第 1 章 DINOv3全景导览与学习指南

DINOv3是新一代视觉模型（Vision Models），由Meta AI于2025年8月15日推出。该模型通过自监督学习（Self-Supervised Learning，SSL）进行训练，致力于构建强大的视觉表征能力，为计算机视觉（Computer Vision，CV）领域的各类下游任务提供坚实基础，并推动该领域的范式革新。

为协助读者系统化地学习并掌握这一前沿技术，本章将从多个维度对DINOv3进行全景式导览，提供清晰的学习路径指导：首先，将分析DINOv3诞生的背景，探讨传统监督学习在计算机视觉领域所面临的若干限制，以及自监督学习如何成为突破这些限制的关键思路，帮助读者理解DINOv3出现的必然性及其重要价值；然后，将系统梳理DINOv3的发展历程，帮助读者构建对模型原理的整体认知；最后，针对不同技术基础的学习者，将推荐从入门到进阶的学习路径，提出关键技术点的学习顺序建议，并提供实践项目指引，助力读者高效且深入地理解DINOv3，为后续章节的原理剖析、应用实践及训练部署打下稳固基础。

1.1 DINOv3：一种自监督学习的思想范式

DINOv3代表了从监督学习到自监督学习的关键技术跨越。它不再依赖大规模人工标注数据来驱动模型训练，而是从无标签数据本身挖掘监督信号，使模型在自我探索与学习中掌握视觉世界的内在规律。这种思想范式的转变，使DINOv3摆脱了对昂贵的人工标注的过度依赖，极大拓展了模型训练数据的来源，从而有望在更广泛的视觉任务和更复杂的应用场景中发挥作用。

1.1.1 产生背景：数据标注的局限与 DINOv3 的出现

DINOv3以自监督学习为主要特征，使视觉模型在预训练阶段就摆脱了对大规模人工标注数据的依赖，从而能够从海量无标签图像数据中自主学习有效的视觉表征。在传统的计算机视觉任务中，基于数据标注的监督学习长期占据主导地位，模型性能高度依赖精确标注的数据集。然而，随着视

觉任务向更复杂、更通用的方向发展，监督学习的局限性日益凸显：

- 首先，构建大规模、高精度的标注数据集成本高昂、耗时费力。尤其在医疗影像、遥感图像等专业领域，标注工作需要大量具备专业知识的人力资源。
- 其次，标注数据往往局限于特定场景或任务，导致模型泛化能力受限。而真实世界中的视觉输入复杂多变，往往需要更具通用性的模型支撑。
- 最后，依靠标注数据的训练方法难以充分利用互联网中海量、无约束的图像资源，无法借助规模优势训练任意规模的强大模型，也难以挖掘无标签图像中蕴含的丰富视觉知识。

在此背景下，DINOv3以自监督学习为核心思想，旨在突破监督学习的桎梏。通过创新的训练机制，使模型能够从无标签数据中自主学习通用的视觉表征，从而为计算机视觉的进一步发展开辟新的道路。

近年来，大语言模型（Large Language Models，LLMs）的成功为这一方向提供了重要启示。大语言模型通过自监督学习在海量文本数据上进行预训练，获取通用知识与能力，随后通过微调适配具体任务，展现出卓越的通用性、数据可得性与扩展能力。

然而，文本数据具有离散特性，而图像像素具有连续特性，二者在数据结构与信息表达形式上存在本质差异。图像数据蕴含更复杂的空间信息、纹理细节与多层次语义信息，直接套用大语言模型的自监督方法难以充分释放视觉数据的潜能。

尽管如此，DINOv3在训练理念上与大语言模型有异曲同工之处。它立足视觉数据的独特属性，提出了一种面向大规模图像的自监督学习框架：模型通过原始图像本身学习判别性特征，在170亿幅图像上进行自监督训练，提取与具体任务无关的多维通用特征表示，再利用小规模标注数据集训练特定任务头，以适配不同的下游任务。

这一方法为后续多种下游视觉任务（如图像分类、目标检测、语义分割等）提供了卓越的预训练模型基础，也标志着视觉领域向“大规模自监督预训练”模式迈出了关键一步。

1.1.2　核心贡献：DINOv3 的技术里程碑与范式突破

在计算机视觉领域的发展历程中，监督学习凭借直观的训练方式以及在特定任务上的出色表现，成为推动技术进步的重要力量。从20世纪末杨立昆（Yann LeCun）利用卷积神经网络改进手写数字识别，到近年来依托李飞飞团队构建的ImageNet数据集开展大规模图像分类竞赛，监督学习模式下训练的模型不断刷新性能纪录，深刻改变了人们对人工智能视觉能力的认知。

然而，随着技术的持续演进和应用场景的不断拓展，监督学习的固有瓶颈逐渐显现，并成为制约计算机视觉迈向更高层次的关键因素。

- 首先，在数据标注层面，成本与规模之间的矛盾日益突出。以ImageNet为例，它的完整数据集包含超过1400万幅标注图像，而这些标注的背后是成千上万名标注人员耗时数年的投入。对于更为细分的应用场景，如自动驾驶中的长尾目标检测，不仅需要标注目标的类别，还需精确标注其边界框、姿态甚至遮挡关系，数据构建成本进一步攀升。

- 其次，在模型泛化能力方面，监督学习训练的模型往往表现出较强的“任务特异性”和“数据集偏见”。在特定数据集上训练的图像分类模型，可能在识别某些目标时表现优异，但当面对训练集中未出现的类别、拍摄角度、光照条件或场景变化时，性能往往显著下降。例如，在标准数据集上训练的模型，可能无法准确识别极端天气下的交通标志，或对医学影像中罕见疾病的微小病灶视而不见。这种对标注数据分布的依赖和过拟合，使模型难以真正理解视觉世界的内在规律，无法像人类一样通过少量样本甚至无样本快速习得新的视觉概念。
- 最后，在知识迁移方面，监督学习获得的模型参数更多编码的是特定任务的判别性特征，而非通用的视觉表征。因此，当模型迁移到新任务时，往往仍需依赖目标任务的标注数据进行微调，迁移成本较高。

正是在这样的背景下，自监督学习作为一种能够从无标注数据中学习通用视觉表征的新范式，逐渐成为研究热点。

DINOv3的诞生，正是对这些瓶颈的直接回应。它通过创新的自监督学习机制，使模型能够在海量无标注图像中自主挖掘视觉数据的内在结构和语义信息，从而摆脱对人工标注的过度依赖，学习更加通用且可迁移的视觉表征，为计算机视觉领域的范式革新奠定坚实基础。

