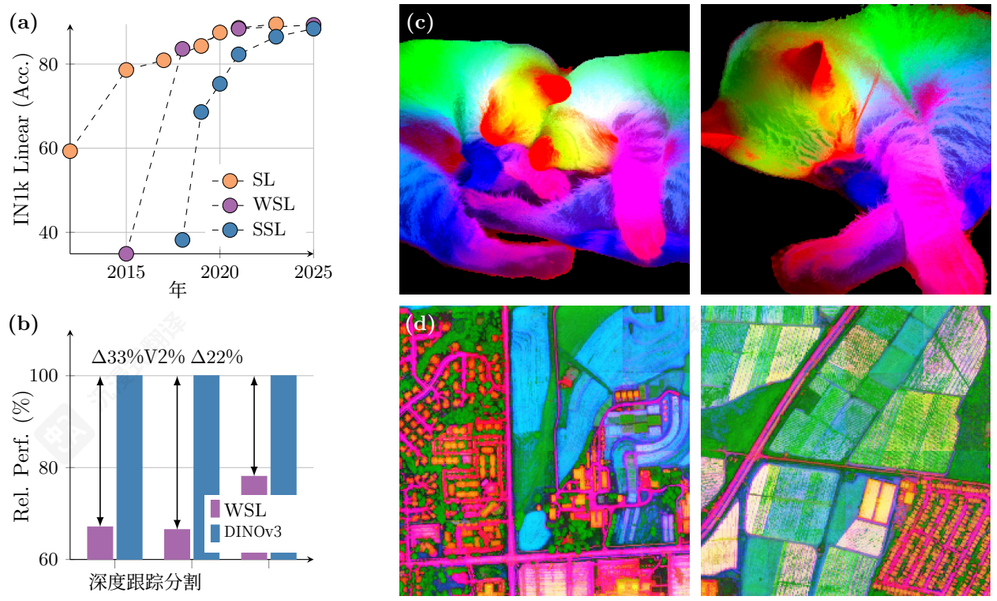
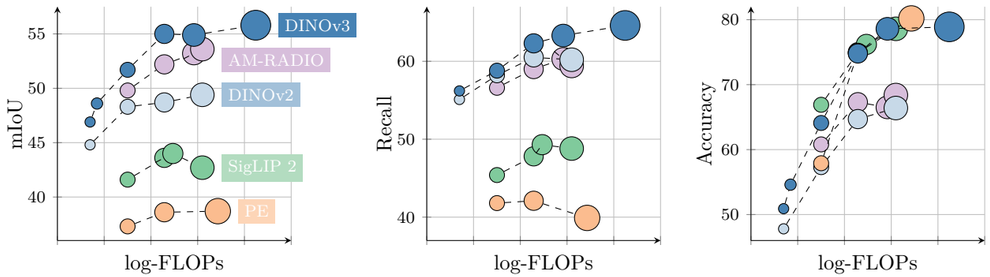


图1: (a) 在ImageNet1k (IN1k) 上多年线性探测结果的演变，比较全监督 (SL)、弱监督 (WSL) 和自监督学习 (SSL) 方法。尽管出现较晚，但SSL已迅速发展，现已达到近年来ImageNet的准确率平台。另一方面，我们证明了SSL提供了高质量密集特征的独特优势。使用DINOv3，我们在密集任务上显著优于弱监督模型，这由顶级WSL模型与DINOv3的相对性能所示 (b)。我们还使用在自然 (c) 和航空图像 (d) 上训练的DINOv3，生成了高分辨率图像特征的PCA图。

像病理学（Vorontsov等人，2024）、生物学（金等人，2025）、医学影像（Pérez- García等人，2025）、遥感（Cong等人，2022；Tolan等人，2024）、天文学（Parker等人，2024）或高能粒子物理（Dillon等人，2022）这样的领域通常缺乏元数据，并且已经证明它们能从DINOv2等基础模型中受益。最后，SSL无需人工干预，非常适合在网页数据不断增长的情况下进行终身学习。

在实践中，SSL的承诺，即通过利用大量无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。虽然模型不稳定性与崩溃可以通过Oquab等人提出的启发式方法来缓解 (2024)，但进一步扩展时更多问题出现。首先，不清楚如何从无标签集合中收集有用数据。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范围，这在训练大型图像语料库时很困难。第三，特征的性能在早期训练后逐渐下降，通过视觉检查补丁相似度图得到证实。这种现象出现在大于ViT-Large大小（300M参数）的更长训练运行中，降低了扩展DINOv2的实用性。

针对上述问题，本研究提出了 *DINOv3*，该工作推动了大规模SSL训练的进展。我们证明，一个单个冻结的*SSL*骨干网络可以作为通用的视觉编码器，在具有挑战性的下游任务上达到最先进的性能，优于有监督和依赖元数据的预训练策略。我们的研究由以下目标指导：(1) 训练一个跨任务和领域通用的基础模型，(2) 改进现有SSL模型在密集特征上的不足，(3) 推广一系列即用型模型。我们将在以下部分讨论这三个目标。



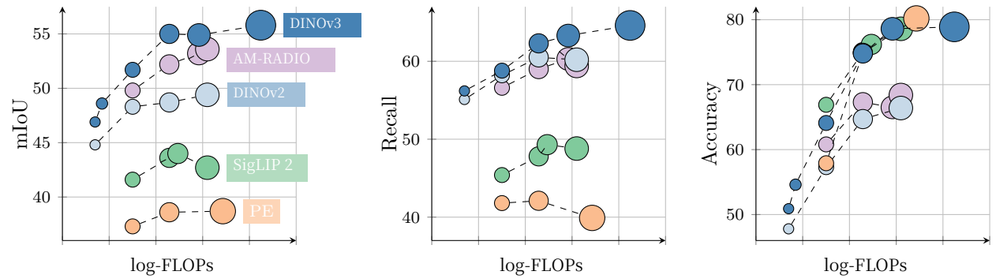
(a) Semantic segmentation (ADE20k) **(b)** 3D keypoint matching (NAVI) **(c)** OOD classif. (ObjectNet)

Figure 2: Performance of the DINOv3 family of models, compared to other families of self- or weakly-supervised models, on different benchmarks. DINOv3 significantly surpasses others on dense benchmarks, including models that leverage mask annotation priors such as AM-RADIO (Heinrich et al., 2025).

Strong & Versatile Foundational Models DINOv3 aims to offer a high level of versatility along two axes, which is enabled by the scaling of the model size and training data. First, a key desirable property for SSL models is to achieve excellent performance while being kept frozen, ideally reaching similar state-of-the-art results as specialized models. In that case, a single forward pass can deliver cutting-edge results across multiple tasks, leading to substantial computational savings—an essential advantage for practical applications, particularly on edge devices. We show the wide breadth of tasks that DINOv3 can successfully be applied to in Sec. 6. Second, a scalable SSL training pipeline that does not depend on metadata unlocks numerous scientific applications. By pre-training on a diverse set of images, whether web images or observational data, SSL models generalize across a large set of domains and tasks. As illustrated in Fig. 1(d), the PCA of DINOv3 features extracted from a high-resolution aerial image clearly allows to separates roads, houses, and greenery, highlighting the model's feature quality.

Superior Feature Maps Through Gram Anchoring Another key feature of DINOv3 is a significant improvement of its dense feature maps. The DINOv3 SSL training strategy aims at producing models excelling at high-level semantic tasks while producing excellent feature maps amenable to solving geometric tasks such as depth estimation, or 3D matching. In particular, the models should produce dense features that can be used off-the-shelf or with little post-processing. The compromise between dense and global representation is especially difficult to optimize when training with vast amounts of images, since the objective of high-level understanding can conflict with the quality of the dense feature maps. These contradictory objectives lead to a collapse of dense features with large models and long training schedules. Our new Gram anchoring strategy effectively mitigates this collapse (see Sec. 4). As a result, DINOv3 obtains significantly better dense feature maps than DINOv2, staying clean even at high resolutions (see Fig. 3).

The DINOv3 Family of Models Solving the degradation of dense feature map with Gram anchoring unlocks the power of scaling. As a consequence, training a much larger model with SSL leads to significant performance improvements. In this work, we successfully train a DINO model with 7B parameters. Since such a large model requires significant resources to run, we apply distillation to compress its knowledge into smaller variants. As a result, we present the *DINOv3 family of vision models*, a comprehensive suite designed to address a wide spectrum of computer vision challenges. This model family aims to advance the state of the art by offering scalable solutions adaptable to diverse resource constraints and deployment scenarios. The distillation process produces model variants at multiple scales, including Vision Transformer (ViT) Small, Base, and Large, as well as ConvNeXt-based architectures. Notably, the efficient and widely adopted ViT-L model achieves performance close to that of the original 7B teacher across a variety of tasks. Overall, the DINOv3 family demonstrates strong performance on a broad range of benchmarks, matching or exceeding the accuracy of competing models on global tasks, while significantly outperforming them on dense prediction tasks, as visible in Fig. 2.



(a) 语义分割 (ADE20k) **(b)** 3D关键点匹配 (NAVI) **(c)** OOD分类 (ObjectNet)

图2： DINOv3模型家族在不同基准测试上的性能，与其他自监督或弱监督模型家族的比较。DINOv3在密集基准测试上显著超越其他模型，包括利用掩码标注先验的AM-RADIO (Heinrich等人， 2025)。

强大且通用的基础模型DINOv3旨在沿两个轴提供高水平的通用性，这得益于模型大小和训练数据的扩展。首先，SSL模型的一个关键期望特性是在保持冻结状态的同时实现优异的性能，理想情况下达到与专用模型相似的先进结果。在这种情况下，单个前向传递即可在多个任务上提供尖端结果，从而带来显著的计算节省——这对于实际应用，尤其是在边缘设备上，是一个重要的优势。我们展示了DINOv3可以成功应用于第6节。其次，一个不依赖元数据的可扩展SSL训练流程为众多科学应用打开了大门。通过在多样化的图像集上进行预训练，无论是网络图像还是观测数据，SSL模型可以在大量领域和任务上进行泛化。如图1(d)所示，从高分辨率航拍图像中提取的DINOv3特征的主成分分析清晰地实现了区分道路、房屋和绿化，突出了模型的特征质量。

通过Gram锚定获得更优特征图 DINOv3的另一个关键特性是其密集特征图的显著改进。DINOv3的SSL训练策略旨在生成在高级语义任务上表现出色的模型，同时生成适合解决几何任务（如深度估计或3D匹配）的优质特征图。特别是，模型应生成可用于即用型或只需少量后处理的密集特征。在用大量图像进行训练时，密集表示和全局表示之间的权衡尤其难以优化，因为高级理解的目标可能与密集特征图的质量相冲突。这些相互矛盾的目标导致在大型模型和长时间的训练计划下密集特征崩溃。我们的新Gram锚定策略有效缓解了这种崩溃（参见 第4节）。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能保持清晰（参见 图3）。

DINOv3模型家族 通过Gram锚定解决密集特征图的退化问题，释放了扩展的潜力。因此，使用SSL训练一个更大的模型可以显著提高性能。在这项工作中，我们成功训练了一个参数量为7B的DINO模型。由于这样一个大型模型需要大量资源来运行，我们应用蒸馏将其知识压缩到更小的变体中。结果，我们提出了*DINOv3*视觉模型家族，这是一套综合解决方案，旨在应对广泛的计算机视觉挑战。该模型家族旨在通过提供可扩展的解决方案来推动当前最佳水平，这些解决方案能够适应不同的资源限制和部署场景。蒸馏过程产生了多个尺度的模型变体，包括视觉Transformer（ViT）小型、基础和大型，以及基于ConvNeXt的架构。值得注意的是，高效且广泛采用的ViT-L模型在各种任务上的性能接近原始7B教师。总体而言，DINOv3模型家族在广泛的基准测试中表现出色，在全局任务上与竞争模型的准确率相当或更高，而在密集预测任务上则显著优于它们，如 图2所示。

Figure 3: High-resolution dense features. We visualize the cosine similarity maps obtained with DINOv3 output features between the patches marked with a red cross and all other patches. Input image at 4096×4096. *Please zoom in*, do you agree with DINOv3?

Overview of Contributions In this work, we introduce multiple contributions to address the challenge of scaling SSL towards a large frontier model. We build upon recent advances in automatic data curation (Vo et al., 2024) to obtain a large "background" training dataset that we carefully mix with a bit of specialized data (ImageNet-1k). This allows leveraging large amounts of unconstrained data to improve the model performance. This contribution **(i)** around data scaling will be described in Sec. 3.1.

We increase our main model size to 7B parameters by defining a custom variant of the ViT architecture. We include modern position embeddings (axial RoPE) and develop a regularization technique to avoid positional artifacts. Departing from the multiple cosine schedules in DINOv2, we train with constant hyperparameter schedules for 1M iterations. This allows producing models with stronger performance. This contribution **(ii)** on model architecture and training will be described in Sec. 3.2.

With the above techniques, we are able to train a model following the DINOv2 algorithm at scale. However, as mentioned previously, scale leads to a degradation of dense features. To address this, we propose a core improvement of the pipeline with a Gram anchoring training phase. This cleans the noise in the feature maps, leading to impressive similarity maps, and drastically improving the performance on both parametric and non-parametric dense tasks. This contribution **(iii)** on Gram training will be described in Sec. 4.

Following previous practice, the last steps of our pipeline consist of a high-resolution post-training phase and distillation into a series of high-performance models of various sizes. For the latter, we develop a novel and

图3： 高分辨率密集特征。我们可视化了使用DINOv3输出特征在标记有红色十字的块与其他所有块之间获得的余弦相似度图。输入图像为4096×4096。 请放大，您同意DINOv3吗?

贡献概述 在这项工作中，我们提出了多个贡献，以应对将SSL扩展到大型前沿模型的挑战。我们基于自动数据管理（Vo 等人， 2024）的最新进展，获得了一个大型“背景”训练数据集，我们小心地将少量专业数据（ImageNet-1k）混合其中。这允许利用大量无约束数据来提高模型性能。这一贡献 **(i)** 围绕数据扩展将在第3.1节中描述。

我们将主模型大小增加到7B参数，通过定义ViT架构的自定义变体。我们包含现代位置嵌入（轴向RoPE），并开发了一种正则化技术来避免位置伪影。与DINOv2中的多个余弦调度不同，我们使用恒定的超参数调度进行1M次迭代训练。这使得能够生成性能更强的模型。这项关于模型架构和训练的贡献将在 **(ii)** Sec. 3.2。

通过上述技术，我们能够在大规模上训练遵循DINOv2算法的模型。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，我们提出了一个带有Gram锚定训练阶段的核心改进流程。这清理了特征图中的噪声，产生了令人印象深刻的相似度图，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训练的贡献将在 **(iii)** Sec. 4。

遵循先前做法，我们管道的最后步骤包括一个高分辨率后训练阶段，以及蒸馏成一系列不同尺寸的高性能模型。对于后者，我们开发了一种新颖的

efficient single-teacher multiple-students distillation procedure. This contribution **(iv)** transfers the power of our 7B frontier model to a family of smaller practical models for common usage, that we describe in Sec. 5.2.

As measured in our thorough benchmarking, results in Sec. 6 show that our approach defines a new standard in dense tasks and performs comparably to CLIP derivatives on global tasks. In particular, *with a frozen vision backbone*, we achieve state-of-the-art performance on longstanding computer vision problems such as object detection (COCO detection, mAP 66.1) and image segmentation (ADE20k, mIoU 63.0), outperforming specialized fine-tuned pipelines. Moreover, we provide evidence of the generality of our approach across domains by applying the DINOv3 algorithm to satellite imagery, in Sec. 8, surpassing all prior approaches.

2 Related Work

Self-Supervised Learning Learning without annotations requires an artificial learning task that provides supervision in lieu for training. The art and challenge of SSL lies in carefully designing these so-called pre-text tasks in order to learn powerful representations for downstream tasks. The language domain, by its discrete nature, offers straightforward ways to set up such tasks, which led to many successful unsupervised pre-training approaches for text data. Examples include word embeddings (Mikolov et al., 2013; Bojanowski et al., 2017), sentence representations (Devlin et al., 2018; Liu et al., 2019), and plain language models (Mikolov et al., 2010; Zaremba et al., 2014). In contrast, computer vision presents greater challenges due to the continuous nature of the signal. Early attempts mimicking language approaches extracted supervisory signals from parts of an image to predict other parts, *e.g.* by predicting relative patch position (Doersch et al., 2015), patch re-ordering (Noroozi and Favaro, 2016; Misra and Maaten, 2020), or inpainting (Pathak et al., 2016). Other tasks involve re-colorizing images (Zhang et al., 2016) or predicting image transformations (Gidaris et al., 2018).

Among these tasks, *inpainting-based* approaches have gathered significant interest thanks to the flexibility of the patch-based ViT architecture (He et al., 2021; Bao et al., 2021; El-Nouby et al., 2021). The objective is to reconstruct corrupted regions of an image, which can be viewed as a form of denoising auto-encoding and is conceptually related to the masked token prediction task in BERT pretraining (Devlin et al., 2018). Notably, He et al. (2021) demonstrated that pixel-based masked auto-encoders (MAE) can be used as strong initializations for finetuning on downstream tasks. In the following, Baevski et al. (2022; 2023); Assran et al. (2023) showed that predicting a *learned latent space* instead of the pixel space leads to more powerful, higher-level features—a learning paradigm called JEPA: "Joint-Embedding Predictive Architecture" (LeCun, 2022). Recently, JEPAs have also been extended to video training (Bardes et al., 2024; Assran et al., 2025).

A second line of work, closer to ours, leverages *discriminative signals between images* to learn visual representations. This family of methods traces its origins to early deep learning research (Hadsell et al., 2006), but gained popularity with the introduction of instance classification techniques (Dosovitskiy et al., 2016; Bojanowski and Joulin, 2017; Wu et al., 2018). Subsequent advancements introduced contrastive objectives and information-theoretic criteria (Hénaff et al., 2019; He et al., 2020; Chen and He, 2020; Chen et al., 2020a; Grill et al., 2020; Bardes et al., 2021), as well as self clustering-based strategies (Caron et al., 2018; Asano et al., 2020; Caron et al., 2020; 2021). More recent approaches, such as iBOT (Zhou et al., 2021), combine these discriminative losses with masked reconstruction objectives. All of these methods show the ability to learn strong features and achieve high performance on standard benchmarks like ImageNet (Russakovsky et al., 2015). However, most face challenges scaling to larger model sizes (Chen et al., 2021).

Vision Foundation Models The deep learning revolution began with the AlexNet breakthrough (Krizhevsky et al., 2012), a deep convolutional neural network that outperformed all previous methods on the ImageNet challenge (Deng et al., 2009; Russakovsky et al., 2015). Already early on, features learned end-to-end on the large manually-labeled ImageNet dataset were found to be highly effective for a wide range of transfer learning tasks (Oquab et al., 2014). Early work on vision *foundation models* then focused on architecture development, including VGG (Simonyan and Zisserman, 2015), GoogleNet (Szegedy et al., 2015), and ResNets (He et al., 2016).

Given the effectiveness of *scaling*, subsequent works explored training larger models on big datasets. Sun et al. (2017) expanded supervised training data with the proprietary JFT dataset containing 300 million

高效单教师多学生蒸馏流程。本工作的贡献（iv）将我们7B前沿模型的功耗转移到一个用于通用场景的更小的实用模型家族中，我们将在第5.2节中对其进行描述。

根据我们详尽的基准测试，第6节的结果表明，我们的方法在密集任务中定义了新的标准，并在全局任务上与CLIP衍生方法表现相当。特别是，使用冻结的视觉主干，我们在目标检测（COCO检测，mAP 66.1）和图像分割（ADE20k，mIoU 63.0）等长期计算机视觉问题上取得了最先进性能，超越了专门微调管道。此外，我们通过将DINOv3算法应用于卫星影像，在第8节中提供了我们方法跨领域普遍性的证据，超越了所有先前方法。

2 相关工作

自监督学习 无标注学习需要一个提供训练监督的人工学习任务。SSL的艺术和挑战在于精心设计这些所谓的预文本任务，以便为下游任务学习强大的表示。语言领域由于其离散性，提供了设置此类任务的直接方法，这导致了许多针对文本数据成功的无监督预训练方法。例如，词嵌入（Mikolov等人，2013；Bojanowski等人，2017）、句子表示（Devlin等人，2018；刘等人，2019）和纯语言模型（Mikolov等人，2010；Zaremba等人，2014）。相比之下，计算机视觉由于信号连续性而面临更大挑战。早期尝试模仿语言方法从图像的一部分提取监督信号来预测其他部分，例如通过预测相对块位置（Doersch等人，2015）、块重排序（Noroozi和Favaro，2016；Misra和Maaten，2020）或图像修复（Pathak等人，2016）。其他任务涉及图像重新着色（Zhang等人，2016）或预测图像变换（Gidaris等人，2018）。

在这些任务中，基于图像修复的方法由于基于块的ViT架构的灵活性而引起了广泛关注（He 等人，2021；Bao 等人，2021；El-Nouby 等人，2021）。其目标是重建图像的损坏区域，这可以被视为一种去噪自编码，并在概念上与BERT预训练中的掩码词预测任务相关（Devlin 等人，2018）。值得注意的是，He 等人（2021）证明了基于像素的掩码自编码器（MAE）可以作为下游任务微调的强初始化。在下文中，Baevski 等人（2022；2023）；Assran 等人（2023）表明，预测学习到的潜在空间而不是像素空间会产生更强大、更高级的特征——一种称为JEPA的学习范式："联合嵌入预测架构"（LeCun，2022）。最近，JEPA也被扩展到视频训练（Bardes 等人，2024；Assran 等人，2025）。

另一条研究线，更接近我们的工作，利用图像之间的判别性信号来学习视觉表示。这一系列方法追溯到早期的深度学习研究（Hadsell等人，2006年），但随着实例分类技术的引入（Dosovitskiy等人，2016年；Bojanowski和Joulin，2017年；Wu等人，2018年）而变得流行。随后的进展引入了对比目标和信息论标准（亨利夫等人，2019年；He等人，2020年；Chen和He，2020年；Chen等人，2020a；Grill等人，2020年；Bardes等人，2021年），以及基于自聚类的策略（卡隆等人，2018年；Asano等人，2020年；卡隆等人，2020年；2021年）。更近期的技术，如iBOT（周等人，2021年），将这些判别性损失与掩码重建目标相结合。所有这些方法都显示出学习强特征并在标准基准测试（如ImageNet（Russakovsky等人，2015年））上实现高性能的能力。然而，大多数方法在扩展到更大的模型大小时面临挑战（Chen等人，2021年）。

视觉基础模型 深度学习革命始于AlexNet的突破（Krizhevsky 等人， 2012），这是一种深度卷积神经网络，在ImageNet挑战（Deng 等人， 2009； Russakovsky 等人， 2015）中超越了所有先前的方法。早在早期，在大规模人工标注的ImageNet数据集上端到端学习到的特征就被发现对广泛的迁移学习任务（Oquab 等人， 2014）非常有效。早期的视觉基础模型 研究主要集中在架构开发上，包括VGG（Simonyan 和 Zisserman，2015），GoogleNet（Szegedy 等人， 2015），和ResNets（He 等人，2016）。

鉴于其有效性，后续研究探索了在大数据集上训练更大模型。 Sun 等人(2017) 使用专有的 JFT 数据集扩展了监督训练数据，该数据集包含 3 亿张

labeled images, showing impressive results. JFT also enabled significant performance gains for Kolesnikov et al. (2020). In parallel, scaling was explored using a combination of supervised and unsupervised data. For instance, an ImageNet-supervised model can be used to produce pseudo-labels for unsupervised data, which then serve to train larger networks (Yalniz et al., 2019). Subsequently, the availability of large supervised datasets such as JFT also facilitated the adaptation of the transformer architecture to computer vision (Dosovitskiy et al., 2020). In particular, achieving performance comparable to that of the original vision transformer (ViT) without access to JFT requires substantial effort (Touvron et al., 2020; 2022). Due to the learning capacity of ViTs, scaling efforts were further extended by Zhai et al. (2022a), culminating in the very large ViT-22B encoder (Dehghani et al., 2023).

Given the complexity of manually labeling large datasets, *weakly-supervised training*—where annotations are derived from metadata associated with images—provides an effective alternative to supervised training. Early on, Joulin et al. (2016) demonstrated that a network can be pre-trained by simply predicting all words in the image caption as targets. This initial approach was further refined by leveraging sentence structures (Li et al., 2017), incorporating other types of metadata and involve curation (Mahajan et al., 2018), and scaling (Singh et al., 2022). However, weakly-supervised algorithms only reached their full potential with the introduction of contrastive losses and the joint-training of caption representations, as exemplified by Align (Jia et al., 2021) and CLIP (Radford et al., 2021).

This highly successful approach inspired numerous *open-source reproductions and scaling efforts*. OpenCLIP (Cherti et al., 2023) was the first open-source effort to replicate CLIP by training on the LAION dataset (Schuhmann et al., 2021); following works leverage pre-trained backbones by fine-tuning them in a CLIP-style manner (Sun et al., 2023; 2024). Recognizing that data collection is a critical factor in the success of CLIP training, MetaCLIP (Xu et al., 2024) precisely follows the original CLIP procedure to reproduce its results, whereas Fang et al. (2024a) use supervised datasets to curate pretraining data. Other works focus on improving the training loss, *e.g.* using a sigmoid loss in SigLIP (Zhai et al., 2023), or leveraging a pre-trained image encoder (Zhai et al., 2022b). Ultimately though, the most critical components for obtaining cutting-edge foundation models are abundant high-quality data and substantial compute resources. In this vein, SigLIP 2 (Tschannen et al., 2025) and Perception Encoder (PE) (Bolya et al., 2025) achieve impressive results after training on more than 40B image-text pairs. The largest PE model is trained on 86B billion samples with a global batch size of 131K. Finally, a range of more complex and natively multimodal approaches have been proposed; these include contrastive captioning (Yu et al., 2022), masked modeling in the latent space (Bao et al., 2021; Wang et al., 2022b; Fang et al., 2023; Wang et al., 2023a), and auto-regressive training (Fini et al., 2024).

In contrast, relatively little work has focused on *scaling unsupervised image pretraining*. Early efforts include Caron et al. (2019) and Goyal et al. (2019) utilizing the YFCC dataset (Thomee et al., 2016). Further progress has been achieved by focusing on larger datasets and models (Goyal et al., 2021; 2022a), as well as initial attempts at data curation for SSL (Tian et al., 2021). Careful tuning of the training algorithms, larger architectures, and more extensive training data lead to the impressive results of DINOv2 (Oquab et al., 2024); for the first time, an SSL model matched or surpassed open-source CLIP variants on a range of tasks. This direction has recently been further pushed by Fan et al. (2025) by scaling to larger models without data curation, or by Venkataramanan et al. (2025) using open datasets and improved training recipes.

