# **DINO与CLIP综合分析报告：范式、性能及在现代计算机视觉中的演进**

### **第一部分：视觉基础模型导论**

本部分旨在为DINO和CLIP的出现建立宏观背景，将它们定位为对传统监督学习范式固有局限性的解决方案，并阐明它们代表了两种截然不同但同样强大的视觉表征学习哲学。

#### **1.1 从监督学习到范式转移**

在深度学习的早期发展阶段，监督学习（Supervised Learning）是计算机视觉领域无可争议的主导范式。然而，其成功建立在一个关键前提之上：海量、高质量的标注数据。这一前提逐渐演变为该领域发展的核心瓶颈。

##### **1.1.1 标注瓶颈**

传统监督学习模型的训练依赖于大规模、劳动密集型且成本高昂的数据标注过程 1。无论是图像分类、目标检测还是语义分割，每个任务都需要人类专家为数百万张图片提供精确的标签。这种对“黄金标签”的依赖极大地限制了模型的通用性和可用性，因为每当需要识别一个新的视觉概念时，都必须重新收集和标注相应的数据 4。这不仅减缓了研发周期，也使得模型难以扩展到标签稀缺或标注成本过高的专业领域（如医学影像、卫星图像分析）。

##### **1.1.2 自监督学习的兴起**

为了突破标注瓶颈，研究界将目光投向了自监督学习（Self-Supervised Learning, SSL）。SSL是一种革命性的方法，它巧妙地利用数据本身来生成监督信号，从而摆脱了对外部人工标签的依赖 2。其核心思想是通过精心设计的“代理任务”（pretext task）让模型学习数据内在的结构和关系 3。例如，模型可能被要求预测图像被遮挡的部分、将图像的不同增强视图识别为同一主体，或判断图像块的相对位置。通过完成这些任务，模型被迫学习到关于视觉世界的通用、可迁移的表征。在当今数据爆炸但绝大多数未被标注的时代，SSL的出现为利用海量原始数据提供了可能 2。

##### **1.1.3 语言监督的力量**

与SSL并行发展的另一条路径是语言监督（Language Supervision）。该范式主张直接从与图像伴随出现的原始文本中学习 4。互联网上存在着数以亿计的“图像-文本”对，这些文本描述为视觉概念提供了丰富、多样且极具表现力的监督来源，远超预先设定的固定类别标签 4。监督信号从数千个离散的类别（如ImageNet）跃升至自然语言所能表达的近乎无限的视觉概念集合 5。这种方法利用人类语言的泛化能力，为模型构建了一个更广阔的语义世界。

##### **1.1.4 定义基础模型**

正是在这样的背景下，“基础模型”（Foundation Models）的概念应运而生。基础模型通常指那些在广泛、海量的数据上（通常通过自监督或语言监督方式）进行预训练的大规模模型，它们可以被适配（adapt）到各种各样的下游任务中 10。其核心价值在于用一个通用的、强大的预训练模型取代过去为每个特定任务单独训练的多个狭隘模型，极大地提升了AI开发的效率和模型的泛化能力 12。

#### **1.2 两大支柱的定义：DINO与CLIP**

在视觉基础模型的浪潮中，DINO和CLIP分别成为了自监督学习和语言监督学习两大范式的杰出代表，它们各自选择了通往通用视觉智能的不同道路。

##### **1.2.1 DINO：从视觉结构中学习**

DINO（Self-**DI**stillation with **NO** Labels，无标签自蒸馏）是自监督学习的典范。它通过强制模型对同一图像的不同增强视图（例如，裁剪、变色、旋转）产生一致的输出来学习视觉表征 1。DINO的哲学根植于一个信念：丰富而深刻的语义与结构信息内在于视觉数据本身，无需任何外部标签或文本，仅通过探索图像内部的一致性与不变性即可被有效提取 17。

##### **1.2.2 CLIP：从语言关联中学习**

CLIP（**C**ontrastive **L**anguage-**I**mage **P**re-training，对比语言-图像预训练）是语言监督学习的里程碑。它通过在一个巨大的“图像-文本”对数据集上进行训练，学习将图像和描述它们的自然语言文本在语义上关联起来 4。CLIP的哲学在于，视觉概念的“意义”可以通过学习它们与人类语言的对应关系来掌握 7。模型通过辨别哪个文本描述与哪个图像匹配，从而构建起一个跨模态的联合表示空间。

##### **1.2.3 根本性的分歧**

DINO与CLIP的对比，揭示了通往通用视觉理解道路上一个根本性的哲学分岔。DINO致力于从物体的结构、纹理和不变性中学习“一个物体看起来是什么样”，它是一种内求的、基于视觉世界内在规律的探索。而CLIP则选择了一条外求的道路，它通过学习人类赋予视觉概念的语言标签来理解“一个物体被称作什么”。这种学习信号的本质差异，导致了它们各自拥有互补的优势与劣势。DINO的优势在于对像素级、结构化信息的深刻理解，例如其能够自发地进行语义分割；而CLIP的优势在于其惊人的零样本分类能力和对抽象语义概念的把握。这并非简单的技术路线差异，而是关于视觉智能本质的两种不同押注：一种相信智能源于对世界内在结构的无监督发现，另一种则相信智能最高效的构建方式是借助人类已有的符号系统——语言。后续模型的发展，特别是融合两者的混合模型的出现，也暗示着最终的答案可能在于统一这两种哲学。本报告的后续部分将深入探讨这两种范式所带来的具体影响和深远后果。

### **第二部分：DINO与DINOv2：通过自监督掌握视觉表征**

本部分将对DINO系列模型进行深入的技术解构，不仅阐述其功能，更探究其各组件的工作原理与设计动机，并追溯其如何演进成为一个真正的视觉基础模型。

#### **2.1 DINO框架：无标签自蒸馏**

DINO框架的设计精妙地将知识蒸馏、动量编码和多种正则化技术结合在一起，专为解决自监督学习中的核心挑战——模型坍塌（model collapse）而生。

##### **2.1.1 核心架构：视觉Transformer（ViT）与自监督的协同作用**

视觉Transformer（ViT）最初被认为比卷积神经网络（ConvNets）需要更多的数据和计算资源 15。然而，DINO的研究表明，自监督学习能够解锁ViT中在监督学习设置下未能显现的新特性 15。ViT的核心机制——自注意力（self-attention），允许模型在图像块（patches）之间建立长距离依赖关系，捕捉全局上下文信息 22。这种能力与自监督学习的目标不谋而合，即通过比较图像的不同视图来学习整体结构。DINO证明了ViT与SSL的结合能够产生高质量的、具有明确语义结构的特征。

##### **2.1.2 师生动态：动量编码器与知识蒸馏**

DINO的核心思想可以被看作一种没有标签的知识蒸馏（knowledge distillation） 1。框架中包含两个网络：一个学生网络（student network,

gθs​​）和一个教师网络（teacher network, gθt​​）。学生网络通过梯度下降进行训练，其目标是匹配教师网络的输出。

与传统知识蒸馏不同，DINO的教师网络不是一个固定的、预先训练好的模型。相反，它是一个**动量更新**的学生网络副本 1。教师网络的参数

θt​ 是学生网络参数 θs​ 的指数移动平均（Exponential Moving Average, EMA）：

θt​←λθt​+(1−λ)θs​

其中，λ 是一个动量系数，通常从一个较高的值（如0.996）开始，并随着训练的进行逐渐增加到1 15。

这种动量编码器机制扮演了关键角色。它提供了一个比学生网络更稳定、更平滑的目标。由于教师网络是过去多个学生网络版本的“集合”或“平均”，其输出特征质量更高，能够有效指导学生网络的学习，防止训练因目标快速变化而陷入不稳定 15。实验观察到，在整个训练过程中，教师网络的性能始终优于学生网络，证明了其作为可靠向导的有效性 15。

##### **2.1.3 防止模型坍塌的关键技术组件**

在自监督学习中，一个常见的失败模式是模型坍塌，即模型对任何输入都输出相同的、无意义的向量。DINO通过引入一系列互补的技术来有效避免这一问题。

* **多尺度裁剪增强（Multi-Crop Augmentation）**：这是DINO性能的关键贡献之一。对于每一张输入图像，系统会生成一组不同的视图：两个高分辨率（如224x224）的“全局”视图和多个低分辨率（如96x96）的“局部”视图 1。训练的核心在于“局部到全局”的对应学习（local-to-global correspondence）：只有两个全局视图被送入教师网络，而所有视图（包括全局和局部）都被送入学生网络 1。学生网络的目标是，即使只看到一个小的局部视图，也要能预测出教师网络从全局视图中看到的表征。这迫使模型学习到物体的组成部分与整体之间的关系，从而获得更丰富的特征 15。
* **中心化（Centering）与锐化（Sharpening）**：这两个操作直接作用于教师网络的输出，以防止其分布坍塌。
  + **中心化**：该操作为教师网络的输出 logits 加上一个偏置项 c。这个中心 c 是通过批次特征的EMA来更新的，确保特征在所有维度上是零中心的 1。其目的是防止某个维度主导输出，从而避免模型总是预测相同的向量。
  + **锐化**：该操作通过在教师网络的Softmax函数中使用一个较低的温度参数 τt​ 来实现。较低的温度会使输出的概率分布更加“尖锐”，即更集中于少数几个维度，从而防止模型输出均匀分布的“平庸”解 1。