自监督学习能够直接从原始图像数据中挖掘规律，借助图像中模式的自然共现关系，逐渐减少对人工标注数据的依赖，使模型得以扩展至海量数据集与大型架构。这种训练方式无须为特定任务或领域单独设计数据标注方案，仅依靠统一算法就能从自然图像、航拍图像等多样数据源中习得稳定的视觉表征。像DINOv3这样的视觉基础模型，已成为现代计算机视觉的关键组成部分，凭借单一可复用的模型实现跨任务、跨领域的泛化能力。

不同于弱监督或完全监督的预训练方法通常依赖高质量元数据或图文配对信息，自监督学习可以直接在海量原始图像集合上进行训练，从而获得几乎无穷无尽的可用训练数据。这对训练大规模视觉编码器尤其关键。

通过自监督学习训练的模型展现出多方面的优势，例如，能高效适应输入分布的变化，提供兼具全局与局部信息的特征表示，并生成有助于理解物理场景的丰富嵌入向量。基于自监督学习方法训练的DINOv3，属于多功能视觉基础模型，在若干情境下无须微调即可超越现有的专业模型。其生成的密集（Dense，图块级）特征在各类视觉任务中表现优异，整体性能也明显优于以往的自监督和弱监督基础模型。

1.2　DINOv3：一种通用的视觉骨干网络

DINOv3的核心贡献在于，通过纯粹且可扩展的自监督学习机制，训练出一个能够生成通用、强大且高分辨率视觉特征的骨干网络（Backbone）。该骨干网络凭借“冻结即用”的特性，正在改变视觉模型的应用方式——从“为每个任务训练一个模型”，转向“以一个通用骨干网络服务于多

种任务”。同时，其工程化的模型套件设计和开源策略，确保了这一技术能广泛应用于学术界和工业界，用以解决真实世界中的复杂问题。

1.2.1　无须标注的自监督学习

在视觉模型训练中，“监督”可以大致分为三个层次：完全监督、弱监督和自监督。

完全监督，是指模型训练所需的监督信号完全来自人工标注的数据。例如，在图像分类任务中，每幅图像对应明确的类别标签；在目标检测任务中，每个目标都需标注精确的边界框和类别信息。典型案例包括使用精确标注的ImageNet数据集进行图像分类训练，或使用COCO2017数据集进行目标检测训练。

弱监督所使用的监督信号精度较低或标注成本相对较低，例如图像级标签、视频标题文本，或者从网络爬取数据时附带的嘈杂标签等。弱监督学习视图在标注信息不完整或不精确的情况下学习有用特征，但仍依赖某种形式的人工间接标注或外部辅助信息。例如，CLIP系列视觉模型依赖互联网上大规模“图像-文本描述对”进行训练，从而获得视觉与语言概念对齐的能力，实现零样本图像分类和图文检索。

自监督学习则不依赖人工标注的数据，而是从数据本身自动挖掘内在的监督信号，使模型在没有任何人工标签的帮助下，从原始图像数据中自主学习视觉世界的内在特征和表示。模型通过挖掘数据的结构特性、统计规律或空间关系，如图像中的颜色通道之间的相关性、局部与整体的几何关系、不同视角下同一物体的一致性，或视频序列中的帧间运动信息等，逐步形成稳定的视觉表征。

需要强调的是，标注与监督并非等价概念，监督并非必须依赖人工标注。完全监督、弱监督与自监督均属于广义的监督学习范畴，但三者在对标注数据的依赖程度上存在显著差异：完全监督依赖完整且准确的人工标注；弱监督虽然也以人工标注为基础，但其标注信息通常并不完整或不够精确；自监督则主要利用数据自身的结构进行学习，不直接依赖外部标注。

然而，在实际应用中，自监督也并非完全脱离标注数据。例如，在DINOv3框架中，尽管骨干网络可通过自监督方式进行预训练，但在下游任务微调时，任务头通常仍需借助ImageNet、COCO2017与SYNTHMIX等带有标注的数据集进行训练。此外，即使在骨干网络预训练阶段，也需要引入少量高质量标注数据以进一步提升模型性能。

总体来说，DINOv3骨干网络的训练本质上是一种自监督学习过程，其主要学习信号来自图像本身的结构和内容，基本上不依赖人工标注或外部元数据。其核心思想在于，使模型在同一图像的不同局部“视角”下学习一致且稳定的特征表示，从而捕捉同一事物在不同视角下的本质特征，进而理解物体部件结构、场景布局以及更高层次的语义信息。

这种机制带来了以下优势：

（1）数据可扩展性：理论上，任何图像都可作为训练数据，而不受是否具备人工标签或描述的限制。这使得训练数据规模能够扩展至170亿幅图像，充分挖掘海量无标签数据所蕴含的潜在价值。

（2）特定领域应用的改进：在遥感、医疗、科学成像等专业领域，高质量标注数据往往难以获取。DINOv3的训练方法使模型能够直接在这些领域的原始图像上进行预训练，生成具有该领域适应性的基础特征表示，并有望改变相关领域的研究与应用范式。

1.2.2　多种视觉任务上的统一骨干网络

“冻结骨干网络权重，仅训练任务头”是DINOv3的一项关键特性。传统迁移学习通常要对预训练骨干网络进行微调，以适应新的任务（如分类、分割、检测），并更新骨干网络的权重。

而从DINOv3的实践与应用结果来看，其骨干网络在多数情况下无须微调，只需保持权重冻结，并在其生成的特征之上叠加一个轻量级的任务特定模块（如简单的分类头或小型解码器），即可在多种任务上达到甚至超越需要微调的专用模型性能。这种“即插即用”的特性显著简化了模型在不同任务之间的迁移过程，既规避了因微调骨干网络参数而可能引发的过拟合风险和额外计算成本，也有效提升了模型开发与部署效率。

在处理各类任务时，视觉模型对图像特征的关注点各有侧重。例如，分类任务注重提取全局语义信息（如基于[CLS] token的表示），要求模型从整体上理解图像的核心内容，如判断图像中主体是猫还是狗；目标检测任务更关注图像中多个局部目标的定位与识别，依赖Patch（图像块）级别的细节特征，需精确捕捉每个目标的边界、类别以及它们之间的空间关系；语义分割任务则要求对图像中的每个像素进行分类，需要模型具备像素级的精细特征理解能力，以区分图像中不同区域的语义类别，如将“道路”“建筑”“植被”等区域逐像素划分出来。

尽管不同视觉任务对特征的需求存在显著差异，但DINOv3通过自监督学习获得的通用视觉表征，仍能够有效满足这些多样化的特征需求。其关键在于，DINOv3从海量无标签图像中学习到的是视觉世界的内在结构与通用规律，而非面向某一特定任务优化的浅层特征。

这些通用结构包含丰富的层次化信息：从底层的边缘与纹理特征，到中层的部件与形状结构，再到高层的语义概念与场景关系，构建起一个多尺度、多粒度的特征空间。在面对不同任务时，只需设计合适的任务头，即可从这一通用特征空间中筛选与组合出该任务所需的特定特征子集。如此一来，一个统一的骨干网络既能胜任图像分类中对全局语义的理解，也能适应语义分割中对每个像素类别的精确解析；既可以在目标检测任务中精准定位物体边界，也可以在深度估计任务中推断场景的三维结构等。这种统一骨干网络、多任务适配的范式，体现了DINOv3在视觉基础模型方向上的重要突破。