Dense Transformer Features A broad range of modern vision applications consume *dense features* of pre-trained transformers, including multi-modal models (Liu et al., 2023; Beyer et al., 2024), generative models (Yu et al., 2025; Yao et al., 2025), 3D understanding (Wang et al., 2025), video understanding (Lin et al., 2023a; Wang et al., 2024b), and robotics (Driess et al., 2023; Kim et al., 2024). On top of that, traditional vision tasks such as detection, segmentation, or depth estimation require accurate local descriptors. To enhance the quality of SSL-trained local descriptors, a substantial body of work focuses on developing *local SSL losses*. Examples include leveraging spatio-temporal consistency in videos, *e.g.* using point track loops as training signal (Jabri et al., 2020), exploiting the spatial alignment between different crops of the same image (Pinheiro et al., 2020; Bardes et al., 2022), or enforcing consistency between neighboring patches (Yun et al., 2022). Darcet et al. (2025) show that predicting clustered local patches leads to improved dense representations. DetCon (Hénaff et al., 2021) and ORL (Xie et al., 2021) perform contrastive learning on

标记图像，并取得了令人印象深刻的结果。JFT 还为 科列斯尼科夫等人(2020) 带来了显著的性能提升。与此同时，研究人员探索了结合监督和无监督数据进行扩展的方法。例如，可以使用 ImageNet 监督模型为无监督数据生成伪标签，这些伪标签随后用于训练更大网络 (Yalniz 等人,2019)。随后，随着 JFT 等大型监督数据集的可用性，Transformer 架构也得以适应计算机视觉 (Dosovitskiy 等人,2020)。特别是，在没有 JFT 的情况下实现与原始视觉 Transformer (ViT) 相当的性能需要大量工作 (Touvron 等人， 2020;2022)。由于 ViTs 具有强大的学习能力，翟等人(2022a) 进一步扩展了扩展工作，最终形成了非常大的 ViT-22B 编码器 (Dehghani 等人， 2023)。

鉴于手动标注大型数据集的复杂性，弱监督训练—其标注来自与图像关联的元数据—为有监督训练提供了一种有效的替代方案。早期，Joulin 等人(2016)证明了网络可以通过简单地预测图像标题中的所有单词作为目标来进行预训练。这种初始方法通过利用句子结构（Li 等人,2017）、结合其他类型的元数据和策展（Mahajan 等人， 2018），以及扩展（Singh 等人， 2022）得到了进一步改进。然而，弱监督算法直到引入对比损失和联合训练标题表示时才充分发挥了其潜力，例如 Align（Jia 等人,2021）和 CLIP（Radford 等人， 2021）。

这种非常成功的方法启发了许多开源复现和扩展工作。Open- CLIP (Cherti 等人，2023) 是第一个通过在 LAION 数据集（Schuhmann 等人，2021）上训练来复制 CLIP 的开源工作；后续工作通过以 CLIP 风格的方式微调预训练主干网络来利用预训练主干网络（Sun 等人，2023； 2024）。认识到数据收集是 CLIP 训练成功的关键因素，MetaCLIP (Xu 等人，2024) 精确地遵循了原始 CLIP 程序以复制其结果，而 Fang 等人 (2024a) 使用监督数据集来策展预训练数据。其他工作专注于改进训练损失，例如在 SigLIP (Zhai 等人，2023) 中使用 Sigmoid 损失，或利用预训练图像编码器（Zhai 等人，2022b）。但最终，获得尖端基础模型的最关键组件是丰富的优质数据和大量的计算资源。在这方面，SigLIP 2 (Tschannen 等人，2025) 和 感知编码器（PE） (Bolya 等人，2025) 在训练超过 40B 图像-文本对后取得了令人印象深刻的结果。最大的 PE 模型在 86B 十亿样本上训练，全局批处理大小为 131K。最后，已经提出了更复杂和原生多模态的方法；这些包括对比标题（Yu 等人，2022）、在潜在空间中的掩码建模（Bao 等人，2021; Wang 等人，2022b； Fang 等人，2023； Wang 等人，2023a），和自回归训练（Fini 等人，2024）。

相比之下，相对较少的工作关注于无监督图像预训练的扩展。早期的努力包括卡隆等人 (2019) 和 高oyal等人 (2019) 利用YFCC数据集 (索米等人, 2016)。通过关注更大的数据集和模型 (高oyal等人,2021;2022a)，以及SSL数据策展的初步尝试 (田等人, 2021)。对训练算法、更大架构和更广泛训练数据的仔细调整带来了 DINOv2 (Oquab 等人,2024); 首次，一个SSL模型在一系列任务上匹配或超越了开源CLIP变体。这一方向最近由 Fan等人 (2025) 通过扩展到更大的模型而不进行数据策展，或由 文卡特拉马南等人 (2025) 使用开放数据集和改进的训练配方进一步推动。

密集Transformer特征 现代视觉应用的广泛范围消耗预训练变换器的密集特征，包括多模态模型（刘等人，2023；贝耶等人，2024）、生成模型（余等人，2025；姚等人，2025）、3D理解（王等人，2025）、视频理解（林等人，2023a；王等人，2024b）和机器人学（德里斯等人，2023；金等人，2024）。在此基础上，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训练的局部描述符的质量，大量研究工作集中于开发局部SSL损失。例如，利用视频中的时空一致性，例如使用点跟踪循环作为训练信号（Jabri等人，2020），利用同一图像的不同裁剪之间的空间对齐（Pinheiro等人，2020; Bardes等人，2022），或强制相邻块之间的一致性（Yun等人，2022）。达尔塞特等人（2025）表明，预测聚集的局部块会导致改进的密集表示。DetCon（亨利夫等人，2021）和ORL（谢等人，2021）在

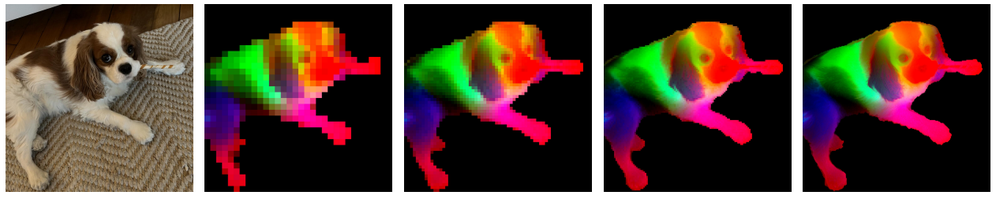


Figure 4: DINOv3 at very high resolution. We visualize dense features of DINOv3 by mapping the first three components of a PCA computed over the feature space to RGB. To focus the PCA on the subject, we mask the feature maps via background subtraction. With increasing resolution, DINOv3 produces crisp features that stay semantically meaningful. We visualize more PCAs in Sec. 6.1.1.

region proposals but assume that such proposals exist *a priori*; this assumption is relaxed by approaches such as ODIN (Hénaff et al., 2022) and SlotCon (Wen et al., 2022). Without changing the training objective, Darcet et al. (2024) show that adding register tokens to the input sequence greatly improves dense feature maps, and recent works find this can be done without model training (Jiang et al., 2025; Chen et al., 2025).

A recent trend are distillation-based, "*agglomerative*" methods that combine information from multiple image encoders with varying in global and local feature quality, trained using different levels of supervision (Ranzinger et al., 2024; Bolya et al., 2025): AM-RADIO (Ranzinger et al., 2024) combines the strengths of the fully-supervised SAM (Kirillov et al., 2023), the weakly-supervised CLIP, and the self-supervised DINOv2 into a unified backbone. The Perception Encoder (Bolya et al., 2025) similarly distills SAM(v2) into a specialized dense variant called PEspatial. They use an objective enforcing cosine similarity between student and teacher patches to be high, where their teacher is trained with mask annotations. Similar losses were shown to be effective in the context of style transfer, by reducing the inconsistency between the Gram matrices of feature dimensions (Gatys et al., 2016; Johnson et al., 2016; Yoo et al., 2024). In this work, we adopt a Gram objective to regularize cosine similarity between student and teacher patches, favoring them being close. In our case, we use earlier iterations of the SSL model itself as the teacher, demonstrating that early-stage SSL models effectively guides SSL training for both global and dense tasks.

Other works focus on post-hoc improvements to the local features of SSL-trained models. For example, Ziegler and Asano (2022) fine-tune a pre-trained model with a dense clustering objective; similarly, Salehi et al. (2023) fine-tune by aligning patch features temporally, in both cases enhance the quality of local features. Closer to us, Pariza et al. (2025) propose a patch-sorting based objective to encourage the student and teacher to produce features with consistent neighbor ordering. Without finetuning, STEGO (Hamilton et al., 2022) learns a non-linear projection on top of frozen SSL features to form compact clusters and amplify correlation patterns. Alternatively, Simoncini et al. (2024) augment self-supervised features by concatenating gradients from different self-supervised objectives to frozen SSL features. Recently, Wysoczańska et al. (2024) show that noisy feature maps are significantly improved through a weighted average of patches.

Related, but not specific to SSL, some recent works generate high-resolution feature maps from ViT feature maps (Fu et al., 2024), which are often low-resolution due to patchification of images. In contrast with this body of work, our models natively deliver high-quality dense feature maps that remain stable and consistent across resolutions, as shown in Fig. 4.

3 Training at Scale Without Supervision

DINOv3 is a next-generation model designed to produce the most robust and flexible visual representations to date by pushing the boundaries of self-supervised learning. We draw inspiration from the success of large language models (LLMs), for which scaling-up the model capacity leads to outstanding *emerging properties*. By leveraging models and training datasets that are an order of magnitude larger, we seek to unlock the full potential of SSL and drive a similar paradigm shift for computer vision, unencumbered by the limitations

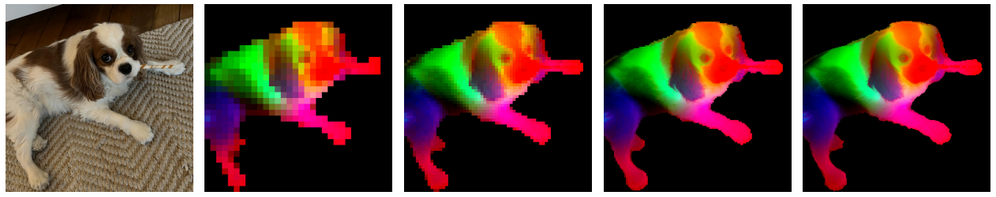


图4： DINOv3在极高分辨率下。我们通过将PCA计算得到的特征空间的前三个分量映射到RGB来可视化DINOv3的密集特征。为了使PCA专注于主体，我们通过背景减法对特征图进行掩码处理。随着分辨率的提高，DINOv3产生的清晰特征仍然保持语义意义。我们在第6.1.1节中展示了更多PCA结果。

区域提议但假设此类提议存在 先验；这种假设通过ODIN（亨利夫等人， 2022）和SlotCon（温等人， 2022）等方法放宽。在不改变训练目标的情况下，达尔塞特等人 ((2024)表明，向输入序列添加注册标记可以大大提高密集特征图，而最近的研究发现这可以在不进行模型训练的情况下完成（蒋等人， 2025；陈等人， 2025）。

近期趋势是蒸馏方法，“聚合”方法，这些方法结合了来自多个图像编码器的信息，这些编码器在全局和局部特征质量上有所不同，并使用不同级别的监督（Ranzinger 等人， 2024； Bolya 等人， 2025）：AM-RADIO (Ranzinger 等人,2024) 结合了全监督SAM (Kirillov 等人,2023) 的优势、弱监督CLIP以及自监督DINOv2，形成一个统一的骨干网络。感知编码器 (Bolya 等人,2025) 类似地将SAM(v2)蒸馏到一个称为PEspatial的专业密集变体中。他们使用一个目标来强制学生和教师块之间的余弦相似度很高，其中他们的教师使用掩码标注进行训练。类似的损失在风格迁移的上下文中已被证明是有效的，通过减少特征维度Gram矩阵之间的一致性（Gatys 等人， 2016； Johnson 等人， 2016； Yoo 等人， 2024）。在这项工作中，我们采用Gram目标来正则化学生和教师块之间的余弦相似度，倾向于使它们接近。在我们的案例中，我们使用SSL模型本身的早期迭代作为教师，证明早期SSL模型有效地指导了全局和密集任务的SSL训练。

其他工作集中于对SSL训练模型的局部特征进行后处理改进。例如，Ziegler 和 Asano (2022) 使用密集聚类目标微调预训练模型；类似地，Salehi 等人 (2023) 通过对齐块特征在时间维度上微调，在两种情况下都提高了局部特征的质量。与我们更接近的是，Pariza 等人 (2025) 提出了一种基于块排序的目标，以鼓励学生和教师生成具有一致邻域排序的特征。在不进行微调的情况下，STEGO (Hamilton 等人， 2022) 在冻结的SSL特征之上学习非线性投影，以形成紧凑的集群并增强相关性模式。或者，Simoncini 等人 (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 等人 (2024) 表明，通过块的重加权平均，噪声特征图得到了显著改进。

相关的，但并非专门针对SSL，一些近期工作从ViT特征图生成高分辨率特征图（Fu 等人， 2024），这些特征图由于图像分块化通常分辨率较低。与此类工作不同，我们的模型原生提供高质量密集特征图，这些特征图在不同分辨率下保持稳定和一致，如图图4所示。

3 大规模训练 无监督

DINOv3 是一种下一代模型，旨在通过突破自监督学习的边界，生成迄今为止最强大和灵活的视觉表示。我们从大型语言模型（LLM）的成功中获得灵感，因为扩大模型容量会导致卓越的 涌现特性。通过利用规模大一个数量级的模型和训练数据集，我们力求充分释放 SSL 的潜力，并为计算机视觉驱动类似的范式转变，不受限于传统监督或特定任务方法的局限性

Table 1: Influence of training data on features quality shown via performance on downstream tasks. We compare datasets curated with *clustering* (Vo et al., 2024) and *retrieval* (Oquab et al., 2024) to *raw* data and to our data mixture. This ablation study is run for a shorter schedule of 200k iterations.

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**

inherent to traditional supervised or task-specific approaches. In particular, SSL produces rich, high-quality visual features that are not biased toward any specific supervision or task, thereby providing a versatile foundation for a wide range of downstream applications. While previous attempts at scaling SSL models have been hindered by issues of instability, this section describes how we harness the benefits of scaling with careful data preparation, design, and optimization. We first describe the dataset creation procedure (Sec. 3.1), then present the self-supervised SSL recipe used for this first training phase of DINOv3 (Sec. 3.2). This includes the choice of architecture, loss functions, and optimization techniques. The second training phase, focusing on dense features, will be described in Sec. 4.

3.1 Data Preparation

Data scaling is one of the driving factors behind the success of large foundation models (Touvron et al., 2023; Radford et al., 2021; Xu et al., 2024; Oquab et al., 2024). However, increasing naively the size of the training data does not necessarily translate into higher model quality and better performance on downstream benchmarks (Goyal et al., 2021; Oquab et al., 2024; Vo et al., 2024): Successful data scaling efforts typically involve careful data curation pipelines. These algorithms may have different objectives: either focusing on improving data *diversity* and *balance*, or data *usefulness*—its relevance to common practical applications. For the development of DINOv3, we combine two complementary approaches to improve both the generalizability and performance of the model, striking a balance between the two objectives.

Data Collection and Curation We build our large-scale pre-training dataset by leveraging a large data pool of web images collected from public posts on Instagram. These images already went through platform-level content moderation to help prevent harmful contents and we obtain an initial data pool of approximately 17 billions of images. Using this raw data pool, we create three dataset *parts*. We construct the first part by applying the automatic curation method based on hierarchical k-means from Vo et al. (2024). We employ DINOv2 as image embeddings, and use 5 levels of clustering with the number of clusters from the lowest to highest levels being 200M, 8M, 800k, 100k, and 25k respectively. After building the hierarchy of clusters, we apply the balanced sampling algorithm proposed in Vo et al. (2024). This results in a curated subset of 1,689 million images (named LVD-1689M) that guarantees a balanced coverage of all visual concepts appearing on the web. For the second part, we adopt a retrieval-based curation system similar to the procedure proposed by Oquab et al. (2024). We retrieve images from the data pool that are similar to those from selected seed datasets, creating a dataset that covers visual concepts relevant for downstream tasks. For the third part, we use raw publicly available computer vision datasets including ImageNet1k (Deng et al., 2009), ImageNet22k (Russakovsky et al., 2015), and Mapillary Street-level Sequences (Warburg et al., 2020). This final part allows us to optimize our model's performance, following Oquab et al. (2024).

Data Sampling During pre-training, we use a sampler to mix different data parts together. There are several different options for mixing the above data components. One is to train with *homogeneous* batches of data that come from a single, randomly selected component in each iteration. Alternatively, we can optimize the model on *heterogeneous* batches that are assembled by data from all components, selected using certain ratios. Inspired by Charton and Kempe (2024), who observed that it is beneficial to have homogeneous batches consisting of very high quality data from a small dataset, we randomly sample in each iteration

表1： 训练数据对特征质量的影响通过下游任务的性能显示。我们比较了使用 聚类 (Vo 等人, 2024)和 检索 (Oquab 等人, 2024)构建的数据集与 原始数据以及我们的数据混合。这项消融研究在200k次迭代的较短调度下运行。

数据集	IN1k k-NN	IN1k线性	ObjectNet	iNaturalist 2021	巴黎检索
Raw	80.1	84.8	70.3	70.1	63.3
聚类	79.4	85.4	72.3	81.3	85.2
检索	84.0	86.7	70.7	86.0	82.7
LVD-1689M（我们）	**84.6**	**87.2**	**72.8**	**87.0**	**85.9**

固有于传统监督或特定任务方法。特别是，SSL 产生丰富、高质量的视觉特征，这些特征不会偏向任何特定的监督或任务，从而为广泛的下游应用提供多功能的基础。虽然以前尝试扩展 SSL 模型受到不稳定性问题的阻碍，但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展的好处。我们首先描述数据集创建程序（第 3.1 节），然后介绍用于 DINOv3 第一次训练阶段的自监督 SSL 配方（第 3.2 节）。这包括架构选择、损失函数和优化技术。第二个训练阶段，专注于密集特征，将在 第 4 节中描述。

3.1 数据准备

数据扩展是大型基础模型成功背后的驱动因素之一（Touvron等人，2023； Radford等人，2021； 徐等人，2024； Oquab 等人， 2024 ）。然而，盲目地增加训练数据的大小并不一定会带来更高的模型质量和在下游基准测试上的更好性能（Goyal等人，2021； Oquab等人，2024； Vo等人，2024）：成功的数据扩展工作通常涉及仔细的数据管理流程。这些算法可能有不同的目标：要么专注于提高数据的多样性和平衡，要么是数据的实用性——它对常见实际应用的相关性。在DINOv3的开发中，我们结合了两种互补的方法来提高模型的泛化能力和性能，在两个目标之间取得了平衡。

数据收集与管理 我们通过利用从Instagram公开帖子收集的大量网络图像数据池来构建我们的大规模预训练数据集。这些图像已经经过平台级别的内容审核，以帮助防止有害内容，我们获得了一个约170亿张图像的初始数据池。使用这个原始数据池，我们创建了三个数据集部分。我们通过应用Vo等人（2024）提出的基于层次 k-means的自动管理方法构建第一部分。我们采用DINOv2作为图像嵌入，并使用5个级别的聚类，从最低到最高级别的簇数量分别为200M、8M、800k、100k和25k。在构建聚类层次结构后，我们应用了Vo等人（2024）提出的平衡采样算法。这产生了一个名为LVD-1689M的1,689百万张图像的管理子集，该子集保证了所有出现在网络上的视觉概念的均衡覆盖。对于第二部分，我们采用了一种类似于Oquab等人（2024）提出的程序的检索式管理系统。我们从数据池中检索与选定的种子数据集相似的图像，创建一个涵盖与下游任务相关的视觉概念的数据集。对于第三部分，我们使用了原始的公开计算机视觉数据集，包括ImageNet1k（Deng等人，2009）、ImageNet22k（Russakovsky等人，2015）和Mapillary街景序列（Warburg等人，2020）。这一最终部分使我们能够根据Oquab等人（2024）的方法优化我们的模型性能。

数据采样 在预训练期间，我们使用采样器将不同的数据部分混合在一起。对于混合上述数据组件，有几种不同的选项。一种是使用 同质的 数据批次进行训练，这些批次来自每个迭代中随机选择的一个组件。或者，我们可以使用来自所有组件的数据（按一定比例选择）组装的 异质的 数据批次来优化模型。受 Charton和Kempe(2024) 启发，他们观察到由小型数据集中非常高质量数据组成的同质批次是有益的，我们在每个迭代中随机采样

Table 2: Comparison of the teacher architectures used in DINOv2 and DINOv3 models. We keep the model 40 blocks deep, and increase the embedding dimension to 4096. Importantly, we use a patch size of 16 pixels, changing the effective sequence length for a given resolution.

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

either a homogeneous batch from ImageNet1k alone or a heterogeneous batch mixing data from all other components. In our training, homogeneous batches from ImageNet1k account for 10% of training.

Data Ablation To assess the impact of our data curation technique, we perform an ablation study to compare our data mix against datasets curated with clustering or retrieval-based methods alone, and the raw data pool. To this end, we train a model on each dataset and compare their performance on standard downstream tasks. For efficiency, we use a shorter schedule of 200k iterations instead of 1M iterations. In Tab. 1, it can be seen that no single curation technique works best across all benchmarks, and that our full pipeline allows us to obtain the best of both worlds.

3.2 Large-Scale Training with Self-Supervision

While models trained with SSL have demonstrated interesting properties (Chen et al., 2020b; Caron et al., 2021), most SSL algorithms have not been scaled-up to larger models sizes. This is either due to issues with training stability (Darcet et al., 2025), or overly simplistic solutions that fail to capture the full complexity of the visual world. When trained at scale (Goyal et al., 2022a), models trained with SSL do not necessarily show impressive performance. One notable exception is DINOv2, a model with 1.1 billion parameters trained on curated data, matching the performance of weakly-supervised models like CLIP (Radford et al., 2021). A recent effort to scale DINOv2 to 7 billion parameters (Fan et al., 2025) demonstrates promising results on global tasks, but with disappointing results on dense prediction. Here, we aim to scale up the model and data, and obtain even more powerful visual representations with both improved global and local properties.

Learning Objective We train the model with a discriminative self-supervised strategy which is a mix of several self-supervised objectives with both global and local loss terms. Following DINOv2 (Oquab et al., 2024), we use an image-level objective (Caron et al., 2021) $\mathcal{L}_{\text{DINO}}$, and balance it with a patch-level latent reconstruction objective (Zhou et al., 2021) $\mathcal{L}_{\text{iBOT}}$. We also replace the centering from DINO with the Sinkhorn-Knopp from SwAV (Caron et al., 2020) in both objectives. Each objective is computed using the output of a dedicated head on top of the backbone network, allowing for some specialization of features before the computation of the losses. Additionally, we use a dedicated layer normalization applied to the backbone outputs of the local and global crops. Empirically, we found this change to stabilize ImageNet kNN-classification late in training (+0.2 accuracy) and improve dense performance (*e.g.* +1 mIoU on ADE20k segmentation, -0.02 RMSE on NYUv2 depth estimation). In addition, a Koleo regularizer $\mathcal{L}_{\text{Koleo}}$ is added to encourage the features within a batch to spread uniformly in the space (Sablayrolles et al., 2018). We use

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

教师模型	DINOv2	DINOv3
骨干网络	ViT-giant	ViT-7B
#Params	1.1B	6.7B
#Blocks	40	40
补丁大小	14	16
位置嵌入	可学习的	RoPE
寄存器	4	4
嵌入. 维度	1536	4096
FFN类型	SwiGLU	SwiGLU
FFN隐藏维度	4096	8192
注意力头	24	32
注意力头维度	64	128
DINO头MLP	4096-4096-256	8192-8192-512
DINO原型	128k	256k
iBOT头部MLP	4096-4096-256	8192-8192-384
iBOT原型	128k	96k

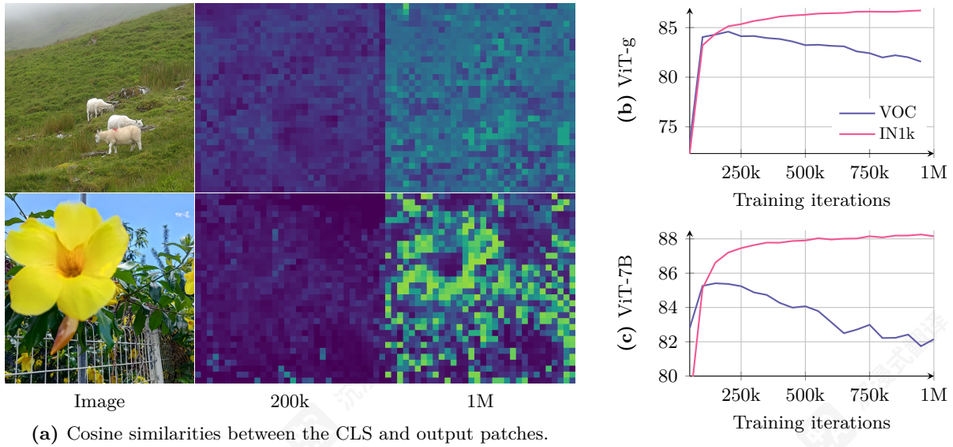
要么仅来自 ImageNet1k 的同质批次，要么混合来自所有其他组件的数据的异质批次。在我们的训练中，来自 ImageNet1k 的同质批次占训练的 10%。

数据消融 为了评估我们的数据管理技术的 影响，我们进行了一项消融研究，以比较我们的数据混合与仅使用聚类或检索方法单独策展的数据集以及原始数据池。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上的性能。为了提高效率，我们使用200k次迭代的较短计划，而不是1M次迭代。在表1中，可以看出没有单一的数据管理技术能在所有基准测试中都表现最佳，而且我们的完整流程使我们能够获得两者的最佳效果。

3.2 大规模自监督训练

虽然使用SSL训练的模型已经展示出一些有趣的特性（陈等人，2020b； 卡隆等人，2021），但大多数SSL算法尚未扩展到更大的模型规模。这是由于训练稳定性问题（达尔塞特等人， 2025），或过于简单的解决方案，这些方案无法捕捉视觉世界的全部复杂性。当在大规模上训练时（高oyal等人， 2022a），使用SSL训练的模型并不一定表现出令人印象深刻的性能。一个值得注意的例外是DINOv2，这是一个拥有11亿参数的模型，在精选数据上进行训练，其性能与弱监督模型如CLIP（Radford等人， 2021）相当。最近的一项努力将DINOv2扩展到70亿参数（Fan等人，2025）在全球任务上取得了有希望的结果，但在密集预测任务上表现令人失望。在这里，我们旨在扩展模型和数据，并获取更强大的视觉表示，同时改进全局和局部特性。

学习目标 我们使用一种判别性自监督策略来训练模型，该策略是几种自监督目标的混合，具有全局和局部损失项。遵循 DINOv2 (Oquab 等人,2024)，我们使用图像级目标 (卡隆等人, 2021)$\mathcal{L}_{\text{DINO}}$，并使用块级潜在重建目标 (周 等人, 2021)$\mathcal{L}_{\text{iBOT}}$进行平衡。我们还用 SwAV 的 Sinkhorn-Knopp (卡隆等人, 2020) 替换了 DINO 中的中心化，在两个目标中都使用了。每个目标都使用骨干网络顶部的专用头的输出进行计算，允许在计算损失之前对特征进行一些专业化。此外，我们使用一个专用层归一化层应用于局部和全局裁剪的骨干网络输出。经验上，我们发现这一变化在训练后期稳定了 ImageNet kNN- 分类 (+0.2 准确率) 并提高了密集性能 (例如。 +1 ADE20k 分割的 mIoU，-0.02 NYUv2 深度估计的 RMSE)。此外，添加了一个 Koleo 正则化器 $\mathcal{L}_{\text{Koleo}}$ 鼓励一批内的特征在空间中均匀分布 (Sablayrolles 等人 , 2018)。我们使用



(a) Cosine similarities between the CLS and output patches.

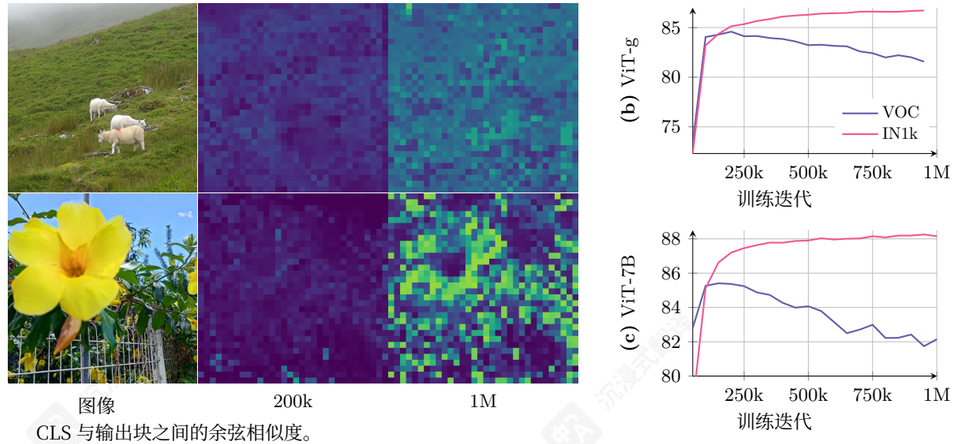
Figure 5: Evolution of the cosine similarities (a) and of the accuracy on ImageNet1k linear (IN1k) and segmentation on VOC for ViT-g (b) and ViT-7B (c). We observe that the segmentation performance is maximal when the cosine similarities between the patch tokens and the class tokens are low. As training progresses, these similarities increase and the performance on dense tasks decreases.

a distributed implementation of Koleo in which the loss is applied in small batches of 16 samples—possibly across GPUs. Our initial training phase is carried by optimizing the following loss:

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

Updated Model Architecture For the model scaling aspect of this work, we increase the size of the model to 7B parameters, and provide in Tab. 2 a comparison of the corresponding hyperparameters with the 1.1B parameter model trained in the DINOv2 work. We also employ a custom variant of RoPE: our base implementation assigns coordinates in a normalized $[-1, 1]$ box to each patch, then applies a bias in the multi-head attention operation depending on the relative position of two patches. In order to improve the robustness of the model to resolutions, scales and aspect ratios, we employ *RoPE-box jittering*. The coordinate box $[-1, 1]$ is randomly scaled to $[-s, s]$, where $s \in [0.5, 2]$. Together, these changes enable DINOv3 to better learn detailed and robust visual features, improving its performance and scalability.

Optimization Training large models on very large datasets represents a complicated experimental workflow. Because the interplay between model capacity and training data complexity is hard to assess *a priori*, it is impossible to guess the right optimization horizon. To overcome this, we get rid of all parameter scheduling, and train with constant learning rate, weight decay, and teacher EMA momentum. This has two main benefits. First, we can continue training as long as downstream performance continues to improve. Second, the number of optimization hyperparameters is reduced, making it easier to choose them properly. For the training to start properly, we still use a linear warmup for learning rate and teacher temperature. Following common practices, we use AdamW (Loshchilov and Hutter, 2017), and set the total batch size to 4096 images split across 256 GPUs. We train our models using the multi-crop strategy (Caron et al., 2020), taking 2 global crops and 8 local crops per image. We use square images with a side length of 256/112 pixels for global/local crops, which, along with the change in patch size, results in the same effective sequence length per image as in DINOv2 and a total sequence length of 3.7M tokens per batch. Additional hyperparameters can be found in App. C and in the code release.



