arXiv:2304.03752v2 [cs.CV] 5 Oct 2023

V3Det：大词汇量视觉检测数据集

Jiaqi Wang*[1]，Pan Zhang*[1]，Tao Chu*[1]，YuhangCao*[2]，YujieZhou[1]，吴彤[2]，BinWang[1]，何聪慧[1]，DahuaLin[1]，[2]，[3]✉[1]上海人工智能实验室，[2]香港中文大学，[3]感知与交互智能中心{wangjiaqi,zhangpan}@pjlab.org.cn, dhlin@ie.cuhk.edu.hk



图1：**V3Det** 是一个**超大词汇量视觉检测**数据集，包含精确标注的 13,204 个类别的边界框，其多样性显著高于现有视觉检测数据集，例如 COCO [38], Objects365 [49], OpenImages [32], 和 LVIS [25]。图中展示了 V3Det 的层次化类别树、各类别的目标分布以及标注图像样本。

摘要

近期，在真实世界中检测任意目标的研究，通常在词汇相对受限的目标检测数据集上进行训练和评估。为了促进更通用的视觉目标检测发展，我们提出了 ***V3Det***，这是一个大规模词汇视觉检测数据集，包含海量图像中经过精确标注的边界框。*V3Det* 具有以下几项引人注目的特性：***1)*** 大规模词汇：它包含真实世界图像中来自 *13,204* 个类别的目标边界框，其规模是现有大规模词汇目标检测数据集（例如*LVIS*）的 *10* 倍。***2)*** 层次化类别组织：*V3Det* 的大规模词汇通过层次化类别树进行组织，该类别树标注了类别之间的包含关系，有助于探索大规模开放词汇目标检中的类别关系。目标检测。***3)*** 丰富的标注：*V3Det* 包含了 *24.3*万张图像中的精确标注对象，以及由人类专家和强大的聊天机器人为每个类别撰写的专业描述。通过提供广阔的探索空间，*V3Det* 支持在广泛词汇和开放词汇目标检测方面开展大量基准测试，从而为未来研究带来新的观察、实践和洞见。它有潜力成为开发更通用视觉感知系统的基石数据集。*V3Det* 可通过以下网址获取 *https://v3det.openxlab.org.cn/*。

1. 引言

目标检测 [23, 22, 47, 3, 39, 37, 5, 51, 56, 54, 57,55, 66] 是各种现实世界应用的基石，例如自动驾驶、机器人技术和增强现实。它以图像作为输入，在给定词汇表中定位并分类对象，其中每个检测到的对象都是

* 同等贡献。

表示为带有类别标签的边界框。

检测任意物体一直是计算机视觉领域长期以来的目标。由于现实世界中的物体具有极大的多样性，分布在无数类别中，理想的视觉检测系统应能够检测来自极大类别集合的物体，并且能够直接应用于开放词汇类别。因此，为该领域配备一个具有大规模词汇的目标检测数据集至关重要，因为这能够加速对更加通用的视觉检测系统的探索。

尽管人们为创建目标检测数据集付出了大量努力，但这些努力仅部分满足了构建一个类别词汇极其丰富的综合数据集的需求。例如，如表 1所示，COCO [38]是应用最广泛的数据集之一，包含 12.3 万张图像和 80 个类别。Objects365 [49]和 Open Images [32] 是旨在用于预训练的大规模数据集，其中 Objects365 标注了 63.8万张图像，包含 365个类别，而 OpenImages 包含 151.5万张图像和 600 个类别。尽管这两个数据集提供了大量经过标注的对象，但其有限的类别数量仍不足以训练具有类别泛化能力的检测器。LVIS [25]是一个包含 12 万张图像和1,203 个类别的大词汇目标检测和实例分割 [28, 6, 59, 52, 10, 63, 64, 6] 数据集。尽管 LVIS 在大词汇量/开放词汇目标检测的近期进展中取得了成功，但LVIS的词汇量（1,203个类别）仍不足以表示现实世界中类别的巨大多样性。因此，具有显著更大词汇量的新目标检测数据集将有利于视觉检测器的训练和评估。

为了促进未来对更通用目标检测的研究，我们提出了**V3Det**：**大**规模 **词**汇 **视**觉检测数据集，其中包含真实世界图像中经过精确标注的大量词汇对象。所提出的数据集具有以下几项突出特性：**1）大规模词汇：**它包含真实世界图像中来自 13,204 个类别的对象边界框，其规模是现有大规模词汇目标检测数据集的10倍，例如LVIS中的1,203个类别。**2）层次化类别组织：**V3Det的大规模词汇通过层次化类别树进行组织，该类别树标注了类别之间的包含关系。如图1所示，这棵树以“实体”类别为起点，它是所有类别的祖先，包含已标注对象的类别及其父类别。这种层次化树丰富了V3Det的词汇，使得探索开放世界识别中的类别关系成为可能。**3）丰富的标注：** V3Det包含24.3万张图像中经过精确标注的对象。大量图像集合确保了对象和场景的多样性。除类别名称外，V3Det还提供了由人类专家和功能强大的聊天机器人撰写的每个类别的丰富描述，

数据集	类别	图像	边界框	类别树	描述
PASCAL VOC [17]	20	11.5k	27k	No	No
COCO [8]	80	123k	896k	No	No
对象365 [49]	365	638k	10,101k	No	No
Open Images [32]	600	1,515k	14,815k	Yes	No
LVIS [25]	1,203	120k	1,525k	No	No
ELEVATER [34]	314	132k	938k	No	Yes
BigDetection [2]	600	3,480k	35,960k	No	No
V3Det	**13,029**	215k	1,540k	Yes	Yes
V3Det（全部）	**13,029**	245k	1,753k	Yes	Yes

表 1：V3Det 与现有目标检测基准的数据集统计对比，基于已发布的标注，包括训练数据和验证数据。V3Det（全部）包含 V3Det 的所有训练、验证和测试数据。

即，ChatGPT，它可以作为开放词汇目标检测的语言提示。

我们在 V3Det 上进行了大量基准测试，以深入了解这一大词汇量目标检测数据集。具体而言，我们针对经典的封闭集目标检测和开放词汇目标检测开展了大量实验。1）在封闭集目标检测中，我们评估了各种广泛采用的目标检测框架，并介绍了数据采样器、优化器和分类器设计方面的最佳实践，以提升 V3Det 上的性能。2）在开放词汇设置下，我们在 V3Det 上对近期具有代表性的方法进行了基准测试，揭示了在大量类别中开展开放词汇目标检测所面临的挑战。此外，我们还表明，在 V3Det 上进行预训练能够显著增强检测器的类别泛化能力，证明了大规模词汇数据集对开放词汇算法具有切实益处。

总而言之，我们提出了 **V3Det**：**海**量词汇**视**觉检测数据集，包含13,204个类别，并通过类别树进行层次化组织。据我们所知，V3Det 是首个类别数量超过一万的目标检测数据集。243k张图像中的丰富检测边界框标注，以及专业的类别描述，确保了该数据集的多样性和通用性。借助 V3Det 所提供的广阔探索空间，大量基准测试揭示了关于海量和开放词汇目标检测的新观察、新实践和新见解，为未来研究提供参考。它有望成为开发更通用视觉感知系统的基石数据集。

2. 相关工作

视觉检测数据集。 大规模且类别广泛的数据集是构建稳健视觉检测系统的关键，社区为此付出了诸多努力。早期的 PASCAL VOC [17] 数据集包含 1.15 万张图像，涵盖 20 个类别。MS COCO [38] 数据集引入了 12.3 万张图像，涵盖 80 个类别，用于描绘复杂的日常场景

中的常见物体。物体365 [49], Open Images [32],和 BigDetection [2] 是用于预训练的大规模数据集，其中规模最大的 BigDetection 包含 348 万张图像，涵盖 600 个类别。ELEVATER [34] 是一个包含 35 个数据集的多领域基准。尽管这些数据集提供了大量带标注的物体，但类别数量有限，仍不足以训练出具有类别泛化能力的检测器。LVIS [25]是迄今词汇量最大的检测数据集，包含 1,203 个类别，并呈现长尾分布。我们的 V3Det 数据集追求更广泛的类别分布，涵盖 13,204 个类别，从而能够训练出更通用的检测器。

目标检测。 目标检测旨在定位并分类物体，近年来取得了显著进展。经典的基于 CNN 的检测器可分为两类。两阶段检测器[23, 22, 47, 3,71] 首先预测一组候选区域，然后在第二阶段对其进行细化。单阶段检测器 [37, 46, 39], 直接对预定义的锚框进行分类和回归，或通过点 [53], 、中心[16], 、角点 [33]等密集搜索几何线索。近期基于 Transformer 的检测器 [5,72, 62] 去除了非极大值抑制和锚等人工设计的组件，性能优于基于 CNN 的检测器。在这项工作中，我们在 V3Det 数据集上测试了不同范式的多种方法。

开放词汇目标检测。 开放词汇目标检测（OVD）旨在检测训练期间未见过的新类。为克服传统检测器只能检测预定义类别这一局限，OVD 方法通常采用视觉语言模型 [44, 13] ，或利用带有弱标注的外部数据 [48, 8, 50]。具体而言，OV-RCNN [61] 使用 BERT[13]在图像-文本对语料上预训练 Faster R-CNN。ViLD [24] 将 CLIP [44]中的通用知识蒸馏到两阶段检测器中。GLIP [35]和 MDETR [29] 将检测转化为基于文本描述的定位任务。Region-CLIP [69] 增强了 CLIP，使其能更好地匹配图像区域与文本概念。Detic [70] 利用来自 ImageNet-21K [48]的图像级标注，将分类能力扩展至两万个类别。尽管这些工作在两个基准测试——MS COCO [38],和 LVIS [25], ——上取得了出色的表现，但我们的实验发现，它们很难应用于类别数量庞大的场景。