1.2.3　模型参数规模与部署的工程化设计

与大语言模型动辄上千亿的参数量（如DeepSeek V3的参数量达6710亿）不同，在DINOv3出现之前，视觉模型鲜有超过10亿参数的案例，其中拥有11亿参数的DINOv2已属于规模较大的模型。这一现象的根源在于，传统视觉模型的性能提升高度依赖高质量人工标注数据，从而在一定程度上制约了模型规模的进一步扩展。

基于自监督学习的DINOv3在规模化方向上进行了积极探索，将参数规模扩展至70亿（7B），发掘了“更大模型 + 更多数据”在视觉自监督学习范式下的潜力，证明了参数规模扩展能持续带来性能提升。这一实践表明，在摆脱对人工标注的依赖后，视觉模型同样可以受益于规模化趋势。

当然，超大模型也意味着更高的算力需求与部署成本。为解决原始大模型在实际应用中难以部署的问题，Meta AI积极响应社区需求，通过知识蒸馏技术，将DINOv3 7B模型中蕴含的“知识”压缩并迁移到一系列更小、更高效的模型中，形成了包含Vision Transformer系列和ConvNeXt系列在内的多种轻量化模型版本（小参数规模）。

这种面向工程落地的系统化设计，使DINOv3在保持高性能的同时，也兼顾了不同算力条件下的实际需求，确保了在从云端服务器到边缘设备的多种应用场景中，用户都能够选择到合适的DINOv3模型。

1.2.4　DINOv3 的实际应用与开放理念

DINOv3推出后，凭借其鲜明的特性，在标注资源稀缺的领域大放异彩。例如，在环境监测领域，Meta AI与世界资源研究所合作，利用在卫星图像上训练的DINOv3骨干网络，更精确地估计树冠高度与监测森林砍伐情况，从而显著降低误差，助力气候融资流程的自动化；在太空探索领域，NASA喷气推进实验室基于其前代模型DINOv2，使得火星探测机器人仅凭单一冻结骨干网络即可同时完成多项视觉任务，有效节省了太空设备中宝贵的算力资源；在医疗影像领域，DINOv3在组织病理学与内窥镜等医学图像分析中，借助无标注数据预训练的骨干网络，显著提升了诊断模型的初始性能。

此外，DINOv3不仅公开了模型权重，还开源了核心训练代码，允许社区复现、研究并改进其自监督学习方法。这种深度的开放策略，使DINOv3不仅是一组模型成果，更成为一个新的基础研究平台，鼓励计算机视觉和多模态领域的研究者在其基础框架上进行创新，共同推动该视觉基础模型的发展。

1.3　DINOv3：从 DINO 开始的发展历程

DINOv3的发展并非一蹴而就，而是在持续研究与技术积累中逐步演进的成果。其成长轨迹经历了从理论探索到关键技术突破，再到体系化完善的多个阶段。其发展脉络可以追溯至自监督学习在视觉领域的早期实践，以及Meta AI研究团队对DINO系列模型的持续迭代与优化。

1.3.1　DINO 模型

早在5年前发布的DINO（self-distillation with no labels）模型中，Meta AI研究团队便开始探索自监督学习在Vision Transformer（ViT）架构上的应用。通过构建教师-学生网络架构并设计相应的

对比学习目标，DINO初步验证了无标签条件下自监督学习训练ViT以获得强大视觉表征的可行性。目前，DINO项目已进入归档状态，代码库和模型权重不再更新，相关研究工作已整体迁移到DINOv2模型。

1.3.2　DINOv2 模型

作为DINO的后续版本，3年前发布的DINOv2在模型规模和训练数据量上实现了显著扩展，进一步提升了自监督视觉表征的性能。该模型展示了在多种下游任务中无须微调即可取得优异表现的强大泛化能力，为视觉基础模型的发展奠定了重要基础。

然而，在向更大规模模型和更长训练周期扩展的过程中，DINOv2逐渐显现了特征性能下降等瓶颈问题。这些问题成为推动DINOv3研发的重要技术动因。目前，DINOv2项目依然处于活跃状态，比如于2025年12月18日更新的代码库合并了xray_dino分支。

1.3.3　DINOv3 模型

通过海量无约束数据训练任意规模的强大模型，是自监督学习的核心优势。然而，在实际应用中，这种训练方法在大规模扩展过程中仍面临诸多挑战：

- 首先，如何从海量无标注数据集中筛选并构建高质量训练数据集，尚缺乏成熟且可直接借鉴的系统化方案。
- 其次，在常规训练实践中广泛采用的余弦学习率调度策略，需要预先确定优化周期范围，这在处理规模庞大、训练周期较长的图像语料库时尤为困难。
- 最后，模型的特征性能在初期训练后出现逐渐退化的趋势。通过对视觉斑块相似度图谱（基于主成分分析法的可视化结果）进行观察，可以验证这一退化趋势。当模型规模超过ViT-Large（约3亿参数）时，这种退化现象在更长的训练周期中愈发明显，从而在一定程度上制约了DINOv2的进一步规模扩展与应用价值提升。

在DINOv3的研发过程中，Meta AI研究团队首先针对DINOv2面临的核心问题进行了深入分析。他们发现，随着训练周期的延长和模型参数规模的扩大，模型学习到的密集特征图质量呈现出明显下降趋势，严重影响了模型对细粒度视觉信息的捕捉能力。

为应对这一关键挑战，研究团队开展了大量实验探索，尝试多种优化策略，最终创新性地提出了Gram锚定（Gram Anchoring）方法。该方法通过引入与Gram矩阵（Gram Matrix）相关的约束机制，有效稳定了特征学习过程，缓解了长期训练中出现的特征质量退化问题，为模型向更大规模和更长训练时间扩展扫清了重要技术障碍。

通过持续的实验验证与问题修正，DINOv3逐步从最初的技术构想演进为能够有效应对大规模训练挑战、性能表现卓越的新一代视觉基础模型，也成为自监督学习领域的又一重要里程碑。

1.4　本书架构与学习路径

本书的架构设计紧密围绕DINOv3的技术原理与学习规律，致力于为不同基础的读者提供一条从理论认知到实践应用的清晰、高效的学习路径。笔者深知，掌握一个前沿且复杂的视觉基础模型，需要系统性的知识构建与循序渐进的能力培养。因此，在架构规划上，笔者既注重知识体系的完整性，也强调学习过程的逻辑性与实践性。