中心化和锐化起着相反但互补的作用，它们的结合足以在动量教师的配合下稳定训练，避免坍塌 15。

##### **2.1.4 损失函数与训练**

DINO的损失函数是一个标准的交叉熵（cross-entropy）函数，用于最小化学生网络和教师网络输出的K维概率分布之间的差异 1。对于多视图输入，总损失是所有学生网络视图与所有教师网络视图之间的交叉熵之和。

一个至关重要的实现细节是，在计算损失时，对教师网络的输出应用了\*\*停止梯度（stop-gradient, sg）\*\*操作（在PyTorch中为 .detach()）15。这意味着梯度只通过学生网络反向传播，而不会流向教师网络。这确保了学习的单向性：学生向教师学习，而不是两者相互影响导致不稳定。

#### **2.2 DINO特征的涌现属性**

当DINO框架与ViT架构结合时，其学习到的特征表现出了一些令人瞩目的“涌现属性”（emerging properties），这些属性在传统的监督学习模型中并不常见。

##### **2.2.1 无监督语义分割**

这是DINO最显著的特性。在没有任何像素级标注的情况下，DINO训练的ViT模型最后一层的自注意力图（self-attention maps）能够直接、清晰地揭示图像的语义分割 15。注意力头能够自动地将焦点集中在图像中的一个或多个语义对象上，其生成的注意力图质量之高，堪比一些有监督的分割模型。这种现象在监督学习训练的ViT或ConvNet中均未被明确观察到 15。这表明，自监督学习迫使模型理解了物体的轮廓和空间范围，而不仅仅是识别其类别。

##### **2.2.2 高保真k-NN分类器**

DINO学习到的特征具有高度的辨别力和良好的空间结构。这一点通过k-近邻（k-NN）分类实验得到了证明。直接使用DINO提取的特征，无需任何微调或训练线性分类器，仅通过简单的k-NN算法，就能在ImageNet等基准上取得非常高的分类准确率。例如，一个小型ViT模型仅凭k-NN就在ImageNet上达到了78.3%的top-1准确率 1。这表明DINO学习到的特征空间中，同类样本在几何上高度聚集，不同类样本则被有效分离开。

#### **2.3 演进至DINOv2：为稳健、通用的特征而扩展**

DINO的成功为自监督学习奠定了坚实的基础，而其后续者DINOv2则将这一范式推向了新的高度。DINOv2的目标不再仅仅是证明自监督的可行性，而是要创造出真正意义上的“通用视觉特征”——一种无需微调即可在各种下游任务和数据分布上表现出色的“基础”特征 8。

##### **2.3.1 LVD-142M数据集：自动化数据管理的决定性作用**

DINOv2的成功在很大程度上归功于其对训练数据的深刻理解和精心构建。研究团队认识到，简单地增加无标签数据的数量并不足以保证模型质量的提升；相反，数据的*质量、多样性和管理*至关重要。这一认识标志着自监督学习从“更多数据”向“更优质数据”的理念成熟。过去，研究者们普遍认为，只要数据量足够大，模型就能学到有用的知识。但DINOv2的实践表明，未经筛选的原始网络数据质量参差不齐，直接用于训练会导致特征质量显著下降 24。

为了解决这个问题，DINOv2团队开发了一个复杂的自动化数据管理流水线，用于构建名为LVD-142M的高质量预训练数据集 24。这个过程本身就是一项重大的研究和工程贡献，其重要性不亚于模型架构本身。

1. **数据源**：流水线从一个包含12亿张图片的原始、未经过滤的网络图片池开始 10。
2. **去重（Deduplication）**：为了增加数据的多样性并减少冗余，系统首先使用一个高效的相似性检测算法去除重复和近乎重复的图片 10。
3. **检索与策划（Retrieval & Curation）**：这是最关键的一步。系统使用一个预训练好的自监督模型（一个在ImageNet-22k上训练的ViT-H/16）为所有图片计算嵌入向量。然后，它以一系列高质量的、已有的策划数据集（如ImageNet、Google Landmarks等）作为“种子”，从12亿的无标签图片池中检索出与这些种子图片在视觉上最相似的图片 10。整个过程完全基于图像内容的相似性，不依赖任何文本或元数据 10。

最终，这个自动化流程筛选出了1.42亿张高质量、多样化的图片，构成了LVD-142M数据集。这个过程本身就体现了一种元学习思想：利用已有的高质量数据分布去引导和塑造一个更大、更通用的无标签数据集。这也预示着，未来基础模型研发的瓶颈将越来越多地从模型设计转向大规模数据工程。

##### **2.3.2 算法与效率的增强**

DINOv2在DINO的基础上融合了多项最新的算法和工程优化，以支持更大规模的稳定训练。

* **混合损失函数**：DINOv2结合了两种互补的自监督损失。除了DINO原有的图像级对比损失，它还引入了**iBOT损失**，这是一种基于掩码图像建模（Masked Image Modeling）的块级（patch-level）损失 24。这迫使模型不仅要理解图像的全局内容，还要学习预测局部图像块的细节。
* **KoLeo正则化器**：为了使学习到的特征在嵌入空间中分布得更均匀，避免特征坍缩到空间的某个小区域，DINOv2引入了KoLeo损失项。这被证明能显著提升模型在图像检索等需要良好特征分布的任务上的性能 24。
* **规模化与效率**：为了在可接受的时间内训练数十亿参数的模型，DINOv2团队实现了一系列尖端的工程优化，包括：使用FlashAttention等快速且内存高效的注意力机制、用于处理不同尺寸输入的序列打包（sequence packing）、高效的随机深度（stochastic depth）以及用于分布式训练的完全分片数据并行（FSDP）技术 10。这些改进使得DINOv2的训练速度比之前的实现快了约2倍，内存占用减少了3倍 10。
* **模型蒸馏**：DINOv2系列中较小的模型（如ViT-S, B, L）并非从头训练，而是通过知识蒸馏的方式，由最大的10亿参数模型（ViT-g）“教导”而成。这种方法被证明比从零开始训练能获得更好的性能 24。

##### **2.3.3 性能分析：超越前代与弱监督模型**

DINOv2的性能表现验证了其设计理念的成功。其冻结的（frozen）骨干网络提取的特征，在众多基准测试中全面超越了之前的自监督模型，甚至在许多任务上匹敌或超过了强大的弱监督模型（如CLIP）10。

以下表格汇总了DINOv2与DINO及其他代表性模型在关键基准上的性能对比，数据主要来源于DINOv2论文中的详细实验结果 24。

**表1：DINO与DINOv2在关键基准上的性能对比**

| 任务 / 指标 | 模型 (架构 / 数据) | 性能 (DINOv2) | 性能 (DINO) | 性能 (OpenCLIP) | 性能 (iBOT) |
| --- | --- | --- | --- | --- | --- |
| **ImageNet-1k 分类 (线性评估)** |  |  |  |  |  |
| Top-1 准确率 (val) | DINOv2 ViT-g/14 (LVD-142M) | 86.5% | 79.2% (ViT-S/8) | 86.2% (ViT-G/14) | 82.3% (ViT-L/16) |
| **深度估计 (RMSE, ↓越好)** |  |  |  |  |  |
| NYUv2 (DPT) | DINOv2 ViT-g/14 (LVD-142M) | 0.279 | 0.492 (ViT-B/8) | 0.414 (ViT-G/14) | 0.358 (ViT-L/16) |
| KITTI (DPT) | DINOv2 ViT-g/14 (LVD-142M) | 2.11 | 2.74 (ViT-B/8) | 2.56 (ViT-G/14) | 2.55 (ViT-L/16) |
| **语义分割 (mIoU, ↑越好)** |  |  |  |  |  |
| ADE20k (+ms) | DINOv2 ViT-g/14 (LVD-142M) | 53.0% | 35.2% (ViT-B/8) | 46.0% (ViT-G/14) | 47.5% (ViT-L/16) |
| CityScapes (+ms) | DINOv2 ViT-g/14 (LVD-142M) | 81.0% | 66.2% (ViT-B/8) | 70.3% (ViT-G/14) | 74.5% (ViT-L/16) |

从表中可以清晰地看到：

* **分类性能**：DINOv2在线性评估上实现了显著飞跃，不仅远超初代DINO，还略微超过了使用更大规模图文对数据训练的OpenCLIP模型。
* **密集预测任务**：在深度估计和语义分割等需要精细空间理解的任务上，DINOv2的优势是压倒性的。其性能远超DINO、OpenCLIP和iBOT，这直接证明了其学习到的特征具有卓越的局部化和空间感知能力。这支持了一个重要观点：基于文本描述的预训练（如CLIP）可能难以学习到这些任务所需的细微像素级信息 10。