3. V3Det 数据集

本节全面介绍 V3Det 数据集，涵盖其数据获取、标注流程和数据集分析。

3.1. 数据采集

V3Det 数据集的构建基于 Bamboo [67] 分类数据集，其中 V3Det 数据集使用的大多数类别和图像均来源于此。

Bamboo 分类数据集。广泛的Bamboo [67]图像分类数据集包含6,900 万条标注，涵盖119,000 个视觉类别，为 V3Det 数据集提供了丰富的潜在图像和类别来源。尽管 Bamboo 分类数据集是一个超大规模数据集，但其中并非所有类别都适用于检测任务，例如“晴天”和“幸福”。此外，许多分类图像对于检测任务而言不够复杂。为了解决这些问题，我们设计了类别构建和图像选择流程，以获取适用于 V3Det 数据集的类别和图像。

类别构建。通过仔细检查现有的 119,000个视觉类别，三位在目标检测方面拥有丰富经验的研究人员手动筛选出适用于目标检测的视觉概念。每个概念都配有一个类别名称、一组描述和分类图像。描述来源于人类专家（例如 WordNet、Wikidata）或由 ChatGPT 生成。所有研究人员一致选定的概念将被纳入 V3Det 数据集，最终形成11,922 个类别。除从Bamboo 分类数据集中进行选择外，我们还从网络数据中精心收集了 1,282 个目标检测类别。因此，V3Det 数据集共包含 13,204 个类别。

图像筛选。 对于从Bam- boo分类数据集中获得的类别，我们采用其对应的图像。对于从网络数据中收集的类别，我们从Flickr中为每个类别抓取120张图像 [1] 并手动移除任何不匹配的图像。为筛选简单分类图像，我们使用一个区域候选网络（RPN） [47] ，该网络在四个检测数据集上联合训练，即COCO [38], LVIS [25], Objects365 [49], 和Open- Images [32], ，以提取每张图像的候选区域。候选区域数量超过4,000的图像将被移除，以避免数据集过度拥挤。每个类别中剩余的20张候选区域数量最多的图像将保留在V3Det 数据集中。

3.2. 标注

V3Det 的整体标注流程遵循由粗到细策略，如图2所示。在粗粒度标注和联合验证阶段，使用边界框标注属于粗粒度类别的对象和细粒度图像分类标签。在细粒度标注和类别内验证阶段，进一步将每个边界框的粗粒度类别细化为相应的细粒度类别。

标注团队。 为提高标注的准确性，我们建立了四个标注团队。一个大型标注员团队负责整个标注流程中的所有标注步骤。两个检查员团队负责审核

[1] https://www.flickr.com



图 2：**我们的标注流程示意图**。a) 每个小组负责标注一个粗粒度类别，以及属于该粗粒度类别的图像级类别。b) 汇总所有小组对同一图像的标注，并一并呈现以进行联合验证。c) 对于每个带有粗粒度标注的框，会向标注人员提供一些建议，以便进行细粒度标注。d) 将标注为相同细粒度类别的对象一并呈现，以进行类别内验证。

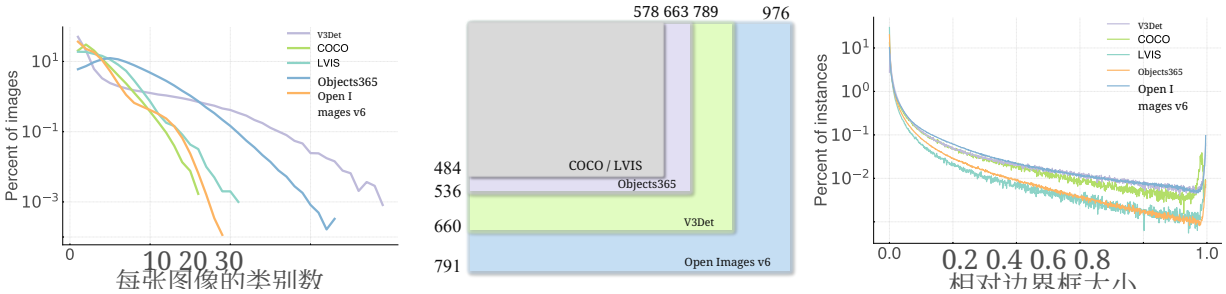
所有已标注的图像。检查团队 I 负责审查标注员团队完成的所有标注。一旦在一个数据包（包含 100 张图像）中发现标注错误，该数据包将被退回重新标注。经检查团队 I 批准的数据包将由检查团队 II 进一步审查。经两个检查团队成功批准后，一个审查团队将对数据包进行最终检查。任何标注错误都会导致数据包被退回重新标注。所述的整个流程将在标注流程的每个阶段执行。

粗粒度标注。 让每位标注者记住并标注全部 13,204个类别是不现实的。因此，我们首先将13,204个类别划分为 53个粗粒度类别，例如“动物和人类”“车辆”和“设备”。随后，对标注者进行培训，使其理解单个粗粒度类别并记住相应的子类别。由于每个子类别集合中的外观较为相似，标注者更容易完成上述任务。为避免错误标注，我们将标注者分为53 组，每组负责在整个数据集中标注对应于单个粗粒度类别的边界框。对于每个小组而言，大量图像可能不包含与所分配粗粒度类别相对应的任何对象，从而导致大量图像被跳过且没有标注。但为了确保标注准确性，进行广泛的标注是必要的。

在标注粗粒度边界框时，标注-者还会识别并标注图像级标签的对象，前提是该对象属于其负责的粗粒度类别。如图2 a）所示，粗粒度组1的标注者（‘标-imal 和 person’）也会标注图像标签（‘boy’）的边界框。图像级分类标签来自 Bamboo 分类数据集，或是在从互联网爬取图像后分配。粗粒度标注阶段完成后，所得数据集包含带有粗粒度边界框和图像级细粒度边界框标注的图像。

联合验证。 尽管“一个标注员对应一个粗粒度类别”的策略显著降低了标注的难度，但仍有一些需要解决的标注问题。其中一个是重复标注，另一个是标注缺失。由于数据集中的类别规模极其庞大，很难确保粗粒度类别彼此互斥。例如，“卡车”这一类别同时属于“车辆”和“重型机械”这两个粗粒度类别。因此，一名标注员可能会将卡车的边界框标注为“车辆”，而另一名标注员可能会将同一辆卡车的边界框标注为“重型机械”。与此同时，一些细粒度类别可能无法很好地归入某个粗粒度类别，这可能导致少数标注员遗漏标注。

为解决上述问题，我们提出了联合验证阶段。如图2 b)所示，每张图像中来自53个标注组的标注会被合并，并提交给在粗粒度标注阶段展现出卓越技能的熟练标注员。标注员负责更正重复或遗漏的标注。检查团队和审查团队会对这一流程进行复核和审查，以确保联合验证阶段的高质量。



a）每张图像的类别数量分布　　b）各数据集的平均分辨率　　c）相对边界框大小分布

图 3：**数据集分析**。a) V3Det 中包含超过 16 个类别的图像占比高于其他数据集。b) V3Det 的平均图像分辨率高于 COCO、LVIS 和 Objects365。c) V3Det 中的物体尺度与 COCO 和 Open Images v6 数据集中的相当。

细粒度标注。完成前述阶段后，数据集中的所有对象均已通过边界框标注为粗粒度类别或细粒度图像级类别。本阶段旨在将每个边界框的粗糙类别细化为其对应的细粒度类别。

为降低从粗糙到精细的标注任务的复杂性，我们利用带有图像级标签的边界框对细粒度类别进行排序，以辅助标注人员。具体而言，首先根据已标注的边界框裁剪图像块，并采用固定的 25 像素填充尺寸。然后，利用大型预训练视觉语言模型 即 CLIP（ViT-L） [44] 提取裁剪图像块的判别性特征。为了获得细粒度类别 i 的表示，我们将类别 i 的所有图像块特征通过 k-means [26] 聚类，得到 k 个原型 c_i^1、c_i^2、……、c_i^k。遵循 CLUE [9], , 粗粒度图像块 s_i^x 与细粒度类别 x 之间的相似度 i 定义如下：

$$s_i^x = \max_{j=1,2,\ldots,k} \frac{e(x) \cdot c_i^j}{\|e(x)\| \|c_i^j\|} \tag{1}$$

其中，$e(x)$ 是图像块 x的提取特征，k 被设置为 3。由于类别规模巨大，仅根据相似度分数 s_i^x对所有细粒度类别进行排序会产生次优推荐。在这里，我们利用粗粒度标注。具体而言，对于粗粒度类别 I，我们构建一个粗粒度类别集合 $\mathcal{I}_m$，其中的类别更可能被误标注为类别 I。对于一个标注为 I的给定裁剪图像块，我们构建一个遍历列表，其中包括属于粗粒度类别 I 的排名前 20的细粒度类别，以及其粗粒度类别属于 $\mathcal{I}_m$ 的排名前 10 的细粒度类别。