1.4.1　设计逻辑：从原理、应用到训练的渐进式学习法

在知识传递的逻辑顺序上，本书遵循“从原理、应用到训练”的渐进式学习法。这一设计逻辑的核心在于，确保读者在动手实践之前，能够初步了解DINOv3模型背后的理论根基，从而在后续的应用与训练中做到“知其然，更知其所以然”。

在开篇部分（第1~3章），首先介绍用标注数据训练视觉模型所面临的瓶颈，以及自监督学习的崛起，从而自然地引出DINOv3诞生的背景及其试图解决的核心问题。接着，深入剖析DINOv3的训练过程，包括它在数据集与模型规模扩展方面的探索，以及通过Gram锚定方法解决特征退化问题等关键技术突破，使读者对DINOv3的先进性和创新性形成宏观而系统的认识。在此基础上，将以DINOv3的技术报告为基础素材，重点解读DINOv3的自监督学习训练过程，包括数据准备策略、Gram锚定的技术细节、损失函数的设计以及模型架构的演变。这部分内容是理解DINOv3工作机制的关键，也是后续学习的重要基础。

在读者对DINOv3的理论原理有了基本了解之后，本书将自然过渡到“应用”层面。这一部分（第4~9章）将聚焦于DINOv3作为多功能视觉基础模型所展现的强大应用能力。笔者将详细介绍DINOv3在无须对骨干网络进行微调的情况下，如何直接应用于图像分类、目标检测、语义分割、实例分割、特征匹配、三维重建等多种经典视觉任务，并通过具体的案例和可验证的实验结果，展示其超越现有专业级模型的卓越性能。

在完成应用部分的学习之后，读者将对DINOv3系列模型所能覆盖的主要任务形成系统性认识。在此基础上，本书将深入“训练”这一更具实践深度与技术挑战的层面，重点解决骨干网络与任务头如何协同适配并应用于具体场景的问题。在相关章节（第10~16章）中，将系统介绍DINOv3的训练框架搭建、参数设置、训练过程以及成果的部署与应用方法。

通过这种从理论原理的深度解析到实际应用的广度拓展，再到模型训练的实践深化的渐进式架构设计，期望读者能够层层递进、稳步提升，最终实现对DINOv3的全面掌握和灵活运用。

1.4.2　要点梳理：实战篇章与关键技能

本书将紧密围绕DINOv3的核心功能与应用场景，为读者提供从基础理论、实验环境搭建、基础操作到训练定制的全流程实践指导。在关键技能的培养方面，首先，读者将掌握DINOv3预训练模型的加载与基础特征提取流程，包括如何高效获取图像的全局特征向量与密集特征图，理解不同

分辨率输入对特征输出的影响，并学会利用这些基础特征完成简单的视觉任务验证，如相似图像检索的初步实现。其次，针对DINOv3无须微调骨干网络即可部署的特性，应用部分将重点讲述如何将其直接集成到各类下游任务中，包括零样本分类、语义分割、目标检测以及DINOv3对3D重建模型的赋能应用等。最后，在训练篇中，针对视觉领域的常见任务，如图像分类、目标检测、语义分割等，将系统阐述基于DINOv3的模型训练全流程关键技能，包括如何构建适用于特定任务的数据集并进行规范化预处理，如何根据任务特性选择与设计合适的任务头结构，以及如何借助各种训练框架完成模型训练与调优，同时还将介绍蒸馏等轻量化处理方法。

通过这些实战内容的系统训练，读者将逐步具备独立运用DINOv3解决实际视觉问题的能力，从模型的基础调用者成长为能够进行定制化开发与性能优化的进阶实践者。

1.5　本章小结

本章系统梳理了DINOv3作为新一代视觉基础模型的核心特性、发展历程以及本书的学习架构。首先，从DINOv3的技术特性出发，阐述了其如何基于自监督学习范式构建通用视觉表征，突破传统视觉模型对人工标注数据的高度依赖；通过Gram锚定等创新技术，有效缓解了大规模训练中的特征退化问题，并在模型参数规模与工程化部署之间取得良好平衡，形成了从云端到边缘设备的完整模型体系。DINOv3在环境监测、太空探索、医疗影像等领域的实践成果，以及Meta AI深度开源的策略，进一步凸显了其作为基础研究平台所具有的社会价值。随后，回顾了DINOv3从DINO的初步探索，到DINOv2的技术积累，再到自身突破核心瓶颈的发展脉络，揭示了其在自监督学习领域持续迭代的技术演进逻辑。最后，本书确立了以“从原理、应用到训练”为路径的渐进式学习架构，通过理论解析、应用拓展与实践深化三个层次，结合实战篇章中预训练模型应用、下游任务集成及训练全流程等关键技能的培养，帮助读者构建对DINOv3的系统认知体系并提升实践能力。

第 2 章

DINOv3的训练原理与核心机制

在DINOv3的技术报告中，详细阐述了其如何在自监督学习框架下，通过一系列创新技术手段实现模型性能的跃升。其中包括DINOv3训练过程中的关键环节和技术创新，如数据准备与扩展策略、模型架构的设计考量、核心自监督学习目标与优化方法、针对DINOv2中特征退化问题提出的Gram锚定技术，以及模型训练过程中的工程化实践和优化技巧。通过对这些内容的系统学习，读者能够深入理解DINOv3背后的理论基础和实现细节，为后续应用与定制化训练打下坚实的理论基础。

2.1 训练数据准备：多种数据集的混合

数据规模化（Data Scaling）是驱动大型基础模型成功的重要因素之一。所谓数据规模化，是指系统性地增加用于模型训练的数据量，即扩大数据规模，并遵循规模定律（Scaling Law）。这也是当前大模型发展的核心驱动力之一。规模定律指出：随着模型参数规模、训练数据量和计算量（算力需求）同步、可预测地扩大，模型性能会按照平滑且可预测的幂律关系持续提升。

然而，在视觉模型领域，单纯增加训练数据规模并不必然转化为更高的模型质量或更优的下游基准测试表现。成功的数据规模化往往依赖精心设计的算法机制和数据处理流程。这些方法的目标大体可分为两类：一类旨在提升数据的多样性和均衡性，以防止模型过拟合于特定类型的数据，从而增强模型对不同场景的适应能力；另一类则侧重于数据的实用性，即数据与常见实际应用场景之间的相关性。这可以理解为确保模型学习到的视觉表征能够更好地迁移到特定任务。在训练DINOv3的过程中，Meta AI团队结合多种互补方法，在兼顾多样性与实用性的前提下，提升模型的泛化能力与整体性能，在两类目标之间取得平衡。

2.1.1 数据收集与筛选

DINOv3的训练数据主要由三部分构成：LVD-1689M、检索数据集，以及少量标注数据集。下面分别进行介绍。

1. LVD-1689M

DINOv3的无标注数据源于Instagram[1]的公开帖子。在构建过程中，数据首先经过筛选（以确保数据质量和多样性），再通过平台级内容审核（以防止有害内容），最终形成了一个包含约170亿幅图像的原始数据池。