CLS 与输出块之间的余弦相似度。

图5： 余弦相似度的演变（a）以及在ImageNet1k线性（IN1k）和VOC上的分割准确率演变，针对ViT-g（b）和ViT-7B（c）。我们观察到，当块标记和类别标记之间的余弦相似度较低时，分割性能达到最大值。随着训练的进行，这些相似度增加，密集任务的性能下降。

一个分布式 Koleo 实现，其中损失以 16 个样本的小批量应用——可能跨 GPU。我们的初始训练阶段是通过优化以下损失来进行的：

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

更新模型架构在模型扩展方面，我们增加了模型大小至7B参数，并在表2中提供了与DINOv2工作中训练的1.1B参数模型对应的超参数比较。我们还采用了一种自定义的RoPE变体：我们的基础实现将归一化框的坐标分配给每个块，然后在多头注意力操作中根据两个块之间的相对位置应用偏差。为了提高模型对分辨率、尺度和宽高比的鲁棒性，我们采用了*RoPE*框抖动。坐标框被随机缩放到 $[-1, 1]$，其中 $[-s, s]$，其中 $s \in [0.5, 2]$。这些变化共同使DINOv3能够更好地学习详细且鲁棒的视觉特征，提升其性能和可扩展性。

优化 在非常大的数据集上训练大型模型代表了一个复杂的实验工作流程。由于模型容量和训练数据复杂度之间的相互作用很难预先评估，因此不可能猜测正确的优化范围。为了克服这一点，我们摒弃了所有参数调度，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，只要下游性能继续提高，我们就可以继续训练。其次，优化超参数的数量减少，使其更容易正确选择它们。为了使训练能够正常开始，我们仍然使用学习率和教师温度的线性预热。遵循常见做法，我们使用 *AdamW* (Loshchilov 和 Hutter， 2017)，并将总批大小设置为4096 图像分布在 256 GPU 上。我们使用多裁剪策略 (卡隆等人， 2020)，每个图像取 2全局裁剪和 8 局部裁剪。我们使用边长为256/112 像素的全局/局部裁剪，这与 DINOv2 中的补丁大小变化一起，导致每张图像的有效序列长度相同，总序列长度为 3。7M tokens 每批。其他超参数可以在 附录 C和代码发布中找到。

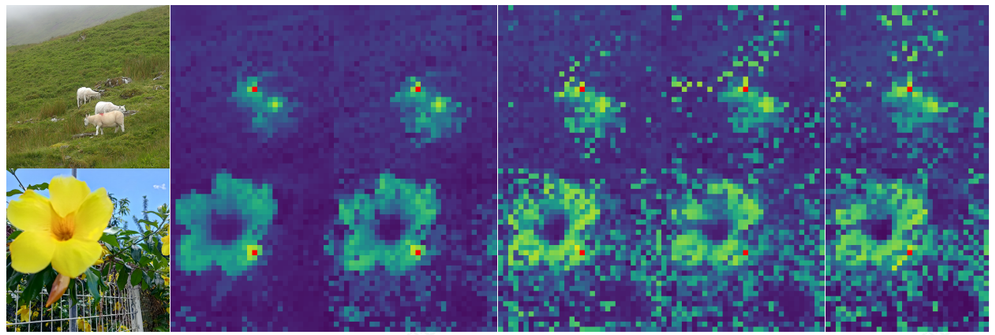


Figure 6: Evolution of the cosine similarity between the patch noted in red and all other patches. As training progresses, the features produced by the model become less localized and the similarity maps become noisier.

4 Gram Anchoring: A Regularization for Dense Features

To fully leverage the benefits of large-scale training, we aim to train the 7B model for an extended duration, with the notion that it could potentially train indefinitely. As expected, prolonged training leads to improvements on global benchmarks. However, as training progresses, the performance degrades on dense tasks (Figs. 5b and 5c). This phenomenon, which is due to the emergence of patch-level inconsistencies in feature representations, undermines the interest behind extended training.[1] In this section, we first analyze the loss of patch-level consistency, then propose a new objective to mitigate it, called *Gram anchoring*. We finally discuss the impact of our approach on both training stability and model performance.

4.1 Loss of Patch-Level Consistency Over Training

During extended training, we observe consistent improvements in global metrics but a notable decline in performance on dense prediction tasks. This behavior was previously observed, to a lesser extent, during the training of DINOv2, and also discussed in the scaling effort of Fan et al. (2025). However, to the best of our knowledge, it remains unresolved to date. We illustrate the phenomenon in Figs. 5b and 5c, which present the performance of the model across iterations on both image classification and segmentation tasks. For classification, we train a linear classifier on ImageNet-1k using the CLS token and report top-1 accuracy. For segmentation, we train a linear layer on patch features extracted from Pascal VOC and report mean Intersection over Union (mIoU). We observe that both for the ViT-g and the ViT-7B, the classification accuracy monotonically improves throughout training. However, segmentation performance declines in both cases after approximately 200k iterations, falling below its early levels in the case of the ViT-7B.

To better understand this degradation, we analyze the quality of patch features by visualizing cosine similarities between patches. Fig. 6 shows the cosine similarity maps between the backbone's output patch features and a reference patch (highlighted in red). At 200k iterations, the similarity maps are smooth and well-localized, indicating consistent patch-level representations. However, by 600k iterations and beyond, the maps degrade substantially, with an increasing number of irrelevant patches with high similarity to the reference patch. This loss of patch-level consistency correlates with the drop in dense task performance.

These patch-level irregularities differ from the high-norm patch outliers described in Darcet et al. (2024). Specifically, with the integration of register tokens, patch norms remain stable throughout training. However, we notice that the cosine similarity between the CLS token and the patch outputs gradually increases during training. This is expected, yet it means that the locality of the patch features diminishes. We visualize this phenomenon in Fig. 5a, which depicts the cosine maps at 200k and 1M iterations. In order to mitigate the

[1]We also observed different types of outliers appearing with continued training; we provide a discussion in App. A.

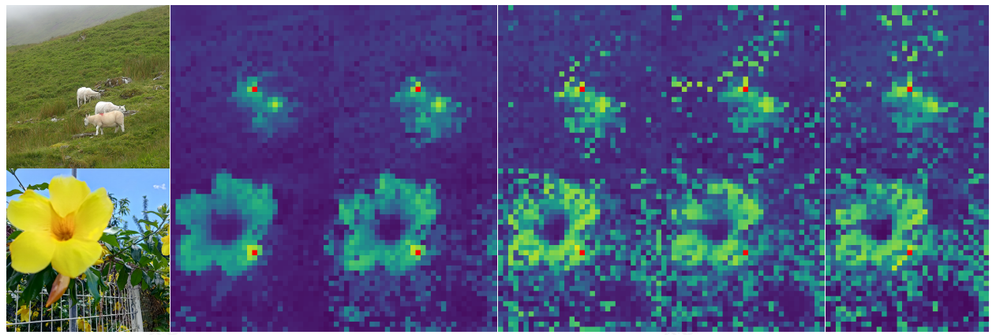


图6： 红色标注的块与其他所有块之间的余弦相似度演化。随着训练的进行，模型产生的特征变得不那么局部化，相似度图变得越加嘈杂。

4 Gram锚定：密集特征的正则化

为了充分利用大规模训练的好处，我们旨在训练 7B 模型更长时间，并认为它有可能无限期地训练下去。正如预期的那样，长时间的训练导致全局基准测试的性能提高。然而，随着训练的进行，密集任务 (图 5b 和 5c) 的性能下降。这种现象是由于特征表示中出现了块级不一致性而导致的，这削弱了延长训练的兴趣。[1] 在本节中，我们首先分析块级一致性的损失，然后提出一个新的目标来减轻它，称为 *Gram* 锚定。最后，我们讨论我们的方法对训练稳定性和模型性能的影响。

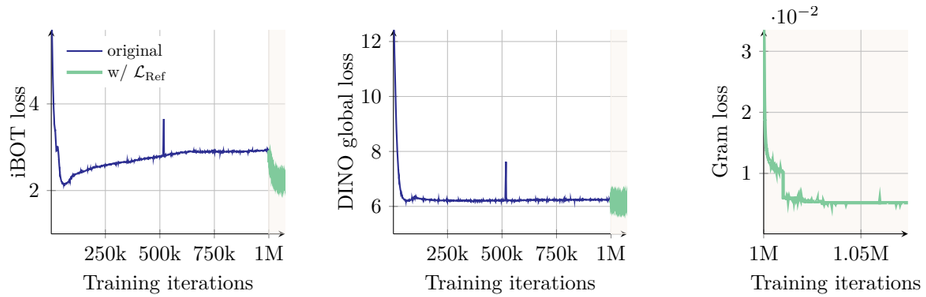
4.1 训练过程中的块级一致性损失

在扩展训练中，我们观察到全局指标持续提升，但密集预测任务的性能显著下降。这种行为之前在DINOv2的训练中程度较轻时被观察到，并在Fan等人⁽²⁰²⁵⁾的扩展工作中讨论过。然而，据我们所知，至今仍未解决。我们在图5b和图5c中展示了该现象，这些图展示了模型在图像分类和分割任务上迭代过程中的性能。对于分类，我们使用CLS标记在ImageNet-1k上训练线性分类器，并报告top-1准确率。对于分割，我们在从Pascal VOC提取的块特征上训练线性层，并报告平均交并比（mIoU）。我们观察到，对于ViT-g和ViT-7B，分类准确率在整个训练过程中单调提升。然而，在两种情况下，分割性能在约200k次迭代后下降，在ViT-7B的情况下低于早期水平。

为了更好地理解这种退化，我们通过可视化补丁之间的余弦相似度来分析补丁特征的质量。 图6 显示了骨干网络输出补丁特征与参考块（红色高亮）之间的余弦相似度图。在200k次迭代时，相似度图平滑且定位良好，表明具有一致的补丁级表示。然而，在600k次迭代及以后，地图显著退化，与参考块具有高相似度的无关补丁数量不断增加。补丁级一致性的丧失与密集任务性能的下降相关。

这些 patch 级别的非规则性不同于在达尔塞特等人(2024)中描述的高范数 patch 离群值。具体来说，随着注册标记的集成，patch 范数在整个训练过程中保持稳定。然而，我们注意到 CLS 标记与 patch 输出之间的余弦相似度在训练过程中逐渐增加。这是可以预料的，但这意味着 patch 特征的局部性减弱。我们在图 5a中可视化了这一现象，该图描绘了在 200k 和 1M 迭代时的余弦图。为了减轻在密集任务上的性能下降，我们提出了一种新的目标，专门设计用于正则化 patch 特征并确保良好的 patch 级一致性，同时保留高全局性能。

[1]我们还观察到在持续训练过程中出现了不同类型的离群值；我们在 附录A中提供了讨论。



(a) iBOT loss **(b)** DINO global loss **(c)** Gram loss

Figure 7: Evolution trough the training iterations of the patch-level iBOT loss, the global loss DINO (applied to the global crops) and the newly introduced Gram loss. We highlight the iterations of the refinement step $\mathcal{L}_{\text{Ref}}$ which uses the Gram objective.

drop on dense tasks, we propose a new objective specifically designed to regularize the patch features and ensure a good patch-level consistency, while preserving high global performance.

4.2 Gram Anchoring Objective

Throughout our experiments, we have identified a relative independence between learning strong discriminative features and maintaining local consistency, as observed in the lack of correlation between global and dense performance. While combining the global DINO loss with the local iBOT loss has begun to address this issue, we observe that the balance is unstable, with global representation dominating as training progresses. Building on this insight, we propose a novel solution that explicitly leverages this independence.

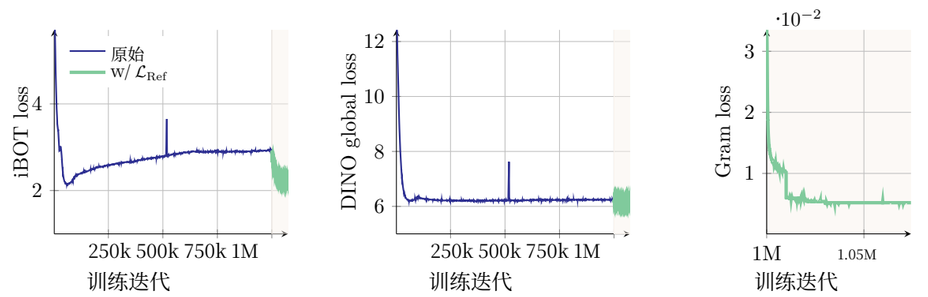
We introduce a new objective which mitigates the degradation of patch-level consistency by enforcing the quality of the patch-level consistency, without impacting the features themselves. This new loss function operates on the Gram matrix: the matrix of all pairwise dot products of patch features in an image. We want to push the Gram matrix of the student towards that of an earlier model, referred to as the *Gram teacher*. We select the Gram teacher by taking an early iteration of the teacher network, which exhibits superior dense properties. By operating on the Gram matrix rather than the feature themselves, the local features are free to move, provided the structure of similarities remains the same. Suppose we have an image composed of P patches, and a network that operates in dimension d. Let us denote by $\mathbf{X}_S$ (respectively $\mathbf{X}_G$) the $P \times d$ matrix of $\mathbf{L}_2$-normalized local features of the student (respectively the Gram teacher). We define the loss $\mathcal{L}_{\text{Gram}}$ as follows:

$$\mathcal{L}_{\text{Gram}} = \left\| \mathbf{X}_S \cdot \mathbf{X}_S^\top - \mathbf{X}_G \cdot \mathbf{X}_G^\top \right\|_{\text{F}}^2 . \tag{2}$$

We only compute this loss on the global crops. Even though it can be applied early on during the training, for efficiency, we start only after 1M iterations. Interestingly, we observe that the late application of $\mathcal{L}_{\text{Gram}}$ still manages to "repair" very degraded local features. In order to further improve performance, we update the Gram teacher every 10k iterations at which the Gram teacher becomes identical to the main EMA teacher. We call this second step of training the *refinement step*, which optimizes the objective $\mathcal{L}_{\text{Ref}}$, with

$$\mathcal{L}_{\text{Ref}} = w_{\text{D}}\mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{\text{DK}}\mathcal{L}_{\text{DKoleo}} + w_{\text{Gram}}\mathcal{L}_{\text{Gram}}. \tag{3}$$

We visualize the evolution of different losses in Fig. 7 and observe that applying the Gram objective significantly influences the iBOT loss, causing it to decrease more rapidly. This suggests that the stability introduced by the stable Gram teacher positively impacts the iBOT objective. In contrast, the Gram objective does not have a significant effect on the DINO losses. This observation implies that the Gram and iBOT objectives impact the features in a similar way, whereas the DINO losses affect them differently.



(a) iBOT损失 **(b)** DINO全局损失 **(c)** Gram损失

图7： 在训练迭代过程中，patch级iBOT损失、全局损失DINO（应用于全局裁剪）以及新引入的Gram损失的演变。我们突出显示了使用Gram目标的精炼步骤$\mathcal{L}_{\text{Ref}}$ 的迭代。

在密集任务上降低性能，我们提出了一种新的目标，专门设计用于正则化 patch 特征并确保良好的 patch 级一致性，同时保留高全局性能。

4.2 Gram锚定目标

在我们的实验中，我们发现学习强判别性特征与保持局部一致性之间存在相对独立性，正如全局性能和密集性能之间缺乏相关性所观察到的那样。虽然将全局DINO损失与局部iBOT损失相结合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，全局表示占主导地位。基于这一见解，我们提出了一种新的解决方案，该方案明确利用了这种独立性。

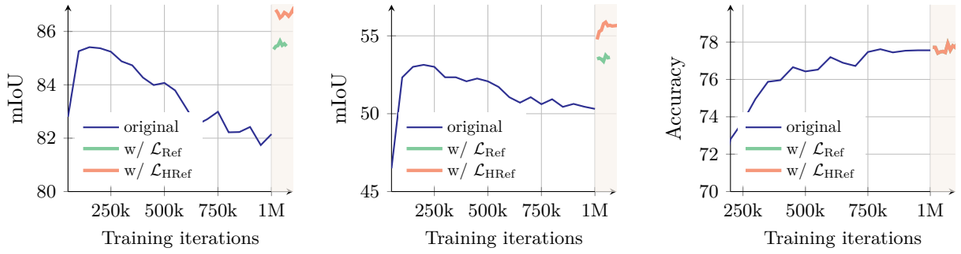
我们引入了一个新的目标，该目标通过强制执行patch级一致性的质量来减轻patch级一致性的退化，而不会影响特征本身。这个新的损失函数作用于Gram矩阵：图像中所有patch特征的两两点积矩阵。我们希望将学生的Gram矩阵推向一个早期模型，该模型称为***Gram***教师。我们通过选择教师网络的早期迭代来选择Gram教师，该迭代表现出优越的密集特性。通过作用于Gram矩阵而不是特征本身，局部特征可以自由移动，前提是相似性的结构保持不变。假设我们有一个由 P 个patch组成的图像，以及一个在维度 d中操作的网络。让我们用$\mathbf{X}_S$ （分别地$\mathbf{X}_G$）表示学生（分别地Gram教师）的 $P \times d$ 个$\mathbf{L}_2$-归一化局部特征的矩阵。我们定义损失$\mathcal{L}_{\text{Gram}}$ 如下：

$$\mathcal{L}_{\text{Gram}} = \left\| \mathbf{X}_S \cdot \mathbf{X}_S^\top - \mathbf{X}_G \cdot \mathbf{X}_G^\top \right\|_{\text{F}}^2 . \tag{2}$$

我们只在全局裁剪上计算这个损失。尽管它可以在训练早期应用，但为了效率，我们只在1M次迭代后开始。有趣的是，我们观察到晚期应用 $\mathcal{L}_{\text{Gram}}$ 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次迭代更新一次Gram教师，此时Gram教师与主EMA教师相同。我们将这个训练的第二步称为 精炼步骤，它优化了目标 $\mathcal{L}_{\text{Ref}}$，与

$$\mathcal{L}_{\text{Ref}} = w_{\text{D}}\mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{\text{DK}}\mathcal{L}_{\text{DKoleo}} + w_{\text{Gram}}\mathcal{L}_{\text{Gram}}. \tag{3}$$

我们可视化不同损失在图7中的演化，并观察到应用Gram目标显著影响了iBOT损失，使其下降得更快。这表明由稳定Gram教师引入的稳定性对iBOT目标产生了积极影响。相比之下，Gram目标对DINO损失没有显著影响。这一观察表明，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它们。



(a) VOC **(b)** ADE20k **(c)** ObjectNet

Figure 8: Evolution of the results on different benchmarks after applying our proposed *Gram anchoring* method. We visualize results when continuing the original training with our refinement step, noted '$\mathcal{L}_{\text{Ref}}$'. We also plot results obtained when using higher-resolution features for the Gram objective as introduced in following Sec. 4.3 and noted '$\mathcal{L}_{\text{HRef}}$'. We highlight the iterations which use the Gram objective.

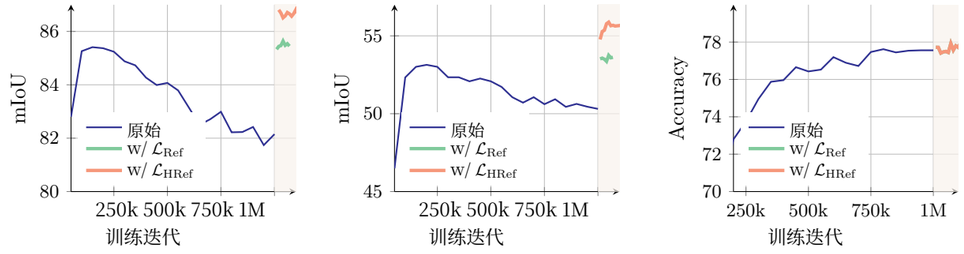
Regarding performance, we observe the impact of the new loss is almost immediate. As shown in Fig. 8, incorporating Gram anchoring leads to significant improvements on dense tasks within the first 10k iterations. We also see notable gains on the ADE20k benchmark following the Gram teacher updates. Additionally, longer training further benefits performance on the ObjectNet benchmark and other global benchmarks show mild impact from the new loss.

4.3 Leveraging Higher-Resolution Features

Recent work shows that a weighted average of patch features can yield stronger local representations by smoothing outlier patches and enhancing patch-level consistency (Wysoczańska et al., 2024). On the other hand, feeding higher-resolution images into the backbone produces finer and more detailed feature maps. We leverage the benefits of both observations to compute high-quality features for Gram teacher. Specifically, we first input images at twice the normal resolution into the Gram teacher, then 2× down-sample the resulting feature maps with the bicubic interpolation to achieve the desired smooth feature maps that match the size of the student output. Fig. 9a visualizes the Gram matrices of patch features obtained with images at resolutions 256 and 512, as well as those obtained after 2× down-sampling features from the 512-resolution (denoted as 'downsamp.'). We observe that the superior patch-level consistency in the higher-resolution features is preserved through down-sampling, resulting in smoother and more coherent patch-level representations. As a side note, our model can seamlessly process images at varying resolutions without requiring adaptation, thanks to the adoption of Rotary Positional Embeddings (RoPE) introduced by Su et al. (2024).

We compute the Gram matrix of the down-sampled features and use it to replace $\mathbf{X}_G$ in the objective $\mathcal{L}_{\text{Gram}}$. We note the new resulting refinement objective as $\mathcal{L}_{\text{HRef}}$. This approach enables the Gram objective to effectively distill the improved patch consistency of smoothed high-resolution features into the student model. As shown in Fig. 8 and Fig. 9b, this distillation translates into better predictions on dense tasks, yielding additional gains on top of the benefit brought by $\mathcal{L}_{\text{Ref}}$ (+2 mIoU on ADE20k). We also ablate the choice of Gram teacher in Fig. 9b. Interestingly, choosing the Gram teacher from 100k or 200k does not significantly impact the results, but using a much later Gram teacher (1M iterations) is detrimental because the patch-level consistency of such a teacher is inferior.

Finally, we qualitatively illustrate the effect of Gram anchoring to patch-level consistency in Fig. 10 which visualizes the Gram matrices patch features obtained with the initial training and high-resolution Gram anchoring refinement. We observe great improvements in feature correlations that our high-resolution refinement procedure brings about.



(a) VOC (b) ADE20k (c) ObjectNet

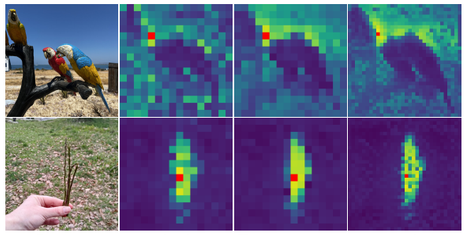
图8：应用我们提出的Gram锚定方法后，在不同基准测试上结果的变化。我们可视化使用我们的精炼步骤继续原始训练时的结果，记为"L_{Ref}"。我们还绘制了使用更高分辨率特征进行Gram目标的结果，如后面第4.3节所述，记为"L_{HRef}"。我们突出了使用Gram目标的迭代。

4.3 利用更高分辨率的特征

近期研究表明，对补丁特征进行加权平均可以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 等人， 2024）。另一方面，将更高分辨率的图像输入骨干网络会产生更精细和更详细的特征图。我们利用这两种观察结果的优势来为Gram教师计算高质量的特征。具体来说，我们首先将正常分辨率两倍的输入图像输入Gram教师，然后 2× 使用双三次插值对生成的特征图进行下采样，以获得与学生输出大小匹配的所需平滑特征图。 图9a 展示了使用分辨率为256和512的图像获得的补丁特征Gram矩阵，以及 2× 下采样512分辨率（记为 'downsamp.' ）后的特征。我们观察到，通过下采样，更高分辨率特征中的优异补丁级一致性得以保留，从而产生了更平滑和更连贯的补丁级表示。作为补充说明，我们的模型可以无缝处理不同分辨率的图像，而无需适应，这得益于 Su 等人(2024）引入的旋转位置嵌入（RoPE）。

我们计算下采样特征的Gram矩阵，并使用它来替换 $\mathbf{X}_G$ 在目标 $\mathcal{L}_{\text{Gram}}$中。我们注意到新的结果细化目标为 $\mathcal{L}_{\text{HRef}}$。这种方法使 Gram 目标能够有效地将平滑高分辨率特征的改进块一致性蒸馏到学生模型中。如图 图8 和 图9b所示，这种蒸馏转化为在密集任务上的更好预测，并在 $\mathcal{L}_{\text{Ref}}$ (+2 ADE20k上的mIoU)带来的益处之上产生了额外增益。我们还消融了 图9b中的Gram教师选择。有趣的是，从100k或200k中选择Gram教师对结果没有显著影响，但使用一个更晚的Gram教师（1M次迭代）是有害的，因为这种教师的块级一致性较差。

最后，我们定性地说明了 Gram 锚定对 patch 级一致性的影响，在图10中可视化了通过初始训练和高分辨率 Gram 锚定细化获得的 Gram 矩阵的补丁特征。我们观察到我们的高分辨率细化程序在特征相关性方面带来了巨大的改进。



(a) Gram matrices at different input resolutions.

Method	Teacher Iteration	Res.	IN1k Linear	ADE mIoU	NYU RMSE
Baseline	—	—	**88.2**	50.3	0.307
GRAM	200k	×1	88.0	53.6	0.285
GRAM	200k	×2	88.0	**55.7**	**0.281**
GRAM	100k	×2	87.9	**55.7**	0.284
GRAM	1M	×2	88.1	54.9	0.290

(b) Ablation of Gram teachers and resolutions.

Figure 9: Quantitative and qualitative study of the impact of high-resolution Gram. We show (a) the improved cosine maps after down-sampling the high-resolution maps into smaller ones, and (b) the quantitative improvements brought by varying the training iteration and the resolution of the Gram teacher.

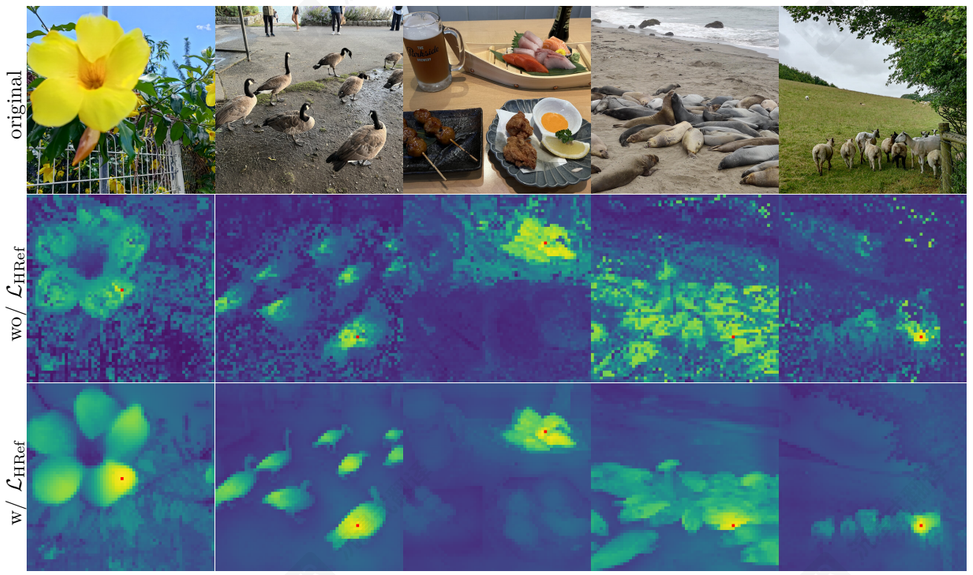


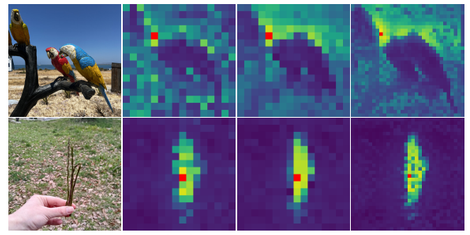
Figure 10: Qualitative effect of Gram anchoring. We visualize cosine maps before and after using the refinement objective $\mathcal{L}_{\text{HRef}}$. The input resolution of the images is 1024×1024 pixels.

5 Post-Training

This section presents *post-training* stages. This includes a high-resolution adaptation phase enabling effective inference at different input resolutions (Sec. 5.1), model distillation producing quality and efficient smaller-sized models (Sec. 5.2), and text alignment adding zero-shot capabilities to DINOv3 (Sec. 5.3).