在细粒度标注阶段，标注人员需要将给定的裁剪图像块与遍历列表中的所有细粒度类别进行匹配。如果遍历列表中的类别均不正确，标注人员则需要根据自身知识，在所有细粒度类别中进行搜索。该搜索平台会根据搜索关键词提供相关类别作为建议。 如图2 c)所示，标注人员根据示例图像和类别描述选择相应的细粒度类别。

完成上述步骤后，仍有少量边界框未能成功匹配。为确保这些边界框的类别不属于这 13,204 个类别，我们从每位标注者标注的边界框中抽取少量样本，并进行全面检查以确认结果。若错误率超过 2%，则需重新标注。

类别内验证。 在细粒度标注阶段之后，13,204 个类别的大部分标注工作已完成。为了提高标注的准确性，我们设计了一个细粒度类别内验证阶段：将所有由边界框裁剪出的、属于同一细粒度类别的图像块并排展示，如图 2 d) 所示。标注人员需要借助示例图像和描述，验证给定细粒度类别中是否存在误分类对象或不一致的标注。误分类对象将在细粒度标注阶段重新标注，而不一致的标注则由标注人员直接更正，并遵循与 Objects365 [49]数据集相同的“最大边界框”规则。该阶段确保了我们的 V3Det 数据集（含 13,204 个类别）的准确性。

3.3. 数据集分析

通过所提出的标注流程，我们已对约 24.3万张图像进行了标注，涵盖 13,204 个类别和 170 万个实例。V3Det 数据集具有以下主要特征：

大规模词汇。 如表 1 所示，V3Det 数据集中的类别数量是以往检测数据集的十倍。丰富的类别有助于推动大规模词汇目标检测的发展。此外，由于真正的开放词汇目标检测器也是大规模词汇目标检测器，词汇量庞大的 V3Det 数据集能够提供更稳

健的基准，用于评估和支持开放词汇目标检测。图 3 a) 展示了完整的每图类别分布，其中 V3Det 中包含超过 16 个类别的图像占比高于其他数据集，这表明所提出的 V3Det 数据集具有复杂性和挑战性。

高分辨率。如图3 b) 所示，与 COCO、LVIS 和 Objects365 相比，V3Det 的图像分辨率更高，能提供更多视觉细节，有助于目标检测。图3 c) 展示了边界框的相对尺寸分布。V3Det 的物体尺度与 COCO 和 Open Images v6 的相近。

高质量。为验证 V3Det 数据集的标注质量，我们邀请了三位在目标检测领域经验丰富的研究人员，分别对从 V3Det 数据集中随机抽取的 100 张图像和从 LVIS 数据集中随机抽取的 100 张图像进行标注。结果表明，就类别特定召回率而言，V3Det 的表现与 LVIS 相当，分别达到 90.9% 和 91.2%。此外，在平均精度均值方面，V3Det 优于 LVIS，达到 86.5%，高于 LVIS 的 81.2%。

4. 基准测试

4.1. 实验设置

数据划分。 V3Det 分为三个子集，即，训练子集、验证子集和测试子集。训练子集包含 183,354 张图像中的 1,299,765 个目标，验证子集包含 29,821 张图像中的 214,416 个目标，测试子集包含 29,863 张图像中的 212,437 个目标。我们手动确保验证子集和测试子集中每个类别至少有一个目标。训练子集和验证子集的图像和标注将公开。我们将发布测试子集的图像，并搭建公共测试服务器，对上传的结果进行评估，以便公平比较。除非另有说明，报告的性能均是在训练子集上训练，并在验证子集上评估所得。我们对闭集大词汇量目标检测和开放词汇目标检测均进行了基准测试。1）在大词汇量设置下，我们使用全部 13,204 个类别进行训练和评估。2）在开放词汇设置下，我们随机抽取 6,709 个类别作为基础类别 C_{base}，其余 6,495 个类别作为新类 C_{novel}。

评估协议。 我们遵循 COCO [8] 数据集的评估指标，报告不同 IoU 阈值（即，0。5 ∼ 0.95）。

实现细节。 1) 在大规模词汇设置下，我们评估两阶段 / 级联检测器 [47, 3, 71],单阶段检测器 [37, 53, 65], 和 DETR 风格检测器 [72, 62]。我们采用其在 mmdetection [7], Detectron2 [60], 和 detrex [12]中的标准实现。所有模型均使用 AdamW [41]优化器、多尺度数据增强和重复因子采样器 [25] 进行训练，作为该.....

故障设置。2）在开放词汇目标检测中，我们遵循官方发布的代码和设置进行训练与推理。有关更多实现细节，请参阅补充材料。

4.2. 大词汇量设置下的实验

与以往的目标检测数据集相比，V3Det包含由13,204 个类别组成的大规模词汇。我们首先在封闭集大规模词汇目标检测上进行基准测试，以揭示有关 V3Det 的见解和最佳实践。

主要结果。我们使用两阶段、级联式、单阶段和 DETR 风格的框架对检测器进行评估。由于资源有限，我们只能测试一些具有代表性的检测器，但我们相信所选框架足以揭示重要的洞见。如表 2所示，我们报告了 Faster R-CNN [47],CenterNet2 [71], 和 Cascade R-CNN [4]在采用 ResNet-50 [27] 和 Swin- [40]骨干网络的两阶段/级联式框架中的性能。实验结果表明，级联式框架（例如 Cascade R-CNN、CenterNet2）显著优于两阶段检测器（例如Faster R-CNN），在ResNet-50上约高出 8%，在 Swin-B 上约高出 10%。单阶段检测器，例如 FCOS [53] 和 ATSS [65]在 V3Det 上的表现远逊于 Faster R-CNN 和 CascadeR-CNN。实验结果表明，在 V3Det 中，回归和分类的级联式细化至关重要。目前领先的 DETR 风格检测器，例如 Deformable DETR [72] 和 DINO [62], 在 V3Det 上取得了强劲的性能。例如，DINO [62]的性能高于 Cascade R-CNN [4]，在 ResNet-50 上具有较大优势（33.5%对 28.6%），在 Swin-B上也具有明显优势（42.0% 对 40.2%）。

此外，我们评估了归一化线性层 [58],，该层在大词汇量目标检测数据集 LVIS 上表现有效。归一化线性层显著提升了 FasterR-CNN、Cascade R-CNN和 FCOS 采用不同骨干网络时在 V3Det 数据集上的性能。**V3Det 数据集上的最佳实践**。 我们探究了不同数据采样器、优化器和分类器在 V3Det 上的有效性。我们以采用 ResNet-50 骨干网络、训练 12 个训练轮次的 Cascade R-CNN 作为基线。在表 3中，我们比较了随机采样器、类别平衡采样器 [43], 和重复因子采样器 [25]。我们观察到，重复因子采样器和类别平衡采样器的表现相当，略优于随机采样器（∼ 0.2%）。结果与 V3Det 数据集的类别分布相对均衡这一事实一致。在表4中，我们对 AdamW 优化器和归一化线性层进行了消融研究。AdamW 优化器和归一化线性层都显著提升了 V3Det 上的性能。因此，在表 2中，我们将 AdamW 优化器和重复因子采样器设为默认配置。**在 V3Det 数据集上评估大型基础模型**。

框架	方法	训练轮次	骨干网络	AP	AP_{50}	AP_{75}
两阶段 / 级联	Faster R-CNN [47]	24	ResNet-50 Swin-B	20.1 30.7	28.2 40.0	22.9 34.7
	使用归一化线性层的 Faster R-CNN [58]	24	ResNet-50 Swin-B	25.4 37.6	32.9 46.0	28.1 41.1
	CenterNet2 [71]	24	ResNet-50 Swin-B	28.3 39.8	33.7 46.1	30.4 42.4
	Cascade R-CNN [4]	24	ResNet-50 Swin-B	28.6 40.2	34.8 47.3	31.1 43.0
	使用归一化线性层的 Cascade R-CNN [58]	24	ResNet-50 Swin-B	31.6 42.5	37.3 49.1	33.7 44.9
单阶段	ATSS [6]	24	ResNet-50 Swin-B	4.4 7.6	5.3 8.9	4.6 8.0
	FCOS [53]	24	ResNet-50 Swin-B	6.5 15.0	8.1 17.9	6.9 16.0
	采用归一化线性层的 FCOS [58]	24	ResNet-50 Swin-B	9.4 21.0	11.7 24.8	10.1 22.3
DETR 风格	Deformable DETR [72]	50	ResNet-50 Swin-B	34.4 42.5	39.9 48.3	36.4 44.7
	DINO [62]	24	ResNet-50 Swin-B	33.5 42.0	37.7 46.8	35.0 43.9

表 2：我们报告了 V3Det 上闭集大词汇量目标检测设置的基准测试结果。本表采用 AdamW [41] 优化器和重复因子采样器 [25] 以获得更好的性能。我们报告了 AP、AP_{50} 和 AP_{75} 进行比较。

数据采样器	AP	AP_{50}	AP_{75}
随机采样器	25.7	31.7	28.1
类别均衡采样器 [43]	25.9	31.8	28.3
重复因子采样器 [25]	25.9	32.0	28.4

表 3：以采用 Cascade R-CNN ResNet-50 并训练 12 个训练轮次为基线，研究不同数据采样器的效果。本表采用 AdamW 优化器。

AdamW	归一化线性层	AP	AP_{50}	AP_{75}
		20.9	25.3	22.6
✓		25.9	31.8	28.3
	✓	27.3	33.8	29.8
✓	✓	28.3	34.5	30.8