### **第三部分：CLIP与OpenCLIP：通过语言监督解锁零样本能力**

本部分将深入剖析CLIP的架构与训练方法，重点阐述其独特的语言监督目标如何催生强大的零样本（zero-shot）能力，并探讨围绕其形成的繁荣开源生态系统。

#### **3.1 CLIP框架：对比语言-图像预训练**

CLIP的设计理念是简洁而强大的：通过对比学习，将视觉和语言两个模态的信息映射到一个统一的、有意义的嵌入空间中。

##### **3.1.1 双编码器架构**

CLIP采用了一个清晰的双塔（two-tower）架构，包含一个图像编码器和一个文本编码器，两者在训练过程中被联合优化 4。

* **图像编码器（Image Encoder）**：研究人员探索了两种主流的骨干网络。一种是经过改进的ResNet-50，例如引入了ResNet-D和抗锯齿模糊池化等技术，并将全局平均池化层替换为更强大的注意力池化机制 30。另一种则是视觉Transformer（ViT），其强大的表征能力在CLIP的规模下得到了充分发挥 30。
* **文本编码器（Text Encoder）**：文本编码器采用标准的Transformer架构 29。它处理的是经过字节对编码（Byte Pair Encoding, BPE）的文本，词汇表大小约为49,000。文本序列被  
  （开始）和（结束）标记包围，而``标记在Transformer最高层的输出激活值被用作整个文本的特征表示，并最终被线性投影到多模态嵌入空间 30。

##### **3.1.2 对比学习目标：在共享嵌入空间中对齐模态**

CLIP的训练任务极其巧妙。在一个包含N个（图像，文本）配对的批次（batch）中，模型的目标是从N×N个所有可能的配对中，准确地找出那N个正确的配对 5。

这个过程通过一个对比损失函数（contrastive loss）来实现。模型被训练来**最大化**N个真实配对的图像嵌入和文本嵌入之间的**余弦相似度**，同时**最小化**其余N2−N个错误配对的嵌入之间的余弦相似度 5。

损失函数是一个对称的交叉熵损失，它同时在“图像到文本”和“文本到图像”两个方向上计算。具体来说，对于一批图像特征If​和文本特征Tf​，它们被投影到共享空间得到Ie​和Te​。计算Ie​和TeT​的点积得到一个N×N的相似度矩阵（logits）。损失函数的目标是让这个矩阵尽可能地接近单位矩阵，即对角线上的值（真实配对）最大，而非对角线上的值（错误配对）最小 32。

这个简洁的对比学习目标被证明在学习零样本能力方面，比之前复杂的图像到文本生成式方法效率高出4到10倍 4。

##### **3.1.3 WIT数据集：网络规模数据的力量与风险**

CLIP的强大能力离不开其海量的训练数据。OpenAI构建了一个名为WIT（WebImageText）的私有数据集，包含了从互联网上收集的4亿个（图像，文本）对 4。

构建这个数据集的动机是，当时已有的公开图文对数据集（如MS-COCO）规模太小，或者元数据质量参差不齐 4。为了真正利用网络上近乎无限的监督信号，必须创建一个更大规模的数据集。WIT的构建过程是自动化的：研究者定义了一个包含50万个查询词/短语的列表，然后在网络上爬取与这些查询相关的图文对，每个查询最多收集2万对，以实现大致的类别平衡 5。

这种从原始、未经过滤的互联网数据中学习的方式，是CLIP惊人泛化能力的直接来源，因为它接触到了极为广泛和多样的视觉概念。然而，这也正是其固有社会偏见的根源。网络数据是现实社会的一面镜子，其中包含了大量的刻板印象和不平衡的表征。CLIP在学习这些图文关联的同时，也忠实地学习并放大了这些偏见 29。

#### **3.2 零样本迁移的力量**

CLIP最革命性的贡献在于其强大的零样本（zero-shot）迁移能力，即在没有见过任何特定任务的标注样本的情况下，直接在该任务上表现出色。

##### **3.2.1 零样本分类的机制**

CLIP实现零样本分类的机制堪称优雅 4：

1. **构建动态分类器**：对于一个下游分类任务（例如，在猫和狗之间分类），CLIP不需要重新训练。取而代之的是，它将目标类别名称（如 "dog", "cat"）转换成描述性的文本提示（prompts），例如 "a photo of a dog", "a photo of a cat" 29。
2. **编码文本提示**：这些文本提示被送入预训练好的文本编码器，生成每个类别的文本嵌入向量。这些文本嵌入向量实质上构成了分类器的“权重”，一个基于文本描述动态生成的分类头 4。
3. **进行分类**：当一张新的待分类图像输入时，图像编码器会计算其图像嵌入向量。然后，计算该图像嵌入与所有类别文本嵌入之间的余弦相似度。相似度最高的那个类别，即为模型的预测结果 19。

这个过程完全不需要任何目标数据集的图像样本进行训练，实现了真正的“零样本”。

##### **3.2.2 提示工程与集成的艺术**

为了最大化CLIP的零样本性能，提示工程（prompt engineering）变得至关重要。研究发现，使用像 "a photo of a {label}." 这样的模板，性能远好于直接使用孤立的类别名 "label" 29。这是因为模板的句式更接近于CLIP在预训练时见过的自然语言描述。

此外，提示集成（prompt ensembling）是另一种有效的提升性能的技巧。研究者可以为同一个类别创建多个不同的提示（例如，"a photo of a big {label}", "a photo of a small {label}", "a {label} in a basket"），分别计算它们的文本嵌入，然后将这些嵌入平均起来，形成一个更鲁棒的类别表征 29。

#### **3.3 OpenCLIP生态系统**

由于OpenAI的CLIP模型和WIT数据集是私有的，社区迅速行动起来，催生了OpenCLIP项目，旨在复现并开源CLIP的训练和模型。

##### **3.3.1 使用公开数据复现和扩展CLIP**

OpenCLIP是一个至关重要的开源项目，它提供了可复现的CLIP训练代码，并使用公开可用的大规模数据集进行模型训练 29。其中，最著名的数据集是LAION，它提供了不同规模的图文对数据集，如LAION-400M（4亿对）和LAION-2B（20亿对）29。

通过在这些公开数据集上进行大规模训练，OpenCLIP社区成功地复现了原始CLIP的性能，并且训练出了许多性能更强、尺寸更多样的新模型，极大地推动了该领域的研究和应用 29。

##### **3.3.2 自定义数据训练的实践考量**

OpenCLIP的开源使得在自定义数据集上训练或微调CLIP模型成为可能。然而，这并非易事。

* **数据准备**：训练需要遵循特定的数据格式，例如提供一个CSV文件，其中包含图像文件的本地路径和对应的文本描述 35。
* **训练挑战**：CLIP的对比学习范式对训练设置要求很高。一个关键挑战是需要非常大的批次大小（batch size），有时甚至达到数万 29。小批次训练会导致模型性能不佳。另一个严峻的问题是“灾难性遗忘”（catastrophic forgetting）：当在一个新的、领域特定的数据集上微调CLIP时，模型很容易忘记它在海量数据上学到的通用知识，导致其强大的零样本能力在仅仅一个训练周期后就急剧下降 40。
* **解决方案**：为了解决灾难性遗忘，社区发展出了一些先进的微调技术，如WiSE-FT（通过加权平均原始模型和微调模型的权重来保留零样本能力）和LoRA（低秩适应，一种参数高效的微调方法）40。

CLIP的成功不仅仅在于其模型本身，更在于它所揭示的学习范式。它证明了，通过巧妙地利用人类在互联网上留下的海量、无组织的图文数据，AI可以学习到一种接近人类的、基于语义关联的视觉理解能力。这种能力的核心，可以看作是对人类集体网络行为的一种“算法蒸馏”。模型在学习图文配对的过程中，实际上是在学习一个压缩的、统计意义上的表征，这个表征反映了人类社会如何通过语言来描述和分类视觉世界。

这种学习方式解释了CLIP为何如此强大。当模型在4亿个样本中反复看到“狗的照片”这类文本与各种狗的图片配对时，它自然学会了这种泛化的关联 33。然而，这也直接解释了它的缺陷。如果网络数据中，“医生”的图片更多地与男性面孔关联，而“护士”的图片更多地与女性面孔关联，那么CLIP会忠实地学习到这种强相关性，并将其固化为模型的“知识” 34。模型本身没有公平或伦理的概念，它只是一个高效的优化器，在给定的数据分布上寻找最强的统计规律。因此，CLIP既是人类集体智慧的结晶，也是其集体偏见的一面镜子。理解这一点，对于我们利用其力量、规避其风险至关重要。