基于该数据池，结合k-均值聚类算法与DINOv2图像嵌入技术，将原始图像数据划分为5级聚类。k-均值是一种经典且高效的无监督聚类算法，其目标是将一组数据自动划分为k个互斥的簇（类别），使同一簇内的数据点高度相似，而不同簇之间的数据点尽可能差异显著。作为自动整理与数据清洗的核心工具，该方法无须人工标注，即可根据视觉语义相似性，将170亿幅来源多样、内容混杂的网络图像自动归类。DINOv2图像嵌入技术将每幅图像映射为固定长度的特征向量，用于在k-均值计算过程中度量图像之间的语义距离，从而实现语义相近图像的有效聚类。

基于上述方法，这170亿幅图像被组织为5级聚类。按照抽象级别由低到高，分别包含约2亿、800万、80万、10万和2.5万个簇。这些聚类的关系如表2-1所示。

表 2-1　DINOv3 训练数据聚类示意

数据阶段	数　量	说　明	举　例
原始数据	170 亿	数据总量	
一级聚类	2 亿	第一层细分，不是图像数，而是类别数	实例级别或超细分类，比如“同一只猫在沙发上睡觉的不同角度的照片”
二级聚类	800 万	更粗一层的类别数	比如“家猫在室内的照片”
三级聚类	80 万	再粗一层的类别数	比如“猫科动物”
四级聚类	10 万	粗粒度类别数	比如“宠物”（可能包含猫、狗、仓鼠等）
五级聚类	2.5 万	最粗粒度类别数	如“动物”

这种分层归类方案能够有效实现去重、保持多样性并获得高质量的图像数据。

在去重方面，借助同一聚类内图像的高度相似性，算法可自动识别并剔除重复或近乎重复的样本（如同一场景在不同光照下的微小差异图像），从而显著减少数据冗余，提升训练效率。

在保持多样性方面，不同聚类之间具有明确区分——从最细粒度的“同一只猫的不同角度”到最粗粒度的“动物”，覆盖了从具体实例到抽象概念的广泛视觉范畴。这种层级结构确保模型能够接触到丰富多样的视觉信息，避免学习到片面或单一的特征表示。

在高质量图像数据获取方面，聚类过程基于语义一致性进行筛选：同一聚类内的图像在内容上具有较高的相关性与纯净度，有效减少了无关或低质量图像对训练的干扰，为后续特征学习提供了坚实的数据基础。

构建聚类层级后，每幅原始图像在每一级聚类中均被明确分配到对应的簇。然后，DINOv3团队采用一种平衡采样算法，基于聚类结构进行配额计算，并按照语义概念进行均衡重采样，从170

[1] Meta公司旗下的社交应用。

亿幅图像中筛选出具有代表性且语义覆盖均衡的16.89亿幅图像。该数据集被命名为LVD-1689M。

2. 检索数据集

第二部分数据集与LVD-1689M一样，同样源于Instagram的170亿幅图像数据池。其生成方法是：首先，构建一个种子数据集。该种子数据集包含一系列精心挑选的高质量图像，涵盖不同视觉领域和语义类别。这些图像通常具有明确的类别标签或特定的视觉属性，能够代表关键的视觉概念。然后，以这些种子图像的嵌入向量为检索依据，在170亿幅原始图像数据池中进行近似搜索，筛选出在语义和视觉特征上与种子图像高度相似的候选图像。通过这一过程，可以有效扩展与下游任务相关的视觉概念覆盖范围，并提升场景多样性。最后，构建出一个涵盖丰富下游任务相关视觉概念的检索数据集。

3. 标注数据集

在DINOv3的训练数据中，除大规模无标注图像数据外，还引入了少量公开的经典标注数据集，为模型训练提供更加系统和深入的视觉知识支持。这些数据集主要包括：用于通用目标识别基准的ImageNet-1k、用于大规模细粒度识别的ImageNet-22k，以及用于真实世界场景理解的Mapillary街景序列等。通过引入这些高质量标注数据，可以在自监督预训练过程中对模型表示进行适度引导，增强其在结构化视觉任务中的表现能力。

2.1.2　数据混合策略

在预训练阶段，DINOv3团队使用采样器将不同数据部分混合。对于LVD-1689M、检索数据集以及标注数据集，混合方式有两种：一种是使用来自单个随机选择数据源的同质数据批次进行训练；另一种则是按特定比例将3种数据组合为异质数据批次进行训练。最终，研究团队选用的方案是在每次迭代中进行随机采样。这样做的原因是，小数据集中的高质量数据构成的同质批次在实践中更为有效。在具体训练中，ImageNet-1k构成的同质批次数据约占整体训练批次的10%。

为评估数据清洗与混合策略的效果，研究团队进行了数据消融实验。在具体操作中，将上述数据混合策略分别与仅采用聚类或检索方法清洗的数据集以及原始数据池进行对比分析，并通过标准下游任务对模型性能进行评估。实验结果表明，没有任何单一的数据清洗技术能够在所有基准测试中表现最优，而数据混合策略在整体指标上表现最佳。

2.2　大规模自监督训练：架构创新与算法优化

从DINOv3所取得的效果反向推导其训练方法，自然会得出“大规模自监督训练是其成功的必要条件”这一结论。然而，在DINOv3出现之前，这种方案在视觉模型领域并不成熟。尽管当时基于自监督学习（Self-Supervised Learning，SSL）的模型已展现出一定潜力的特征学习能力，但大多数SSL方法仍难以有效扩展至更大规模的模型。主要原因有两点：一是训练过程中的稳定性不足；

二是方法过于简化，不能充分建模视觉世界的复杂结构。在此背景下，拥有11亿参数的DINOv2通过SSL取得了显著进展。该模型在经过筛选的数据集上进行训练，其性能与基于弱监督的CLIP模型相差不大。

受到DINOv3的鼓舞，Meta AI未停止对DINOv2的研究。近期，研究团队将DINOv2的参数规模扩展至与DINOv3骨干网络相同的70亿，在全局视觉任务中取得了良好效果。然而，美中不足的是，在侧重图像局部细节的密集预测任务上，其表现不尽如人意。

基于这一现实背景，本书将重点讲解DINOv3在同步扩展模型和数据规模的过程中，如何在增强全局语义理解能力的同时，兼顾局部细节表征能力。

2.2.1　创新的学习目标设计

在对DINO系列模型的训练过程中，研究团队发现，随着模型参数规模的扩大，全局指标和密集预测能力往往难以兼顾。所谓全局指标，可以简单地理解为图像分类任务，它关注的是对整个图像内容的高层语义理解，如判断一幅图像是否包含猫或汽车。而密集预测能力则侧重对图像中每个像素或局部区域进行精细的语义划分。例如，在语义分割任务中，需要精确标注出图像中每个物体的轮廓和类别；在目标检测任务中，需要定位并识别图像中多个物体的具体位置及其类别。