表 4：以采用 Cascade R-CNN ResNet-50 并训练 12 个训练轮次为基线，研究 AdamW 优化器和归一化线性层。

数据集	AP	AP_{50}	AP_{75}
COCO [38]	64.2	81.9	70.6
LVIS [25]	62.2	76.2	65.4
V3Det	49.4	54.8	51.4

表 5：强基础模型（即 EVA [18] ）在 COCO、LVIS 和 V3Det 上的检测性能比较。

基础模型的最新进展极大地提升了现有目标检测基准的性能，例如 COCO 和 LVIS。以最近的基础模型EVA [18] 为例，其 ViT-G [14] 版本首先在 Objects365 [49] 数据集上进行预训练，然后在目标数据集上进行微调，在COCO上取得了 64.2% 的平均精度，在LVIS上取得了 62.2% 的平均精度。值得注意的是，由于其庞大的词汇量和长尾巴分布，LVIS 一直被认为是比 COCO 更具挑战性的数据集。然而，EVA 在两个数据集上都取得了相当的性能。我们探究这样一个强大的模型在 V3Det 上是否仍能取得如此出色的性能。如表 5所示，我们报告了 EVA 搭配ViT-G 在 COCO、LVIS和 V3Det 上的性能。为进行公平比较，实验中的所有训练设置和模型架构均保持一致。与 60+% 在COCO 和 LVIS 上的平均精度相比，V3Det 的性能仍有很大的提升空间。

验证集与测试集之间的性能差异。我们在**V3Det 的验证集和测试集划分上，使用 24 个训练轮次和 Swin-B 主干网络**评估 ***Cascade R-CNN*** [4] 和 DINO [62] 。测试集的表现仅略高于验证集，两种方法的差异分别为0.6% 和 0.4%，分别对应两种方法，表明这两个划分之间具有一致性。

4.3. 开放词汇设置下的实验。

检测任意类别的物体一直是目标检测的长期目标。近年来，开放词汇目标检测越来越受到关注，其旨在类别封闭集上训练后，检测任意类别的物体。在本节中，我们在 V3Det 上评估现有的开放词汇目标检测（OVD）工作，以更全面地衡量其类别泛化能力。

方法	框架	训练数据	AP bAP nAP
PBL [20]	MRCNN-R50	COCO,CCCap,VG,SBU	0.24 0.22 0.26
VL-PLM [68]	FRCNN-R50	COCO	0.10 0.08 0.19
PromptDet [19]	MRCNN-R50	LVIS,LAION	0.61 0.58 0.63
ViLD [15]	MRCNN-R50	LVIS	0.52 0.49 0.55
DetPro [15]	MRCNN-R50	LVIS	0.53 0.52 0.55
GLIP [35]	DyHead[11]-SWT	O365,GoldG,CC3M,SBU	- - 0.91
OC-OVD [45]	FRCNN-R50	LVIS,ImageNet	0.97 0.92 1.03
RegionCLIP [69]	FRCNN-R50	LVIS	0.72 0.71 0.72
Detic [70]	Centernet2-R50	LVIS, ImageNet	1.00 0.95 1.04

表 6：在 V3Det 上直接测试预训练开放词汇目标检测器的评估结果。同时报告了框架和训练数据。 AP、 bAP和 nAP分别表示 V3Det 上所有类别、基础类别 C_{base}和新类 C_{novel} 的性能。由于类别数量较大时推理成本较高，GLIP 仅在 nAP 上进行评估。对于框架，MR-CNN、FRCNN、R-50 和 SWT 分别表示 Mask R-CNN、Faster R-CNN、ResNet-50 和 Swin Transformer Tiny Model。对于训练数据，CCCap、VG、SBU、O365 和 CC3M 分别表示 COCO Caption [8], Visual Genome [31], SBU Caption [42], Objects365 [49], 和 Conceptual Caption 3M [50], 。GoldG 表示由 MDETR 整理的金标准 grounding 数据 [29]。

方法	框架	训练数据	AP	bAP	nAP
Detic [70]	Centernet2-R50	V3Det-C_{base}	17.7	30.2	6.7
RegionClip [69]	FR-CNN-R50	V3Det-C_{base}	12.6	22.1	3.1

表 7：在 C_{base} 中的训练划分上进行训练的两项代表性开放词汇目标检测工作的评估。

评估预训练的开放词汇检测模型。我们首先通过在 V3Det 验证划分上进行直接推理，测试各种开放词汇方法。如表 6所示，我们报告了 AP 所有类别的结果、 bAP基础类别的结果 C_{base}、 nAP新类的结果，均在C_{novel}上进行。实验结果表明，在现有目标检测数据集（如COCO、LVIS和Objects365）上训练的所有开放词汇检测方法，在 V3Det上都未能取得良好表现。其中，Detic 的性能最高，但仅为1% AP。根据表 2，完全监督的检测器在 V3Det 上可以获得显著更优的性能。这一观察表明，开放词汇检测方法仍有很大的探索空间来提升自身性能。

评估在 V3Det 上训练的 OVD 方法。我们进一步探究当前 OVD 方法在 V3Det 上训练后的有效性。具体而言，我们在 V3Det 的即，Detic [70] 和 RegionCLIP [69] 基础类别 C_{base} 的训练划分上训练这两种知名 OVD 方法，并在 V3Det 的 C_{base} 和 C_{novel} 的验证划分上进行测试。如表7所示，在 V3Det 的 C_{base} 上训练后，两种方法在 bAP 和 nAP上均取得了显著提升。值得注意的是，nAP Detic 和 RegionCLIP 的指标分别从 1.04% 提升至 6.7%，以及从 0.72% 提升至 3.1%。尽管性能仍远不及

训练数据	LVIS-nAP	Objects365-AP
LVIS* [25]	7.0	1.6
LVIS* +ImageNet-22k* [48]	13.2	3.4
V3Det*	14.5	4.4

表 8：类别泛化能力对比。采用 Detic 作为 OVD 框架。1) LVIS*: LVIS训练划分，不包含 LVIS-稀有类别和 Objects365 类别；2) LVIS* + ImageNet-22k*: LVIS训练划分，不包含 LVIS-稀有类别，以及带有 LVIS 和 Objects365 类别标签的 Objects365 + ImageNet-22k 图像；3) V3Det*: V3Det训练划分，不包含 LVIS-稀有类别和 Objects365 类别。LVIS-nAP和 Objects365-AP 分别表示 LVIS 稀有类别和全部 Objects365 类别的性能。

与监督模型相比，这些实验结果证实，大词汇量数据集将带来更强的类别泛化能力。

V3Det 的类别泛化。为了进一步验证 V3Det在训练开放词汇目标检测器中的有效性，我们在LVIS-稀有类别和 Objects365 类别上评估使用 V3Det 训练的 Detic。我们构建了一个新的类别子集 $C^*_{v3det} = C_{v3det} \setminus (C_{lvis\ rare} \cup C_{objects365})$，该子集从 V3Det 的词汇表中移除了 LVIS 稀有类别和 Objects365 类别，以及 $C^*_{lvis} = C_{lvis} \setminus (C_{lvis\ rare} \cup C_{objects365})$，该子集从 LVIS 的词汇表中移除了 LVIS 稀有类别和 Objects365 类别。在从 C^*_{v3det} 和 C^*_{lvis}中学习后，我们报告了 LVIS-nAP（LVIS 稀有类别上的平均精度）和 Objects365-AP（Objects365 上的平均精度）在两个数据集的 验证划分上的性能。如表8所示，在 C^*_{v3det} 上训练的 Detic 显著优于在 C^*_{lvis}上训练的模型，即使是在 LVIS 自身的稀有类别上也是如此。此外，我们还尝试利用带有 LVIS 和 Objects365 类别标签的 ImageNet 图像，这意味着“LVIS*+ImageNet-22k*”设置包含目标类别，即 LVIS-稀有类别和Objects365。令人惊讶的是，V3Det 仍然表现更好，展现了其作为数据集服务于开放词汇目标检测进一步研究的能力。

5. 结论

我们介绍了**V3Det**，这是一个在大量真实世界图像上精确标注边界框的大规模词汇视觉检测数据集。V3Det包含24.3万张图像，涵盖13,204个类别。大规模词汇按照层级类别树进行组织。每个类别均提供由人类专家和强大的聊天机器人撰写的专业描述。在大规模词汇目标检测和开放词汇目标检测方面均开展了广泛的基准测试，从而为未来研究带来了新的观察、实践和洞见。

致谢。本项目由上海人工智能实验室（P23KS00010, 2022ZD0160201）、创新科技署（ITC）InnoHK 项目下的感知与交互智能中心（CPII）有限公司，以及香港中文大学跨学科人工智能研究院资助。

在补充材料中，我们在附录 A中介绍了 V3Det 数据集的许可证，在附录 B中介绍了更多实现细节，并在附录 C中展示了更多实验结果。在附录D中，我们展示了 V3Det 层级类别组织的更详细可视化；在附录 E中，列出了标注过程中使用的粗糙类别；在附录中，我们展示了一些由人类专家和强大的聊天机器人 ChatGPT 编写的类别描述示例[2] ；在附录F中，我们展示了更多 V3Det 数据集的可视化内容，见附录 G。

A. V3Det 数据集许可与下载。

V3Det 图像。 V3Det 中约 90% 的图像选自 Bamboo 数据集 [67], ，来源于 Flickr 网站。其余 10% 直接从 Flickr 抓取。我们不拥有这些图像的版权。图像的使用必须遵守 Flickr 使用条款[3]。我们仅提供图像 URL 列表，不提供图像再分发。**V3Det 标注。** V3Det 标注、类别关系树及相关工具均依据知识共享署名 4.0 许可协议授权 [4]。