5.1 Resolution Scaling

We train our model at a relatively small resolution of 256, which gives us a good trade-off between speed and effectiveness. For a patch size of 16, this setup leads to the same input sequence length as DINOv2, which was trained with resolution 224 and patch size 14. However, many contemporary computer vision



(a) 在不同输入分辨率下的 Gram 矩阵。

方法	教师迭代	Res.	IN1k 线性	ADE mIoU	NYU RMSE
基线	—	—	**88.2**	50.3	0.307
GRAM	200k	×1	88.0	53.6	0.285
GRAM	200k	×2	88.0	**55.7**	**0.281**
GRAM	100k	×2	87.9	**55.7**	0.284
GRAM	1M	×2	88.1	54.9	0.290

(b) Gram教师的消融和分辨率。

图9：对高分辨率GRAM影响的定量和定性研究。我们展示了（a）将高分辨率图下采样为更小的图后改进的余弦图，以及（b）通过改变训练迭代和高分辨率GRAM教师的分辨率所带来的定量改进。

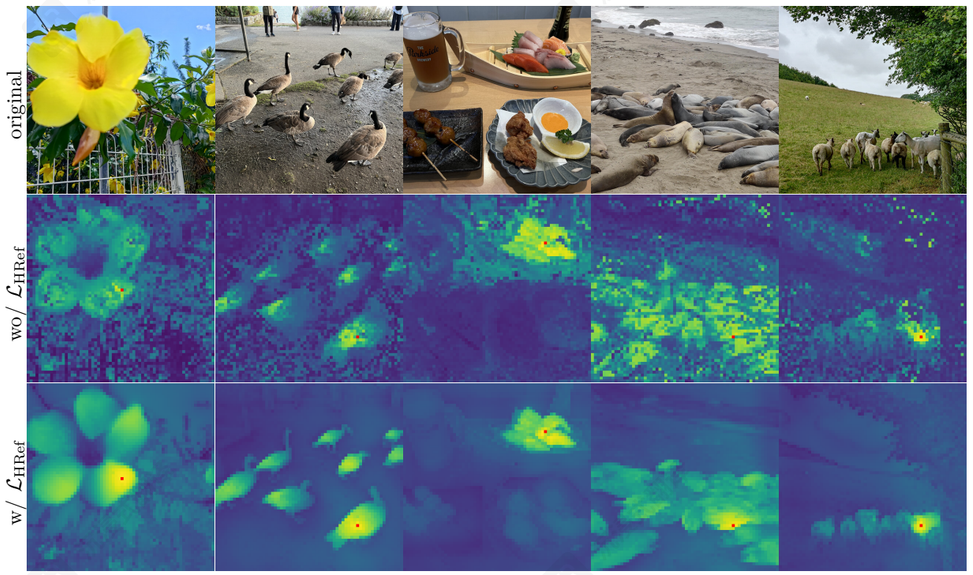


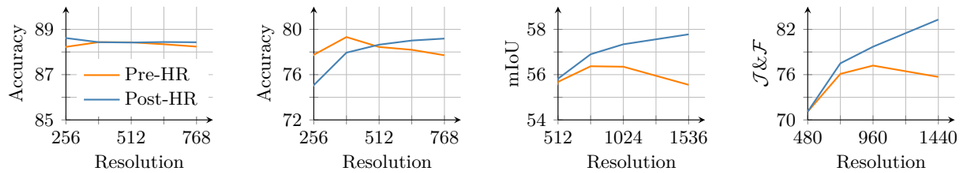
图10：Gram锚定的定性效果。我们展示了在使用细化目标<样式id='3'>L</样式>$_{\text{HRef}}$前后余弦图的变化。图像的输入分辨率为 1024×1024 像素。

5 后训练

本节介绍了后训练阶段。这包括一个高分辨率适应阶段，能够在不同的输入分辨率下进行有效推理（第5.1节），模型蒸馏产生高质量且高效的较小尺寸模型（第5.2节），以及文本对齐为 DINOv3 添加零样本能力（第5.3节）。

5.1 分辨率缩放

我们在相对较低的 256 分辨率下训练我们的模型，这让我们在速度和效果之间取得了良好的平衡。对于 16 的补丁大小，这种设置导致输入序列长度与 DINOv2 相同，DINOv2 是在 224 分辨率和 14 补丁大小下训练的。然而，许多当代计算机视觉



(a) IN1k linear **(b)** ObjectNet **(c)** ADE20k **(d)** DAVIS

Figure 11: Effect of high resolution adaptation. Results before ('Pre-HR') and after ('Post-HR') resolution scaling (Sec. 5.1) on (a) linear classification on ImageNet, (b) applied OOD to ObjectNet, (c) linear semantic segmentation on ADE20k, and (d) segmentation tracking on DAVIS at different evaluation resolutions.

applications require processing images at significantly higher resolutions, often 512×512 pixels or greater, to capture intricate spatial information. The inference image resolution is also not fixed in practice and varies depending on specific use cases. To address this, we extend our training regime with a high-resolution adaptation step (Touvron et al., 2019). To ensure high performance across a range of resolutions, we utilize *mixed resolutions*, sampling differently-sized pairs of global and local crops per mini-batch. Specifically, we consider global crop sizes from $\{512, 768\}$ and local crop sizes from $\{112, 168, 224, 336\}$ and train the model for 10k additional iterations.

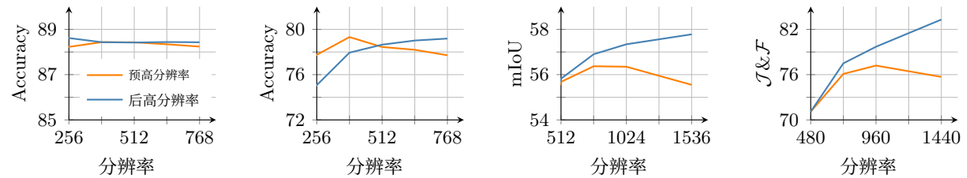
Similar to the main training, a key component of this high-resolution adaptation phase is the addition of Gram anchoring, using the 7B teacher as Gram teacher. We found this component to be essential: without it, the model performance on dense prediction tasks degrades significantly. The Gram anchoring encourages the model to maintain consistent and robust feature correlations across spatial locations, which is crucial when dealing with the increased complexity of high-resolution inputs.

Empirically, we observe that this relatively brief but targeted high-resolution step substantially enhances the overall model's quality and allows it to generalize across a wide range of input sizes, as shown visually in Fig. 4. In Fig. 11, we compare our 7B model before and after adaptation. We find that resolution scaling leads to a small gain on ImageNet classification (a) with relatively stable performance w.r.t. resolution. However, in ObjectNet OOD transfer (b), we observe that the performance tends to degrade slightly for lower resolutions, while improving for higher resolutions. This is largely compensated by the improvement in the quality of local features at high resolution, shown by the positive trend in segmentation on ADE20k (c) and tracking on DAVIS (d). Adaptation leads to local features that *improve with image size*, leveraging the richer spatial information available at larger resolutions and effectively enabling high-resolution inference. Interestingly, the adapted model supports resolutions way beyond the maximum training resolution of 768—we visually observe stable feature maps at resolutions above 4k (*c.f.* Fig. 4).

5.2 Model Distillation

A Family of Models for Multiple Use-Cases We perform knowledge distillation of the ViT-7B model into smaller Vision Transformer variants (ViT-S, ViT-B, and ViT-L), which are highly valued by the community for their improved manageability and efficiency. Our distillation approach uses the same training objective as in the first training phase, ensuring consistency in learning signals. However, instead of relying on an exponential moving average (EMA) of model weights, we use the 7B model directly as the teacher to guide the smaller student models. In this case, the teacher model is fixed. We do not observe patch-level consistency issues and therefore do not apply the Gram anchoring technique. This strategy enables the distilled models to inherit the rich representational power of the large teacher while being more practical for deployment and experimentation.

Our ViT-7B model is distilled into a series of ViT models with sizes covering a broad range of compute budgets, and allowing proper comparison with concurrent models. They include the standard ViT-S (21M params), B (86M), L (0.3B), along with a custom ViT-S+ (29M) and a custom ViT-H+ (0.8B) model to close the performance gap with the self-distilled 7B teacher model. Indeed, we observe in DINOv2 that smaller



(a) IN1k线性 **(B)** ObjectNet **(c)** ADE20k **(d)** DAVIS

图11： 高分辨率适应的效果。分辨率缩放前（'Pre-HR'）和后（'Post-HR'）（第5.1节）对（a）ImageNet上的线性分类、（b）应用于ObjectNet的OOD、（c）ADE20k上的线性语义分割以及（d）DAVIS上的分割跟踪在不同评估分辨率下的结果。

应用程序需要处理更高分辨率的图像，通常 512×512 像素或更高，以捕捉复杂的空间信息。推理图像分辨率在实际中也不是固定的，并且根据具体用例而变化。为了解决这个问题，我们扩展了我们的训练机制，增加了高分辨率适应步骤（Touvron等人， 2019）。为了确保在一系列分辨率下都能获得高性能，我们利用混合分辨率，每个小批量中采样不同大小的全局和局部裁剪对。具体来说，我们考虑全局裁剪尺寸为 $\{512, 768\}$ 和局部裁剪尺寸为 $\{112,168, 224, 336\}$，并训练模型额外 10k 次迭代。

与主要训练类似，高分辨率适应阶段的一个关键组件是添加Gram锚定，使用7B教师作为Gram教师。我们发现这个组件至关重要：没有它，模型在密集预测任务上的性能会显著下降。Gram锚定鼓励模型在空间位置上保持一致和鲁棒的特征相关性，这在处理高分辨率输入的复杂性时至关重要。

经验上，我们观察到这一相对简短但目标明确的高分辨率步骤显著提升了总体模型的质量，并使其能够泛化到广泛的输入尺寸范围，如图所示Fig. 4。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，且性能相对于分辨率相对稳定。然而，在ObjectNet OOD迁移（b）中，我们观察到对于低分辨率，性能略有下降，而对于高分辨率则有所改善。这主要得益于高分辨率下局部特征质量的提升，如图像分割在ADE20k（c）和DAVIS（d）上的积极趋势所示。适应过程使局部特征随着图像尺寸的增大而提升，利用了更大分辨率下更丰富的空间信息，有效支持了高分辨率推理。有趣的是，适应后的模型支持远超768（最大训练分辨率）的分辨率——我们观察到在4k以上的分辨率下，特征图保持稳定（*c.f.* Fig. 4）。

5.2 模型蒸馏

一个适用于多种用例的模型家族 我们对 ViT-7B 模型进行知识蒸馏，将其转化为更小的视觉Transformer变体（ViT-S、ViT-B 和 ViT-L），这些变体因其改进的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阶段训练相同的训练目标，确保学习信号的一致性。然而，我们没有依赖模型权重的指数移动平均（EMA），而是直接使用 7B 模型作为教师来指导较小的学生模型。在这种情况下，教师模型是固定的。我们没有观察到 patch 级一致性问题，因此不应用 Gram 锚定技术。这种策略使蒸馏后的模型能够继承大型教师模型的丰富表示能力，同时更便于部署和实验。

我们的ViT-7B模型被蒸馏成一系列ViT模型，其大小覆盖了广泛的计算预算范围，并允许与同期模型进行适当比较。它们包括标准的ViT-S（21M参数）、B（86M）、L（0.3B），以及定制的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。确实，我们在DINOv2中观察到，较小的

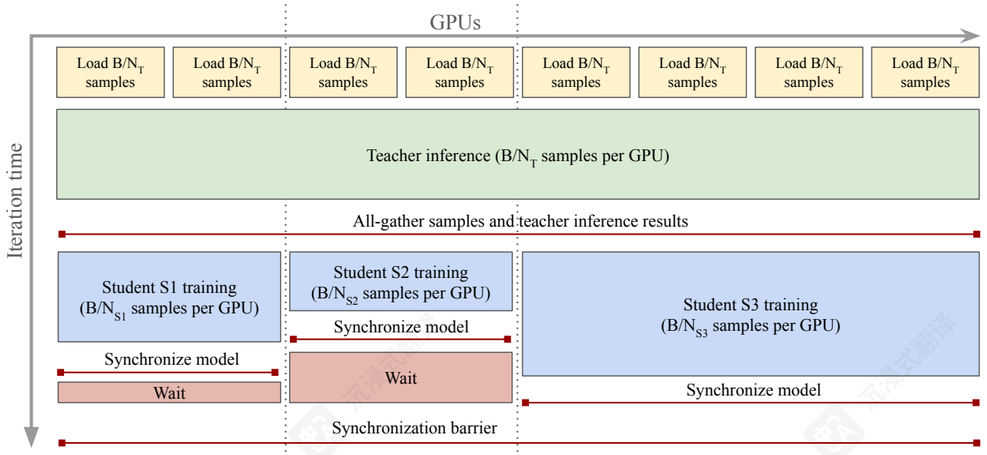


Figure 12: Multi-student distillation procedure. In this diagram, we distill 3 students in parallel: we first share teacher inference across all T nodes to save compute, and gather inputs and results on all GPUs. Then, smaller groups perform student training. We adjust the size of these groups such that the training step has the same duration across all students Si, minimizing idle time waiting at the synchronization barrier.

student models can reach a performance on par with their teacher as the distillation. As a result, the distilled models deliver frontier-level performance for a fraction of the inference compute as we see in Tab. 14. We train the models for 1M iterations then perform 250k iterations of learning-rate cooldown following a cosine schedule before applying the high-resolution phase described in Sec. 5.1 above without Gram anchoring.

Efficient Multi-Student Distillation As the inference cost for a large teacher can be orders of magnitude higher than for students (see Fig. 16a), we design a parallel distillation pipeline that allows training multiple students at the same time and sharing the teacher inference across all nodes involved in the training (see Fig. 12 for a diagram). Let C_T and C_S be respectively the cost of running the teacher inference and the student training on a single sample, in single-teacher/single-student distillation with batch-size B where each of the N GPUs processes a B/N slice of the data, the teacher inference costs $B/N \times C_T$ and the student training costs $B/N \times C_S$ per GPU. In multi-student distillation, we proceed as follows. Each student Si is assigned a set of N_{Si} GPUs for training, and all $N_T = \sum N_{Si}$ GPUs are part of the global inference group. At each iteration, we first run the teacher inference on the global group for a $B/N_T \times C_T$ compute cost per GPU. We then run an *all-gather* collective operation to share the input data and inference result with all compute nodes. Finally, each student group separately performs student training for a $B/N_{Si} \times C_{Si}$ cost.

The above calculations shows that adding an additional student to the distillation pipeline will (1) reduce the per-GPU compute at each iteration, thus globally improving distillation speed, and (2) increase the overall compute only by the training cost of the new student, since the total teacher inference cost is now fixed. The implementation only requires setting up GPU process groups carefully, adapting data-loaders and teacher inference to ensure inputs and outputs are synchronized across groups using NCCL collectives. As the groups are synchronized at each iteration, in order to maximize speed, we adapt the number of GPUs for each student such that their iteration times are roughly the same. With this procedure, we seamlessly train multiple students, and produce a whole family of distilled models from our flagship 7B model.

5.3 Aligning DINOv3 with Text

Open-vocabulary image-text alignment has received significant interest and enthusiasm from the research community, thanks to its potential to enable flexible and scalable multimodal understanding. A large body

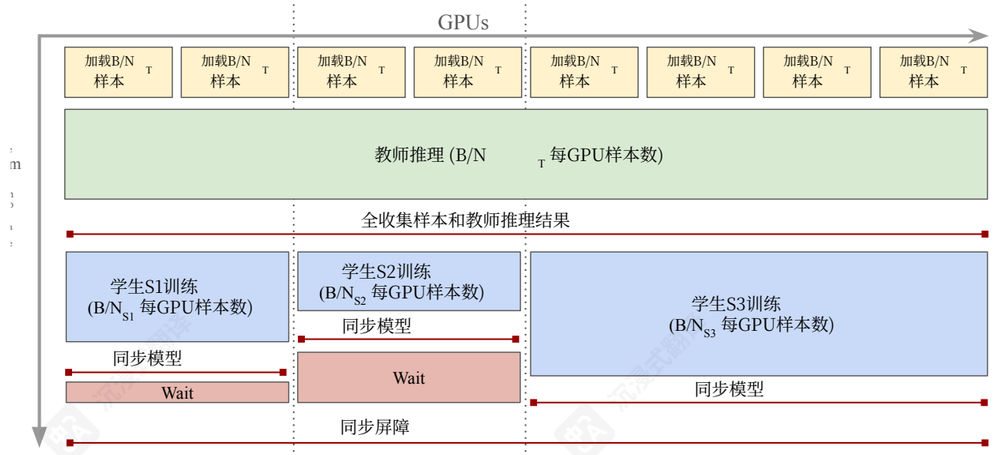


图12： 多学生蒸馏流程。在这个图中，我们并行蒸馏3个学生：我们首先在所有T节点上共享教师推理以节省计算量，并在所有GPU上收集输入和结果。然后，较小的组执行学生训练。我们调整这些组的大小，使得所有学生的训练步骤具有相同的持续时间 Si，从而最小化在同步屏障处的等待空闲时间。

学生模型可以达到与教师模型相当的性能，就像蒸馏一样。因此，蒸馏模型以少量推理计算就实现了前沿级性能，正如我们在表14中看到的那样。我们训练模型1M次迭代，然后按照余弦调度进行250k次迭代的学习率冷却，最后应用上述第5.1节中描述的高分辨率阶段，但不使用Gram锚定。

高效多学生蒸馏 由于大型教师的推理成本可能比学生高出数量级（参见图16a），我们设计了一个并行蒸馏管道，允许同时训练多个学生，并在所有参与训练的节点上共享教师推理（参见图12 的示意图）。令 C_T 和 C_S 分别为在单教师单学生蒸馏中，对单个样本运行教师推理和学生训练的成本，其中批处理大小为 B，每个 N GPU处理数据的一个切片，教师推理成本为 $B/N \times C_T$，每个GPU的学生训练成本为 $B/N \times C_S$。在多学生蒸馏中，我们按以下步骤进行。每个学生 Si 被分配一组 N_{Si} GPU用于训练，所有 $N_T = \sum N_{Si}$ GPU 都属于全局推理组。在每个迭代中，我们首先在全局组上运行教师推理，每个 GPU 的 $B/N_T \times C_T$ 计算成本。然后，我们运行一个*all-gather* 集体操作，以在所有计算节点之间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，向蒸馏流程中添加一个额外的学生将（1）减少每次迭代的每GPU计算量，从而全局提高蒸馏速度，以及（2）仅增加总体计算量，即新学生的训练成本，因为总教师推理成本现在是固定的。该实现只需要仔细设置GPU进程组，调整数据加载器和教师推理以确保输入和输出在组间使用NCCL集体同步。由于组在每次迭代中都是同步的，为了最大化速度，我们调整每个学生的GPU数量，使其迭代时间大致相同。通过此流程，我们无缝地训练多个学生，并从我们的旗舰7B模型生成一系列蒸馏模型。

5.3 与文本对齐DINOv3

开放词汇图像文本对齐因其能够实现灵活且可扩展的多模态理解潜力，已引起研究社区的高度关注和热情。大量研究工作

of work has focused on improving the quality of CLIP (Radford et al., 2021), which originally learned only a global alignment between image and text representations. While CLIP has demonstrated impressive zero-shot capabilities, its focus on global features limits its ability to capture fine-grained, localized correspondences. More recent works (Zhai et al., 2022b) have shown that effective image-text alignment can be achieved with pre-trained self-supervised visual backbones. This makes it possible to leverage these powerful models in multi-modal settings, facilitating richer and more precise text-to-image associations that extend beyond global semantics while also reducing computational costs, since the visual encoding is already learned.

We align a text encoder with our DINOv3 model by adopting the training strategy previously proposed in Jose et al. (2025). This approach follows the LiT training paradigm (Zhai et al., 2022b), training a text representation from scratch to match images to their captions with a contrastive objective, while keeping the vision encoder frozen. To allow for some flexibility on the vision side, two transformer layers are introduced on top of the frozen visual backbone. A key enhancement of this method is the concatenation of the mean-pooled patch embeddings with the output CLS token before matching to the text embeddings. This enables aligning both global and local visual features to text, leading to improved performance on dense prediction tasks without requiring additional heuristics or tricks. Furthermore, we use to the same data curation protocol as established in Jose et al. (2025) to ensure consistency and comparability.

6 Results

In this section, we evaluate our flagship DINOv3 7B model on a variety of computer vision tasks. Throughout our experiments, unless otherwise specified, *we keep DINOv3 frozen* and solely use its representations. We demonstrate that with DINOv3, finetuning is not necessary to obtain strong performance. This section is organized as follows. We first probe the quality of DINOv3's dense (Sec. 6.1) and global (Sec. 6.2) image representations using lightweight evaluation protocols and compare it to the strongest available vision encoders. We show that DINOv3 learns exceptional dense features while offering robust and versatile global image representations. Then, we consider DINOv3 as a basis for developing more complex computer vision systems (Sec. 6.3). We show with little effort on top of DINOv3, we are able to achieve results competitive with or exceeding the state of the art in tasks as diverse as object detection, semantic segmentation, 3D view estimation, or relative monocular depth estimation.

6.1 DINOv3 provides Exceptional Dense Features

We first investigate the raw quality of DINOv3's dense representations using a diverse set of lightweight evaluations. In all cases, we utilize the frozen patch features of the last layer, and evaluate them using (1) qualitative visualizations (Sec. 6.1.1), (2) dense linear probing (Sec. 6.1.2: segmentation, depth estimation), (3) non-parametric approaches (Sec. 6.1.3: 3D correspondence estimation, Sec. 6.1.4: object discovery, Sec. 6.1.5: tracking), and (4) lightweight attentive probing (Sec. 6.1.6: video classification).

Baselines We compare the dense features of DINOv3 with those of the strongest publicly available image encoders, both weakly- and self-supervised ones. We consider the weakly-supervised encoders Perception Encoder (PE) Core (Bolya et al., 2025) and SigLIP 2 (Tschannen et al., 2025), which use CLIP-style image-text contrastive learning. We also compare to the strongest self-supervised methods: DINOv3's predecessor DINOv2 (Oquab et al., 2024) with registers (Darcet et al., 2024), Web-DINO (Fan et al., 2025), a recent scaling effort of DINO, and Franca (Venkataramanan et al., 2025), the best open-data SSL model. Finally, we compare to the agglomerative models AM-RADIOv2.5 (Heinrich et al., 2025), distilled from DINOv2, CLIP (Radford et al., 2021), DFN (Fang et al., 2024a), and Segment Anything (SAM) (Kirillov et al., 2023), and to PEspatial, distilling SAM 2 (Ravi et al., 2025) into PEcore. For each baseline, we report the performance of the strongest model available and specify the architecture in the tables.

6.1.1 Qualitative Analysis

We start by analyzing DINOv3's dense feature maps qualitatively. To this end, we project the dense feature space into 3 dimensions using principal component analysis (PCA), and map the resulting 3D space into RGB. Because of the sign ambiguity in PCA (eight variants) and the arbitrary mapping between principal

集中于提升CLIP（Radford等人， 2021）的质量，该模型最初仅学习图像和文本表示之间的全局对齐。尽管CLIP展示了令人印象深刻的零样本能力，但其对全局特征的侧重限制了其捕捉细粒度、局部对应关系的能力。更近期的作品（翟等人， 2022b）表明，通过预训练自监督视觉骨干网络可以实现有效的图像文本对齐。这使得在多模态环境中利用这些强大的模型成为可能，从而促进超越全局语义的更丰富、更精确的文本到图像关联，同时降低计算成本，因为视觉编码已经学习完成。

我们通过采用先前在Jose 等人(2025)中提出的训练策略，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 等人， 2022b），从零开始训练文本表示，以对比目标将图像与其标题匹配，同时冻结视觉编码器。为了在视觉方面提供一些灵活性，在冻结的视觉骨干网络上引入了两个 Transformer 层。此方法的关键增强是将均值池化的块嵌入与匹配文本嵌入之前的输出 CLS 标记连接起来。这使得能够将全局和局部视觉特征与文本对齐，从而在不需要额外的启发式方法或技巧的情况下，在密集预测任务上提高了性能。此外，我们使用与 Jose 等人 (2025)中建立的相同的数据管理协议，以确保一致性和可比性。

6 结果

在本节中，我们在各种计算机视觉任务上评估了我们的旗舰 DINOv3 7B 模型。在我们的实验中，除非另有说明，我们保持 *DINOv3* 冻结 并仅使用其表示。我们证明，使用 DINOv3，无需微调即可获得强大的性能。本节的结构如下。我们首先使用轻量级评估协议探测 DINOv3 的密集（第 6.1 节）和全局（第 6.2 节）图像表示的质量，并将其与最强的可用视觉编码器进行比较。我们表明，DINOv3 学习了卓越的密集特征，同时提供了稳健和通用的全局图像表示。然后，我们将 DINOv3 作为开发更复杂计算机视觉系统的基础（第 6.3 节）。我们表明，在 DINOv3 之上只需少量工作，我们就能在目标检测、语义分割、3D 视角估计或相对单目深度估计等多样化的任务中实现与当前最佳相当或超越当前最佳的结果。

6.1 DINOv3 提供了优异的密集特征

我们首先使用一系列轻量级评估方法研究 DINOv3 密集表示的原始质量。在所有情况下，我们使用最后一层的冻结块特征，并使用 (1) 定性可视化 (Sec. 6.1.1), (2) 密集线性探测 (Sec. 6.1.2: 分割、深度估计), (3) 非参数方法 (Sec. 6.1.3: 3D 对应关系估计, Sec. 6.1.4: 对象发现,Sec. 6.1.5: 跟踪), 以及 (4) 轻量级注意力探测 (Sec. 6.1.6: 视频分类).

基线 我们将 DINOv3 的密集特征与当前最强的公开可用的图像编码器进行比较，包括弱监督和自监督的编码器。我们考虑弱监督编码器感知编码器（PE）核心 (Bolya 等人，2025) 和 SigLIP 2 (Tschannen 等人，2025)，它们使用 CLIP 风格的图像-文本对比学习。我们还与最强的自监督方法进行比较：DINOv3 的前身 DINOv2 (Oquab 等人，2024) 配合寄存器 (Darcet 等人，2024)，Web-DINO (Fang 等人，2025)，DINO 的一项最近扩展工作，以及 Franca (文卡特拉马南等人，2025)，即最佳开源数据 SSL 模型。最后，我们与聚合模型 AM-RADIOv2.5 (Heinrich 等人，2025) 进行比较，该模型源自 DINOv2，CLIP (Radford 等人，2021)，DFN (Fang 等人，2024a)，以及 Segment Anything (SAM) (Kirillov 等人，2023)，以及 PEspatial，将 SAM 2 (Ravi 等人，2025) 蒸馏到 PEcore 中。对于每个基线，我们报告最强模型的性能，并在表格中指定架构。

6.1.1 定性分析

我们首先通过定性分析来分析 DINOv3 的密集特征图。为此，我们使用主成分分析（PCA）将密集特征空间投影到 3 维，并将生成的 3D 空间映射到 RGB。由于 PCA 中存在符号模糊性（8 种变体）以及主成分和颜色（6 种变体）之间的任意映射，我们探索了所有组合，并报告了视觉上最引人注目的组合。生成的可视化结果显示在</style><style id='3'>图 13</style><style id='5'>。与其他视觉主干相比，可以看出 DINOv3 的特征更清晰，包含的噪声更少，并表现出更优越的语义一致性。</style>

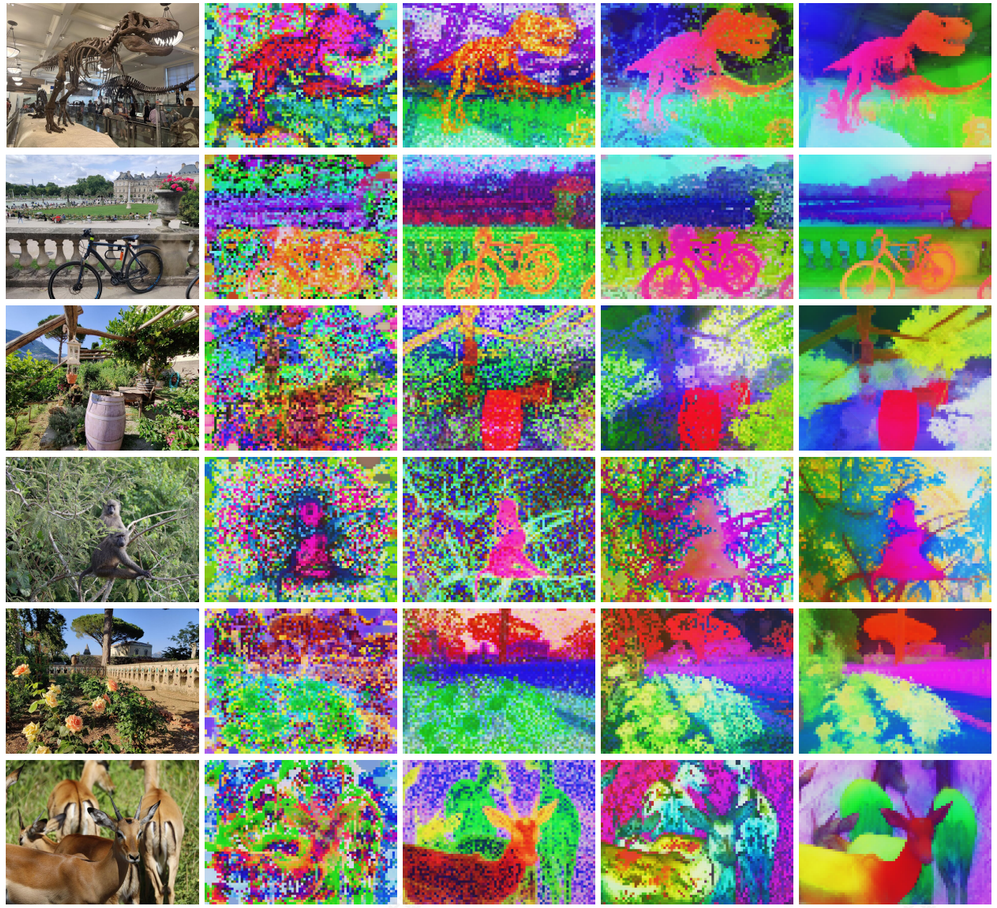


Figure 13: Comparison of dense features. We compare several vision backbones by projecting their dense outputs using PCA and mapping them to RGB. From left to right: SigLIP 2 ViT-g/16, PEspatial ViT-G/14, DINOv2 ViT-g/14 with registers, DINOv3 ViT-7B/16. Images are forwarded at resolution 1280×960 for models using patch 16 and 1120×840 for patch 14, *i.e.* all feature maps have size 80×60.

components and colors (six variants), we explore all combinations and report the visually most compelling one. The resulting visualization is shown in Fig. 13. Compared to other vision backbones, it can be seen that the features of DINOv3 are sharper, containing much less noise, and showing superior semantical coherence.