### **第四部分：DINO与CLIP的比较解剖**

本部分将对这两种模型进行直接的、基于证据的比较，从它们的基础原理到实际性能，再到它们所学习的表征特征，进行全方位的剖析。

#### **4.1 学习范式：正面交锋**

DINO和CLIP代表了两种截然不同的学习哲学，这体现在它们设计的方方面面。

* **监督来源**：DINO的监督信号是**内生的**，完全来自于数据本身。它通过数据增强和视图一致性来生成学习目标，不需要任何外部信息 1。相比之下，CLIP的监督信号是  
  **外生的**，完全依赖于人类生成的文本描述 4。
* **数据要求**：DINO可以从任何图像集合中学习，即使这些图像没有任何元数据或标签 17。而CLIP则严格要求成对的（图像，文本）数据 5。
* **目标函数**：DINO使用自蒸馏框架，通过交叉熵损失来匹配学生和教师网络的输出分布 15。CLIP则使用对称的对比损失函数，在一个共享的嵌入空间中，最大化正样本对的相似性，同时最小化负样本对的相似性 30。

**表2：DINO与CLIP架构与训练参数对比**

| 特性 | DINO / DINOv2 | CLIP / OpenCLIP |
| --- | --- | --- |
| **核心架构** | 学生-教师网络 (Student-Teacher) | 双编码器 (Dual-Encoder) |
| **图像编码器** | ViT, ResNet | ViT, ResNet |
| **文本编码器** | 无 | Transformer |
| **训练目标** | 自蒸馏 (Self-Distillation) + 交叉熵损失 | 对比学习 (Contrastive Learning) + 对称交叉熵损失 |
| **数据集类型** | 仅图像 (Image-only) | 图像-文本对 (Image-Text Pairs) |
| **数据集规模** | DINO: ImageNet (1.2M)  DINOv2: LVD-142M (142M) | CLIP: WIT (400M)  OpenCLIP: LAION (400M - 2B+) |
| **关键模块** | 动量编码器, 多尺度裁剪, 中心化/锐化, iBOT损失 (DINOv2) | 共享嵌入空间, 文本提示工程 |
| **计算量 (示例)** | DINOv2-g: 22,016 A100 GPU-hours | ViT-L/14: 12天 on 256 V100 GPUs |

#### **4.2 特征空间分析：局部与全局的二分法**

DINO和CLIP训练目标上的根本差异，直接导致了它们学习到的特征空间在几何结构和性质上呈现出鲜明的对比。这种差异并非偶然，而是其各自学习范式可预测的必然结果。

##### **4.2.1 DINO的优势：空间感知、定位与密集特征**

DINO的训练目标，特别是其“局部到全局”的多尺度裁剪策略，本质上是一种强大的正则化器，它迫使模型学习精细的、部件级别的对应关系 15。为了让一个小的局部图像块（例如，一只猫的耳朵）的表征能够匹配整个图像（一只完整的猫）的表征，模型必须理解这个“耳朵”是“猫”这个整体的一部分，并学习它们在空间上的关联。这个过程自然而然地催生了具有卓越空间感知能力的特征 15。

这就是为什么DINO的自注意力图能够神奇地对应于语义分割区域 15，也是为什么DINOv2在语义分割、深度估计等密集预测任务上表现卓著的原因 24。其特征被描述为捕获了“细粒度的语义信息” 45 和具有强大的“空间对应性” 46。

相比之下，CLIP的全局对齐目标使其缺乏这种空间意识。CLIP只关心整个图像的全局嵌入是否与文本描述的嵌入相匹配，而不在乎图像中“猫”的具体位置或轮廓 47。因此，若不经过大量修改或微调，CLIP难以胜任密集预测任务。

##### **4.2.2 CLIP的优势：全局语义、抽象概念与文本对齐**

CLIP的对比学习目标则塑造了一个完全不同的特征空间。通过将一个图像与正确的文本标题拉近，同时与批次内所有其他错误的标题推开，模型被迫学习能够区分高层语义概念的特征。这个空间是按“概念”组织的，而不是按“像素布局”组织的。

因此，CLIP的特征非常擅长捕捉图像的全局、整体性语义内容，即“图像的主旨” 50。更重要的是，由于其与自然语言深度绑定，CLIP能够理解和处理DINO无法企及的抽象概念（如“孤独感”）、艺术风格（如“立体主义风格”）或复杂的场景描述（如“一只宇航员骑着马的逼真照片”）52。而DINO学习到的纯视觉特征，若不经过额外的对齐训练，则无法用于任何基于文本的零样本任务 53。

#### **4.3 下游任务性能：量化比较**

两种模型在不同下游任务上的性能表现，具体地印证了它们特征空间的差异。

**表3：DINOv2与OpenCLIP在下游任务上的正面对决**

| 任务 / 指标 | DINOv2 (ViT-g/14) | OpenCLIP (ViT-G/14) | 优胜者 |
| --- | --- | --- | --- |
| **ImageNet-1k 分类 (线性评估)** | **86.5%** | 86.2% | DINOv2 (微弱优势) |
| **ImageNet-1k 分类 (零样本)** | N/A | **78.0% (ViT-H/14)** | CLIP (压倒性优势) |
| **ADE20k 语义分割 (mIoU)** | **53.0%** | 46.0% | DINOv2 (显著优势) |
| **NYUv2 深度估计 (RMSE)** | **0.279** | 0.414 | DINOv2 (显著优势) |
| **图像检索 (实例/拷贝检测)** | **极强** | 较弱 | DINOv2 |
| **图像检索 (文本-图像)** | N/A | **极强** | CLIP |

数据来源: 24。OpenCLIP ViT-G/14 数据来自DINOv2论文的对比实验。CLIP零样本数据来自其原始论文，使用不同但可比的架构。

这张表格清晰地描绘了两者之间的权衡：

* **零样本能力**：CLIP拥有DINO无法比拟的零样本分类和跨模态检索能力，这是其核心优势 33。
* **密集预测**：DINOv2在需要精细空间理解的任务上是无可争议的王者，其冻结特征的性能远超CLIP 24。
* **有监督微调/线性评估**：当有少量标签可用时（如线性评估），DINOv2的特征质量非常高，甚至略优于CLIP，表明其底层视觉表征的强大。

#### **4.4 鲁棒性与泛化能力**

* **分布外（OOD）泛化**：在ImageNet-ReaL和ImageNet-V2等更具挑战性的分布外测试集上，DINOv2表现出比CLIP更强的泛化能力 24。这可能是因为DINO的特征根植于更普适的视觉基本规律，而不是特定于网络文本词汇的模式。
* **数据增强/变换鲁棒性**：研究表明，DINO对图像变换（如模糊、压缩）的鲁棒性更强，性能保持稳定；而CLIP的性能在面对这些变换时可能会有显著波动 54。这再次印证了DINO学习到的是物体更本质、更不变的属性。

### **第五部分：生态系统影响与应用扩散**

本部分将视角从模型本身转向它们在整个计算机视觉生态系统中引发的连锁反应，展示DINO和CLIP如何从单纯的先进模型，演变为驱动新一代工具和应用创新的基础构建模块。

#### **5.1 密集预测任务：语义分割与深度估计**

DINO系列模型，特别是DINOv2，已经成为密集预测任务领域的首选特征提取器。

* **DINO的主导地位**：DINO最初展现出的“涌现”分割能力，使其自然而然地被应用于无监督和少样本语义分割任务 15。DINOv2的出现则彻底巩固了这一地位。其强大的冻结骨干网络被直接用作各种先进分割和深度估计算法的编码器，并取得了业界领先的成果 10。
* **混合方法**：当前的研究趋势是将DINO的强大视觉特征与其他模型的能力相结合。例如，研究者们将DINOv2的特征与Meta AI的分割一切模型（Segment Anything Model, SAM）结合，以进一步提升在医学影像等专业领域的分割性能 18。

#### **5.2 机器人学与具身智能：抓取、功能启示与导航**

DINOv2强大的细粒度特征使其在机器人感知领域大放异彩，尤其是在需要理解物体几何与功能的任务中。

* **DINOv2用于几何与功能理解**：机器人的一个核心挑战是理解物体的“功能启示”（affordance），即一个物体可以被如何使用。DINOv2的特征被用于功能启示建模，通过在不同物体实例（如不同种类的锤子）之间找到视觉上的对应关系，来预测“应该在哪里抓取” 57。
* **在抓取中的应用**：DINOv2提取的特征被用作机器人抓取生成策略的强大感知输入，常常与深度估计相结合 58。在一些先进的框架中，机器人直接使用DINOv2编码器处理摄像头图像，以在杂乱的场景中实现灵巧的抓取。其特征对背景、光照等环境变化具有很强的鲁棒性 60。
* **从图像到3D**：DINO的特征甚至被用作从单张图像重建三维模型的模型的输入，重建出的3D模型随后被用于抓取规划 61。