V3Det 下载。 图像 URL 元文件、用于轻松下载图像的脚本、标注、类别关系树及其他相关工具均可在 https://v3det.openxlab.org.cn/ 获取。

B. 实现细节。

B.1. 大词汇量目标检测。

两阶段与级联检测器。我们使用 ResNet-50[27] 和 Swin-B [40]骨干网络，对三个具有代表性的两阶段或级联检测器进行测试，即 Faster R-CNN [47],Cascade R-CNN [4], 和 CenterNet2 [71]。对于 Faster R-CNN 和 Cascade R-CNN，我们采用 mmdetection [7] 中的标准实现。对于 CenterNet2，我们采用其基于 Detectron 2 [60]的官方实现。为进行公平比较，实验采用 FPN [36], AdamW 优化器[41],多尺度增强和重复因子采样器 [25]训练。具体而言，我们使用 AdamW 优化器 [30] , 学习率为 10^{-4}， β_1=0.9，β_2=0.999，批量大小为 32。遵循 LVIS [25] 数据集上的最佳实践，我们采用多尺度增强训练：从 (640, 672, 704, 736, 768, 800) 像素中随机选择图像的较短边，并将图像的较长边限制在 1333 像素以内。遵循 LVIS [25], , 我们将重复因子采样器 [25]中的重复因子设为 10^{-3}。检测器训练 24 个训练轮次，并在第 16 和第 22 个训练轮次时将学习率降低 $10\times$ 。

单阶段检测器。 我们采用 ATSS[6]和 FCOS [53]在 mmdetection [7] 中的标准实现，作为具有代表性的单阶段检测器。其他设置遵循两阶段与级联检测器实验。

DETR 风格检测器。我们测试了两种 DETR 风格检测器，即使用两种不同骨干网络的 DeformableDETR [72], 和 DINO [62]，分别为 ResNet-50 [27], 和Swin-B [40]。我们使用 mmdetection [7] 实- 现Deformable DETR，并使用批量大小为 32、学习率为2×10^{-4}训练 50 个训练轮次。我们使用 detrex [12] 实- 现 DINO，并使用批量大小为 16、学习率为10^{-4}训练24 个训练轮次。两个模型均使用 AdamW 优化器 [30]，参数为 β_1=0.9、 β_2=0.999，并使用重复因子采样器，其因- 子为 10^{-3}。我们遵循相应框架所采用的默认数据增强方法。

B.2. 开放词汇目标检测（OVD）。

在本节中，我们提供了在 V3Det 上训练Detic [70] 和 RegionCLIP [69] 的实现细节。

对于 Detic，我们遵循原论文的设置，使用搭配 ResNet-50 [27] 的 Cen- terNet2 [71] 。我们采用大规模抖动 [21] ，输入分辨率为 640x640。批量大小为 64，使用 AdamW 优化器，学习率为 2×10^{-4} 。对于正文表 6 中的结果，我们首先在 V3Det 的基础类别 C_{base} 上训练模型 90k 次迭代。随后，我们创建了 ImageNet-21K 的一个子集，命名为 IN-V3Det，其中包含与 V3Det 重叠的 4197 个类别。接着，我们在 V3Det 和 IN-V3Det 上进行多数据集微调，训练 90k 次迭代。对于正文表 8 中的结果，V3Det* 的 Detic 仅在 V3Det* 上训练，不使用 ImageNet 图像，训练 180k 次迭代。

对于 RegionCLIP，我们遵循原论文的设置，使用以 ResNet-50 [27] 为骨干网络的 Faster R-CNN [47] ，并通过 CLIP 引导的区域-文本对齐进行微调。首先，在基础类别上训练离线区域提议网络（RPN） [47] ，共 180k 次迭代。随后，进行微调阶段，将学到的区域-文本对齐知识迁移到开放词汇检测中，训练 180k 次迭代。训练期间，批量大小为 16 张图像，初始学习率为 2×10^{-3} 。

C. 更多实验结果。

用于预训练的 V3Det。 V3Det 旨在成为海量词汇和开放词汇目标检测的基准。有趣的是，我们的研究结果表明，V3Det 还可有效用作预训练数据集。我们对在数据集 V3Det 和 Objects365 上预训练 CenterNet 2 [71] 和 R-50 的多种策略进行了比较分析。随后在 LVIS 数据集上进行微调，以评估初始预训练的效果。表A1显示，V3Det 和 Objects365 的表现相当，且可在预训练中互为补充。

* 贡献相同。[2] https://openai.com/blog/chatgpt [3] https://www.flickr.com/creativecommons/ [4] https://creativecommons.org/licenses/by/4.0/

预训练（第 1 阶段）	预训练（第 2 阶段）	微调	AP
-	-	LVIS	35.1
V3Det	-	LVIS	36.2
Objects365	-	LVIS	36.9
V3Det	Objects365	LVIS	37.3
Objects365	V3Det	LVIS	37.7

表 A1：使用 R-50 主干网络和归一化线性层在V3Det 和 Objects365 上对CenterNet2进行预训练，随后在LVIS上进行微调时不同策略的比较。对于每个阶段，训练计划均为批量大小为 64 的 90k 次迭代。

τ	AP	AP_{50}	AP_{75}
30	27.4	32.9	29.7
40	27.8	34.0	30.3
50	**28.3**	**34.5**	**30.8**
60	27.8	33.9	30.3

表 A2：归一化线性层中不同温度因子 τ 的比较。本表所用框架为使用 AdamW 优化器训练 12 个训练轮次的 Cascade R-CNN ResNet-50。

阶段	AP	AP50	AP75
Faster R-CNN [47]	21.2	29.5	24.1
Cascade R-CNN 第 1 阶段	22.7	31.1	25.8
Cascade R-CNN 第 2 阶段	27.0	33.7	29.7
Cascade R-CNN 所有阶段	**28.3**	**34.5**	**30.8**

表 A3：Cascade R-CNN 在不同阶段的性能。

方法	划分	AP	AP_{50}	AP_{75}
Cascade R-CNN [4] + 归一化线性层 [58]	验证集	42.5	49.1	44.9
	测试集	43.1	49.7	45.6
DINO [62]	验证集	42.0	46.8	43.9
	测试集	42.4	47.2	44.3

表 A4：验证集 与测试集 划分之间的性能差异。

归一化线性层消融实验。在正文表 2 和表 4 的实验中，我们展示了归一化线性层 [58]，$z = \tau \widetilde{\mathcal{W}}^T \widetilde{x} + b$，其中$\widetilde{\mathcal{W}}_i = \frac{\mathcal{W}_i}{\|\mathcal{W}_i\|_2}$，$i \in C$，$\widetilde{x} = \frac{x}{\|x\|_2}$。如表 A2所示，我们探究了其温度因子 τ，并发现当 τ 为 50 时性能最佳。因此，我们在实验中采用τ 50 作为默认设置。

级联 R-CNN 的不同阶段。 如表 A3所示，我们展示了 Cascade R-CNN [4] ResNet- 50 在不同阶段的性能。随着阶段数增加，平均精度也随之提升，证明了级联细化级联的有效性。另一方面，Cascade R-CNN 在第 1 阶段的平均精度为 22.7，仍高于 Faster R-CNN [47]，的 21.2，表明阶段级联结构有利于模型优化。

Val 与 Test 划分之间的性能差异。 在表 A4中，我们评估了 Cascade R-CNN [4] +归一化线性层 [58] 和 DINO [62] 使用 24 个训练轮次、AdamW 优化器和 Swin-B 主干网络，在 V3Det 的 *val* 与 *test* 划分上进行测试，结果表明这两个划分的标注具有一致性。

数据集	AP	分类错误	定位错误
LVIS [25]	**62.2**	7.69	**4.78**
V3Det	49.4	**25.32**	0.75

表 A5：EVA [18] 在 LVIS [25] 和 V3Det 上的分类错误（CLS ER）与定位错误（LOC ER）比较。

EVA 错误分析。 正文中的表 5 显示，EVA 在 V3Det 上的平均精度 [18]为 49.4，比其在 LVIS 上的平均精度低 12.8 [25]，，后者为 62.2。本节将探究平均精度差异的误差来源。表 A5比较了通过 TIDE 计算的 EVA 在 LVIS 和 V3Det 上的分类错误和定位错误 [1]。分类错误指定位正确（交并比 >0.5）但分类错误；定位错误指分类正确但定位错误（交并比 <0.5）。可以看到，V3Det 的定位错误（0.75）低于 LVIS（4.78），但 V3Det 的分类错误（25.32）远高于 LVIS（7.69）。这证实，与 LVIS 相比，V3Det 展现出更具挑战性的大规模词汇分类问题，因此带来了更广阔的探索空间。