在DINOv2等先前的模型中，当通过扩大参数规模来追求更强的全局语义理解能力时，其在密集预测任务中的表现往往会停滞甚至下降。正如前文所述，DINOv2在扩展参数规模后，出现了全局与局部预测能力此长彼消的现象。为应对这一核心瓶颈，DINOv3在学习目标设计上展开了创新性探索，提出了一种兼顾全局语义一致性和局部特征判别性的复合学习目标。具体而言，模型采用判别式自监督策略，引入名为iBOT的损失函数，与DINO损失函数相结合，构建包含全局与局部损失项的多重自监督目标进行联合训练。

2.2.2　模型架构的改进与优化

DINOv3通过数据规模化进一步释放了模型能力，其参数量从DINOv2的11亿增至70亿，相当于为模型增加了更多学习和记忆的单元。为了使这一大模型能够稳定、高效地训练，研究团队不仅调整了相关训练设置，还改进了模型对图像局部区域位置关系的建模方法。简单来说，研究团队为图像中的每个局部区域赋予坐标信息，使模型在分析图像时能够显式考虑各区域之间的相对位置关系。为了增强模型对不同尺寸和长宽比图像的适应能力，训练过程中还会随机调整坐标范围。这相当于让模型在多种缩放视角下进行学习，从而提升其适应能力。这些改进使模型能够学习到更精细、更稳定的视觉特征，最终在性能和扩展性方面取得提升。

训练一个超大规模模型是一个复杂的探索过程。由于难以准确预估模型需要多少数据、训练多长时间才能达到最佳效果，DINOv3研究团队采用了一种相对简化的训练策略：在整个过程中基本保持学习率恒定，仅在初始阶段对学习率以及教师模型的更新温度进行短暂预热。这种做法带来两方面的好处：其一，只要模型在下游任务上的表现仍在持续提升，便可以继续训练，而无须频繁

担心超参数设置是否合理；其二，由于需要调节的超参数数量减少，研究者可以更集中、更审慎地选择其余关键参数。

在实际训练中，DINOv3使用AdamW优化器[1]，将4096幅图像作为一个批次，分配到256个GPU上并行训练。每幅图像会被裁剪成2幅较大尺寸和8幅小尺寸的视图，从而兼顾整体语义信息与局部细节特征。通过精心调整裁剪尺寸和图像块大小，研究团队确保每次输入模型的序列总长度与DINOv2保持一致，从而在扩大模型规模的同时维持训练过程的稳定性。

2.3　Gram 锚定：提升密集特征一致性

DINOv3研究团队尝试对这个具有70亿参数的模型进行更长时间的训练，初衷是希望该方法能赋予模型持续学习的能力。然而，与DINOv2在扩大参数后出现的整体理解与局部感知之间的矛盾类似，DINOv3在延长训练时间的过程中也暴露出类似问题：虽然更长的训练时间确实提升了模型在图像整体理解任务上的性能，但在处理需要精细感知的任务（如语义分割、深度估计）时，其效果反而随着训练推进而逐渐下降。其原因在于，模型内部不同区域（图像块）学习到的特征逐渐变得不一致、不协调。为缓解这一问题，研究团队提出了一种名为Gram锚定（Gram Anchoring）的新训练目标。

2.3.1　密集特征退化问题分析

在对大规模视觉模型进行长期训练时，模型的整体图像分类能力会随着训练持续提升，而需要精细理解的密集预测任务性能却明显下降。这一现象被称为“密集特征退化问题”。具体表现为：无论是中等规模模型ViT-g（谷歌开发的视觉模型，约25亿参数），还是70亿参数的大模型ViT-7B（DINOv3所采用的Vision Transformer架构模型），其ImageNet分类准确率都随着训练稳步上升，但在Pascal VOC分割任务上的性能却在训练约20万步后开始下滑，其中ViT-7B的性能退化尤为明显。

为探究其原因，研究团队分析了模型内部图像块特征的一致性。在训练初期，不同图像块的特征分布清晰、局部关联性强；但到训练后期，特征分布逐渐变得散乱，大量原本不相关的图像块在特征空间中被错误关联，导致模型对局部结构的理解能力下降。这种“块级特征不一致”现象与特征之间的方向关系紊乱有关：随着训练进行，用于整体分类的特征（[CLS] token）与局部图像块特征之间的关联性逐渐增强，削弱了局部特征自身的区分度与局部一致性。这就好比在训练到某一阶段后，块级特征逐步向图像整体判别力靠拢，而放弃了自身的专长，呈现出个体服从整体的趋势。

为解决密集特征退化问题，DINOv3提出了一种新的训练目标，旨在对图像块特征施加额外约束，增强其局部一致性与判别力，使模型在保持强大全局分类能力的同时，也能够在密集预测任务上维持稳健表现。

---

[1] AdamW是一种广泛应用于深度学习的优化算法。它在经典Adam优化器的基础上，将权重衰减与梯度更新过程解耦，从而更有效地控制模型复杂度，通常能带来更稳定的训练和更好的泛化性能。

2.3.2　Gram 锚定目标设计

在对DINOv3的训练过程中，研究团队发现，模型学习强判别性特征和保持局部特征一致性这两个目标在一定程度上是相对独立的。这也解释了为何模型在整体分类任务上表现越来越好，却在需要精细感知的任务上逐步退化。尽管此前已结合全局的DINO损失和局部的iBOT损失尝试实现平衡，但这种平衡并不稳定；随着训练推进，全局表征往往逐渐占据主导地位。

基于上述观察，DINOv3研究团队提出了一种名为Gram锚定的解决方案。该方案正是利用两个目标之间的相对独立性，引入一个新的优化目标——Gram目标。它并不直接改变特征本身，而是通过约束特征之间的相互关系来防止局部特征一致性退化。具体而言，操作的是Gram矩阵（即一幅图像中所有局部特征之间两两点积构成的矩阵），通过约束学生模型的特征关系矩阵，使其不断接近一个中期版本的教师模型（Gram教师）的对应矩阵。由于这个中期教师模型在密集任务上表现更优，因此能够为当前模型提供更合理的局部结构参考。

在这一架构中，Gram教师和Gram学生是Gram锚定这个新训练方法中的核心概念，延续了自监督学习中经典的“教师-学生”机制。但其目标具有独特性：它们并非用于直接学习特征本身，而是用于“校准”特征之间的关系模式。

实验结果表明，加入Gram目标后，iBOT损失下降更快，说明稳定的特征关系结构有助于提升局部特征一致性的学习效率；而Gram目标对DINO损失的影响很小，表明Gram约束与iBOT目标在影响特征的方式上相似，而与全局DINO目标作用于不同层面。这一设计使模型在保持强大全局判别力的同时，有效恢复并提升了模型在密集任务上的表现。