6.1.2 Dense Linear Probing

We perform linear probing on top of the dense features for two tasks: semantic segmentation and monocular depth estimation. In both cases, we train a linear transform on top of the frozen patch outputs of DINOv3. For semantic segmentation, we evaluate on the ADE20k (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and PASCAL VOC 2012 (Everingham et al., 2012) datasets and report the mean intersection-over-union

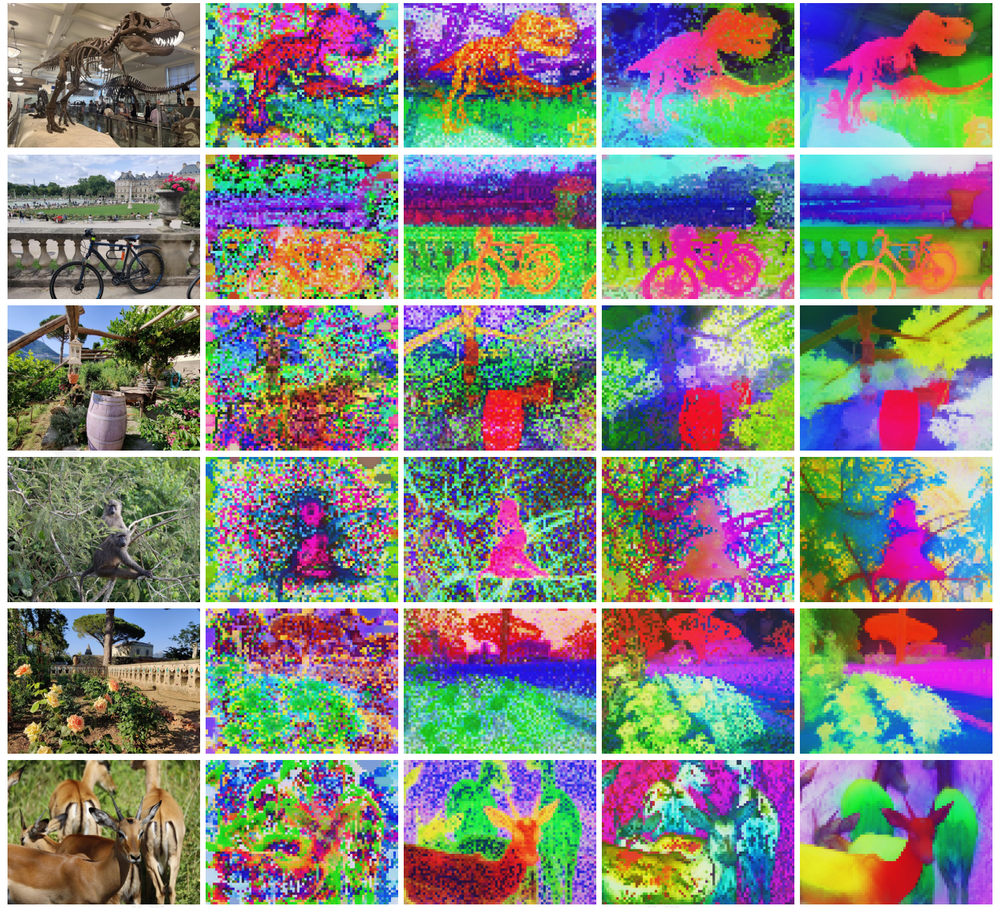


图13： 密集特征的比较。我们通过PCA将几个视觉主干密集输出投影并映射到RGB进行比较。从左到右：SigLIP 2 ViT-g/16、PEspatial ViT-G/14、DINOv2 ViT-g/14 带注册、DINOv3 ViT-7B/16。对于使用块16的模型，图像在分辨率 1280×960 处传递，对于块14，在 1120×840 处传递<code style id='5'>即</code>。所有特征图的大小为 80×60。

组件和颜色（六种变体），我们探索了全部组合并报告了视觉上最引人注目的组合。生成的可视化结果显示在图13。与其他视觉主干相比，可以看出 DINOv3 的特征更清晰，包含的噪声更少，并表现出更优越的语义一致性。

6.1.2 密集线性探测

我们在密集特征之上进行线性探测，用于两个任务：语义分割和单目深度估计。在这两种情况下，我们在冻结的 DINOv3 块输出之上训练线性变换。对于语义分割，我们在 ADE20k (周 等人, 2017), Cityscapes (Cordts等人, 2016), 和 PASCAL VOC 2012 (Everingham等人, 2012) 数据集上进行评估，并报告均值交并比

Table 3: Dense linear probing results on semantic segmentation and monocular depth estimation with frozen backbones. We report the mean Intersection-over-Union (mIoU) metric for the segmentation benchmarks ADE20k, Cityscapes, and VOC. We report the Root Mean Squared Error (RMSE) metric for the depth benchmarks NYUv2 and KITTI. For segmentation, all models are evaluated with input resolution adapted to 1024 patch tokens (*i.e.* 448×448 for patch size 14, 512×512 for patch size 16).

		Segmentation			Depth	
Method	ViT	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**

(mIoU) metric. For depth estimation, we use the NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2013) datasets and report the root mean squared error (RMSE).

Results (Tab. 3) The segmentation results demonstrate the superior quality of our dense features. On the general ADE20k dataset, DINOv3 outperforms the self-supervised baselines by more than 6 mIoU points, and the weakly supervised baselines by more than 13 points. Furthermore, DINOv3 surpasses PEspatial by more than 6 points, and AM-RADIOv2.5 by nearly 3 points. These results are remarkable as both are strong baselines, being distilled from the heavily supervised segmentation model SAM (Kirillov et al., 2023). Similar results are observed on the self-driving benchmark Cityscapes, with DINOv3 achieving the best mIoU of 81.1, surpassing AM-RADIOv2.5 by 2.5 points, and all other backbones by at least 5.5 points.

On monocular depth estimation, DINOv3 again outperforms all other models by significant margins: the weakly-supervised models PEcore and SigLIP 2 are still lagging, with DINOv2 and the more advanced models derived from SAM are the closest competitors. Interestingly, while PEspatial and AM-RADIO show strong performance on NYU, their performance is lower than DINOv2's on KITTI. Even there, DINOv3 outperforms its predecessor DINOv2 by 0.278 RMSE.

Both sets of evaluations show the outstanding representation power of the dense features of DINOv3 and reflect the visual results from Fig. 13. With only a linear predictor, DINOv3 allows robust prediction of object categories and masks, as well as physical measurements of the scene such as relative depth. These results show that the features are not only visually sharp and properly localized, they also represent many important properties of the underlying observations in a linearly separable way. Finally, the absolute performance obtained with a linear classifier on ADE20k (55.9 mIoU) is itself impressive, as it is not far from the absolute the state-of-the-art (63.0 mIoU) on this dataset.

6.1.3 3D Correspondence Estimation

Understanding the 3D world has always been an important goal of computer vision Image foundation models have recently fueled research in 3D understanding by offering *3D-aware features*. In this section, we evaluate the *multi-view consistency* of DINOv3—that is, whether patch features of the same keypoint in different views of an object are similar—following the protocol defined in Probe3D (Banani et al., 2024). We distinguish between *geometric* and *semantic* correspondence estimation. The former refers to matching keypoints for the *same object instance* while the latter refers to matching keypoints for different instances of the *same object class*. We evaluate geometric correspondence on the NAVI dataset (Jampani et al., 2023) and semantic

表3： 冻结主干网络在语义分割和单目深度估计上的密集线性探测结果。我们报告了分割基准测试 ADE20k、Cityscapes 和 VOC 的均值交并比 (mIoU) 指标。我们报告了深度基准测试 NYUv2 和 KITTI 的均方根误差 (RMSE) 指标。对于分割，所有模型均使用输入分辨率调整为 1024 块标记进行评估 (即。448×448 对于块大小 14，512×512 对于块大小 16)。

		分割			深度	
方法	ViT	ADE20k	Citysc.	VOC	NYUv2 ↓	KITTI ↓
聚合主干						
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
Weakl y-supervised 骨干网络						
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 骨干网络						
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**

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

Results (Tab. 3) 分割结果展示了我们密集特征的优越性。在通用的 ADE20k 数据集上，DINOv3 比自监督基线高出 6 个 mIoU 点，比弱监督基线高出 13 个点。此外，DINOv3 比PEspatial高出 6 个点，比 AM-RADIOv2.5 高出近 3 个点。这些结果非常显著，因为它们都是强基线，源自重度监督的分割模型 SAM (Kirillov 等人。， 2023)。在自动驾驶基准 Cityscapes 上观察到类似的结果，DINOv3 实现了最佳的 mIoU 为 81。1，比 AM-RADIOv2.5 高出 2.5 个点，比所有其他骨干网络高出至少 5.5 个点。

在单目深度估计方面，DINOv3 再次以显著优势超越所有其他模型：弱监督模型 PEcore 和 SigLIP 2 仍然落后，DINOv2 以及从 SAM 衍生出的更先进的模型是最近的竞争对手。有趣的是，虽然 PEspatial 和 AM-RADIO 在 NYU 上表现出色，但它们在 KITTI 上的性能低于 DINOv2，甚至在那时，DINOv3 也比其前身 DINOv2 超出 0.278 RMSE。

两组评估都显示了 DINOv3 密集特征的突出表示能力，并反映了来自 图 13 的视觉效果。仅使用线性预测器，DINOv3 允许对物体类别和掩码进行稳健预测，以及对场景的物理测量，如相对深度。这些结果表明，这些特征不仅视觉上清晰且定位准确，而且还以线性可分的方式表示底层观察的许多重要属性。最后，在 ADE20k 上使用线性分类器获得的绝对性能（55.9 mIoU）本身就很令人印象深刻，因为它接近该数据集上的绝对最先进技术（63.0 mIoU）。

6.1.3 3D对应估计

理解3D世界一直是计算机视觉的一个重要目标 图像基础模型最近通过提供 *3D*感知特征。在本节中，我们评估了DINOv3的 多视图一致性，即物体在不同视图中的相同关键点的补丁特征是否相似——遵循在 Probe3D（Banani等人,2024）中定义的协议）。我们区分 几何和语义对应估计。前者指的是匹配 同一对象实例的关键点，而后者指的是匹配 同一对象类的不同实例的关键点。我们在NAVI数据集（Jampani等人, 2023）上评估几何对应，并在语义

Table 4: Evaluation of 3D consistency of dense representations. We estimate 3D keypoint correspondences across views following the evaluation protocol of Probe3D (Banani et al., 2024). To measure performance, we report the correspondence recall, *i.e.* the percentage of correspondences falling into a specified distance.

		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**

correspondence on the SPair dataset (Min et al., 2019), and measure performance with correspondence recall in both cases. Please refer to App. D.3 for more experimental details.

Results (Tab. 4) For geometric correspondences, DINOv3 outperforms all other models and improves over the second best model (DINOv2) by 4.3% recall. Other SSL scaling endeavors (Franca and WebSSL) lag behind DINOv2, showing that it is still a strong baseline. Weakly-supervised models (PEcore and SigLIP 2) do not fare well on this task, indicating a lack of 3D awareness. For models with SAM distillation, AM-RADIO nearly reaches the performance of DINOv2, but PEspatial still lags behind it (−11.6% recall), and even falls behind Franca (−0.8% recall). This suggests that self-supervised learning is a key component for strong performance on this task. For semantic correspondences, the same conclusions apply. DINOv3 performs best, outperforming both its predecessor (+2.6% recall) and AM-RADIO (+1.9% recall). Overall, these impressive performance on keypoint matching are very promising signals for downstream use of DINOv3 in other 3D-heavy applications.

6.1.4 Unsupervised Object Discovery

Powerful self-supervised features facilitate discovering object instances in images without requiring *any* annotations (Vo et al., 2021; Siméoni et al., 2021; Seitzer et al., 2023; Wang et al., 2023c; Siméoni et al., 2025). We test this capability for different vision encoders via the task of unsupervised object discovery, which requires class-agnostic segmentation of objects in images (Russell et al., 2006; Tuytelaars et al., 2010; Cho et al., 2015; Vo et al., 2019). In particular, we use the non-parametric graph-based TokenCut algorithm (Wang et al., 2023c), which has shown strong performance on a variety of backbones. We run it on three widely used datasets: VOC 2007, VOC 2012 (Everingham et al., 2015), and COCO-20k (Lin et al., 2014; Vo et al., 2020). We follow the evaluation protocol defined by Siméoni et al. (2021) and report the CorLoc metric. To properly compare backbones with different feature distributions, we perform a search over the main TokenCut hyperparameter, namely the cosine similarity threshold applied when constructing the patch graph used for partitioning. Originally, the best object discovery results were obtained with DINO (Caron et al., 2021) using the keys of the last attention layer. However, this hand-crafted choice does not consistently generalize to other backbones. For simplicity, we always employ the output features for all models.

Results (Fig. 14) The original DINO has set a very high bar for this task. Interestingly, while DINOv2 has shown very strong performance for pixel-wise dense tasks, it fails at object discovery. This can in part be attributed to the artifacts present in the dense features (*c.f.* Fig. 13). DINOv3, with its clean and precise output feature maps outperforms both its predecessors, with a 5.9 CorLoc improvement on VOC 2007, and all other backbones, whether self-, weakly-supervised or agglomerative. This evaluation confirms that

表4： 对密集表示的3D一致性评估。我们遵循Probe3D（Banani等人,2024）的评估协议，跨视图估计3D关键点对应。为了衡量性能，我们报告对应召回率，即。落入指定距离的对应百分比。

		几何	语义
方法	ViT	NAVI	SPair
*Agg*聚合骨干网络			
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
自监督骨干网络			
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**

对应在SPair数据集（Min等人, 2019），并在这两种情况下使用对应召回率来衡量性能。请参考 附录D.3 以获取更多实验细节。

Results (表4) 对于几何对应，DINOv3优于所有其他模型，并在第二好的模型（DINOv2）上提升了4.3%召回率。其他SSL扩展尝试（Franca和WebSSL）落后于DINOv2，表明它仍然是一个强基线。弱监督模型（PEcore和SigLIP 2）在这个任务上表现不佳，表明缺乏3D感知。对于具有SAM蒸馏的模型，AM-RADIO几乎达到了DINOv2的性能，但PEspatial仍然落后于它（−11.6%召回率），甚至落后于Franca（−0.8%召回率）。这表明自监督学习是这个任务取得强性能的关键组成部分。对于语义对应，同样的结论也适用。DINOv3表现最佳，优于其前身（+2.6%召回率）和AM-RADIO（+1.9%召回率）。总体而言，这些在关键点匹配上的出色性能为DINOv3在其他3D密集应用中的下游应用提供了非常有希望的信号。

6.1.4 无监督对象发现

强大的自监督特征能够无需任何标注即可在图像中发现对象实例（Vo 等人，2021；Siméoni 等人，2021；Seitzer 等人，2023；王等人，2023c；Siméoni 等人，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一能力，该任务要求对图像中的物体进行类无关分割（Russell 等人，2006；Tuytelaars 等人，2010；Cho 等人，2015；Vo 等人，2019）。特别是，我们使用非参数图模型TokenCut算法（王等人，2023c），该算法在各种骨干网络上都表现出强大的性能。我们在三个广泛使用的数据集上运行它：VOC 2007、VOC 2012（Everingham 等人，2015）和COCO-20k（Lin 等人，2014；Vo 等人，2020）。我们遵循Siméoni 等人（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骨干网络，我们对主要的TokenCut超参数进行搜索，即在构建用于分区的补丁图时应用的余弦相似度阈值。最初，使用DINO（Caron 等人，2021）的最后一个注意力层的键获得了最佳对象发现结果。然而，这种手工选择并不始终如一地泛化到其他骨干网络。为简化起见，我们始终对所有模型使用输出特征。

结果（图14）原始DINO为这项任务设定了非常高的标准。有趣的是，虽然DINOv2在像素级密集任务中表现出非常强的性能，但在对象发现方面却失败了。这部分可以归因于密集特征中存在的伪影（参见。图13）。DINOv3凭借其干净且精确的输出特征图，在VOC 2007上比其前两个版本提高了5。9 CorLoc，并且在所有其他骨干网络（无论是自-监督、弱监督还是聚合）上均表现优异。这项评估证实了

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: Unsupervised object discovery. We apply TokenCut (Wang et al., 2022c) on the output patch features of different backbones and report CorLoc metric. We also visualize predicted masks obtained with DINOv3 (red overlay on input images at res. 1024), obtained *with no annotation and no post-processing*.

DINOv3's dense features are both semantically strong and well localized. We believe that this will pave the way for more class-agnostic object detection approaches, especially in scenarios where annotations are costly or unavailable, and where the set of relevant classes is not confined to a predefined subset.

6.1.5 Video Segmentation Tracking

Beyond static images, an important property of visual representations is their *temporal consistency*, *i.e.* whether the features evolve in a stable manner through time. To test for this property, we evaluate DINOv3 on the task of video segmentation tracking: given ground-truth instance segmentation masks in the first frame of a video, the goal is to propagate these masks to subsequent frames. We use the DAVIS 2017 (Pont-Tuset et al., 2017), YouTube-VOS (Xu et al., 2018), and MOSE (Ding et al., 2023) datasets. We evaluate performance using the standard $\mathcal{J}\&\mathcal{F}$-mean metric, which combines region similarity ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) (Perazzi et al., 2016). Following Jabri et al. (2020), we use a non-parametric label propagation algorithm that considers the similarity between patch features across frames. We evaluate at three input resolutions, using a short side length of 420/480 (S), 840/960 (M), and 1260/1440 (L) pixels for models with patch size 14/16 (matching the number of patch tokens). The $\mathcal{J}\&\mathcal{F}$ score is always computed at the native resolution of the videos. See App. D.5 for more detailed experimental settings.

Results (Tab. 5) Aligned with all previous results, weakly-supervised backbones do not deliver convincing performance. PEspatial, distilled from the video model SAMv2, provides satisfactory performance, surpassing DINOv2 on smaller resolutions, but falling short on larger ones. Across resolutions, DINOv3 outperforms all competitors, with a staggering 83.3 $\mathcal{J}\&\mathcal{F}$ on DAVIS-L, 6.7 points above DINOv2. Furthermore, performance as a function of resolution follows a healthy trend, confirming that our model is able to make use of more input pixels to output precise, high-resolution feature maps (*c.f.* Figs. 3 and 4). In contrast, performance at higher resolutions stays almost flat for SigLIP 2 and PEcore, and degrades for PEspatial. Interestingly, our image model, without any tuning on video, allows to properly track objects in time (see Fig. 15). This makes it a great candidate to embed videos, allowing to build strong video models on top.

6.1.6 Video Classification

The previous results have shown the low-level temporal consistency of DINOv3's representations, allowing to accurately track objects in time. Going beyond, we evaluate in this section the suitability of its dense features for high-level video classification. Similar to the setup of V-JEPA 2 (Assran et al., 2025), we train an *attentive probe*—a shallow 4-layer transformer-based classifier—on top of patch features extracted from each frame. This enables reasoning over temporal and spatial dimensions as the features are extracted independently per

方法	ViT	VOC07	VOC12	COCO
聚合主干				
AM-RADIOv2.5	g/14	55.0	59.7	45.9
PEspatial	G/14	51.2	56.0	43.9
*Weakly-supervised*有监督的骨干网络				
SigLIPv2	g/16	20.5	24.7	18.6
PEcore	G/14	14.2	18.2	13.5
自监督骨干网络				
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**

图14： 无监督对象发现。我们对不同骨干网络的输出块特征应用TokenCut（王等人， 2022c）并报告CorLoc指标。我们还使用DINOv3（输入图像在res. 1024上的红色叠加）可视化预测掩码，该掩码 未使用标注和后处理。

DINOv3的密集特征兼具语义强大和良好定位的特点。我们相信这将推动更多类无关目标检测方法的发展，特别是在标注成本高昂或不可用，且相关类别的集合不局限于预定义子集的场景中。

6.1.5 视频分割跟踪

超越静态图像，视觉表示的一个重要特性是它们的 时间一致性， 即。特征是否随时间稳定演变。为了测试这一特性，我们在视频分割跟踪任务上评估了 DINOv3：给定视频中第一帧的真实标签实例分割掩码，目标是将这些掩码传播到后续帧。我们使用了 DAVIS 2017 (Pont-Tuset 等人， 2017)， YouTube-VOS (徐等人， 2018)，和 MOSE (丁等人,2023) 数据集。我们使用标准的 $\mathcal{J}\&\mathcal{F}$-均值指标来评估性能，该指标结合了区域相似性 ($\mathcal{J}$) 和轮廓准确度 ($\mathcal{F}$) (Perazzi 等人,2016)。遵循 Jabri 等人 (2020)，我们使用了一种非参数标签传播算法，该算法考虑了跨帧的补丁特征之间的相似性。我们在三个输入分辨率下进行评估，对于补丁大小为 14/16 (匹配补丁标记数量) 的模型，使用短边长度为 420/480 (S)，840/960 (M)，和 1260/1440 (L) 像素的输入。$\mathcal{J}\&\mathcal{F}$ 分数始终在视频的原生分辨率下计算。更多详细的实验设置请参见 附录 D.5。

结果（表5）与所有先前结果一致，弱监督主干网络并未提供令人信服的性能。来自视频模型SAMv2的PEspatial提供了令人满意的性能，在小分辨率上超越了DINOv2，但在大分辨率上有所不足。在所有分辨率下，DINOv3均优于所有竞争对手，以惊人的83.3 $\mathcal{J}\&\mathcal{F}$ 在DAVIS-L上，6.7 分高于DINOv2。此外，性能随分辨率的变化呈现健康趋势，证实我们的模型能够利用更多输入像素输出精确、高分辨率的特征图（参见 图3和4）。相比之下，SigLIP 2和PEcore在高分辨率下的性能几乎保持不变，而PEspatial则有所下降。有趣的是，我们的图像模型在没有视频调优的情况下，能够正确地跟踪时间中的物体（参见图15）。这使得它成为嵌入视频的理想选择，允许在之上构建强大的视频模型。

6.1.6 视频分类

之前的结果已经表明DINOv3的表示具有低级时间一致性，这使得能够在时间上准确跟踪物体。更进一步，在本节中，我们评估了其密集特征用于高级视频分类的适用性。类似于V-JEPA 2（Assran等人， 2025）的设置，我们在从每帧提取的补丁特征之上训练了一个注意力探针——一个浅层4层基于Transformer的分类器。这使得能够在时间和空间维度上进行推理，因为特征是独立于每

Table 5: Video segmentation tracking evaluation. We report the $\mathcal{J}\&\mathcal{F}$-mean on DAVIS, YouTube-VOS, and MOSE at multiple resolutions. For models with patch size 14/16, the small, medium and large resolutions correspond to a video short side of 420/480, 840/960, 1260/1140 pixels.

		DAVIS			YouTube-VOS			MOSE		
Method	ViT	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**

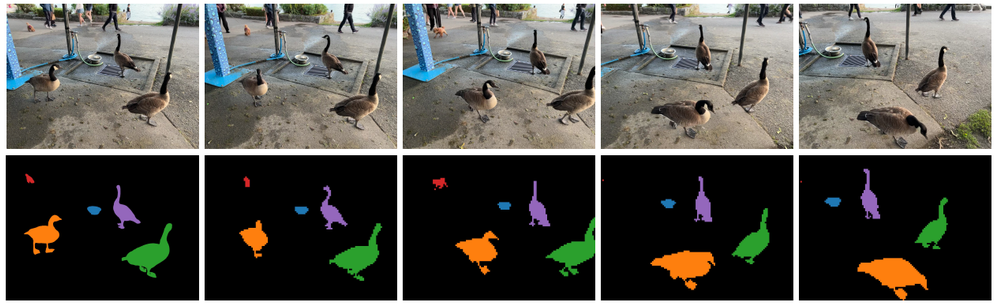


Figure 15: Segmentation tracking example. Given the ground-truth instance segmentation masks for the initial frame, we propagate the instance labels to subsequent frames according to patch similarity in the feature space of DINOv3. The input resolution is 2048×1536 pixels, resulting in 128×96 patches.

frame. During evaluation, we either take a single clip per video, or use test-time augmentation (TTA) by averaging the predictions of 3 spatial and 2 temporal crops per video. See App. D.6 for experimental details. We run this evaluation on three datasets: UCF101 (Soomro et al., 2012), Something-Something V2 (Goyal et al., 2017), and Kinetics-400 (Kay et al., 2017), and report top-1 accuracy. As an additional baseline, we report the performance of V-JEPA v2, a state-of-the-art SSL model for video understanding.

Results (Tab. 6) In line with the conclusion of the previous experiment, we find that DINOv3 can be successfully used for extracting strong video features. As this evaluation involves training several layers of self-attention, the differences between models are less visible. However, DINOv3 lands in the same range as PEcore and SigLIP 2, and clearly outperforms other models (DINOv2, AM-RADIO) across datasets. UCF101 and K400 are appearance-focused, where strong category-level understanding of objects gives most of the performance. SSv2 on the other hand, requires better understanding of motion—the dedicated video model V-JEPA v2 shines on this dataset. Interestingly, the gap between DINOv3 and the weakly-supervised models is slightly bigger on this dataset. This again confirms the suitability of DINOv3 to video tasks.

表5： 视频分割跟踪评估。我们在多个分辨率下报告了DAVIS、YouTube-VOS和MOSE上的$\mathcal{J}\&\mathcal{F}$-均值。对于补丁大小为14/16的模型，小、中和大分辨率分别对应视频短边为420/480、840/960、1260/1140像素。

		DAVIS			YouTube-VOS			MOSE		
方法	ViT	S	M	L	S	M	L	S	M	L
聚合主干										
AM-RADIOv2.5	g/14	66.5	77.3	81.4	70.1	78.1	79.2	44.0	52.6	54.3
PEspatial	G/14	68.4	74.5	70.5	68.5	67.5	55.6	39.3	40.2	34.0
弱监督主干										
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
自监督骨干网络										
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**

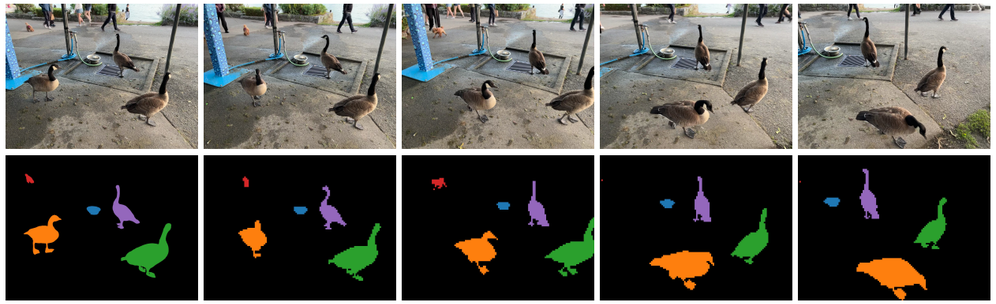


图15： 分割跟踪示例。给定初始帧的真实标签实例分割掩码，我们根据DINOv3特征空间中的块相似度将实例标签传播到后续帧。输入分辨率为 2048×1536 像素，产生 128×96 个块。

帧提取的。在评估过程中，我们对每个视频取单个片段，或者使用测试时增强（TTA），通过平均每个视频的3个空间和2个时间裁剪的预测。参见附录D.6 以获取实验细节。我们在三个数据集上运行此评估：UCF101（Soomro等人，2012），Something-Something V2（Goyal等人，2017），和Kinetics-400（Kay等人，2017），并报告top-1准确率。作为另一个基线，我们报告了V-JEPA v2的性能，它是一个用于视频理解的最先进的SSL模型。

结果（表6）与先前实验的结论一致，我们发现DINOv3可以成功用于提取强视频特征。由于此评估涉及训练多层自注意力机制，模型之间的差异不太明显。然而，DINOv3落在PEcore和SigLIP 2的同一范围内，并且在所有数据集上都明显优于其他模型（DINOv2、AM-RADIO）。UCF101和K400是外观导向的，其中对物体进行强类别级理解提供了大部分性能。另一方面，SSv2需要更好的运动理解——专用的视频模型V-JEPA v2在这个数据集上表现突出。有趣的是，DINOv3与弱监督模型之间的差距在这个数据集上略大。这再次证实了DINOv3适用于视频任务。

Table 6: Video classification evaluation using attentive probes. We report top-1 accuracy on UCF101, Something-Something V2 (SSv2), and Kinetics-400 (K400). For each model, we report performance for evaluating a single clip per video, or applying test-time augmentation (TTA) by averaging the predicted probabilities from multiple clips.

		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

6.2 DINOv3 has Robust and Versatile Global Image Descriptors

In this section, we evaluate DINOv3's ability to capture global image statistics. To this end, we consider classic classification benchmarks using linear probes (Sec. 6.2.1) and instance retrieval benchmarks (Sec. 6.2.2). Again, we compare to the strongest publicly available image encoders. In addition to the models from the previous section, we evaluate the two weakly supervised models AIMv2 (Fini et al., 2024), trained using joint auto-regressive pixel and text prediction, and the massive EVA-CLIP-18B (Sun et al., 2024).

6.2.1 Image Classification with Linear Probing

We train a linear classifier on top of DINOv3's output CLS token to evaluate the model on classification benchmarks. We consider the ImageNet1k (Deng et al., 2009) dataset and its variants to evaluate out-of-distribution robustness, and a suite of datasets from different domains to understand DINOv3's ability to distinguish fine-grained classes. See App. D.7 for evaluation details.