#### **5.3 开放词汇与生成式AI**

CLIP的出现，为计算机视觉带来了前所未有的“开放性”，并成为第一代大规模生成式AI模型的关键技术。

##### **5.3.1 CLIP赋能开放词汇目标检测（OVOD）**

CLIP的零样本分类能力是开放词汇目标检测（OVOD）的核心引擎 62。OVOD系统通常遵循一个两阶段流程：一个标准的检测器（如Faster R-CNN或DETR）负责提出“可能存在物体”的候选区域（region proposals），然后CLIP负责对这些区域进行分类，其类别可以来自一个任意定义的、开放的文本词汇表。

然而，将CLIP应用于区域级别也面临挑战：

* **分布不匹配**：CLIP在训练时看到的是完整的图像，而检测时面对的是裁剪出的物体区域，存在分布差异 63。
* **背景处理不佳**：CLIP的训练数据都是有对应文本的图像，它对不包含任何明确物体的“背景”区域分类能力很差，容易产生误报 66。

为了解决这些问题，研究者提出了多种方法，如DST-Det（利用CLIP为新物体生成伪标签进行自训练）62 和CORA（通过区域提示来弥合分布差异）63。

##### **5.3.2 CLIP在文生图生成模型中的角色**

CLIP是第一代大规模文生图模型（如DALL-E 2、Stable Diffusion）能够成功的基石之一 67。

* **文本编码**：在这些模型中，CLIP的文本编码器被用作将用户的文本提示（prompt）转换成一个富含语义的嵌入向量。这个向量作为条件，引导整个图像生成过程 68。
* **图文对齐（DALL-E 2）**：DALL-E 2（也称为unCLIP）的工作流程深刻体现了CLIP的作用。它首先训练一个“先验”（prior）模型，该模型学习如何将一个CLIP文本嵌入映射到一个对应的CLIP图像嵌入。然后，一个扩散解码器（diffusion decoder）再从这个图像嵌入中生成最终的图像 67。CLIP为这个过程提供了统一的、跨模态的语义空间。
* **重排序与引导**：生成模型一次可能产生多张图片，CLIP也被用来对这些生成结果进行打分和重排序，选出与文本提示在语义上最匹配的图片 52。

#### **5.4 视频分析：动作识别与密集跟踪**

将这些强大的图像模型应用于视频领域是一个活跃的研究方向，但也面临着模型缺乏时间维度理解的挑战 46。

* **DINO用于密集跟踪**：DINO强大的语义化块级特征使其成为密集点跟踪任务的理想基础。DINO-Tracker通过在单个视频上对DINO特征进行测试时微调，实现了业界领先的长期跟踪效果，尤其是在物体被长时间遮挡的情况下 45。
* **为视频蒸馏DINO**：为了让模型具备时间感知能力，FRAME模型通过知识蒸馏，将冻结的DINO（提供空间特征）和CLIP（提供语义特征）的知识“教给”一个专门的视频编码器，使其在密集视频任务上的性能超越了原始的两个图像模型 46。
* **DINO用于动作识别**：DINOv2的特征能够有效捕捉由动作引起的外观变化，这使得它在少样本动作识别任务中表现出色 73。

#### **5.5 更广泛的技术影响**

DINO和CLIP的影响力已经远远超出了模型本身，它们作为一种“使能技术”，催生并重塑了邻近的技术领域。这种从模型创新到产业变革的“三阶涟漪效应”深刻地展示了基础模型的颠覆性力量。

##### **5.5.1 彻底改变数据标注与管理工具**

首先，基础模型直接冲击了计算机视觉流程中最昂贵、最耗时的环节——数据标注 74。

* **模型辅助标注**：现代数据标注平台（如Roboflow, Kili, Picsellia）已经深度集成了Grounding DINO、SAM和CLIP等基础模型，实现了“自动标注”（auto-labeling）功能 75。用户只需提供一个文本提示（如“安全帽”），模型就能自动在图像中定位并生成所有安全帽的边界框或分割掩码。
* **从“劳工”到“监工”**：这极大地改变了人类标注员的角色。他们不再需要从零开始逐一像素地进行繁琐的手动标注，而是转变为一个监督和修正模型自动标注结果的“监工”角色。这一转变将标注效率提升了数倍甚至数十倍，并有望在未来完全取代某些任务的手动标注环节 74。

##### **5.5.2 驱动向量数据库在多模态检索中的普及**

其次，CLIP的成功直接催生了一个全新的数据库市场。

* **新数据库的需求**：传统的数据库（如SQL数据库）专为结构化数据和关键词搜索设计，它们无法处理像CLIP生成的那种高维、稠密的语义向量嵌入 77。
* **CLIP创造市场**：CLIP能够将图像和文本映射到同一个向量空间，这为实现真正的语义搜索提供了可能，同时也创造了一个巨大的、亟待满足的市场需求：一个能够高效存储、索引和查询数十亿高维向量的数据库。这直接推动了向量数据库（Vector Database）的兴起和商业化，代表产品包括Faiss、Milvus、Pinecone、Weaviate等 78。
* **工作原理**：向量数据库的核心是为高维向量上的近似最近邻（Approximate Nearest Neighbor, ANN）搜索进行优化 52。在应用中，整个图像数据集可以被预先编码成向量并存入数据库。当用户输入一个文本查询时，系统在运行时将其编码成一个查询向量，然后数据库利用高效的ANN算法，在海量数据中快速找到与查询向量在语义上最相似的图像向量，并返回对应的图像 80。这正是现代语义图像搜索引擎的底层技术。

这种从模型创新（一阶效应）到应用赋能（二阶效应），再到重塑邻近技术产业（三阶效应）的传导路径，清晰地表明了基础模型的深远影响。它们不仅是学术界的突破，更是推动整个AI技术栈演进的结构性力量。

### **第六部分：协同的未来：DINO与CLIP的融合**

随着研究的深入，社区开始超越将DINO和CLIP作为独立工具使用的阶段，积极探索如何将它们各自的优势结合起来，创造出性能更全面的混合模型。这标志着一个更高层次的“集成学习”范式的出现：研究者们不再满足于在推理时集成多个模型的预测结果，而是尝试在训练或微调阶段，直接“集成”不同基础模型的特征空间和知识表征，以期创造出一个通过“站在巨人肩膀上”而变得更强大的单一模型。

#### **6.1 融合的理论基础：结合空间精度与语义灵活性**

融合的动机非常明确，它源于对DINO和CLIP核心能力互补性的深刻认识。

* DINO系列模型拥有卓越的细粒度视觉编码能力和空间定位精度，但其特征与语言无关，缺乏语义的灵活性 82。它能精准地回答“在哪里”，但无法回答“是什么”。
* CLIP则拥有强大的语义理解和文本对齐能力，能够处理开放词汇，但其特征缺乏空间意识，定位能力差 47。它能模糊地回答“是什么”，但无法回答“在哪里”。

因此，研究的终极目标是创造一个统一的模型，既拥有DINO的精确定位能力，又具备CLIP的灵活语义理解能力 50。

#### **6.2 新兴的混合架构：融合方法分类**

当前，融合DINO和CLIP的尝试主要可以分为以下几类：

##### **6.2.1 蒸馏与精炼（Distillation and Refinement）**

这类方法利用一个模型的知识来“指导”或“改善”另一个模型的特征。

* **CLIP-DINOiser**：该方法旨在解决CLIP特征定位能力差的问题。它将DINO的块级相关性（patch correlations）和无监督对象显著性（objectness）先验知识作为监督信号，来训练几个轻量级的卷积层。这些卷积层的作用是学习如何对CLIP的密集特征进行加权池化，从而“教给”CLIP一种类似DINO的定位能力，而无需重新训练庞大的CLIP模型 47。通过这种方式，CLIP-DINOiser在单次前向传播中就能生成高质量的开放词汇语义分割结果，达到了业界领先水平。

##### **6.2.2 特征空间对齐与合并（Feature Space Alignment and Merging）**

这类方法尝试直接将两个模型的特征进行组合。

* **COMM (Combining CLIP and DINO with Multi-level feature Merging)**：这项工作首先发现，即使是纯视觉的DINOv2，只需配备一个简单的MLP对齐层，就能成为多模态大语言模型（MLLM）中一个出人意料的有效视觉分支。基于此，COMM提出了一种特征融合策略，将CLIP的全局语义特征和DINOv2的细粒度局部特征进行合并，为MLLM提供一个更强大、更全面的视觉输入 50。
* **Talk2DINO**：此方法更为激进，它学习一个非线性的映射函数，将CLIP的**文本嵌入**直接映射到DINOv2的**块级视觉特征空间**中。这相当于在不改动任何一个模型骨干网络的情况下，为DINOv2赋予了“听懂”语言的能力，使其能够根据文本提示进行精细的分割，同样在开放词汇分割任务上取得了顶尖性能 82。