2.3.3　高分辨率特征增强

近期的一些研究表明，通过对图像局部特征进行加权平均，可以平滑异常值、增强特征一致性，从而获得更稳健的局部表征；与此同时，向模型输入更高分辨率的图像能够生成更精细的特征图。为了使Gram教师提供更高质量的特征关系模板，DINOv3团队结合了这两种思路：首先，将分辨率提高1倍的图像输入Gram教师，以提取富含细节的特征；然后，通过双三次插值将特征图下采样至原尺寸。这一过程在分辨率变换中自然实现了特征的平滑与融合，使局部特征之间的关系更加稳定和一致。

在计算这些平滑特征的Gram矩阵，并将其作为高质量关系模板用于训练后，学生模型能够有效吸收其中更优的局部一致性，从而在密集预测任务上获得显著的性能提升。值得注意的是，高清特征经下采样后，其优越的局部一致性仍得以保留。而Gram教师的选择至关重要——实验发现，训练中期（如10万步或20万步）的教师模型效果最佳，训练后期（如100万步）的教师模型因自身局部特征一致性已经发生退化，反而会对最终结果产生负面影响。得益于旋转位置编码（RoPE）技术，DINOv3模型能够自然适配不同分辨率的图像输入，而无须额外调整结构。定性分析结果显示，经过这一高分辨率精炼过程后，特征之间的相关性得到明显改善，直观验证了该方法在提升局部特征一致性方面的有效性。

通俗来讲，为了让学生模型学到更优质的局部细节，研究人员为教师模型开了一个“外挂”：先给它看一幅超高清的图像，让它观察到最细微的特征；然后，对这些高清特征图进行整体平滑处理，让它们更加稳定、连贯。这个经过平滑处理的版本，就成了一份理想的“局部特征关系说明书”。学生模型在训练过程中参照这份说明书进行校准，因而能够在需要精细处理的任务中表现更佳。需要强调的是，所选用的教师模型不应是训练过久、已出现“偏科”的模型，而应是训练中期、全局与局部能力相对均衡的模型。

2.4　模型蒸馏：多场景模型家族的形成

DINOv3的“满血”骨干网络包括在LVD-1689M数据集上训练的ViT-7B/16模型，以及一个在SAT-493M数据集上训练的同名模型。从模型的命名方式可以看出，ViT-7B/16模型采用Vision Transformer架构，拥有6716M（约7B，即70亿）个参数，其中的“/16”表示模型使用16×16像素的图像块作为基本输入单元。

正如前文所述，DINOv3凭借大模型训练集、超大参数量以及优化的训练方法，获得了强大的特征表达能力。通过冻结骨干网络参数，仅训练与之匹配的特定任务头，便可在相应的场景下实现高效的推理。然而，如此参数规模的模型在实际部署时对算力资源的要求较高，在企业级特定视觉任务中往往难以兼顾成本与收益。解决此问题的方法是将ViT-7B/16模型的知识蒸馏到参数规模较小的ViT变体模型中，以获得性能损失可控、算力需求更低、推理效率更高的轻量化模型。

2.4.1　算力需求：DINOv3 多模型家族算力估算

与大多数大语言模型采用半精度（16位，每个参数占2字节）权重不同，DINOv3使用的是全精度浮点型参数（32位，每个参数占4字节）。因此，7B模型的权重至少需要28GB的GPU内存用于加载和推理。相应地，实际部署通常需要两张16GB或24GB显存的显卡（或等效推理加速卡）协同运行。

相比之下，许多常见视觉模型通常针对特定任务设计，在4GB显存的消费级显卡上即可稳定运行。由此可见，DINOv3 7B模型的部署成本相当高，这在一定程度上限制了它的推广应用，因为多数研究机构、企业或开发者难以承担如此规模的硬件投入。

为了使DINOv3强大的视觉表示能力能够覆盖更广泛的应用场景，降低使用门槛，研究团队采用了模型蒸馏（Model Distillation）技术，将70亿参数大模型中的知识有效迁移到更小、更高效的轻量模型中，从而构建出一个能够适配不同算力预算和应用场景的DINOv3模型家族。

该模型家族覆盖了广泛的算力预算范围，也便于与同期主流模型进行对比。其成员包括标准的ViT-S（2100万参数）、ViT-B（8600万参数）、ViT-L（3亿参数）模型，以及自定义的ViT-S+（2900万参数）和ViT-H+（8亿参数）模型。后两种模型的设计目的在于进一步缩小与70亿参数教师模型之间的性能差距。即便是蒸馏模型中参数规模最大的ViT-H+，也可以部署到4GB显存的显

卡上运行，从而显著降低DINOv3模型的应用门槛。

2.4.2　知识迁移：DINOv3 蒸馏模型的基本原理

模型蒸馏的核心思想是，利用一个性能强大但计算成本较高的教师模型，来指导结构更简单、参数规模更小的学生模型进行学习。在DINOv3的蒸馏过程中，教师模型即为训练充分、兼具卓越全局语义理解能力和局部细节表征能力的70亿参数模型ViT-7B。学生模型则可以是一系列不同参数规模与架构配置的轻量级模型，例如参数规模从数千万到数亿不等的各种ViT变体，如ViT-Small、ViT-Base、ViT-Large等，甚至可以是针对特定硬件平台优化的模型结构。

具体而言，DINOv3的蒸馏策略不仅仅是简单地让学生模型模仿教师模型的输出类别概率，更重要的是传递教师模型所学习到的丰富特征表示。通过特征层面的知识迁移，学生模型不仅能够继承教师模型在图像分类等任务中的高精度，还能获得对下游密集预测任务至关重要的、具有良好一致性和判别性的视觉特征。

蒸馏后的小模型在保持与大模型性能接近的同时，显著降低了计算资源消耗和推理延迟。例如，一个经过蒸馏的ViT-Base模型，其参数规模不到1亿，仅为70亿参数大模型的几十分之一，却能够在大多数实际应用场景中提供令人满意的性能，并可在普通消费级GPU甚至部分边缘计算设备上高效运行。

由此，DINOv3模型家族得以形成。从参数规模庞大、性能顶尖、适用于云端服务器等算力充沛场景的“旗舰版”70亿模型，到轻量化、高效率、适用于移动端、嵌入式设备或实时应用场景的“精简版”模型，DINOv3能够满足不同用户与应用场景的多样化需求。这种多场景模型家族的形成，显著扩展了DINOv3的应用范围，使其先进的视觉理解能力能够在更广泛领域落地生根，进而推动计算机视觉技术在工业、医疗、自动驾驶、智慧城市等行业中的实际应用。