Domain Generalization from ImageNet (Tab. 7) In this experiment, we train on ImageNet-*train*, use ImageNet-*val* as a *validation set* to select hyperparameters, and transfer the best found classifier to different test datasets: ImageNet-**V2** (Recht et al., 2019) and **ReaL** (Beyer et al., 2020) are alternative sets of images and labels for ImageNet, used to test overfitting on the ImageNet validation set; **R**endition (Hendrycks et al., 2021a) and **S**ketch (Wang et al., 2019) show stylized and artificial versions of the ImageNet classes; **A**dversarial (Hendrycks et al., 2021b) and **Obj**ectNet (Barbu et al., 2019) contain deliberately-chosen difficult examples; **C**orruptions (Hendrycks and Dietterich, 2019) measures robustness to common image corruptions. For reference, we also list linear probing results from Dehghani et al. (2023) for ViTs trained using supervised classification on the massive JFT dataset (3B–4B images). Note that these results follow a slightly different evaluation protocol and are not directly comparable to our results.

DINOv3 significantly surpasses all previous self-supervised backbones, with gains of +10% on ImageNet-R, +6% on -Sketch, +13% on ObjectNet over the previously strongest SSL model DINOv2. We note that the strongest weakly-supervised models, SigLIP 2 and PE, are now better than the strongest supervised ones (ViT-22B) on hard OOD tasks like ImageNet-A and ObjectNet. DINOv3 reaches comparable results on ImageNet-R and -Sketch, and, on the hard tasks ImageNet-A and ObjectNet, is closely behind PE, while exceeding SigLIPv2. On ImageNet, while validation scores are 0.7–0.9 points behind SigLIPv2 and PE, the performance on the "cleaner" test sets -V2 and -ReaL is virtually the same. Notably, DINOv3 achieves the best robustness to corruptions (ImageNet-C). All in all, *this is the first time that a SSL model has reached*

表 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一个片段的性能，或通过平均多个片段的预测概率应用测试时增强 (TTA) 的性能。

		UCF101		SSv2		K400	
方法	ViT	单个	TTA	单个	TTA	单个	TTA
聚合主干							
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
Weakl y-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**
自监督骨干网络							
DINOv2	g/14	93.5	93.8	67.4	68.4	84.4	85.6
V-JEPA 2	g/16	**94.0**	93.8	**73.8**	**75.4**	83.3	84.3
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

6.2 DINOv3 具有鲁棒和通用的全局图像描述符

在本节中，我们评估DINOv3捕获全局图像统计的能力。为此，我们考虑使用线性探针的经典分类基准（第6.2.1节）和实例检索基准（第6.2.2节）。同样，我们与最强的公开可用的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等人。，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等人。，2024）。

6.2.1 使用线性探针的图像分类

我们在 DINOv3 的输出 CLS 标记上训练一个线性分类器，以在分类基准测试上评估模型。我们考虑 ImageNet1k (邓等人， 2009) 数据集及其变体，以评估分布外鲁棒性，并考虑来自不同领域的套件数据集，以了解 DINOv3 区分细粒度类别的能力。参见附录 D.7 以获取评估详细信息。

从 ImageNet 到领域泛化 (表 7) 在此实验中，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迁移到不同的测试数据集：ImageNet-V2 (Recht 等人，2019) 和 ReaL (贝耶等人，2020) 是 ImageNet 的替代图像和标签集，用于测试在 ImageNet 验证集上的过拟合；Rendition (Hendrycks 等人，2021a) 和 Sketch (王等人，2019) 显示了 ImageNet 类的样式化和人工版本；Adversarial (Hendrycks 等人，2021b) 和 ObjectNet (Barbu 等人，2019) 包含故意选择的困难示例；Corruptions (Hendrycks 和 Dietterich，2019) 衡量了对常见图像损坏的鲁棒性。为参考，我们还列出了 Dehghani 等人 (2023) 对 ViTs 在使用监督分类在庞大的 JFT 数据集 (3B–4B 图像) 上训练的结果。请注意，这些结果遵循略微不同的评估协议，并且不能直接与我们的结果进行比较。

DINOv3显著超越了所有之前的自监督骨干网络，在ImageNet-R上提升了 +10%，在-Sketch上提升了+6%，在ObjectNet上提升了 +13%，超过了之前最强的SSL模型DINOv2。我们注意到，最强的弱监督模型SigLIP 2和PE，在ImageNet-A和ObjectNet这类困难OOD任务上已经优于最强的监督模型（ViT-22B）。DINOv3在ImageNet-R和-Sketch上取得了可比的结果，而在ImageNet-A和ObjectNet这类困难任务上，其表现紧随PE之后，同时超过了SigLIPv2。在ImageNet上，虽然验证分数比SigLIPv2和PE低0.7–0.9分，但在“更干净”的测试集-V2和-ReaL上，性能几乎相同。值得注意的是，DINOv3对corruptions（ImageNet-C）的鲁棒性最佳。总而言之，这是*SSL*模型首次达到

Table 7: Classification accuracy of linear probes trained on ImageNet1k with frozen backbones. Weakly- and self-supervised models are evaluated with image resolution adapted to 1024 patch tokens (*i.e.* 448×448 for patch size 14, 512×512 for patch size 16). For reference, we also list results from Dehghani et al. (2023) using a different evaluation protocol (marked with *).

		ImageNet			Rendition		Hard		
Method	ViT	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

Table 8: Finegrained classification benchmarks. Fine-S averages over 12 datasets, see Tab. 22 for full results.

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: Instance recognition benchmarks. See Tab. 23 for additional metrics.

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**

comparable results to weakly- and supervised models on image classification—a domain which used to be the strong point of (weakly-)supervised training approaches. This is a remarkable result, given that models like ViT-22B, SigLIP 2, and PE are trained using massive human-annotated datasets. In contrast, DINOv3 learns purely from images, which makes it feasible to further scale/improve the approach in the future.

Finegrained Classification (Tab. 8) We also measure DINOv3's performance when training linear probes on several datasets for fine-grained classification. In particular, we report the accuracy on 3 large datasets, namely Places205 (Zhou et al., 2014) for scene recognition, and iNaturalist 2018 (Van Horn et al., 2018) and iNaturalist 2021 (Van Horn et al., 2021)) for detailed plant and animal-species recognition, as well as the average over 12 smaller datasets covering scenes, objects, and textures (as in Oquab et al. (2024), here termed Fine-S). See also Tab. 22 for individual results on those datasets.

表 7： 在 ImageNet1k 上使用冻结主干网络训练的线性探针的分类准确率。弱监督和自监督模型使用图像分辨率调整为 1024 块标记（即。 448×448对于块大小 14， 512×512 对于块大小 16）。为参考，我们还列出了 Dehghani 等人 (2023) 使用不同评估协议的结果（用 *标记）。

		ImageNet			渲染		Hard		
方法	ViT	Val	V2	ReaL	R	S	A	C ↓	Obj.
监督主干									
翟等人 (2022a)*	G/14	89.0	81.3	90.6	91.7	—	78.8	—	69.6
Chen 等人 (2023)*	e/14	89.3	82.5	90.7	94.3	—	81.6	—	71.5
Dehghani 等人 (2023)*	22B/14	89.5	83.2	90.9	94.3	—	83.8	—	74.3
聚合主干									
AM-RADIOv2.5	g/14	88.0	80.2	90.3	83.8	67.1	81.3	27.1	68.4
Weakly-supervised 聚合主干									
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
自监督骨干网络									
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

表8： 细粒度分类基准。Fine-S 在12个数据集上的平均值为，参见表22以获取完整结果。

方法	ViT	Fine-S	地点	iNat18	iNat21
聚合主干					
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
自监督骨干网络					
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**

表9： 实例识别基准。参见表23以获取附加指标。

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**

在图像分类领域取得与弱监督和监督模型可比的结果——这一领域曾是（弱-)监督训练方法的优势所在。这是一个显著的结果，考虑到ViT-22B、SigLIP 2和PE等模型都是使用大规模人工标注数据集训练的。相比之下，DINOv3纯粹从图像中学习，这使得未来进一步扩展/改进该方法成为可能。

细粒度分类 (表8) 我们还测量了 DINOv3 在多个数据集上训练线性探针进行细粒度分类时的性能。具体来说，我们报告了在 3 个大型数据集上的准确率，即 Places205 (周 等人， 2014) 用于场景识别，以及 iNaturalist 2018 (范·霍恩等人,2018) 和 iNaturalist 2021 (范·霍恩等人， 2021)) 用于详细植物和动物物种识别，以及 12 个较小数据集上的平均值，这些数据集涵盖场景、物体和纹理（如 Oquab 等人 (2024)，此处称为 Fine-S）。另见表22在这些数据集上的单个结果。

We find that, again, DINOv3 surpasses all previous SSL methods. It also shows competitive results compared to the weakly-supervised methods, indicating its robustness and generalization capability across diverse finegrained classification tasks. Notably, DINOv3 attains the highest accuracy on the difficult iNaturalist21 dataset at 89.8%, outperforming even the best weakly-supervised model PEcore with 87.0%.

6.2.2 Instance Recognition

To evaluate the instance-level recognition capabilities of our model, we adopted a non-parametric retrieval approach. Here, database images are ranked by their cosine similarity to a given query image, using the output CLS token. We benchmark performance across several datasets: the Oxford and Paris datasets for landmark recognition (Radenović et al., 2018), the Met dataset featuring artworks from the Metropolitan Museum (Ypsilantis et al., 2021), and AmsterTime, which consists of modern street view images matched to historical archival images of Amsterdam (Yildiz et al., 2022). Retrieval effectiveness is quantified using mean average precision for Oxford, Paris, and AmsterTime, and global average precision for Met. See App. D.8 for more evaluation details.

Results (Tabs. 9 and 23) Across all evaluated benchmarks, DINOv3 achieves the strongest performance by large margins, *e.g.* improving over the second best model DINOv2 by +10.8 points on Met and +7.6 points on AmsterTime. On this benchmark, weakly-supervised models are lagging far behind DINOv3, with the exception of AM-RADIO, which is distilled from DINOv2 features. These findings highlight the robustness and versatility of DINOv3 for instance-level retrieval tasks, spanning both traditional landmark datasets and more challenging domains such as art and historical image retrieval.

6.3 DINOv3 is a Foundation for Complex Computer Vision Systems

The previous two sections already provided solid signal for the quality of DINOv3 in both dense and global tasks. However, these results were obtained under "model probing" experimental protocols, using lightweight linear adapters or even non-parametric algorithms to assess the quality of features. While such simple evaluations allowed to remove confounding factors from involved experimental protocols, they are not enough to evaluate the full potential of DINOv3 as a foundational component in a larger computer vision system. Thus, in this section, we depart from the lightweight protocols, and instead train more involved downstream decoders and consider stronger, task-specific baselines. In particular, we use DINOv3 as a basis for (1) object detection with Plain-DETR (Sec. 6.3.1), (2) semantic segmentation with Mask2Former (Sec. 6.3.2), (3) monocular depth estimation with Depth Anything (Sec. 6.3.3), and (4) 3D understanding with the Visual Geometry Grounded Transformer (Sec. 6.3.4). These tasks are only intended as explorations for what is possible with DINOv3. Still, we find that building on DINOv3 unlocks competitive or even state-of-the-art results with little effort.

6.3.1 Object Detection

As a first task, we tackle the long-standing computer vision problem of object detection. Given an image, the goal is to provide bounding boxes for all instances of objects of pre-defined categories. This task requires both precise localization and good recognition, as boxes need to match the object boundaries and correspond to the correct category. While performance on standard benchmarks like COCO (Lin et al., 2014) is mostly saturated, we propose to tackle this task with a *frozen* backbone, only training a small decoder on top.

Datasets and Metrics We evaluate DINOv3 on object detection capabilities with the COCO dataset (Lin et al., 2014), reporting results on the COCO-VAL2017 split. Additionally, we evaluate out-of-distribution performance on the COCO-O evaluation dataset (Mao et al., 2023). This dataset contains the same classes but provides input images under six distribution shift settings. For both datasets, we report mean Average Precision (mAP) with IoU thresholds in $[0.5 : 0.05 : 0.95]$. For COCO-O, we additionally report the effective robustness (ER). Since COCO is a small dataset, comprising only 118k training images, we leverage the larger Objects365 dataset (Shao et al., 2019) for pre-training the decoder, as is common practice.

我们发现，DINOv3 再次超越了所有之前的 SSL 方法。它在与弱监督方法的比较中也表现出有竞争力的结果，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 在困难的数据集 iNaturalist21 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（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。在此方法中，数据库图像根据其与给定查询图像的余弦相似度，使用CLS标记进行排序。我们在多个数据集上基准测试性能：用于地标识别的牛津数据集和巴黎数据集（Radenović等人，2018），包含大都会艺术博物馆艺术品的Met数据集（Ypsilantis等人，2021），以及由现代街道视图图像与阿姆斯特丹历史档案图像匹配而成的AmsterTime（Yildiz等人，2022）。检索有效性使用牛津、巴黎和AmsterTime的平均精度均值进行量化，使用全局平均精度进行Met的量化。更多评估细节请参见附录D.8。

结果（标签页。9 和 23） 在所有评估基准测试中，DINOv3 以巨大优势取得了最佳性能，例如。在 Met 上比第二好的模型 DINOv2 提高了 +10。8 分，在 AmsterTime 上提高了 +7。6 分。在这个基准测试中，弱监督模型远远落后于 DINOv3，唯一的例外是 AM-RADIO，它是由 DINOv2 特征蒸馏而来的。这些发现突出了 DINOv3 在实例级检索任务中的鲁棒性和多功能性，涵盖了传统地标数据集以及更具挑战性的领域，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石

前两节已经为 DINOv3 在密集和全局任务中的质量提供了可靠的信号。然而，这些结果是在“模型探测”实验协议下获得的，使用轻量级线性适配器甚至非参数算法来评估特征的质量。虽然这种简单的评估允许从涉及的实验协议中消除混杂因素，但它们不足以评估 DINOv3 作为更大计算机视觉系统中基础组件的全面潜力。因此，在本节中，我们脱离轻量级协议，转而训练更复杂的下游解码器，并考虑更强的、任务特定的基线。特别是，我们使用 DINOv3 作为 (1) 使用 Plain-DETR 的目标检测（第 6.3.1 节）、(2) 使用 Mask2Former 的语义分割（第 6.3.2 节）、(3) 使用 Depth Anything 的单目深度估计（第 6.3.3 节），以及 (4) 使用视觉几何基础 Transformer 的 3D 理解（第 6.3.4 节）。这些任务仅旨在探索 DINOv3 的可能性。尽管如此，我们发现基于 DINOv3 可以以少量努力获得具有竞争力的甚至最先进的结果。

6.3.1 目标检测

作为一项首要任务，我们着手解决计算机视觉领域长期存在的目标检测问题。给定一张图像，目标是为预定义类别中所有物体的实例提供边界框。这项任务需要精确的位置定位和良好的识别能力，因为边界框需要与物体边界匹配并对应到正确的类别。虽然标准基准测试如COCO（Lin 等人， 2014）的性能已基本饱和，但我们提出使用一个冻结的骨干网络，仅在顶部训练一个小型解码器。

数据集和指标 我们在COCO数据集（Lin等人， 2014）上使用COCO-VAL2017划分评估DINOv3的目标检测能力，并报告结果。此外，我们在COCO-O评估数据集（Mao等人， 2023）上评估分布外性能。该数据集包含相同的类别，但提供六种分布偏移设置下的输入图像。对于这两个数据集，我们使用IoU阈值为 $[0.5: 0.05: 0.95]$. 对于COCO-O，我们额外报告有效鲁棒性（ER）。由于COCO是一个小数据集，仅包含118k个训练图像，我们利用更大的Objects365数据集（Shao等人， 2019）预训练解码器，这符合常见做法。

Table 10: Comparison with state-of-the-art systems on object detection. We train a detection adapter on top of a *frozen* DINOv3 backbone. We show results on the validation set of the COCO and COCO-O datasets, and report the mAP across IoU thresholds, as well as the effective robustness (ER). Our detection system based on DINOv3 sets a new state of the art. As the InternImage-G detection model has not been released, we were unable to reproduce their results or compute COCO-O scores.

			Parameters			COCO		COCO-O	
Model	Detector	FT	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**

Implementation We build upon the Plain-DETR (Lin et al., 2023b), but make the following modification: We do not fuse the transformer encoder into the backbone, but keep it as a separate module, similar to the original DETR (Carion et al., 2020), which allows us to keep the DINOv3 backbone completely frozen during training and inference. To the best of our knowledge, this makes it *the first competitive detection model to use a frozen backbone*. We train the Plain-DETR detector on Objects365 for 22 epochs at resolution 1536, then one epoch at resolution 2048, followed by 12 epochs on COCO at resolution 2048. At inference time, we run at resolution 2048. Optionally, we also apply test-time augmentation (TTA) by forwarding the image at multiple resolutions (from 1536 to 2880). See App. D.9 for full experimental details.

Results (Tab. 10) We compare our system with four models: EVA-02 with a Cascade detector (Fang et al., 2024b), EVA-02 with Co-DETR (Zong et al., 2023), InternImage-G with DINO (Wang et al., 2023b), and PEspatial with DETA (Bolya et al., 2025). We find that our lightweight detector (100M parameters) trained on top of a frozen DINOv3 backbone manages to reach state-of-the-art performance. For COCO-O, the gap is pronounced, showing that the detection model can effectively leverage the robustness of the DINOv3. Interestingly, our model outperforms all previous models with much fewer trained parameters, with the smallest comparison point still using more than 300M trainable parameters. We argue that achieving such strong performance without specializing the backbone is an enabler for various practical applications: A single backbone forward can provide features that support multiple tasks, reducing compute requirements.

6.3.2 Semantic Segmentation

Following the previous experiment, we now evaluate on semantic segmentation, another long-standing computer vision problem. This task also requires strong, well localized representations, and expects a dense per-pixel prediction. However, opposed to object detection, the model does not need to differentiate instances of the same object. Similar to detection, we train a decoder on top of a *frozen* DINOv3 model.

Datasets and Metrics We focus our evaluation on the ADE20k dataset (Zhou et al., 2017), which contains 150 semantic categories across 20k training images and 2k validation images. We measure performance using the mean Intersection over Union (mIoU). To train the segmentation model, we additionally use the COCO-Stuff (Caesar et al., 2018) and Hypersim (Roberts et al., 2021) datasets. Those contain 164k images with 171 semantic categories, and 77k images with 40 categories respectively.

Implementation To build a decoder that maps DINOv3 features to semantic categories, we combine ViT-Adapter (Chen et al., 2022) and Mask2Former (Cheng et al., 2022), similar to prior work (Wang et al., 2022b; 2023b;a). However, in our case, the DINOv3 backbone remains frozen during training. In order to avoid altering the backbone features, we further modify the original ViT-Adapter architecture by removing the injector component. Compared to baselines, we also increase the embedding dimensions from 1024 to 2048, to support processing the 4096-dimensional output of the DINOv3 backbone. We start by pre-training the

表10: 与当前最佳系统在目标检测上的比较。我们在<code>DINOv3主干网络</code>上训练了一个<code>检测适配器</code>。我们在COCO和COCO-O数据集的验证集上展示结果，并报告了跨IoU阈值的mAP以及有效鲁棒性（ER）。我们的基于DINOv3的<code>检测系统</code>设立了新的当前最佳。由于<code>InternImage-G检测模型</code>尚未发布，我们无法复现他们的结果或计算COCO-O分数。

			参数			COCO		COCO-O	
模型	检测器	FT	编码器	解码器	可训练	简单	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**

实现 我们基于Plain-DETR（Lin等人， 2023b）进行构建，但进行了以下修改：我们不将变换器编码器与骨干网络融合，而是将其保留为独立模块，类似于原始DETR（Carion等人， 2020），这使我们能够在训练和推理过程中完全冻结DINOv3骨干网络。据我们所知，这使其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1536分辨率训练Plain-DETR检测器22个epoch，然后以2048分辨率训练1个epoch，接着在COCO上以2048分辨率训练12个epoch。推理时，我们运行在2048分辨率。可选地，我们还通过在1536到2880的多个分辨率上传递图像应用测试时增强（TTA）。参见附录D.9 以获取完整实验细节。

Results (Tab. 10) 我们比较了我们的系统与四个模型：EVA-02配合级联检测器（Fang 等人， 2024b），EVA-02配合Co-DETR（Zong 等人， 2023），InternImage-G配合DINO（王等人， 2023b），以及PEspatial配合DETA（Bolya 等人， 2025）。我们发现，我们轻量级检测器（100M参数）在冻结的DINOv3主干上训练，成功达到了最先进性能。对于COCO-O，差距非常明显，表明检测模型能有效利用DINOv3的鲁棒性。有趣的是，我们的模型在训练参数远少的情况下优于所有先前模型，最小的对比点仍使用超过300M可训练参数。我们认为，在不专门设计主干的情况下实现如此强大的性能，是多种实际应用的基础：单个主干前向可以提供支持多任务的特征，从而降低计算需求。

6.3.2 语义分割

在之前的实验之后，我们现在评估语义分割，这是另一个长期存在的计算机视觉问题。这个任务也需要强大且良好定位的表示，并期望像素级的密集预测。然而，与目标检测不同，模型不需要区分相同对象的实例。类似于检测，我们在一个<code style id='1'>冻结的</code>DINOv3模型之上训练一个解码器。

<code style id='1'>数据集和指标</code><code style id='3'>我们的评估集中在ADE20k数据集（</code><code style id='5'>周 等人</code><code style id='7'>， </code><code style id='9'>2017</code><code style id='11'>），它包含150个语义类别，跨越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，我们额外使用了COCO- Stuff（</code><code style id='13'>Caesar 等人</code><code style id='15'>，</code><code style id='17'>2018</code><code style id='19'>）和Hypersim（</code><code style id='21'>Roberts 等人</code><code style id='23'>， </code><code style id='25'>2021</code><code style id='27'>）数据集。这些数据集分别包含164k张图像和171个语义类别，以及77k张图像和40个类别。

实现 为了构建一个将 DINOv3 特征映射到语义类别的解码器，我们将 ViT-Adapter（Chen 等人， 2022）和 Mask2Former（Cheng 等人， 2022）结合起来，类似于先前的工作（王等人， 2022b;2023b;a）。然而，在我们的情况下，DINOv3 骨干网络在训练期间保持冻结。为了防止改变骨干网络特征，我们进一步修改了原始 ViT-Adapter 架构，通过移除注入器组件。与基线相比，我们还将嵌入维度从 1024 增加到 2048，以支持处理 DINOv3 骨干网络的 4096-维输出。我们从对

Table 11: Comparison with state-of-the-art systems for semantic segmentation on ADE20k. We evaluate the model in a single- or multi-scale setup (respectively Simple and TTA). Following common practice, we run this evaluation at resolution 896 and report mIoU scores. BEIT3, ONE-PEACE and DINOv3 use a Mask2Former with ViT-Adapter architecture, and the decoder parameters take into account both. We report results on further datasets in Tab. 24

		Parameters			mIoU	
Model	FT	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**

segmentation decoder on COCO-Stuff for 80k iterations, followed by 10k iterations on Hypersim (Roberts et al., 2021). Finally, we train for 20k iterations on the training split of ADE20k and report results on the validation split. All training is done at an input resolution of 896. At inference time we consider two setups: single-scale, *i.e.* we forward images at training resolution, or multi-scale, *i.e.* we average predictions at multiple image ratios between $\times 0.9$ and 1.1 the original training resolution. We refer to App. D.10 for more experimental details.

Results (Tab. 11) We compare our model's performance with several state-of-the-art baselines, including BEIT-3 (Wang et al., 2022b), InternImage-H (Wang et al., 2023b) and ONE-PEACE (Wang et al., 2023a), and report results on additional datasets in Tab. 24. Our segmentation model based on the frozen DINOv3 backbone reaches state-of-the-art performance, equaling that of ONE-PEACE (63.0 mIoU). It also improves over all prior models on the COCO-Stuff (Caesar et al., 2018) and VOC 2012 (Everingham et al., 2012) datasets. As semantic segmentation requires accurate per-pixel predictions, vision transformer backbones pose a fundamental problem. Indeed, the 16 pixel-wide input patches make the granularity of the prediction relatively coarse—encouraging solutions like ViT-Adapter. On the other hand, we have shown that we can obtain high-quality feature maps, even at very high resolutions up to 4096 (*c.f.* Figs. 3 and 4); this corresponds to dense feature maps 512-tokens wide. We hope that future work will be able to leverage these high-resolution features to reach state-of-the-art performance without having to rely on heavy decoders like ViT-Adapter with Mask2Former.

6.3.3 Monocular Depth Estimation

We now consider building a system for monocular depth estimation. To do so, we follow the setup of Depth Anything V2 (DAv2) (Yang et al., 2024b), a recent state-of-the-art method. The key innovation of DAv2 is to use a large collection of synthetically generated images with ground truth depth annotations. Critically, this relies on DINOv2 as a feature extractor that is able to bridge the *sim-to-real* gap, a capability that other vision backbones like SAM (Kirillov et al., 2023) do not show (Yang et al., 2024b). Thus, we swap DINOv2 with DINOv3 in the DAv2 pipeline to see if we can achieve similar results.

Implementation Like DAv2, we use a Dense Prediction Transformer (DPT) (Ranftl et al., 2021) to predict a pixelwise depth field, using features from four equally spaced layers of DINOv3 as input. We train the model using the set of losses from DAv2 on DAv2's synthetic dataset, increasing the training resolution to 1024×768 to make use of DINOv3's high resolution capabilities. In contrast to DAv2, we *keep the backbone frozen* instead of finetuning it, testing the out-of-the-box capabilities of DINOv3. We also found it beneficial to scale up the DPT head to obtain the full potential DINOv3 7B's larger features. See App. D.11 for details.

Datasets and Metrics We evaluate our model on 5 real-world datasets (NYUv2 (Silberman et al., 2012), KITTI (Geiger et al., 2013), ETH3D (Schöps et al., 2017), ScanNet (from Ke et al. (2025)) and DIODE (Vasil-jevic et al., 2019)) in the zero-shot scale-invariant depth setup, similar to Ranftl et al. (2020); Ke et al. (2025);

表11： 在ADE20k上的语义分割方面与最先进系统的比较。我们以单尺度或多尺度设置（分别对应简单和TTA）评估模型。遵循常见做法，我们在分辨率896的情况下运行此评估，并报告mIoU分数。BEIT3、ONE-PEACE和DINOv3使用Mask2Former与ViT-Adapter架构，解码器参数考虑了这两者。我们在其他数据集上的结果报告在表24

		参数			mIoU	
模型	FT	编码器	解码器	可训练	简单	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**

在 COCO-Stuff 上的分割解码器上 80k 次迭代，然后 10k 次迭代在 Hypersim（Roberts 等人， 2021）上进行。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。所有训练均在输入分辨率 896 下进行。在推理时，我们考虑两种设置：单尺度， 即。我们在训练分辨率前向传播图像，或多尺度， 即。我们在 $\times 0.9$ 和 1之间的多个图像比例处平均预测。1 原始训练分辨率。我们参考 附录 D.10 以获取更多实验细节。

Results (表11) 我们比较了我们的模型的性能与几个最先进基线，包括BEIT-3 (王等人， 2022b)，InternImage-H (王等人， 2023b) 和ONE-PEACE (王等人， 2023a)，并在 表24中报告了在附加数据集上的结果。我们的基于冻结的DINOv3主干网络的分割模型达到了最先进性能，与ONE-PEACE (63.0 mIoU) 相当。它在COCO-Stuff (Caesar 等人， 2018) 和VOC 2012 (Everingham 等人， 2012) 数据集上均优于所有先前模型。由于语义分割需要精确的逐像素预测，视觉Transformer主干网络提出了一个基本问题。实际上，16像素宽的输入块使得预测的粒度相对粗糙——这鼓励了像ViT-Adapter这样的解决方案。另一方面，我们已经证明，即使在高分辨率高达4096 (参见。 图3 和 4)；这对应于512-token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，在不依赖像ViT-Adapter与Mask2Former这样重型解码器的情况下达到最先进性能。

6.3.3 单目深度估计

我们现在考虑构建一个单目深度估计系统。为此，我们遵循深度 anything V2 (DAv2) (杨等人， 2024b) 的设置，这是一种最近的最先进技术。DAv2 的关键创新是使用大量具有真实值深度标注的合成图像。关键在于，这依赖于 DINOv2 作为特征提取器，它能够弥合 仿真到真实 的差距，而其他视觉主干如 SAM (Kirillov 等人， 2023) 并不显示 (杨等人， 2024b)。因此，我们在 DAv2 管道中用 DINOv3 替换 DINOv2，看看我们是否能获得类似的结果。

实现 与 DAv2 类似，我们使用密集预测 Transformer (DPT) (Ranftl 等人， 2021) 来预测像素级的深度场，使用 DINOv3 的四个等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，并将训练分辨率增加到1024×768 以利用 DINOv3 的高分辨率能力。与 DAv2 不同的是，我们 冻结主干 而不是微调它，测试 DINOv3 的开箱即用能力。我们还发现将 DPT 头扩展到获得 DINOv3 7B 更大的特征是有益的。见附录 D.11 的详细信息。

数据集和指标 我们在 5 个真实世界数据集（NYUv2 (Silberman等人， 2012)，KITTI (Geiger等人,2013)，ETH3D (Schöps等人， 2017)，ScanNet (来自 Ke等人 (2025)) 和 DIODE (Vasil-jevic等人， 2019)) 上使用零样本尺度不变深度设置进行评估，类似于 Ranftl等人 (2020)； Ke等人 (2025)；

Table 12: Comparison with state-of-the-art systems for relative monocular depth estimation. By combining DINOv3 with Depth Anything V2 (Yang et al., 2024b), we obtain a SotA model for relative depth estimation.