##### **6.2.3 为自监督模型解锁语言能力**

这类方法以一个强大的自监督视觉模型为起点，为其添加语言对齐能力。

* **dino.txt**：该方法采用了LiT（Locked-image Text tuning，锁定图像的文本调优）范式。它将一个预训练好的、性能强大的DINOv2视觉编码器完全**冻结**，然后只训练一个文本编码器，使其输出的文本嵌入与DINOv2的图像嵌入对齐 53。为了同时对齐全局和局部信息，dino.txt提出了一种改进的池化策略，将图像的标记和所有块标记的平均池化结果拼接起来作为图像的最终表征。这种方法以远低于从零开始训练CLIP的计算成本，成功地训练出了一个具备CLIP式零样本能力、同时保留了DINOv2卓越视觉特征的模型，在多项基准测试上取得了SOTA结果 23。

##### **6.2.4 基于跨模型监督的提示调优**

这类方法利用多个模型的优势，以一种无监督的方式来优化目标模型。

* **NoLA (No Labels Attached)**：这是一个精巧的三步流程，旨在利用无标签图像来提升CLIP的性能。首先，它使用一个大语言模型（LLM）为类别生成丰富的文本描述，以获得比标准提示更好的文本嵌入。然后，它利用这些文本嵌入和DINO强大的视觉特征生成伪标签，并训练一个对齐模块。最后，这个基于DINO的对齐模块被用作“教师”，以无监督的方式指导对CLIP视觉编码器的提示调优（prompt-tuning）83。这种方法巧妙地集成了LLM的语言知识、DINO的视觉精度和CLIP的跨模态能力。

这些融合方法的涌现表明，未来的基础模型开发可能不再是单一模型的单打独斗，而是一个更加模块化、更高效的生态系统。通过聪明地组合、蒸馏和对齐现有的、强大的但各有专长的基础模型，研究界有望以更低的成本、更快的速度推动AI能力的边界。

### **第七部分：批判性评估与未来轨迹**

本部分将对DINO和CLIP进行平衡的批判性审视，承认它们的局限性，深入探讨社会偏见这一关键问题，提供实用的故障排除建议，并展望未来的发展方向。

#### **7.1 已知的局限性**

尽管DINO和CLIP取得了巨大成功，但它们并非没有缺点。

* **CLIP的弱点**：CLIP的原始论文坦诚地指出了其多项弱点。它在细粒度分类任务（如区分不同型号的汽车、不同品种的花卉）上表现不佳；对于更抽象、更系统的任务（如数出图像中物体的数量）几乎无能为力；对于真正分布外的、新颖的数据（如手写数字MNIST），其性能可能接近随机猜测 30。此外，CLIP的成功依赖于海量数据，其数据效率并不高 30。
* **DINO的弱点**：DINO的主要局限性在于其特征与文本天然不相关，这使得它无法直接用于任何基于文本的零样本任务，必须经过一个额外的对齐阶段 53。初代DINO的训练对超参数也较为敏感。
* **DINOv2的注意力伪影**：研究发现，使用类DINO目标训练的大规模ViT（包括DINOv2甚至OpenCLIP）的注意力图中，有时会出现异常的“伪影”（artifacts）。一些与图像背景对应的高范数（high-norm）令牌会错误地获得极高的注意力权重，干扰模型对前景对象的理解。后续研究提出的一个简单而有效的解决方案是，在模型输入序列中加入几个额外的“寄存器”（registers）令牌。这些寄存器令牌充当一个“草稿本”，让模型可以将全局信息存储在其中，从而释放图像令牌，使其更专注于局部内容，有效减少了伪影 86。

#### **7.2 语言监督模型的社会偏见挑战**

社会偏见是CLIP及其衍生模型面临的最严峻的挑战之一。从纯粹的技术角度来看，CLIP学习到的“偏见”并非其学习算法的“缺陷”（bug），而是其学习目标在有偏见的网络数据上成功执行的“特性”（feature）。对比损失函数非常擅长发现并编码数据中最强烈的统计相关性。在互联网上，社会刻板印象（如性别与职业、种族与特定概念的关联）恰恰是极其强大且在统计上显著的信号。因此，模型为了最小化其损失函数，会“正确地”学习到这些偏见，并将其视为有用的预测模式。

这个观点重塑了我们对偏见问题的理解。我们不能简单地要求模型“不要有偏见”，因为这与其核心训练目标相悖。我们必须从根本上解决问题，要么（a）改变数据，使我们希望模型学习的统计相关性是公平的；要么（b）改变学习目标或架构，明确地惩罚模型学习与受保护属性相关的关联。这解释了为什么去偏见如此困难：它往往是在与模型的主要训练目标作斗争。

* **偏见的来源**：问题的根源在于CLIP的训练数据——来自互联网的、未经过滤的图文对。这些数据忠实地反映了现实世界中存在的、有害的社会刻板印象 29。
* **量化偏见**：为了系统地研究偏见，研究人员开发了专门的基准测试（如FairFace, PATA）和探测方法，通过测量不同人群的“最大偏斜”（Max Skew）或与负面概念的关联率来量化偏见 34。这些研究证实，CLIP的嵌入空间中编码了与性别、种族、年龄相关的显著偏见 34。
* **规模与偏见的复杂关系**：模型/数据规模与偏见之间的关系是复杂的、非线性的。“更大”并不总是意味着“更公平”。例如，一项研究发现，扩大编码器尺寸**减少**了OpenAI版CLIP的性别偏见，但却**加剧**了OpenCLIP的性别和种族偏见。将LAION数据集从4亿扩展到20亿，使得OpenCLIP的性别偏见翻了一番 34。
* **下游模型的偏见放大**：CLIP的偏见并不会止于其自身。使用CLIP文本编码器的下游生成模型（如Stable Diffusion）会继承甚至放大这些偏见，在生成的图像中复现和强化刻板印象 42。
* **去偏见技术概览**：
  + **数据中心方法**：通过对数据进行平衡或解耦敏感属性与代理变量，可以减轻表征偏见，但可能对关联偏见产生负面影响 43。
  + **模型中心（后处理）方法**：这类方法在冻结的模型嵌入上应用线性投影或其他变换，以“擦除”与偏见相关的成分。这类方法计算成本低，如FairerCLIP 90。
  + **模型中心（微调）方法**：在平衡的数据集上微调模型是有效的，但面临灾难性遗忘的风险 43。
  + **提示工程方法**：使用去偏见的提示前缀或对抗性提示可以在一定程度上减轻偏见，但无法根除 34。
  + **纯文本去偏见**：TOD框架提出了一种新颖的思路，它利用LLM生成一个平衡的**文本**数据集，然后仅对文本编码器进行提示调优，完全避免了对图像数据的需求 92。

#### **7.3 故障排除与常见问题：实践指南**

在实际应用中，用户可能会遇到一系列问题。

* **DINOv2**：根据GitHub上的讨论，常见问题包括：在自定义数据集上复现深度估计或分割结果时遇到困难；在单机或自定义分辨率下训练时出现配置问题；以及由于Python路径设置不当导致的ModuleNotFoundError: No module named 'dinov2'错误 93。此外，关于  
  xFormers库不可用的警告也很常见，但这通常不影响模型的核心功能 93。
* **CLIP/OpenCLIP**：最突出的问题是在微调过程中遭遇灾难性遗忘，即零样本分类准确率在第一个训练周期后就暴跌至接近于零 40。这通常是由于模型在新数据分布上训练不稳定造成的，可以通过增大批次大小、使用梯度累积或采用专门的微调策略（如WiSE-FT, LoRA）来缓解 40。此外，用户也常常对对称交叉熵损失函数的工作原理感到困惑 32。

#### **7.4 结论与未来研究方向**

##### **7.4.1 综合回顾**

从监督学习的瓶颈出发，我们见证了计算机视觉领域向两大基础模型范式的演进。DINO代表了通过自监督学习探索视觉世界内在结构的道路，其产出的特征在空间感知和密集预测任务上无与伦比。CLIP则代表了借助语言这一强大的人类符号系统来组织视觉概念的道路，其零样本泛化能力开启了开放词汇识别的新纪元。它们之间的核心权衡——空间精度 vs. 语义灵活性——是各自学习目标不可避免的产物。

##### **7.4.2 混合的未来**

展望未来，最激动人心的方向无疑是这两种范式的协同与融合。如第六部分所述，研究界正在积极地将DINO的精细视觉基础与CLIP的灵活语义接口结合起来。无论是通过知识蒸馏、特征融合还是跨模型监督，目标都是创造一个集两者之长的统一视觉基础模型。这样的模型将能够同时理解“一个物体是什么”以及“它在空间中的具体形态”，为更高级的视觉任务奠定基础。