层级开放词汇测试。为全面评估开放词汇检测器，我们提出了一种层级开放词汇测试，用于评估检测器的层级能力，该测试基于我们的层级类别组织构建。首先，我们采用两种不同的方法评估非叶节点的平均精度指标。一种是基于词汇表的非叶节点平均精度 AP^v，定义如下：

$$AP^v = \frac{1}{N} \sum_{i \in \{\text{non-leaf nodes}\}} AP(P_i, f_i(Y)), \tag{2}$$

$$f_i(Y) = \{y | y \in Y \text{ and } y \in \text{descendants of node } i\},$$

其中，N 是层级树中的非叶节点总数。 P_i 是非叶节点 i的预测框。 Y是所有叶节点的真实标注。 $f_i(Y)$ 是节点 i的所有后代节点的真实标注。 $AP(\alpha, \beta)$ 是框集合 α 和 β之间的平均精度。另一种是基于层级的非叶节点平均精度 AP^h，定义如下：

$$AP^h = \frac{1}{N} \sum_{i \in \{\text{non-leaf nodes}\}} AP(f_i(P), f_i(Y)), \tag{3}$$

$$f_i(P) = \{p | p \in P \text{ and } p \in \text{descendants of node } i\},$$

其中， P 是所有叶节点的预测框。 $f_i(P)$ 是属于节点 i 的后代节点的合并预测结果，并经过交并比阈值为0.5的非极大值抑制处理。

基于 AP^v和 AP^h，我们设计了层级分数，简称 H 分数，其定义为

$$\text{H-score} = AP^v / (AP^h + \epsilon) \times 100, \tag{4}$$

其中 ϵ 设为 1e-6。H 分数越高，表示检测器的层级能力越强。表 A6 给出了层级开放词汇

方法	AP^v	AP^h	H 分数
Detic [70]	7.25	16.90	42.90
RegionCLIP [69]	5.58	11.70	47.69

表 A6：层级开放词汇测试的结果。

测试。尽管 RegionCLIP 的 AP^v 和 AP^h 低于 Detic，但与 Detic 相比，RegionCLIP 较高的 H 分数表明其层级能力更强，这得益于对 CLIP[44]参数的利用具有鲁棒性，而非仅仅依赖提取的文本嵌入。

D. 层级类别组织。

图 1 展示了 V3Det 更详细的层级类别组织A1。

E. 粗粒度类别列表。

表 A7 列出了标注过程中使用的粗糙类别的详细信息，包括名称、每个粗糙类别中的细粒度类别总数，以及每个粗糙类别的一些细粒度示例。

F. 类别描述示例。

在表格中A8，我们展示了 V3Det 中类别描述的示例，这些描述由人类专家和 ChatGPT 撰写。

G. 更多数据集可视化。

图A2 和图A3 提供了一些带标注的图像样本以供可视化。

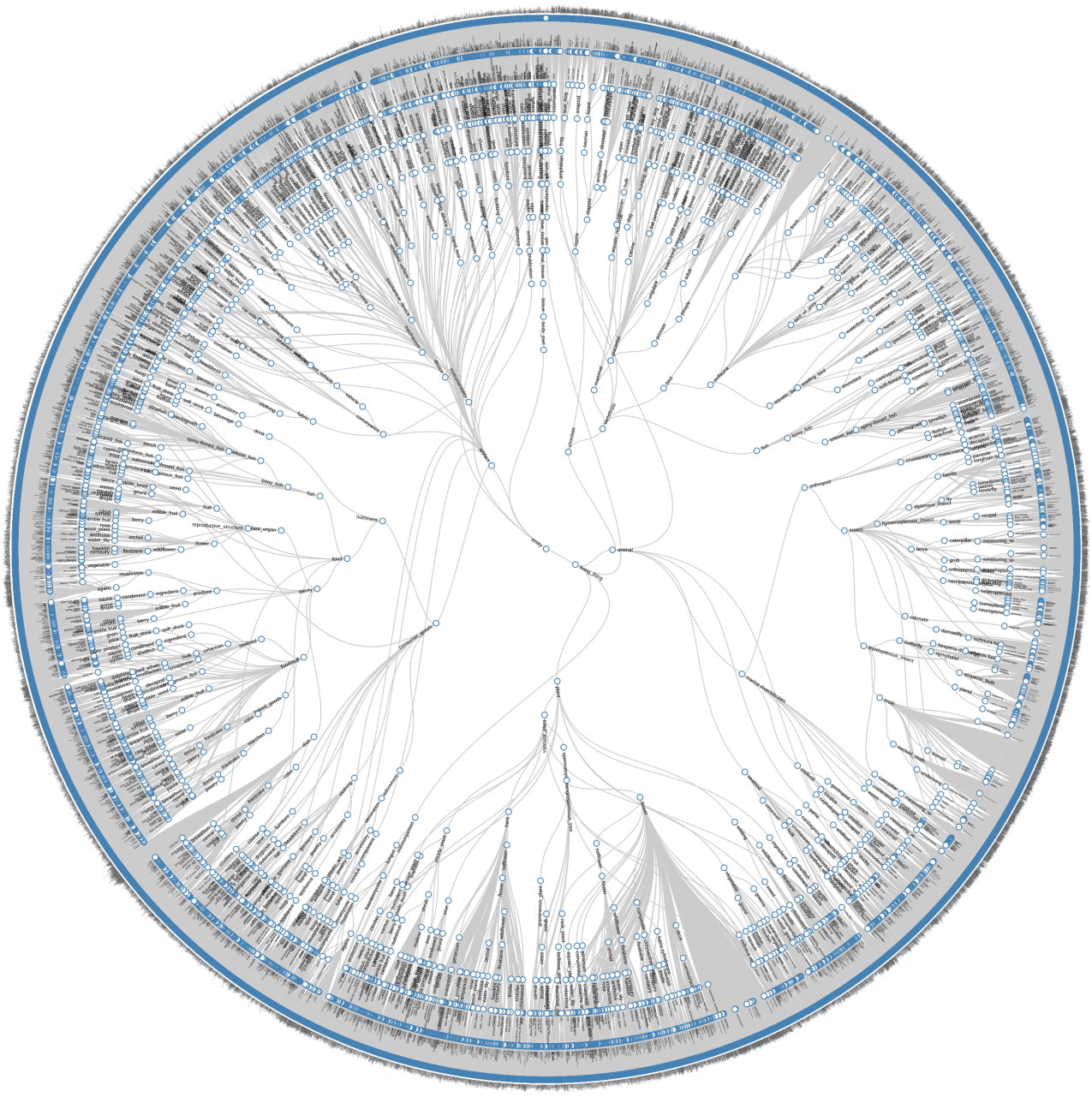


图 A1：**V3Det 中层级类别组织的详细可视化。**

	粗粒度类别	数字	细粒度类别列表
粗糙 1	动物与人类	7485	西伯利亚虎；马赛狮；网翅目昆虫；橄榄球运动员；男性；女性；...
粗糙 2	设备	241	吊扇；平板电脑；显示设备；移动设备；iPad；游戏控制器；.....
粗糙 3	桌椅	30	扶手椅；会议桌；餐桌；折叠椅；折叠桌；.....
粗糙 4	花卉	1911	杂交茶香月季；丰花月季；万寿菊；高山紫菀；夏威夷扶桑；橙色百合；...
粗类别5	蔬菜与豆类 & 水果与果皮	696	南瓜；猕猴桃；萨巴香蕉；四季橘；马托基香蕉；东部仙人掌梨；...
粗类别6	菜肴与肉类 & 主食与蛋类 豆制品与蛋类	602	鸡肉；炒面；火鸡肉；热狗面包；切片面包；荞麦面；.....
粗类别7	可穿戴物品	352	棒球服；针织帽；武术服；狗狗服饰；潜水面罩；.....
粗类别8	真菌	301	药用蘑菇；灵芝；杏鲍菇；脆红菇；.....
粗糙 9	车辆	223	微型面包车；通用汽车；蒸汽汽车；太阳能汽车；悬浮滑板；救护车；.....
粗糙 10	体育器材 不含球	168	滑板桥；Freebord；训练凳；鞍马；飞镖；.....
粗糙 11	饮料 & 调味料 & 油类和乳制品 & 液态巧克力	113	维也纳混合咖啡；摩卡奇诺；奶油利口酒；冰淇淋；酱油；.....
粗糙 12	乐器	109	鼓槌；塔布拉鼓；班卓琴；钹；鼓；吉他；竖琴；钢琴；谢凯雷；.....
粗糙 13	手工具和点火工具	98	手扇；鳄鱼扳手；断线钳；瓶盖起子；开瓶器；镊子；.....
粗糙 14	五金器具	66	钥匙扣；管道配件；穿线针；锚；砧；锥子；引针；...
粗糙 15	无武器 黑客攻击与切割	62	炮座；攻城槌；炸弹；黄铜指虎；子弹；加农炮； ...
粗分类16	卡片与纸制品 与板	62	支付卡；学历证书；黑板；书签；剪贴板；门牌；.....
粗分类17	切割工具 & 劈砍类冷兵器	54	斧头；弓锯；横切锯；绿篱修剪机；刀；奶酪切刀；削皮器；.....
粗分类18	非电动餐具	53	葡萄酒杯；高脚酒杯；烧烤架；蒸锅；杯子；筷子；叉子；...
粗糙 19	建筑与帐篷	51	住宅；水塔；便携式厕所；凉亭；钟形帐篷；纪念碑；...
粗糙 20	个人护理	50	洁面乳；婴儿爽身粉；瓶刷；避孕套；卷发器；发刷；...
粗糙 21	Ball	48	篮球；保龄球；板球；槌球；高尔夫球；手球；.....
粗糙 22	仙人掌	39	圣佩德罗仙人掌；大花仙人掌；柱状仙人掌；.....
粗糙 23	窗户与门 & 交付缺口与良好方面	38	窗户遮挡物；水井；自流井；门把手；老虎窗；喷油井；...
粗糙 24	布料与布料材质	36	折叠餐巾；浴室地垫；沙滩巾；清洁垫；抹布；装饰小垫；.....
粗糙 25	办公设备	37	板擦；算盘；粉笔；钢笔；记号笔；铅笔；砚台；.....
粗糙 26	清洁工具与喷嘴	35	浴室洗手池；厕纸架；灌溉喷头；浴缸；扫帚；垃圾箱；.....
粗糙 27	测量设备	28	烧杯；指南针；检测器；圆规；铅垂线；量角器；三角尺；.....
粗糙 28	架子和储物柜	28	酒架；鞋架；壶架；书架；衣帽架；挂衣架；.....
粗糙 29	宝石与喷泉	28	喷泉；水晶；钻石；红宝石；珍珠；祖母绿；金绿宝石；翡翠；.....
粗糙 30	宗教相关物品	27	护身符；基督教十字架；旗帜；肩章；图腾柱；奖章；.....
粗类别31	医疗相关物品	26	吸引器；导管；皮下注射针；药片；膏药；听诊器；.....
粗类别32	杆与管	24	吹箭筒；烟囱；雪茄；消防水带；扶手杆；海泡石烟斗；试管；.....
粗类别33	救生物品	21	婴儿游泳圈；气垫；裤形救生圈；消防栓；救生圈；游泳臂圈；.....
粗类别34	瓶与袋 & 桶与盒子	22	编织篮；袋子；桶；盆；瓶子；公文包；壶；雨水桶；...
粗类别35	糖果与固体巧克力	21	脆糖；口香糖；糖苹果；可可粉；棒棒糖；果冻；.....
粗糙36	行星与卫星	18	卫星；陨石；月球；太阳；黑洞；地球；火星；水星；金星；.....