		NYUv2		KITTI		ETH3D		ScanNet		DIODE	
Method	FT	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**

Yang et al. (2024b). We report the standard metrics absolute relative error (ARel) (lower is better) and δ_1 (higher is better). We refer to Yang et al. (2024a) for a description of those metrics.

Results (Tab. 12) We compare to the state of the art for relative depth estimation: MiDaS (Ranftl et al., 2020), LeReS (Yin et al., 2021), Omnidata (Eftekhar et al., 2021), DPT (Ranftl et al., 2021), Marigold in the ensemble version (Ke et al., 2025) and DAv2. Our depth estimation model reaches a new state-of-the-art on all datasets, only lacking behind in ARel on DIODE compared to DPT. Remarkably, this is possible using a *frozen backbone*, whereas all other baselines need to finetune the backbone for depth estimation. In addition, this validates that DINOv3 inherits DINOv2's *strong sim-to-real capabilities*, a desirable property that opens up the possibility for downstream tasks to use synthetically generated training data.

6.3.4 Visual Geometry Grounded Transformer with DINOv3

Finally, we consider 3D understanding with the recent Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025). Trained on a large set of 3D-annotated data, VGGT learns to estimate all important 3D attributes of a scene, such as camera intrinsics and extrinsics, point maps, or depth maps, in a single forward pass. Using a simple, unified pipeline, it reaches state-of-the-art results on many 3D tasks while being more efficient than specialized methods—constituting a major advance in 3D understanding.

Implementation VGGT uses a DINOv2-pretrained backbone to obtain representations for different views of a scene, before fusing them with a transformer. Here, we simply swap the DINOv2 backbone with DINOv3, using our ViT-L variant (see Sec. 7) to match DINOv2 ViT-L/14 in the original work. We run the same training pipeline as VGGT, including finetuning of the image backbone. We switch the image resolution from 518×518 to 592×592 to accommodate DINOv3's patch size 16 and keep the the results comparable to VGGT. We additionally adopt a small number of hyperparameter changes detailed in App. D.12.

Datasets and Metrics Following Wang et al. (2025), we evaluate on camera pose estimation on the Re10K (Zhou et al., 2018) and CO3Dv2 (Reizenstein et al., 2021) datasets, dense multi-view estimation on DTU (Jensen et al., 2014), and two-view matching on ScanNet-1500 (Dai et al., 2017). For camera pose estimation and two-view matching, we report the standard area-under-curve (AUC) metric. For multi-view estimation, we report the smallest L2-distance between prediction to ground truth as "Accuracy", the smallest L2-distance from ground truth to prediction as "Completeness" and their average as 'Overall". We refer to Wang et al. (2025) for details about method and evaluation.

Results (Tab. 13) We find that VGGT equipped with DINOv3 *further improves over the previous state-of-the-art* set by VGGT on all three considered tasks—using DINOv3 leads to clear and consistent gains. This is encouraging, given that we only applied minimal tuning for DINOv3. These tasks span different levels of visual understanding: high-level abstraction of scene content (camera pose estimation), dense geometric prediction (multi-view depth estimation), and fine-grained pixel-level correspondence (view matching). To-

表12： 与相对单目深度估计的最先进系统进行比较。通过将DINOv3与Depth Anything V2（杨等人， 2024b）相结合，我们获得了一个用于相对深度估计的最先进模型。

		NYUv2		KITTI		ETH3D		ScanNet		DIODE	
方法	FT	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**

杨等人 (2024b)。我们报告了标准指标绝对相对误差 (ARel)（越低越好）和 δ_1(越高越好)。我们参考 杨等人 (2024a) 了解这些指标的描述。

结果（表12）我们将相对深度估计与当前最佳技术进行比较：MiDaS（Ranftl等人， 2020），LeReS（Yin等人， 2021），Omnidata（Eftekhar等人， 2021），DPT（Ranftl等人， 2021），集成版本中的Marigold（Ke等人， 2025）和DAv2。我们的深度估计模型在所有数据集上都达到了新的最先进水平，仅在DIODE上的ARel方面落后于DPT。值得注意的是，这是通过使用冻结主干实现的，而所有其他基线都需要对主干进行微调以进行深度估计。此外，这验证了DINOv3继承了DINOv2的强大的仿真到真实能力，这是一个理想特性，为下游任务使用合成生成的训练数据打开了可能性。

6.3.4 视觉几何基础Transformer与DINOv3

最后，我们考虑使用最近的视觉几何基础Transformer（VGGT）进行3D理解（王等人， 2025）。VGGT在大量3D标注数据上进行训练，学习在单个前向传递中估计场景的所有重要3D属性，例如相机内参和外参、点图或深度图。使用一个简单、统一的流程，它在许多3D任务上达到最先进技术成果，同时比专用方法更高效——构成了3D理解方面的重大进步。

实现 VGGT使用一个DINOv2预训练的骨干网络来获取场景不同视角的表示，然后将其与一个变换器融合。在这里，我们简单地将DINOv2骨干网络替换为DINOv3，使用我们的ViT-L变体（见第7节）来匹配原始工作中的DINOv2 ViT-L/14。我们运行与VGGT相同的训练流程，包括图像骨干网络的微调。我们将图像分辨率从 518×518 调整为 592×592 以适应DINOv3的16像素补丁大小，并保持结果与VGGT的可比性。我们额外采用了一些超参数变化，详细信息见附录D.12。

数据集和指标 遵循 王等人 (2025)，我们在 Re10K (周等人， 2018) 和 CO3Dv2 (Reizenstein 等人， 2021) 数据集上评估相机位姿估计，在 DTU (Jensen 等人， 2014) 上进行密集多视图估计，以及在 ScanNet-1500 (Dai 等人， 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，我们报告预测与真实值之间的最小 L2 距离作为“准确率”，真实值到预测的最小 L2 距离作为“完整性”，以及它们的平均值为‘总体”。我们参考王等人 (2025) 了解方法和评估的详细信息。

Results (Tab. 13) 我们发现，配备 DINOv3 的 VGGT 进一步超越了之前 VGGT 在所有三个所考虑任务上的最佳表现——使用 DINOv3 带来了清晰且一致的提升。鉴于我们对 DINOv3 仅进行了最小化调优，这一结果令人鼓舞。这些任务涵盖了不同级别的视觉理解：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）和细粒度的像素级对应（视图匹配）。To-

Table 13: 3D understanding using Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025). Simply by swapping DINOv2 for DINOv3 ViT-L as the image feature extractor in the VGGT pipeline, we are able to obtain state-of-the-art results on various 3D geometry tasks. We reproduce baseline results from Wang et al. (2025). We also report methods using ground truth camera information, marked with *. Camera pose estimation results are reported with AUC@30.

(a) Camera pose estimation.

Method	Re10K	CO3Dv2
DUSt3R	67.7	76.7
MASt3R	76.4	81.8
VG GSfM v2	78.9	83.4
CUT3R	75.3	82.8
FLARE	78.8	83.3
VGGT	85.3	88.2
DINOv3	**86.3**	**89.6**

(b) Multi-view estimation on DTU.

Method	Acc.↓	Comp.↓	Overall↓
Gipuma*	0.283	0.873	0.578
CIDER*	0.417	0.437	0.427
MASt3R*	0.403	0.344	0.374
GeoMVSNet*	0.331	0.259	0.295
DUSt3R	2.677	0.805	1.741
VGGT	0.389	0.374	0.382
DINOv3	**0.375**	**0.361**	**0.368**

(c) View matching on ScanNet-1500.

Method	AUC@5	AUC@10
SuperGlue	16.2	33.8
LoFTR	22.1	40.8
DKM	29.4	50.7
CasMTR	27.1	47.0
Roma	31.8	53.4
VGGT	33.9	55.2
DINOv3	**35.2**	**56.1**

gether with the previous results on correspondence estimation (Sec. 6.1.3) and depth estimation (Sec. 6.3.3), we take this as further empirical evidence for the strong suitability of DINOv3 as a basis for 3D tasks. Additionally, we anticipate further improvements from using the larger DINOv3 7B model.

7 Evaluating the Full Family of DINOv3 Models

In this section, we provide quantitative evaluations on the family of models distilled from our 7B-parameters model (See Sec. 5.2). This family includes variants based on the Vision Transformer (ViT) and the ConvNeXt (CNX) architectures. We provide the detailed parameter counts and inference FLOPs for all models in Fig. 16a. These models cover a wide range of computational budgets to accommodate a broad spectrum of users and deployment scenarios. We conduct a thorough evaluation of all ViT (Sec. 7.1) and ConvNeXt variants to assess their performance across tasks.

Figure 2 provides an overview comparison of the DINOv3 family versus other model collections. The DINOv3 family significantly outperforms all others on dense prediction tasks. This includes specialized models distilled from supervised backbones like AM-RADIO and PEspatial. At the same time, our models achieve similar results on classification tasks, making them the optimal choice across compute budgets.

In Sec. 7.1 detail our ViT models and compare them to other open-source alternatives. Then, in Sec. 7.2, we discuss the ConvNeXt models. Finally, following Sec. 5.3, we trained a text encoder aligned with the output of our ViT-L model. We present multi-modal alignment results for this model in Sec. 7.3.

7.1 A Vision Transformer for Every Use Case

Our ViT family spans architectures from the compact ViT-S to the larger 840 million parameter ViT-H+ models. The former is designed to run efficiently on resource-constrained devices such as laptops, the latter delivers state-of-the-art performance for more demanding applications. We compare our ViT models to the best open-source image encoders of corresponding size, namely DINOv2 (Oquab et al., 2024), SigLIP 2 (Tschannen et al., 2025) and Perception Encoder (Bolya et al., 2025). For a fair comparison, we ensure that the input sequence length is equivalent across models. Specifically, for model with a patch size of 16 we input images of size 512×512 versus 448×448 when models are using patch size 14.

Our empirical study clearly demonstrates that DINOv3 models consistently outperform their counterparts on dense prediction tasks. Most notably, on the ADE20k benchmark, the DINOv3 ViT-L model achieves an improvement of over 6 mIoU points compared to the best competitor DINOv2. The ViT-B variant shows a gain of approximately 3 mIoU points against the next best competitor. These substantial improvements highlight the effectiveness of DINOv3's local features in capturing fine-grained spatial details. Furthermore, evaluations on depth estimation tasks also reveal consistent performance gains over competing approaches.

表13： 使用视觉几何基础Transformer (VGGT) 进行的3D理解（王等人， 2025）。只需将VGGT流程中的图像特征提取器DINOv2替换为DINOv3 ViT-L，我们就能在各种3D几何任务上获得最先进技术成果。我们复现了王等人(2025)的基线结果。我们还报告了使用真实值相机信息的方法，标记为 *。相机位姿估计结果使用AUC@30报告。

(a) 相机位姿估计。

方法	Re10K	CO3Dv2
DUSt3R	67.7	76.7
MASt3R	76.4	81.8
VG GSfM v2	78.9	83.4
CUT3R	75.3	82.8
FLARE	78.8	83.3
VGGT	85.3	88.2
DINOv3	**86.3**	**89.6**

(b) DTU上的多视图估计。

方法	Acc.↓	Comp.↓	总体↓
Gipuma*	0.283	0.873	0.578
CIDER*	0.417	0.437	0.427
MASt3R*	0.403	0.344	0.374
GeoMVSNet*	0.331	0.259	0.295
DUSt3R	2.677	0.805	1.741
VGGT	0.389	0.374	0.382
DINOv3	**0.375**	**0.361**	**0.368**

(c) 在ScanNet-1500上查看匹配结果。

方法	AUC@5	AUC@10
SuperGlue	16.2	33.8
LoFTR	22.1	40.8
DKM	29.4	50.7
CasMTR	27.1	47.0
Roma	31.8	53.4
VGGT	33.9	55.2
DINOv3	**35.2**	**56.1**

gether with the previous results on correspondence estimation (Sec. 6.1.3) 和深度估计 (Sec.6.3.3)，我们将其视为 DINOv3 作为 3D 任务基础的强适用性的进一步经验证据。此外，我们预计使用更大的 DINOv3 7B 模型将带来进一步改进。

7 评估完整的 DINOv3 模型家族

在本节中，我们对从我们的 7B-参数模型（参见 第5.2节）中提取的模型家族进行了定量评估。该家族包括基于视觉Transformer（ViT）和 ConvNeXt（CNX）架构的变体。我们提供了所有模型在图16a中的详细参数计数和推理FLOPs。这些模型涵盖了广泛的计算预算，以适应各种用户和部署场景。我们对所有 ViT（第7.1节）和 ConvNeXt 变体进行了全面评估，以评估它们在不同任务上的性能。

图2 提供了 DINOv3 家族与其他模型集合的概述比较。DINOv3 家族在密集预测任务上显著优于所有其他模型。这包括从监督主干（如 AM-RADIO 和 PEspatial）中提取的专用模型。同时，我们的模型在分类任务上取得了相似的结果，使其成为不同计算预算下的最佳选择。

In 第7.1节 详细介绍了我们的ViT模型，并将它们与其他开源替代方案进行了比较。然后，在 第7.2节，我们讨论了ConvNeXt模型。最后，遵循 第5.3节，我们训练了一个与我们的ViT-L模型输出对齐的文本编码器。我们在 第7.3节中展示了该模型的多模态对齐结果。

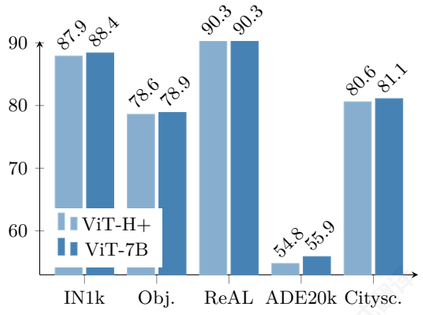
7.1 每种用例的视觉Transformer

我们的ViT家族涵盖了从紧凑型ViT-S到更大型的840亿参数ViT-H+ 模型。前者设计用于在资源受限的设备（如笔记本电脑）上高效运行，后者为更苛刻的应用提供了最先进性能。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了比较，即DINOv2 (Oquab等人， 2024)，SigLIP 2 (Tschannen等人， 2025)和感知编码器 (Bolya等人， 2025)。为了公平比较，我们确保所有模型之间的输入序列长度相同。具体来说，对于补丁大小为16的模型，我们输入图像的尺寸 512×512 与 448×448 当模型使用补丁大小14时。

我们的实证研究表明，DINOv3模型在密集预测任务上始终优于其同类模型。最值得注意的是，在ADE20k基准测试中，DINOv3 ViT-L模型相比最佳竞争对手DINOv2提高了超过6个mIoU点。ViT-B变体则对次优竞争对手提升了约3个mIoU点。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对竞争方法的持续性能提升。

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 family of models.



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

Figure 16: (a) Presentation of the distilled models' characteristics. CNX stands for ConvNeXT. We present per model the number of parameters and the GFLOPs estimated on images of size 256×256 and 512×512. (b) We compare DINOv3 ViT-H+ to its 7B-sized teacher; despite having almost $10\times$ less parameters, the ViT-H+ is close to DINOv3 7B in performance.

Table 14: Comparison of our family of models against open-source alternatives of comparable size. We showcase our ViT-{S, S+, B, L, H+} models on a representative set of global and dense benchmarks: classification (IN-ReAL, IN-R, ObjectNet), retrieval (Oxford-H), segmentation (ADE20k), depth (NYU), tracking (DAVIS at 960px), and keypoint matching (NAVI, SPair). We match the number of patch tokens for a fair comparison across models of different patch size.

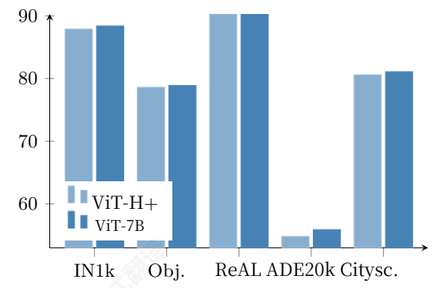
Size	Model	Global Tasks IN-ReaL	IN-R	Obj.	Ox.-H	Dense Tasks ADE20k	NYU↓	DAVIS	NAVI	SPair
S	DINOv2	87.3	54.0	47.8	39.5	45.5	0.446	73.6	53.4	51.6
S	DINOv3	87.0	60.4	50.9	49.5	47.0	0.403	72.7	56.3	50.4
S+	DINOv3	88.0	68.8	54.6	50.0	48.8	0.399	75.5	57.1	55.2
B	PEcore	87.5	68.4	57.9	20.2	37.4	0.641	44.5	41.8	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

This underscores the versatility of the DINOv3 family across different dense vision problems. Importantly, our models achieve competitive results on global recognition benchmarks such as ObjectNet and ImageNet-1k. This indicates that the enhanced dense task performance does not come at the expense of global task accuracy. This balance confirms that DINOv3 models provide a robust and well-rounded solution, excelling across both dense and global vision tasks without compromise.

On another note, we want to also validate if the largest models that we distill capture all the information from the teacher. To this end, we run a comparison of our largest ViT-H+ with the 7B teacher. As shown in Fig. 16b, the largest student achieves performance that is on par with the 8 times larger ViT-7B model.

模型	参数数量	推理 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模型家族。



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

图16： (a) 混合模型的特性展示。CNX代表ConvNeXT。我们按模型展示参数数量和在大小为 256×256 和 512×512的图像上估计的GFLOPs。(b) 我们比较DINOv3 ViT-H+ 与其7B大小的教师模型；尽管参数数量几乎少了 $10\times$ ，ViT-H+ 在性能上接近DINOv3 7B。

表14： 我们与同等规模的开放源代码替代方案比较我们的模型家族。我们在一组具有代表性的全球和密集基准测试上展示了我们的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 960px）和关键点匹配（NAVI, SPair）。我们匹配块标记数量，以在不同块大小的模型之间进行公平比较。

Size	模型	全局任务 IN-ReaL	IN-R	Obj.	Ox.-H	密集任务 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

这突显了DINOv3系列在不同密集视觉问题上的多功能性。重要的是，我们的模型在ObjectNet和ImageNet-1k等全局识别基准测试上取得了具有竞争力的结果。这表明增强的密集任务性能并未以全局任务准确率为代价。这种平衡证实了DINOv3模型提供了一个稳健且全面的解决方案，在密集和全局视觉任务上均表现出色且无妥协。

在另一方面，我们也想验证我们蒸馏的最大的模型是否捕获了教师的所有信息。为此，我们运行了我们的最大ViT-H+ 与7B教师进行比较。如图16b所示，最大的学生达到了与8倍大的ViT-7B模型相当的性能。

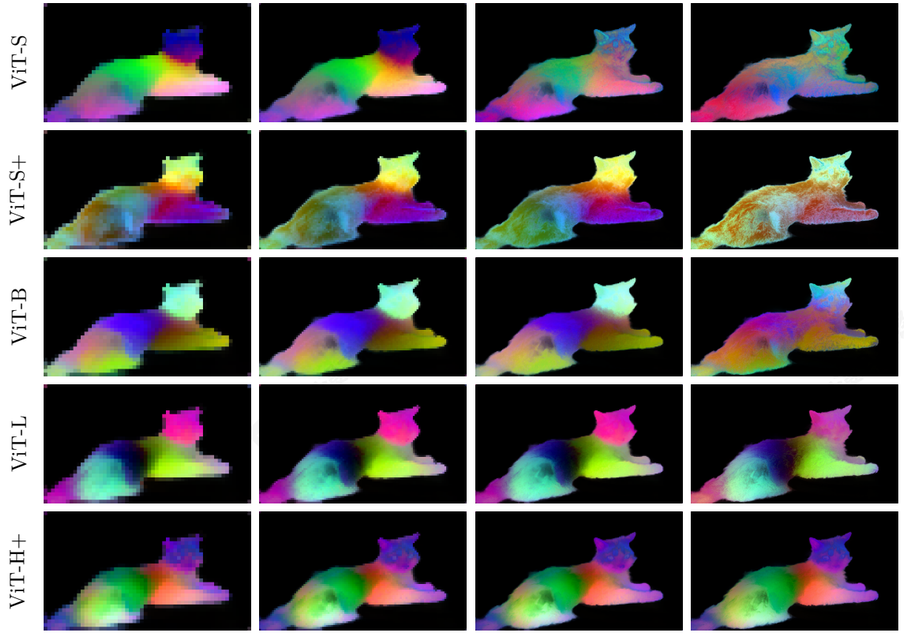


Figure 17: Stability of the features at multiple resolutions for the DINOv3 ViT family of models. Top-to-bottom: ViT-S, S+, B, L, H+. We run inference on an image at multiple resolutions, then perform principal component analysis on the features computed on a 1792×1024 image (112×64 image tokens). We then project features at all resolutions onto the principal components 5–7 that we map to the RGB space for visualization. While the models are functional at all resolutions, we observe that the features remain consistent across a large range of resolutions before drifting: for example, ViT-S+ features are stable between 896×512 and 3584×2048 inputs, while ViT-L barely starts drifting on the largest resolution 7168×4096. ViT-H+ remains stable throughout the whole tested range.

This result not only validates the effectiveness of our distillation process but also demonstrates that, when guided by a high-quality teacher, smaller models can learn to deliver comparable levels of performance. This finding reinforces our belief that *training very large models benefits the broader community.* The strength of larger models can be successfully distilled into more efficient, smaller models with little or no loss of quality.

7.2 Efficient ConvNeXts for Resource-Constrained Environments

In this section, we evaluate the quality of our ConvNeXt (CNX) models distilled from the 7B teacher. ConvNeXt models are highly efficient in terms of FLOPs and are well-suited for deployment on devices optimized for convolutional computations. Furthermore, transformer models often do not lend themselves well to quantization (Bondarenko et al., 2021), whereas quantization of convolutional nets is a well explored subject. We distill CNX architectures of size T, S, B, and L (see Fig. 16a) and compare them to the original ConvNeXt models (Liu et al., 2022). These baselines achieve high performance on ImageNet-1k as they were trained in a supervised fashion using ImageNet-22k labels, and thus represent a strong competitor. For this experiment, we provide results for global tasks at input resolutions 256 and 512, for ADE20k at resolution 512, and for NYU at resolution 640.

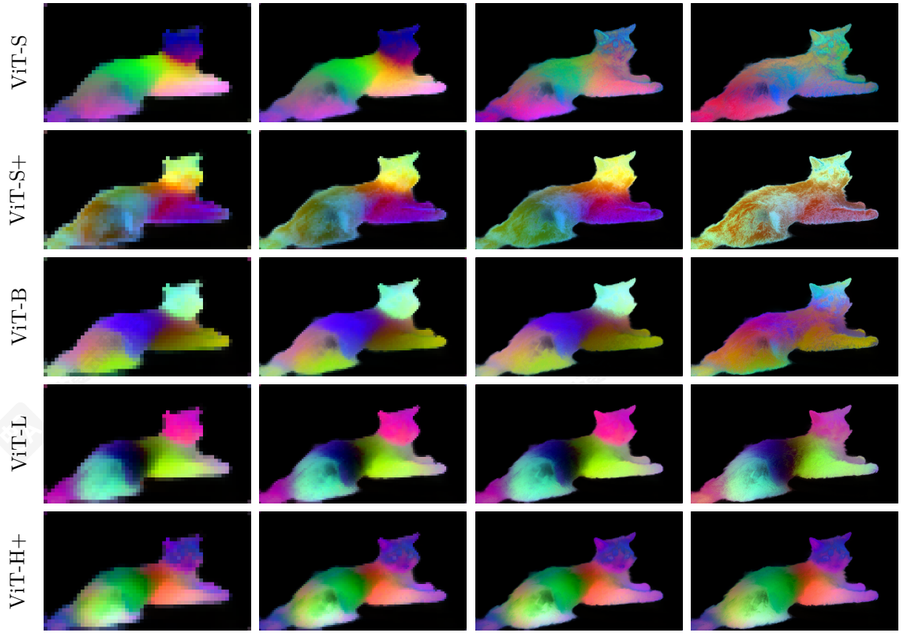


图17： DINOv3 ViT 家族模型在多个分辨率下的特征稳定性。从上到下：ViT-S、S+、B、L、H+。我们在多个分辨率下对图像进行推理，然后在 1792×1024 图像（112×64 图像token）上计算特征，并执行主成分分析。接着，我们将所有分辨率的特征投影到我们映射到RGB空间以进行可视化的主成分5–7上。虽然模型在所有分辨率下都能正常工作，但我们观察到特征在较大分辨率范围内保持一致，然后开始漂移：例如，ViT-S+ 特征在896×512 和 3584×2048 输入之间保持稳定，而ViT-L在最大分辨率 7168×4096上才开始漂移。ViT-H+在整个测试范围内保持稳定。

这一结果不仅验证了我们的蒸馏过程的有效性，而且还表明，在高质量教师的指导下，较小的模型可以学习到相当的性能水平。这一发现加强了我们的信念，即训练非常大的模型有利于整个社区。大型模型的强大优势可以成功地蒸馏到更高效、更小的模型中，几乎没有或没有损失质量。

7.2 资源受限环境下的高效ConvNeXts

在本节中，我们评估了从7B教师蒸馏出的ConvNeXt (CNX) 模型的质量。ConvNeXt模型在FLOPs方面具有很高的效率，并且非常适合部署在针对卷积计算进行优化的设备上。此外，变换器模型通常不太适合量化（Bondarenko等人， 2021），而卷积网络的量化是一个研究得比较充分的主题。我们蒸馏了大小为T、S、B和L的CNX架构（参见图16a），并将它们与原始的ConvNeXt模型（刘等人，2022）进行了比较。这些基线在ImageNet-1k上实现了高性能，因为它们是以有监督的方式使用ImageNet-22k标签进行训练的，因此代表了强大的竞争对手。对于这个实验，我们提供了在输入分辨率256和512的全局任务、在分辨率512的ADE20k以及分辨率640的NYU的结果。

Table 15: Evaluation of our distilled DINOv3 ConvNeXt models. We compare our models to off-the-shelf ConvNeXts trained supervised on ImageNet-22k (Liu et al., 2022). For global tasks, we give results at input resolutions 256 and 512, as we found the supervised models to significantly degrade at resolution 512.

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

Results (Tab. 15) We find that on in-distribution image classification, our models slightly lag behind the supervised ones at resolution 256 (*e.g.* −0.7 IN-ReAL for CNX-T). However, the trend is reversed at resolution 512, with the supervised ConvNeXts significantly degrading, whereas our models scale with increased input resolution. For out-of-distribution classification (IN-R, ObjectNet), there are significant gaps between the two model families for all sizes—a testament to the robustness of the DINOv3 CNX models. Furthermore, the DINOv3 models offer very large improvement on dense tasks. Indeed, for CNX-T, our model yields a +17.9 mIoU (42.7 versus 24.8) improvement, and for CNX-L, our model gets +14.5 mIoU (47.8 versus 33.3). The combination of high performance and computational efficiency makes the distilled ConvNeXt models especially promising for real-world applications where resource constraints are critical. Aside from that, the distillation of the ViT-7B model into smaller ConvNeXt models is particularly exciting, as it bridges two fundamentally different architectures. While ViT-7B is based on transformer blocks with a CLS token, ConvNeXt relies on convolutional operations without a CLS token, making this transfer of knowledge non-trivial. This achievement highlights the versatility and effectiveness of our distillation process.

7.3 Zero-shot Inference with DINOv3-based `dino.txt`

As detailed in Sec. 5.3, we train a text encoder to align both the CLS token and the output patches of the distilled DINOv3 ViT-L model to text, following the recipe of dino.txt Jose et al. (2025). We evaluate the quality of the alignment both at the global- and patch-level on standard benchmarks. We report the zero-shot classification accuracy using the CLIP protocol (Radford et al., 2021) on the ImageNet-1k, ImageNet-Adversarial, ImageNet-Rendition and ObjectNet benchmarks. For image-text retrieval, we evaluate on the COCO2017 dataset (Tsung-Yi et al., 2017) and report Recall@1 on both image-to-text (I → T) and text-to-image (T → I) tasks. To probe the quality of patch-level alignment, we evaluate our model on the open-vocabulary segmentation task using the common benchmarks ADE20k and Cityscapes, for which we report the mIoU metric.

Results (Tab. 16) We compare our text-aligned DINOv3 ViT-L with competitors in the same size class. Compared to Jose et al. (2025), which aligns DINOv2 to text, DINOv3 leads to significantly better performance on all benchmarks. On global alignment tasks, we compare favorably to the original CLIP (Radford et al., 2021) and strong baselines such as EVA-02-CLIP (Sun et al., 2023) but slightly behind SigLIP2 (Tschannen et al., 2025) and Perception Encoder (Bolya et al., 2025). On dense alignment tasks, our text-aligned model shows excellent performance on two challenging benchmarks ADE20K and Cityscapes thanks to clean feature maps of DINOv3.