##### **7.4.3 开放的挑战**

尽管取得了巨大进展，但通往通用视觉智能的道路上仍有许多开放的挑战：

* **真正的抽象推理**：当前的模型主要停留在识别和描述层面。实现更高级的推理，如稳健地计数物体数量、理解复杂的因果关系和物理交互，仍然是一个遥远的目标。
* **设计内置的公平性**：与其依赖后处理的去偏见方法，未来的研究需要探索如何在模型设计和训练阶段就内置公平性约束，例如通过因果学习或构建真正无偏的预训练数据集。
* **数据与计算效率**：开发出能够用更少的数据和计算资源达到同等甚至更高性能的方法，是实现基础模型普惠化、降低其环境成本的关键。
* **超越静态图像**：将这些强大的表征学习技术更有效地扩展到视频、3D和交互式环境等更复杂的动态领域，是具身智能发展的核心。

DINO和CLIP不仅是两个成功的模型，它们更是两种思想的载体，为我们探索机器如何感知世界提供了两条截然不同但又最终交汇的路径。它们的演进与融合，将继续定义未来十年计算机视觉的面貌。

#### 引用的著作

1. Emerging Properties in Self-Supervised Vision Transformers (DINO) — Paper Summary | by Anuj Dutt | Medium, 访问时间为 六月 26, 2025， <https://medium.com/@anuj.dutt9/emerging-properties-in-self-supervised-vision-transformers-dino-paper-summary-4c7a6ed68161>
2. A Survey of the Self Supervised Learning Mechanisms for Vision Transformers, 访问时间为 六月 26, 2025， <https://www.researchgate.net/publication/383648357_A_Survey_of_the_Self_Supervised_Learning_Mechanisms_for_Vision_Transformers>
3. A Survey on Contrastive Self-Supervised Learning - MDPI, 访问时间为 六月 26, 2025， <https://www.mdpi.com/2227-7080/9/1/2>
4. Learning Transferable Visual Models From Natural Language Supervision - arXiv, 访问时间为 六月 26, 2025， <http://arxiv.org/pdf/2103.00020>
5. Learning Transferable Visual Models From Natural Language Supervision, 访问时间为 六月 26, 2025， <https://proceedings.mlr.press/v139/radford21a/radford21a.pdf>
6. A Survey of the Self Supervised Learning Mechanisms for Vision Transformers - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2408.17059v1>
7. The Guide to Foundation Models in Modern AI Development - Viso Suite, 访问时间为 六月 26, 2025， <https://viso.ai/deep-learning/foundation-models/>
8. DINOv2: Learning Robust Visual Features without Supervision | Request PDF, 访问时间为 六月 26, 2025， <https://www.researchgate.net/publication/370058767_DINOv2_Learning_Robust_Visual_Features_without_Supervision>
9. Exploring Intrinsic Properties of Medical Images for Self-Supervised Binary Semantic Segmentation - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2402.02367v2>
10. DINOv2: Learning Robust Visual Features without Supervision - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2304.07193v2>
11. DINOv2: Learning Robust Visual Features without Supervision - SciSpace, 访问时间为 六月 26, 2025， <https://scispace.com/papers/dinov2-learning-robust-visual-features-without-supervision-36k8f7am>
12. Foundation Models Defining a New Era in Vision: A Survey and Outlook, 访问时间为 六月 26, 2025， <https://www.computer.org/csdl/journal/tp/2025/04/10834497/23mYUeDuDja>
13. How Have Foundation Models Redefined Computer Vision Using AI? - Encord, 访问时间为 六月 26, 2025， <https://encord.com/blog/foundation-models-redefining-computer-vision/>
14. Foundation Models - NVIDIA Docs Hub, 访问时间为 六月 26, 2025， <https://docs.nvidia.com/tao/tao-toolkit/text/foundation_models/overview.html>
15. [2104.14294] Emerging Properties in Self-Supervised Vision ..., 访问时间为 六月 26, 2025， <https://ar5iv.labs.arxiv.org/html/2104.14294>
16. DINO and PAWS: Advancing the state of the art in computer vision - Meta AI, 访问时间为 六月 26, 2025， <https://ai.meta.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training/>
17. DINOv2: State-of-the-art computer vision models with self-supervised learning - Meta AI, 访问时间为 六月 26, 2025， <https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/>
18. DSU-Net：An Improved U-Net Model Based on DINOv2 and SAM2 with Multi-scale Cross-model Feature Enhancement - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/pdf/2503.21187>
19. CLIP: Learning Transferable Visual Models From Natural Language Supervision | by David Cochard | axinc-ai | Medium, 访问时间为 六月 26, 2025， <https://medium.com/axinc-ai/clip-learning-transferable-visual-models-from-natural-language-supervision-4508b3f0ea46>
20. Emerging Properties in Self-Supervised Vision Transformers - CVF Open Access, 访问时间为 六月 26, 2025， <https://openaccess.thecvf.com/content/ICCV2021/papers/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.pdf>
21. [2104.14294] Emerging Properties in Self-Supervised Vision Transformers - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/abs/2104.14294>
22. Liveness Detection in Computer Vision: Transformer-based Self-Supervised Learning for Face Anti-Spoofing - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2406.13860v1>
23. PyTorch code and models for the DINOv2 self-supervised learning method. - GitHub, 访问时间为 六月 26, 2025， <https://github.com/facebookresearch/dinov2>
24. DINOv2: Learning Robust Visual Features without Supervision, 访问时间为 六月 26, 2025， <https://arxiv.org/abs/2304.07193>
25. DINOv2 - Steps by steps explanations - Picsellia, 访问时间为 六月 26, 2025， <https://www.picsellia.com/post/dinov2-steps-by-steps-explanations-picsellia>
26. DINOv2: Learning Robust Visual Features without Supervision - OpenReview, 访问时间为 六月 26, 2025， <https://openreview.net/forum?id=a68SUt6zFt>
27. Understanding DINOv2: Engineer's Deep Dive - Lightly.ai, 访问时间为 六月 26, 2025， <https://www.lightly.ai/blog/dinov2>
28. DINOv2 - Hugging Face, 访问时间为 六月 26, 2025， <https://huggingface.co/docs/transformers/model_doc/dinov2>
29. CLIP : Learning Transferable Visual Models From Natural Language Supervision - MYRIAD, 访问时间为 六月 26, 2025， <https://creatis-myriad.github.io/2023/06/05/CLIP.html>
30. Learning Transferable Visual Models From Natural ... - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/abs/2103.00020>
31. CLIP Model and The Importance of Multimodal Embeddings | Towards Data Science, 访问时间为 六月 26, 2025， <https://towardsdatascience.com/clip-model-and-the-importance-of-multimodal-embeddings-1c8f6b13bf72/>
32. [D] Understanding the loss in the CLIP model : r/MachineLearning - Reddit, 访问时间为 六月 26, 2025， <https://www.reddit.com/r/MachineLearning/comments/1g42hnu/d_understanding_the_loss_in_the_clip_model/>
33. CLIP: Connecting text and images - OpenAI, 访问时间为 六月 26, 2025， <https://openai.com/index/clip/>
34. Size ≠ Fairness: A Comprehensive Social Bias Audit of Contrastive Vision–Language Models - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2501.13223v3>
35. ov-seg/open\_clip\_training/README.md at main - GitHub, 访问时间为 六月 26, 2025， <https://github.com/facebookresearch/ov-seg/blob/main/open_clip_training/README.md>
36. mlfoundations/open\_clip: An open source implementation of CLIP. - GitHub, 访问时间为 六月 26, 2025， <https://github.com/mlfoundations/open_clip>
37. robgon-art/open-clip: Test out OpenCLIP for Image Search and Automatic Captioning - GitHub, 访问时间为 六月 26, 2025， <https://github.com/robgon-art/open-clip>
38. A beginner's guide to fine-tuning the CLIP model for your downstream tasks using OpenClip · mlfoundations open\_clip · Discussion #812 - GitHub, 访问时间为 六月 26, 2025， <https://github.com/mlfoundations/open_clip/discussions/812>
39. Emad on Twitter: Happy to announce the release of new state of the art open CLIP models to drive image classification and generation forward : r/StableDiffusion - Reddit, 访问时间为 六月 26, 2025， <https://www.reddit.com/r/StableDiffusion/comments/xf6wqf/emad_on_twitter_happy_to_announce_the_release_of/>
40. [D] Fine Tuning Open CLIP model causes it zero shot accuracy to drastically drop after 1st epoch : r/MachineLearning - Reddit, 访问时间为 六月 26, 2025， <https://www.reddit.com/r/MachineLearning/comments/193gug2/d_fine_tuning_open_clip_model_causes_it_zero_shot/>
41. openclip · GitHub Topics, 访问时间为 六月 26, 2025， <https://github.com/topics/openclip?o=asc&s=updated>
42. arXiv:2503.23398v1 [cs.CV] 30 Mar 2025, 访问时间为 六月 26, 2025， <https://arxiv.org/pdf/2503.23398>
43. CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2403.04547v1>
44. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning | Montreal AI Ethics Institute, 访问时间为 六月 26, 2025， <https://montrealethics.ai/a-prompt-array-keeps-the-bias-away-debiasing-vision-language-models-with-adversarial-learning/>
45. DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2403.14548v1>
46. FRAME: Pre-Training Video Feature Representations via Anticipation and Memory - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2506.05543v1>
47. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2312.12359v2>