	粗粒度类别	数字	细粒度类别列表
粗糙 37	床上用品	18	婴儿床；床；床头枕；双层床；便携式婴儿床；婴儿小床；日式床垫；吊床；.....
粗糙 38	国际象棋	16	国际象棋棋子；国际象棋套装；中国象棋；围棋；国际跳棋；日本将棋；...
粗糙 39	Bell	12	电铃；定时器；风向标；风铃；火灾报警器；风铃；自行车铃；...
粗糙 40	路标与道路障碍物	12	停止标志；人行横道标志；广告牌；人行横道；庭院标记；...
粗糙 41	控制设备与供暖 家用电器与热水器 无电袋子	12	火盆；变速杆；手轮；热水瓶；散热器；烤炉；...
粗糙 42	轮状物体	11	自行车车轮；线轴；摩天轮；齿轮；内胎；滑轮；汽车轮胎；.....
粗糙 43	Lens	10	放大镜；显微镜；望远镜；步枪瞄准镜；三棱镜；.....
粗糙 44	货币与哨子	8	哨子；黄金；钱；现金；纸币；硬币；银行钞票；.....
粗糙 45	动物巢穴	9	鸟笼；鸡舍；兔笼；巢穴；黄蜂窝；笼子；.....
粗糙 46	雨伞与梯子	9	降落伞；鸡尾酒伞；铝合金梯子；木梯；...
粗糙 47	电子元件	6	电池；电容器；线圈；电阻器；太阳能电池；电子元件；...
粗糙 48	插座与插头	6	塞子；软木塞；电源插座；插座；墙壁插座；...
粗糙 49	诱饵与水族箱 & 邮筒	6	水族箱；鱼缸；邮筒；纳米水族箱；勺形拟饵；Penfold 邮筒；...
粗糙 50	工业机器 & 喷漆	5	混凝土搅拌机；起重机；发电机；喷漆；抽油机；.....
粗糙 51	弹簧与磁铁 与指南针	4	指南针；橡皮筋；弹簧；冰箱磁铁；.....
粗糙 52	爆米花机 自动取款机 & 烟灰缸 & 香炉	5	烟灰缸；自动取款机；香炉；爆米花机；香炉；.....
粗类别53	模型	93	玩具车；魔方；双耳陶罐；浑仪；木马；玩偶；.....

表 A7：**粗粒度类别详情**。所有细粒度类别分为 53 个粗粒度类别。“数字”表示每个粗粒度类别中的细粒度类别总数。

Name	图像	Type	描述
platycercus adscitus		专家 ChatGPT	platycercus adscitus 的头部为棕色，腹部为黄色，翅膀为绿色 绿色身体、黄色头部、红色斑纹（淡头玫瑰鹦鹉）
阿德利企鹅		专家 chatgpt	生活在南极洲阿德利海岸大型群落中的中型企鹅 阿德利企鹅的头部和背部为黑色，前胸为白色，眼睛周围有一道独特的白色环纹。
北极熊		专家 chatgpt	与其他熊相比，头部较小，耳朵圆，脖子纤细。皮肤呈黑色，体型庞大，性情凶猛。 大型白色毛皮熊，脖子长，身躯粗壮，腿部有力，爪子锋利
彭布罗克		专家 ChatGPT	体型较小、腿部笔直的柯基犬品种，耳朵尖，尾巴短 体型小巧结实的犬种，耳朵直立，脸像狐狸，尾巴经过断尾处理（或天生短尾）
孟加拉老虎		专家 chatgpt	黄橙色的皮毛上带有深色条纹，腹部和四肢为白色，橙色尾巴上有黑色环纹 橙棕色的皮毛上带有黑色条纹，腹部为白色，眼睛上方有独特的白色斑纹
北极野兔		专家 chatgpt	一种生活在北美洲北部的大型野兔；冬季几乎全身雪白 一种生活在北极地区的大型野兔，毛色洁白，耳朵长，后腿强壮
大猩猩		专家 ChatGPT	大猩猩强壮，脸部和耳朵没有毛发。它额头高耸，下颌骨突出。 体型庞大、肌肉发达的灵长类动物，长有黑色或深棕色皮毛、宽阔的胸膛和突出的眉脊

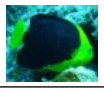
Name	图像	Type	描述
非洲象		专家：	一种原产于非洲的大象，拥有巨大、会扇动的耳朵和象牙
		ChatGPT：	一种体型庞大、灰色的哺乳动物，有长鼻子、大耳朵、弯曲的象牙和褶皱的皮肤
夏洛莱牛		这种牛体型较大，体重为 700-1650 千克。它们长着白色或奶油色的被毛，鼻子呈粉红色。	
		ChatGPT，	大型白色牛品种，体格健壮，额头宽阔，短角
格氏斑马的斑马		专家，	它看起来像一头长着大脑袋、鼻孔细长、耳朵又大又圆且呈圆锥形、鬃毛高高竖起的骡子
		ChatGPT	大型，野生鱿鱼，身上有黑白条纹、白色腹部和高耸直立的鬃毛
箱龟		专家：	它们的龟壳呈拱形，底部有铰链，可以紧紧闭合以躲避捕食者
		ChatGPT	小型，龟壳呈拱形，皮肤为棕色或橄榄色，带有黄色或橙色斑点或条纹
喙头蜥		专家：	这是一种爬行动物，外形像灰绿色蜥蜴，仅分布于新西兰附近的某些小岛上
		ChatGPT	是一种有棘状冠、两排牙齿，并且在其头部顶部有第三只“眼睛”的爬行动物
北方海马		北方海马专家有长长的吻部和卷曲的尾巴，最长可长到8英寸。	
		ChatGPT	这种北方海马有一副覆盖着骨质板的小身体、长吻和卷曲的尾巴
低地掘穴树蛙		专家：	一种陆生、掘穴的夜行性蛙，栖息于草地和灌木林，头部上表面非常坚硬
		chatgpt：	一种体型小巧、身体圆润、皮肤光滑的蛙，适于掘穴，腿短，长有尖吻
岩美神仙鱼		专家：	主要呈黑色。头部、身体前半部分和尾鳍呈亮黄色
		chatgpt	亮黄色头部、尾巴、胸鳍；黑色身体、背鳍、臀鳍；颜色过渡突兀
鳚鱼		小型（长至3英寸）扁平海鱼专家，腹部有吸盘，可吸附在岩石等物体上。	
		chatgpt	该鳚鱼体形扁平，具有吸盘状腹鳍盘，颜色从棕色到绿色不等。
鼠海豚		专家鲸与海豚的区别在于其紧凑体型、体型较小，以及带有匙形牙齿的弯曲钝吻。	
		ChatGPT	小型灰色海洋哺乳动物，头部圆润，背鳍较小，嘴呈弯曲状。
蝠鲼		专家极其大型的远洋热带鳐鱼，具有三角形胸鳍、角状头鳍和巨大的朝前嘴巴	
		ChatGPT	体型较大、身体扁平，有翼状鳍，无尾巴，上侧呈黑色或深色，下侧呈白色或浅色
aglais		aglais 的翅膀呈锈红色，翅尖各有一个由黑色、蓝色和黄色组成的独特眼斑。	
		ChatGPT	蝴蝶翅膀上的眼状斑，蓝绿色和棕色调（欧洲孔雀蝶）
林地向日葵		专家，	一种原产于美国的野生花卉，拥有明亮的黄色花瓣，围绕着深棕色的花心
		ChatGPT	高大的黄色花卉，许多细长花瓣环绕着深色的中央花盘，叶片呈绿色。
粉花豆荚木		专家的花卉长在圆柱形或球形花序中，并具有许多细长的雄蕊。	
		ChatGPT，	一种小型灌木，叶片呈蕨类状，开着鲜艳的粉色绒球状花卉。
岩蔷薇 Salviifolius		专家是一种灌木，具有芬芳的银绿色叶片和白色花朵。	
		chatgpt	一种高大的、亮黄色或橙色的花卉，常因其可食用的种子而种植，也常用作观赏植物。
电钻		专家：	一种手持工具，带有由电动机驱动的旋转夹头，用于钻孔
		ChatGPT：	一种手持工具，带有电动机和旋转夹头，可用于钻孔和拧紧螺钉，通常采用手枪式握柄
单簧管		专家：	单簧管是一种单簧片乐器，具有近乎圆柱形的管膛和喇叭形的铃口。
		ChatGPT：	一种细长的圆柱形木管乐器，配有吹嘴、簧片和多个按键，可演奏不同音符
越野车辆		专家：	越野车辆通常配备四轮驱动、加高的悬架和大尺寸轮胎
		ChatGPT：	具有高离地间隙、宽轮胎且通常配备四轮驱动的大型坚固车辆，适用于非铺装地形

表 A8：V3Det 类别描述示例。



图 A 2: **复杂场景中标注的可视化**。每个框都配有类别名称和框尺寸 e.