表15： 对蒸馏的DINOv3 ConvNeXt模型的评估。我们将我们的模型与在ImageNet-22k上监督训练的即用型ConvNeXts进行比较（刘等人， 2022）。对于全局任务，我们在输入分辨率256和512处给出结果，因为我们发现监督模型在分辨率512时显著退化。

Size	模型	全局任务						密集任务	
		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

结果（表15） 我们发现，在分布内图像分类中，我们的模型在分辨率256时略微落后于有监督模型（例如。−0.7 IN-ReAL for CNX-T ）。然而，在分辨率 512 时，趋势发生了逆转，有监督的 ConvNeXts 显著退化，而我们的模型随着输入分辨率的增加而扩展。对于分布外分类（IN-R，ObjectNet），对于所有大小，两种模型系列之间存在显著差距——这是DINOv3 CNX模型鲁棒性的证明。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（42.7与24.8）的提升，而对于CNX-L，我们的模型获得了 +14.5 mIoU（47.8与33.3）。高性能和计算效率的结合使得蒸馏的ConvNeXt模型在资源限制关键的实时应用中特别有前景。除此之外，将ViT-7B模型蒸馏成更小的ConvNeXt模型尤其令人兴奋，因为它连接了两种根本不同的架构。虽然ViT-7B基于带有CLS标记的Transformer块，而ConvNeXt依赖于没有CLS标记的卷积操作，因此这种知识迁移并不简单。这一成就突出了我们蒸馏过程的通用性和有效性。

7.3 零样本推理with DINOv3-based `dino.txt`

As detailed in 第5.3节，我们训练一个文本编码器来对齐CLIP标记和蒸馏的DINOv3 ViT-L模型的输出块，遵循dino.txt的配方Jose 等人(2025).我们在标准基准上评估对齐的质量，包括全局级别和块级别。我们使用CLIP协议（Radford等人,2021）在ImageNet-1k、ImageNet-对抗、ImageNet-渲染和ObjectNet基准上报告零样本分类准确率。对于图像-文本检索，我们在COCO2017数据集（Tsung-Yi等人， 2017）上评估，并报告图像到文本（I → T)和文本到图像（T → I)任务的Recall@1。为了探测块级别对齐的质量，我们在开放词汇分割任务上评估我们的模型，使用常见的ADE20k和Cityscapes基准，并报告mIoU指标。

结果（表16） 我们将与同类大小的竞争对手比较我们的文本对齐DINOv3 ViT-L。与 Jose等人 (2025)，它将DINOv2对齐到文本，DINOv3在所有基准测试上的性能都显著更好。在全局对齐任务上，我们与原始CLIP（Radford等人， 2021）和强基线（如EVA-02-CLIP（Sun等人， 2023）相比表现良好，但略逊于SigLIP2（Tschannen等人， 2025）和感知编码器（Bolya等人， 2025）。在密集对齐任务上，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。

Table 16: Comparing our text-aligned DINOv3 ViT-L to the state-of-the-art. Our model achieves excellent dense alignment performance while staying competitive in global alignment tasks. All compared models are of ViT-L size and operate on the same sequence length of 576.

	Classification				Retrieval		Segmentation	
Method	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**

8 DINOv3 on Geospatial Data

Our self-supervised learning recipe is generic and can be applied to any image domain. In this section, we showcase this universality by building a DINOv3 7B model for satellite images, which have very different characteristics (*e.g.* object texture, sensor noise, and focal views) than the web images on which DINOv3 was initially developed.

8.1 Pre-Training Data and Benchmarks

Our satellite DINOv3 7B model is pre-trained on SAT-493M, a dataset of 493 millions of 512×512 images sampled randomly from Maxar RGB ortho-rectified imagery at 0.6 meter resolution. We use the exact same set of hyper-parameters that are used for the web DINOv3 7B model, except for the RGB mean and std normalization that are adapted for satellite images, and the training length. Similar to the web model, our training pipeline for the satellite model consists of 100k iterations of initial pre-training with global crops (256×256), followed by 10k iterations using Gram regularization, and finalized with 8k steps of high resolution fine-tuning at resolution 512. Also similar to the web model, we distill our 7B satellite model into a more manageable ViT-Large model to facilitate its use in low-budget regime.

We evaluate DINOv3 satellite and web models on multiple earth observation tasks. For the task of global canopy height mapping, we use the Satlidar dataset described in App. D.13, which consists of one million 512×512 images with LiDAR ground truths split into train/val/test splits with ratios 8/1/1. The splits include the Neon and São Paulo dataset used by Tolan et al. (2024). For national-scale canopy height mapping, we evaluate on Open-Canopy (Fogel et al., 2025), which combines SPOT 6-7 satellite imagery and aerial LiDAR data over 87,000 km^2 across France. Since images in this dataset have 4 channels including the additional infra-red (IR) channel, we adapt our backbone by taking the average of the three channels in the weights of the patch embed module and adding it to the weights as the fourth channel. We trained a DPT decoder on 512×512 crops of images resized to 1667 to match the Maxar ground sample resolution.

Semantic geospatial tasks are assessed with GEO-Bench (Lacoste et al., 2023), which comprises six classification and six segmentation tasks spanning various spatial resolutions and optical bands. The GEO-Bench tasks are diverse, including the detection of rooftop-mounted photovoltaic systems, classifying local climate zones, measuring drivers of deforestation, and detecting tree crowns. For high-resolution semantic tasks, we consider the land cover segmentation dataset LoveDA (Wang et al., 2022a), the object segmentation dataset iSAID (Zamir et al., 2019), and the horizontal detection dataset DIOR (Li et al., 2020).

8.2 Canopy Height Estimation

Estimating canopy height from satellite imagery is a challenging metric task, requiring accurate recovery of continuous spatial structure despite random variations in slope, viewing geometry, sun angle, atmospheric scattering, and quantization artifacts. This task is critical for global carbon monitoring and for forest and agriculture management (Harris et al., 2021). Following Tolan et al. (2024), the first work to leverage a SSL

表16： 将我们的文本对齐DINOv3 ViT-L与最先进技术进行比较。我们的模型在密集对齐性能方面取得了优异的成绩，同时在全局对齐任务中保持竞争力。所有比较的模型都是ViT-L尺寸，并在相同的序列长度576上运行。

	分类				检索		分割	
方法	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**

8 DINOv3在地理空间数据上

我们的自监督学习配方是通用的，可以应用于任何图像领域。在本节中，我们通过构建一个用于卫星图像的DINOv3 7B模型来展示这种通用性，这些卫星图像具有非常不同的特征（例如。物体纹理、传感器噪声和焦距视图），而DINOv3最初是在网络图像上开发的。

8.1 预训练数据和基准测试

我们的卫星DINOv3 7B模型在SAT-493M上预训练，这是一个包含4.93亿张 512×512 图像的数据集，这些图像随机采样自Maxar RGB正射影像，分辨率为0.6米。我们使用与网络DINOv3 7B模型完全相同的超参数集，除了RGB均值和标准差归一化是针对卫星图像进行调整的，以及训练长度。与网络模型类似，我们卫星模型的训练流程包括100k次初始预训练的全局裁剪(256×256)，然后使用Gram正则化进行10k次迭代，最后以512分辨率进行8k步的高分辨率微调。与网络模型类似，我们将我们的7B卫星模型蒸馏成一个更易于管理的ViT-Large模型，以方便其在低预算环境下使用。

我们在多个地球观测任务上评估了DINOv3卫星和网页模型。对于全球冠层高度制图任务，我们使用了在 附录D.13中描述的 512×512 Satlidar数据集，该数据集包含一百万张图像，LiDAR地面真值被分成8/1/1的训练/验证/测试集。这些集包括 Tolan等人 (2024)使用的 Neon和圣保罗数据集。对于国家尺度的冠层高度制图，我们在 Open-Canopy (<style id='26'>Fogel等人<style id='28'>,2025<style id='32'>)上进行了评估，该数据集结合了覆盖法国87,000 km^2的SPOT 6-7卫星影像和航空LiDAR数据。由于该数据集中的图像具有4个通道，包括额外的红外(IR)通道，我们通过将patch嵌入模块权重的三个通道的平均值添加到权重中作为第四个通道来调整我们的骨干网络。我们在 512×512 图像裁剪上训练了一个DPT解码器，这些图像被调整为1667以匹配Maxar地面采样分辨率。

语义地理空间任务通过GEO-Bench（Lacoste等人， 2023）进行评估，该基准包含六个分类任务和六个分割任务，涵盖不同的空间分辨率和光学频段。GEO-Bench任务多样，包括检测屋顶光伏系统、分类局部气候区、测量森林砍伐驱动因素以及检测树冠。对于高分辨率语义任务，我们考虑土地覆盖分割数据集LoveDA（王等人， 2022a），目标分割数据集iSAID（Zamir等人， 2019），以及水平检测数据集DIOR（李等人， 2020）。

8.2 冠层高度估计

从卫星影像中估算冠层高度是一项具有挑战性的度量任务，需要在坡度、观测几何、太阳角度、大气散射和量化伪影等随机变化的情况下，准确恢复连续的空间结构。这项任务对于全球碳监测和森林及农业管理（Harris等人， 2021）至关重要。在 Tolan等人 (2024) 的研究之后，首次利用SSL

Table 17: Evaluation of different backbones for high-resolution canopy height prediction. All models are trained with a DPT decoder. Results are presented either for experiments with the decoder trained on SatLidar and evaluated on IID samples (SatLidar Val) and OOD test sets (SatLidar Test, Neon and São Paulo), or for experiments with the decoder trained and evaluated on the Open-Canopy dataset. We list mean absolute error (MAE) and the block R^2 metric from Tolan et al. (2024). For completeness, we additionally evaluate the original decoder of Tolan et al. (2024) that was trained on Neon dataset (denoted by *).

Method	Arch.	SatLidar								Open Canopy
		SatLidar Val		SatLidar Test		Neon Test		São Paulo		
		MAE↓	R^2↑	MAE↓	R^2↑	MAE↓	R^2↑	MAE↓	R^2↑	MAE↓
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**

backbone trained on satellite images for this task, we train a DPT head on top of frozen DINOv3 on the SatLidar1M training set, then evaluate it on i.i.d. samples on SatLidar1M validation set as well as out-of-distribution test sets including SatLidar1M test, Neon and Sao Paulo. We additionally train and evaluate on the Open-Canopy dataset.

Results (Tab. 17) We compare different SSL backbones, denoting with "DINOv3 Sat" the model trained the SAT-493M dataset, and with "DINOv3 Web" the model trained on LVD-1689M (see Sec. 3.1). It can be seen that DINOv3 satellite models yield state-of-the-art performance on most benchmarks. Our 7B satellite model sets the new state of the art on SatLidar1M val, SatLidar1M test and Open-Canopy, reducing MAE from 2.4 to 2.2, from 3.4 to 3.2 and from 2.42 to 2.02 respectively. These results show that DINOv3 training recipe is generic and can be effectively applied out-of-the-box to other domains. Interestingly, our distilled ViT-L satellite model performs comparably to its 7B counterpart, achieving comparable results on SatLidar1M and Open-Canopy while faring surprisingly better on Neon test set, reaching the lowest MAE of 2.4 compared to 2.6 of the 7B model and 2.9 of Tolan et al. (2024). Our DINOv3 7B web model reaches decent performance on the benchmarks, outperforming Tolan et al. (2024) on SatLidar1M val, Neon and Open-Canopy but stays behind the satellite model. This highlights the strength of domain-specific pretraining for physically grounded tasks like canopy height estimation, where sensor-specific priors and radiometric consistency are important.

8.3 Comparison to the Earth Observation State of the Art

We compare the performance of different methods for Earth observation tasks in Tab. 18 and Tab. 19. The frozen DINOv3 satellite and web models set new state-of-the-art results on 12 out of 15 classification, segmentation, and horizontal object detection tasks. Our Geo-Bench results surpass prior models, including Prithvi-v2 (Szwarcman et al., 2024) and DOFA (Xiong et al., 2024), which use 6+ bands for Sentinel-2 and Landsat tasks, as well as task-specific fine-tuning (Tab. 18). Despite using a frozen backbone with RGB-only input, the DINOv3 satellite model outperforms previous methods on the three unsaturated classification tasks and on five of six segmentation tasks. Interestingly, the DINOv3 7B web model is very competitve on these benchmarks. It achieves comparable or stronger performance on many GEO-Bench tasks as well as on large-scale, high-resolution remote sensing benchmarks for segmentation and detection. As shown in Tab. 18 and Tab. 19, the frozen DINOv3 web model establishes new leading results Geo-Bench tasks as well as for segmentation and detection tasks on the LoveDA and DIOR datasets.

These findings have broader implications for the design of geospatial foundation models. Those have recently emphasized heuristic techniques such as multitemporal aggregation, multisensor fusion, or incorporating satellite-specific metadata (Brown et al., 2025; Feng et al., 2025). Our results show that general-purpose SSL can match or exceed satellite-specific approaches for tasks that depend on precise object boundaries (seg-

表17： 对用于高分辨率冠层高度预测的不同骨干网络的评估。所有模型均使用DPT解码器进行训练。结果分别针对使用在SatLidar上训练并在IID样本上评估的解码器（SatLidar Val）和OOD测试集（SatLidar Test、Neon和圣保罗）的实验，或针对使用在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 R^2 Tolan等人 (2024)的 块 (2024)的指标。为了完整性，我们额外评估了 Tolan等人 (2024)的原始解码器，该解码器在Neon数据集上训练（用 *表示）。

方法	架构	SatLidar								Open Canopy
		SatLidar Val		SatLidar 测试		Neon 测试		圣保罗		
		MAE↓	R^2↑	MAE↓	R^2↑	MAE↓	R^2↑	MAE↓	R^2↑	MAE↓
Tolan等人 (2024)*	ViT-L	2.8	0.86	4.0	0.61	2.7	0.73	**5.4**	0.42	—
Tolan等人 (2024)	ViT-L	2.4	0.90	3.4	0.81	2.9	0.69	**5.4**	0.48	2.42
DINOv3 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**

用于此任务的卫星图像训练的骨干网络，我们在SatLidar1M训练集上在冻结的DINOv3顶部训练了一个DPT头，然后在SatLidar1M验证集以及包括SatLidar1M测试、Neon和圣保罗的分布外测试集上使用独立同分布样本进行评估。我们还在Open-Canopy数据集上进行了额外的训练和评估。

结果（表17）我们比较了不同的SSL骨干网络，用"DINOv3 Sat"表示在SAT-493M数据集上训练的模型，用"DINOv3 Web"表示在LVD-1689M上训练的模型（见第3.1节）。可以看出，DINOv3卫星模型在大多数基准测试上达到了最先进性能。我们的7B卫星模型在SatLidar1M验证集、SatLidar1M测试和Open-Canopy上设定了新的最先进水平，将MAE从2.4降至2.2，从3.4降至3.2，从2.42降至2.02。这些结果表明DINOv3训练配方是通用的，可以有效地开箱即用地应用于其他领域。有趣的是，我们的蒸馏ViT-L卫星模型表现与其7B对应模型相当，在SatLidar1M和Open-Canopy上取得了可比结果，而在Neon测试集上表现出惊人的优势，达到了最低的MAE 2.4，相比之下7B模型的MAE为2.6，Tolan等人（2024）的为2.9。我们的DINOv3 7B网络模型在基准测试上达到了不错的性能，在SatLidar1M验证集、Neon和Open-Canopy上优于Tolan等人（2024），但在卫星模型之后。这突出了领域特定预训练在冠层高度估计等物理基础任务中的优势，其中传感器特定先验和辐射一致性非常重要。

8.3 与地球观测最先进技术的对比

我们比较了不同方法在地球观测任务上的性能，见表18和见表19。冻结的DINOv3卫星和网页模型在15项分类、分割和水平目标检测任务的12项上创造了最先进的结果。我们的Geo-Bench结果超越了先前模型，包括Prithvi-v2（Szwarcman 等人， 2024）和DOFA（Xiong 等人， 2024），它们使用 6+ 频段用于Sentinel-2和Landsat任务，以及任务特定微调（见表18）。尽管使用冻结主干和仅RGB输入，DINOv3卫星模型在三项非饱和分类任务和六项分割任务的五项上优于先前方法。有趣的是，DINOv3 7B网页模型在这些基准测试上非常有竞争力。它在许多Geo-Bench任务以及大规模、高分辨率的遥感分割和检测基准测试上实现了相当或更强的性能。如图表18和表19所示，冻结的DINOv3网页模型在Geo-Bench任务以及LoveDA和DIOR数据集上的分割和检测任务上建立了新的领先结果。

这些发现对地理空间基础模型的设计具有更广泛的意义。最近的研究强调启发式技术，例如多时相聚合、多传感器融合或结合卫星特定元数据（Brown 等人， 2025； Feng 等人， 2025）。我们的结果表明，通用SSL在依赖精确对象边界的任务中可以匹配或超越卫星特定方法（seg-

Table 18: Comparison of our DINOv3 models against strong baselines DOFA (Xiong et al., 2024), Prithvi-v2 (Szwarcman et al., 2024), and Tolan et al. (2024) in Geo-Bench tasks. While Privthi-v2 and DOFA leverage all available optical bands, our models achieve significantly better performance with only RGB inputs.

(a) Classification tasks.

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) Segmentation tasks.

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).*

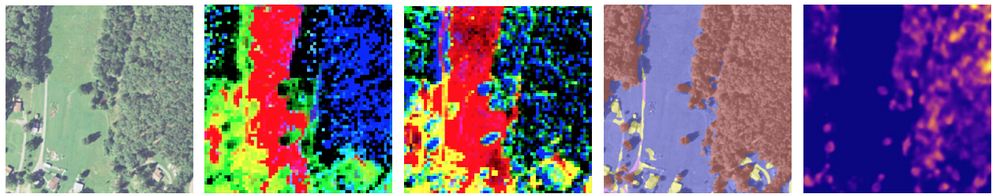


Figure 18: Illustration of versatile applications in remote sensing made possible by a single DINOv3 model. The PCA on DINOv3 features shows finer details than DINOv2. The segmentation map was computed using only GEO-Bench chesapeake labels. The canopy height model decoder was trained on the Open-Canopy dataset using 4 channels (RGB + InfraRed), while inference was performed on RGB channels only.

mentation or object detection). This supports emerging evidence finding that domain-agnostic pretraining can offer strong generalization even in specialized downstream domains (Lahrichi et al., 2025).

Collectively, our results suggest task-dependent benefits of domain-specific pretraining. The DINOv3 satellite model excels in metric tasks like depth estimation, leveraging satellite-specific priors. In contrast, the DINOv3 web model achieves state-of-the-art results on semantic geospatial tasks through diverse, universal representations. The complementary strengths of both models illustrate the broad applicability and effectiveness of the DINOv3 SSL paradigm.

9 Environmental Impact

To estimate the carbon emission of our pre-training, we follow the methodology used in previous work in natural language processing (Strubell et al., 2019; Touvron et al., 2023) and SSL (Oquab et al., 2024). We fix the value of all exogenous variables, *i.e.* the Power Usage Effectiveness (PUE) and carbon intensity factor of a power grid to the same value as used by Touvron et al. (2023), *i.e.* we assume a PUE of 1.1 and a carbon intensity factor of the US average of 0.385 kg CO_2eq/KWh. For the power consumption of GPUs, we take

表18： 我们的DINOv3模型与强基线DOFA（Xiong等人， 2024），Prithvi- v2（Szwarcman等人， 2024），以及Tolan等人 (2024)在Geo-Bench任务中的性能比较。虽然Privthi-v2和DOFA利用了所有可用的光学频段，但我们的模型仅使用RGB输入就实现了显著更好的性能。

(a) 分类任务。

方法	架构	FT	频段	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
Prithvi-v2最佳	ViT-L/H	🔥	all	71.2	**98.8**	96.4	54.1	98.1	59.1	79.6
Tolan等人(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。

方法	架构	FT	频段	m-cashew*	m-chesapeake	m-NeonTree	m-nz-cattle	m-pv4ger-seg	m-SA-作物	Mean
DOFA	ViT-L	🔥	all	81.2	61.6	58.5	77.4	95.1	35.7	68.3
Prithvi-v2最佳	ViT-L/H	🔥	all	90.2	69.4	59.1	81.0	95.3	**41.9**	72.8
Tolan等人(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**

**转换为6个类别，遵循 Szwarcman 等人(2024).*

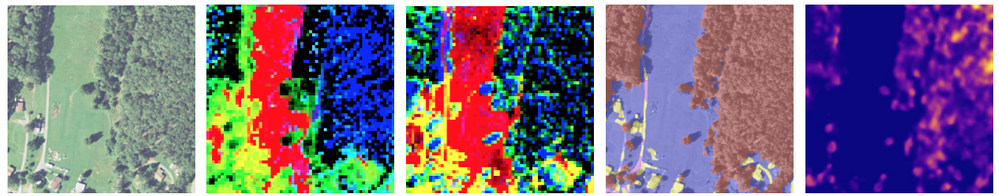


图18： 由单个DINOv3模型实现的遥感多功能应用说明。DINOv3特征的PCA显示比DINOv2更精细的细节。分割图使用仅GEO-Bench chesapeake标签计算。冠层高度模型解码器使用Open-Canopy数据集和4个通道（RGB +红外）进行训练，而推理仅在RGB通道上执行。

分割或目标检测）。这支持新兴的证据，发现领域无关预训练即使在专业下游领域也能提供强大的泛化（Lahrichi 等人， 2025）。

总体而言，我们的结果表明领域特定预训练具有任务相关的优势。DINOv3卫星模型在深度估计等度量任务中表现出色，利用卫星特定先验。相比之下，DINOv3网络模型通过多样化、通用表示在语义地理空间任务上达到最先进技术成果。两种模型的互补优势说明了DINOv3 SSL范式的广泛适用性和有效性。

9 环境影响

为了估算我们预训练的碳排放，我们遵循自然语言处理先前工作中使用的方法论（Strubell 等人， 2019；Touvron 等人， 2023）和 SSL（Oquab 等人， 2024）。我们固定所有外生变量的值， 即。电力系统的电源使用效率 (PUE) 和碳强度因子与 Touvron 等人 (2023)， 即。我们假设 PUE 为 1.1，碳强度因子为美国平均的 0.385 kg CO_2eq/KWh。对于 GPU 的功耗，我们采用

Table 19: We compare the performance of DINOv3 to state-of-the-art models Privthi-v2 (Szwarcman et al., 2024), BillionFM (Cha et al., 2024) and SkySense V2 (Zhang et al., 2025) for high resolution semantic geospatial tasks. We report mIoU for the segmentation datasets LoveDA (1024×) and iSAID (896×), and mAP for the detection dataset DIOR (800×).

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.*

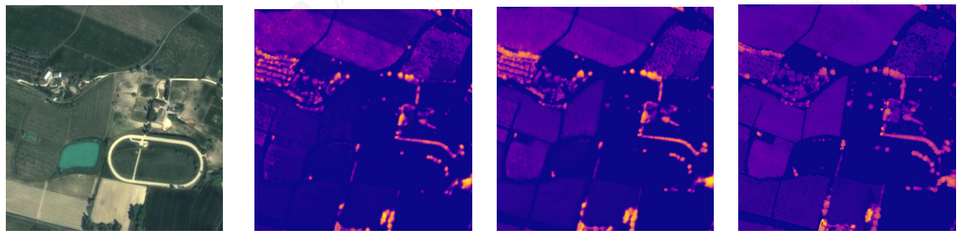


Image from OpenCanopy | Improved Tolan et al. | DINOv3 SAT-493M | Ground Truth

Figure 19: A qualitative comparison of the DINOv3 7B satellite model to Tolan et al. (2024) on the Open Canopy dataset. For both models, the decoder is trained on 448×448 input images. It can be seen that DINOv3 produces more accurate maps, for example the accurate height for the trees on the field.

their thermal design power: 400W for A100 GPUs and 700W for H100 GPUs. We report the details of the computation for the pre-training of our ViT-7B in Tab. 20. For reference, we provide the analogous data for DINOv2 and MetaCLIP. As another point of comparison, the energy required to train one DINOv3 model (47 MWh) is roughly equivalent to that required for 240,000 km of driving with an average electric vehicle.

Carbon Footprint of the Whole Project In order to compute the carbon footprint of the whole project, we use a rough estimate of a total 9M GPU hours. Using the same grid parameters as presented above, we estimate the total footprint to be roughly 2600 tCO_2eq. For comparison, a full Boeing 777 return flight between Paris and New York corresponds to approximately 560 tCO_2eq. Supposing 12 such flights per day, the environmental impact of our project represents half of all flights between these two cities for one day. This estimate only considers the electricity for powering the GPUs and ignores other emissions, such as cooling, manufacturing, and disposal.

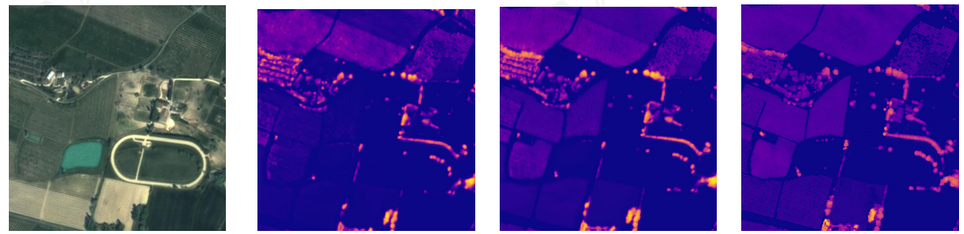
Table 20: Carbon footprint of model training. We report the potential carbon emission of reproducing a full model pre-training, computed using a PUE of 1.1 and carbon intensity factor of 0.385kg CO_2eq/KWh.

Model	Arch.	GPU type	Power (W)	Steps	GPU hours	PUE	Total power (MWh)	Emission (tCO_2eq)
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

表19： 我们将DINOv3的性能与最先进模型Privthi-v2（Szwarcman 等人，2024）、BillionFM（Cha 等人， 2024）和SkySense V2（张等人， 2025）在高分辨率语义地理空间任务上进行比较。我们报告LoveDA（1024×）和iSAID（896×）的分割数据集的mIoU，以及DIOR（800×）的检测数据集的mAP。

方法	架构	FT	LoveDA	iSAID	DIOR
先前SotA		🔥	BillionFM, ViT-g 54.4	SkySense V2, Swin-G* **71.9**	SkySense V2, Swin-G* 79.5
解码器架构			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**

**使用改进的*DINOv2 SSL*，并在*OpenStreetMap*上进行有监督的预训练对齐，报告 *+0*。在*iSAID*上达到*8 mIoU*。*



来自OpenCanopy的图像 | 改进的Tolan等人 | DINOv3 SAT-493M | 真实值

图19： DINOv3 7B卫星模型与Tolan等人(2024)在开放冠层数据集上的定性比较。对于这两个模型，解码器都在 448×448 输入图像上进行训练。可以看出，DINOv3生成的地图更准确，例如田野上树木的准确高度。

它们的散热设计功耗：A100 GPU 为 400W，H100 GPU 为 700W。我们报告了我们 ViT-7B 预训练计算的详细信息在表 20中。作为参考，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（47 MWh）所需的能量大致相当于一辆平均电动汽车行驶 240,000 公里所需的能量。

整个项目的碳足迹为了计算整个项目的碳足迹，我们使用了大约9M GPU小时的粗略估计。使用与上述相同的电网参数，我们估计总足迹约为2600 tCO_2eq。为了比较，一次从巴黎到纽约的波音777往返飞行相当于大约560 tCO_2eq。假设每天有12次这样的航班，我们项目的环境影响相当于这两个城市之间一天内所有航班的一半。这个估计只考虑了为GPU供电的电力，忽略了其他排放，例如冷却、制造和报废。

表20： 模型训练的碳足迹。我们报告了复制完整模型预训练的潜在碳排放，使用PUE为1.1和碳强度因子为 0.385kg CO_2eq/KWh进行计算。

模型	架构	GPU类型	功耗 (W)	步骤	GPU小时	PUE	总功率 (兆瓦时)	排放 (tCO_2eq)
MetaCLIP	ViT-G	A100-40GB	400W	390k	368640	1.1	160	62
DINOv2	ViT-g	A100-40GB	400W	625k	22,016	1.1	9.7	3.7
DINOv3	ViT-7B	H100-SXM5	700W	1,000k	61,440	1.1	47	18

10 Conclusion

DINOv3 represents a significant advancement in the field of self-supervised learning, demonstrating the potential to revolutionize the way visual representations are learned across various domains. By scaling dataset and model size through meticulous data preparation, design, and optimization, DINOv3 showcases the power of self-supervised learning to eliminate the dependency on manual annotations. The introduction of the Gram anchoring method effectively mitigates the degradation of dense feature maps over extended training periods, ensuring robust and reliable performance.

Together with the implementation of post-hoc polishing strategies, such as high-resolution post-training and distillation, we achieve state-of-the-art performance across a wide range of visual tasks with no fine-tuning of the image encoder. The DINOv3 suite of vision models not only sets new benchmarks but also offers a versatile solution across various resource constraints, deployment scenarios, and application use cases. The progress made with DINOv3 is a testament to the promise of self-supervised learning in advancing the state of the art in computer vision and beyond.

10 结论

DINOv3在自监督学习领域代表了一次重大进步，展示了其在跨多个领域学习视觉表示方面的潜力，有可能彻底改变学习方式。通过精心准备、设计和优化数据集和模型大小，DINOv3展示了自监督学习的力量，可以消除对人工标注的依赖。Gram锚定方法的引入有效地减轻了密集特征图在长时间训练过程中的退化，确保了稳健和可靠的性能。

结合后处理优化策略（如高分辨率后训练和蒸馏），我们在无需对图像编码器进行微调的情况下，在广泛的视觉任务中实现了最先进性能。DINOv3视觉模型套件不仅设立了新的基准测试，还提供了适用于各种资源限制、部署场景和应用用例的通用解决方案。DINOv3取得的进展证明了自监督学习在推动计算机视觉及其他领域的当前最佳技术方面的潜力。