48. CLIP-DINOiser: Teaching CLIP a few DINO tricks - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2312.12359v1>
49. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation - European Computer Vision Association, 访问时间为 六月 26, 2025， <https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/07850.pdf>
50. From CLIP to DINO: Visual Encoders Shout in Multi-modal Large Language Models - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2310.08825v3>
51. What makes CLIP or any other vision model better than regular model? : r/LangChain - Reddit, 访问时间为 六月 26, 2025， <https://www.reddit.com/r/LangChain/comments/1hy0mma/what_makes_clip_or_any_other_vision_model_better/>
52. How CLIP Works: The Future of Multimodal Search and Reasoning - GoCodeo, 访问时间为 六月 26, 2025， <https://www.gocodeo.com/post/how-clip-works-the-future-of-multimodal-search-and-reasoning>
53. DINOv2 Meets Text: A Unified Framework for Image- and Pixel-Level Vision-Language Alignment - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2412.16334v1>
54. Extra Paper - Using of DINO Instead of CLIP | PDF | Data Compression - Scribd, 访问时间为 六月 26, 2025， <https://www.scribd.com/document/701828790/Extra-Paper-Using-of-DINO-Instead-of-CLIP>
55. DINOv2-powered Few-Shot Semantic Segmentation: A Unified Framework via Cross-Model Distillation and 4D Correlation Mining - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2504.15669v1>
56. SAM-aware Test-time Adaptation for Universal Medical Image Segmentation - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2506.05221v1>
57. Dexterous Functional Grasping - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2312.02975v1>
58. Spatial RoboGrasp: Generalized Robotic Grasping Control Policy - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2505.20814v1>
59. Bring Your Own Grasp Generator: Leveraging Robot Grasp Generation for Prosthetic Grasping - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2503.00466v1>
60. DexGraspVLA: A Vision-Language-Action Framework Towards General Dexterous Grasping, 访问时间为 六月 26, 2025， <https://dexgraspvla.github.io/>
61. Triplane Grasping: Efficient 6-DoF Grasping with Single RGB Images - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2410.15879v2>
62. Open-Vocabulary Object Detection via Dynamic Self-Training - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2310.01393v2>
63. [2303.13076] CORA: Adapting CLIP for Open-Vocabulary Detection with Region Prompting and Anchor Pre-Matching - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/abs/2303.13076>
64. CORA: Adapting CLIP for Open-Vocabulary Detection With Region Prompting and Anchor Pre-Matching, 访问时间为 六月 26, 2025， <https://openaccess.thecvf.com/content/CVPR2023/papers/Wu_CORA_Adapting_CLIP_for_Open-Vocabulary_Detection_With_Region_Prompting_and_CVPR_2023_paper.pdf>
65. DetCLIPv3: Towards Versatile Generative Open-vocabulary Object Detection, 访问时间为 六月 26, 2025， <https://openaccess.thecvf.com/content/CVPR2024/papers/Yao_DetCLIPv3_Towards_Versatile_Generative_Open-vocabulary_Object_Detection_CVPR_2024_paper.pdf>
66. Boosting Open-Vocabulary Object Detection by Handling Background Samples - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2410.08645v1>
67. Text-to-image Diffusion Models in Generative AI: A Survey - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2303.07909v3>
68. Auditing Internal Bias Dynamics of Text-To-Image Generative Models - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2410.07884v1>
69. Robust CLIP-Based Detector for Exposing Diffusion Model-Generated Images - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/pdf/2404.12908>
70. (PDF) Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2 - ResearchGate, 访问时间为 六月 26, 2025， <https://www.researchgate.net/publication/364126542_Generated_Faces_in_the_Wild_Quantitative_Comparison_of_Stable_Diffusion_Midjourney_and_DALL-E_2>
71. DALL·E: Creating images from text - OpenAI, 访问时间为 六月 26, 2025， [https://openai.com/blog/dall-e/?utm\_campaign=The%20Batch&utm\_source=hs\_email&utm\_medium=email&\_hsenc=p2ANqtz-9OVvx5ZsqALrG5nfEUWweka\_xnjrdeHLfaULPO9dmLR7vEXiJWwmkKf51Yn-\_iYsOgIjbv](https://openai.com/blog/dall-e/?utm_campaign=The+Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-9OVvx5ZsqALrG5nfEUWweka_xnjrdeHLfaULPO9dmLR7vEXiJWwmkKf51Yn-_iYsOgIjbv)
72. Comparison of FRAME and DINO on feature propagation across video... - ResearchGate, 访问时间为 六月 26, 2025， <https://www.researchgate.net/figure/Comparison-of-FRAME-and-DINO-on-feature-propagation-across-video-frames-FRAME_fig2_392515258>
73. Trajectory-aligned Space-time Tokens for Few-shot Action Recognition - European Computer Vision Association, 访问时间为 六月 26, 2025， <https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05362.pdf>
74. Computer Vision Pipeline v2.0. How Foundation Models are transforming… | by The Tenyks Blogger | Medium, 访问时间为 六月 26, 2025， <https://medium.com/@tenyks_blogger/computer-vision-is-already-evolving-3cd0e63e805b>
75. Launch: Model Prompting for Automated Labeling with Autodistill - Roboflow Blog, 访问时间为 六月 26, 2025， <https://blog.roboflow.com/automated-image-labeling/>
76. Use Cases for Computer Vision Foundation Models - Roboflow Blog, 访问时间为 六月 26, 2025， <https://blog.roboflow.com/use-cases-for-computer-vision-foundation-models/>
77. Vector Databases: The Engine Powering Modern AI Applications - Dotzlaw, 访问时间为 六月 26, 2025， <https://dotzlaw.com/articles/vector-databases-the-engine-powering-modern-ai-applications/>
78. What are the advantages of using CLIP for multimodal search? - Milvus, 访问时间为 六月 26, 2025， <https://milvus.io/ai-quick-reference/what-are-the-advantages-of-using-clip-for-multimodal-search>
79. Vector Embeddings Explained - Weaviate, 访问时间为 六月 26, 2025， <https://weaviate.io/blog/vector-embeddings-explained>
80. Building Large-Scale Image Search using VectorDB & OpenAI CLIP: From 120 Hours to 1 Hour, From $$$ to $ | SkyPilot Blog, 访问时间为 六月 26, 2025， <https://blog.skypilot.co/large-scale-vector-database/>
81. Build a multimodal image retrieval system using KNN search and CLIP embeddings - Elastic, 访问时间为 六月 26, 2025， <https://www.elastic.co/search-labs/blog/multimodal-image-retrieval-with-roboflow>
82. Talking to DINO: Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2411.19331v1>
83. (PDF) CLIP meets DINO for Tuning Zero-Shot Classifier using Unlabeled Image Collections, 访问时间为 六月 26, 2025， <https://www.researchgate.net/publication/386336292_CLIP_meets_DINO_for_Tuning_Zero-Shot_Classifier_using_Unlabeled_Image_Collections>
84. CLIP meets DINO for Tuning Zero-Shot Classifier using Unlabeled Image Collections - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2411.19346v1>
85. CLIP meets DINO for Tuning Zero-Shot Classifier using Unlabeled Image Collections - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/abs/2411.19346>
86. Fixing a Bug in DINOv2 for Vision Transformers - Toolify.ai, 访问时间为 六月 26, 2025， <https://www.toolify.ai/gpts/fixing-a-bug-in-dinov2-for-vision-transformers-314431>
87. Social perception of faces in a vision-language model - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2408.14435v1>
88. A Unified Debiasing Approach for Vision-Language Models across Modalities and Tasks - NIPS, 访问时间为 六月 26, 2025， <https://proceedings.neurips.cc/paper_files/paper/2024/file/254404d551f6ce17bb7407b4d6b3c87b-Paper-Conference.pdf>
89. A Large Scale Analysis of Gender Biases in Text-to-Image Generative Models - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/html/2503.23398v1>
90. FairerCLIP: Debiasing CLIP's Zero-Shot Predictions using Functions in RKHSs, 访问时间为 六月 26, 2025， <https://openreview.net/forum?id=HXoq9EqR9e>
91. Debiasing Vision-Language Models via Biased Prompts - OpenReview, 访问时间为 六月 26, 2025， <https://openreview.net/forum?id=Gu1t2ar96S>
92. [2410.09365] Debiasing Vison-Language Models with Text-Only Training - arXiv, 访问时间为 六月 26, 2025， <https://arxiv.org/abs/2410.09365>
93. Issues · facebookresearch/dinov2 - GitHub, 访问时间为 六月 26, 2025， <https://github.com/facebookresearch/dinov2/issues>
94. No module named 'dinov2 · Issue #12 - GitHub, 访问时间为 六月 26, 2025， <https://github.com/facebookresearch/dinov2/issues/12>