2.4.3　并行蒸馏：高效的学生模型蒸馏流程

使用教师模型逐个蒸馏不同规模的小模型，是最稳妥且直观的做法，也符合常理。通常可以理解为，ViT-S/16、ViT-S+/16、ViT-B/16、ViT-L/16和ViT-H+/16等学生模型，均由ViT-7B/16教师模型逐一训练得到，分别对应不同规模的蒸馏结果。然而，这种串行蒸馏方式在算力成本上并不经济，因为在整个蒸馏过程中，教师模型的前向推理（即知识提取）是主要的算力消耗来源。

为更高效地实现多学生模型蒸馏，DINOv3研究团队设计了一种并行蒸馏流水线，使多个学生模型能够同时训练，并共享同一次教师推理的输出结果。也就是说，在每次训练迭代中，仅需执行一次教师模型的前向推理，所有学生模型即可共享该推理结果。这一机制类似于现实中的课堂教学：老师通常不会一对一串行授课，而是讲授一次，全班学生同步学习。

不过，与现实课堂中学生能力相对相近不同，DINOv3蒸馏体系中的学生模型因参数量和架构差异，训练耗时各不相同。通常情况下，规模较小的模型完成一次学生训练迭代所需的时间更短。因此，在并行蒸馏过程中，需要根据模型规模合理分配GPU资源：为较小模型分配相对较少的计算

资源，为较大模型分配更多算力资源，以尽量使各模型的训练进度协调一致，避免部分设备空闲等待，从而最大化整体算力利用率。

通过这一并行蒸馏机制，研究团队得以高效并行地训练多个学生模型，从旗舰级70亿参数的DINOv3模型中构建出完整的蒸馏模型家族，实现性能与效率之间的良好平衡。

2.5　多模态理解：图像-开放词汇文本对齐训练

图像-开放词汇文本对齐是DINOv3支持多模态能力的重要特性之一。在dino.txt数据集上训练得到的“零样本”分类任务头，可以在无须任何图像-文本标签配对训练的情况下，执行开放词汇（即非固定的文本标签集合）的分类任务。这一能力在效果上与CLIP系列模型的零样本分类能力相当。

2.5.1　泛化过程：图像-开放词汇文本对齐的基本原理

传统的“图像-文本标签”训练与“图像-开放词汇文本对齐”训练之间的核心区别在于：前者通常局限于模型训练阶段所使用的特定类别标签，只能识别并输出这些已知标签对应的文本类别；后者则强调模型能够理解并关联任意自然语言描述与图像内容，而无须依赖预定义的、固定的标签集合。例如，在面对一幅包含“一只正在追逐彩色蝴蝶的棕色小狗”的图像时，经过图像-文本标签训练的模型，可能只能区分出“狗”“蝴蝶”这些基础类别；而经过良好图像-开放词汇文本对齐训练的模型，不仅能识别基础类别，还能够理解并关联“追逐”“彩色”“棕色”等更细致的描述性词汇，甚至可以对“一只毛茸茸的、活泼的小家伙在花园里嬉戏”这样的自然语言描述做出准确的图像匹配。

图像-开放词汇文本对齐训练的核心目标是赋予模型更强的泛化能力，使其能够在图像中的视觉概念与人类语言中的任意词语或短语之间建立联系，即便这些词语在训练数据中并未以显式的标签形式出现过。这种能力使模型能够胜任更复杂、更灵活的多模态任务，如零样本图像分类（识别训练阶段未见过的长文本类别描述）、图像的自由文本描述生成、基于文本查询的图像检索等，从而显著拓展模型在真实世界场景中的应用潜力。

为了实现这一目标，DINOv3在训练过程中需要学习图像视觉特征与文本语义特征之间的深层对齐机制。通常的做法是，引入大规模图像-文本对数据，通过对比学习等方式，将图像和文本映射到共享的语义嵌入空间。在这个空间中，语义相近的图像和文本（即便具体词语并未完全匹配）会具有相近的嵌入表示。这一过程既涉及对图像中局部视觉特征与文本中词语、短语等语义单元之间的精细关联建模，也需要对全局语义结构进行统一建构，使模型既能理解“图像内容是什么”，也能理解“如何用语言描述它”，并进一步根据语言描述“找到对应的图像内容”。

2.5.2　图像与标题匹配：DINOv3 图像-开放词汇文本对齐训练方法

图像-开放词汇文本对齐技术因具备实现灵活且可扩展的多模态理解能力，受到学术界与产业界的广泛关注。大量研究工作集中于提升CLIP模型性能，该模型最初仅聚焦于学习图像与文本表示之间的全局对齐关系。近期的研究表明，基于预训练的自监督视觉骨干网络，可以有效实现图像-文本对齐。这意味着，即便视觉骨干网络已经完成训练，仍可在“图像 + 文本”的多模态场景中复用，从而构建丰富、精确的图像-文本跨模态关联能力。

在DINOv3中，针对图像-开放词汇文本对齐场景，训练方式为：在保持视觉编码器参数冻结的前提下，对文本编码器从零开始进行训练，使其输出表示能够与DINOv3的视觉特征对齐，从而实现图像与标题之间的匹配。图像与标题之间的对应关系相对容易获取，例如在知乎、微信公众号等平台中，图片通常配有与其内容相关的标题或说明文字。

由于视觉骨干网络参数被冻结，为增强视觉侧的表达灵活性，研究团队在冻结的视觉骨干网络上添加了两个Transformer层。该方法的一个重要改进在于，在与文本嵌入进行匹配之前，将均值池化后的图像块嵌入与最终输出的[CLS] token进行拼接。该设计使得全局与局部视觉特征能同时参与跨模态对齐，从而无须依赖额外的启发式策略，便可在提升图像-文本匹配效果的同时，进一步增强模型在密集预测任务中的性能。

2.6　本章小结

本章以DINOv3的技术报告为基础，系统讲解了这一强大视觉骨干网络的关键训练步骤。从数据准备、自监督训练到Gram锚定机制等方面，详细分析了模型在不同训练阶段所面临的问题、相应的解决方案和具体实现方法。

随后，本章深入探讨了DINOv3模型家族的构建策略，重点分析了70亿参数“满血版”骨干网络在算力成本方面的挑战，以及如何通过高效的模型蒸馏技术，将其知识迁移至不同规模的学生模型，从而形成覆盖广泛算力预算的模型系列。同时，还介绍了并行蒸馏流水线在提升多学生模型训练效率方面的创新设计。

最后，本章介绍了DINOv3在多模态理解领域的重要进展，重点阐述了图像-开放词汇文本对齐的基本原理与具体训练方法。通过对上述核心技术的系统梳理与分析，本章全面展示了DINOv3在模型训练机制、效率优化策略以及多模态扩展方面的关键创新，使读者对DINOv3的整体技术架构与创新要点形成了较为系统的认识，也为后续章节中DINOv3的环境搭建、实际应用开发和模型训练实践提供了必要的理论基础和技术铺垫。