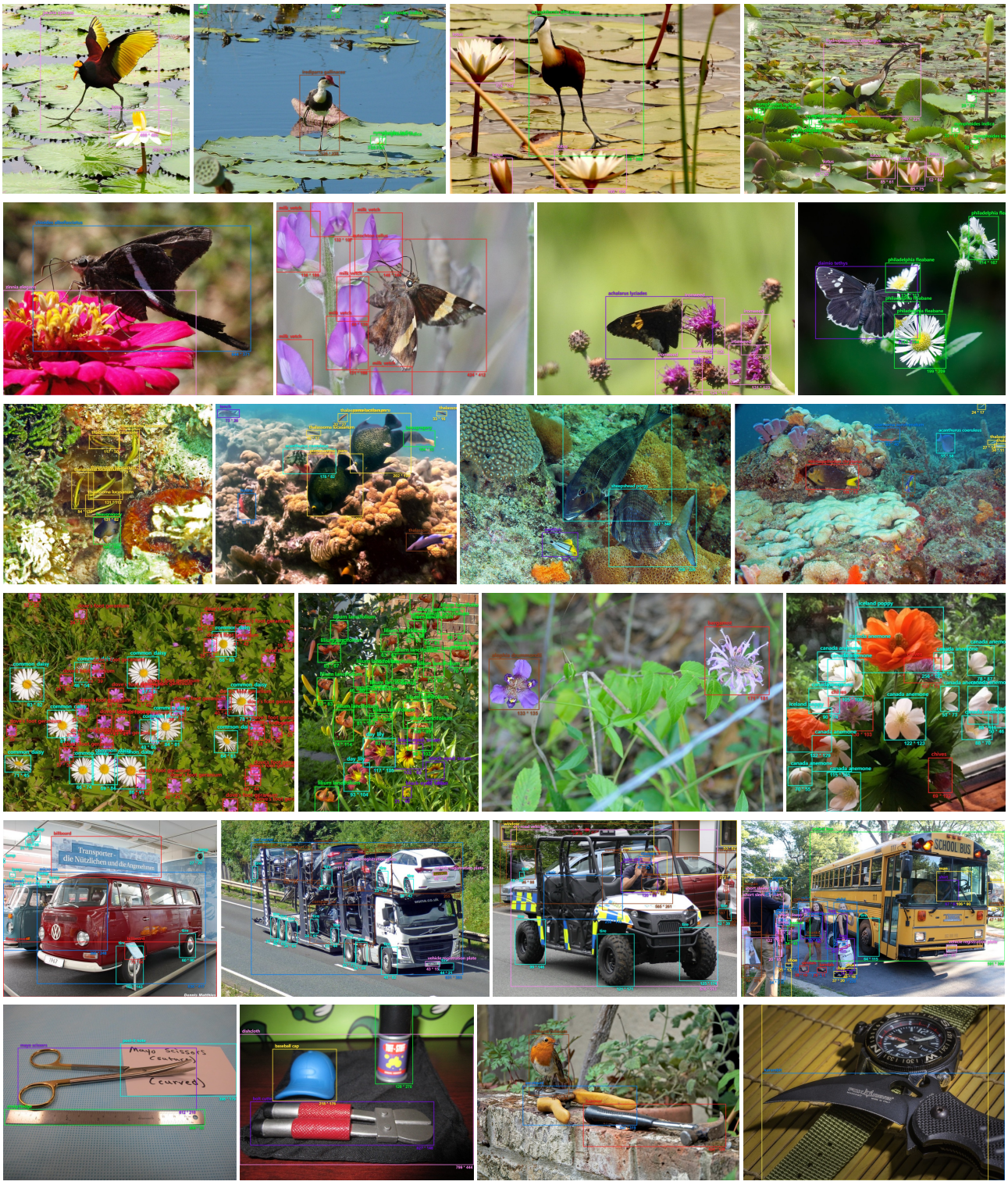


图 A3：**细粒度类别标注的可视化**。 每一行展示了一组容易混淆的视觉相似类别标注。

参考文献

[1] Daniel Bolya、SeanFoley、James Hays和 Judy Hoffman。TIDE：用于识别目标检测错误的通用工具箱。在 ECCV，2020 年。10[2] Likun Cai、Zhi Zhang、Yi Zhu、Li Zhang、MuLi和 Xiangyang Xue。BigDetection：用于改进目标检测器预训练的大规模基准。在 CVPR，2022 年。2、3[3] Zhaowei Cai 和 Nuno Vasconcelos。CascadeR-CNN：深入探究高质量目标检测。在 CVPR，2018 年。1、3、6[4] Zhao wei Cai 和Nuno Vasconcelos。CascadeR-CNN：高质量目标检测与实例分割。CVPR，2018 年。6、7、9、10[5] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用Transformer网络进行端到端目标检测。在 ECCV，2020 年。1、3[6] Kai Chen、Jiangmiao Pang、Jiaqi Wang、Yu Xiong、Xiaoxiao Li、Shuyang Sun、Wansen Feng、Ziwei Liu、Jianping Shi、Wanli Ouyang、ChenChange Loy 和 DahuaLin。用于实例分割的混合任务级联。在 CVPR，2019年。2、7、9[7] KaiChen、Jiaqi Wang、Jiangmiao Pang、曹宇航、Yu Xiong、Xiaoxiao Li、Shuyang Sun、Wansen Feng、Ziwei Liu、Jiarui Xu、ZhengZhang、Dazhi Cheng、Chenchen Zhu、TianhengCheng、Qijie Zhao、Buyu Li、Xin Lu、Rui Zhu、Yue Wu、Jifeng Dai、Jingdong Wang、Jianping Shi、Wanli Ouyang、Chen Change Loy 和 DahuaLin。MMDetection：开放的 mmlab 检测工具箱与基准。arXiv预印本 arXiv:1906.07155，2019 年。6、9[8] Xinlei Chen、Hao Fang、Tsung-Yi Lin、Ramakrishna Vedantam、Saurabh Gupta、PiotrDoll´ar 和 C LawrenceZitnick。Microsoft COCO Captions：数据收集与评估服务器。arXiv，2015 年。2、3、6、8[9] Yixin Chen、JamesZ Wang和 Robert Krovetz。通过聚类进行基于内容的图像检索。在第5 届ACM SIGMM多媒体信息检索国际研讨会论文集，第 193–200 页，2003 年。5[10] Tao Chu、Pan Zhang、Qiong Liu和JiaqiWang。Buol：一种用于从单幅图像进行全景三维场景重建的自底向上框架，结合了占用感知提升。在 CVPR，2023年。2[11] Xiyang Dai、Yinpeng Chen、BinXiao、Dongdong Chen、Mengchen Liu、Lu Yuan 和 Lei Zhang。Dynamic Head：通过注意力机制统一目标检测头。在 CVPR，2021 年。8[12] detrex 贡献者。detrex：用于基于 Transformer 的目标检测算法的研究平台。https: // github.com/IDEA-Research/detrex，2022 年。6、9[13] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和Kristina Toutanova。BERT：用于语言理解的深度双向Transformer网络预训练。arXiv，2018 年。3[14] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、J akob Uszkoreit 和 Neil Houlsby。一幅图像是值得拥有 16×16 个词：用于大规模图像识别的Transformer网络。ICLR，2021。7[15] Yu Du、Fangyun Wei、Zihe Zhang、Miaojing Shi、Yue Gao 和 Guoqi Li。借助视觉语言模型学习开放词汇目标检测的提示方法。见 CVPR，2022。8[16] Kaiwen Duan、Song Bai、Lingxi Xie、Honggang Qi、Qingming Huang 和 Qi Tian。Centernet：用于目标检测的关键点三元组。见 ICCV，2019。3[17] M. Everingham、L. Van Gool、C. K. I. Williams、J. Winn 和 A. Zisserman。PASCAL 视觉目标类别（VOC）挑战赛。IJCV，2010。2[18] Yuxin Fang、Wen Wang、Binhui Xie、Quan Sun、Ledell Wu、Xinggang Wang、Tiejun Huang、Xinlong Wang 和 Yue Cao。EVA：探索大规模掩码视觉表征学习的极限。arXiv，2022。7，10[19] Chengjian Feng、Yujie Zhong、Zequn Jie、Xiangxiang Chu、Haibing Ren、Xiaolin Wei、Weidi Xie 和 Lin Ma。Prompt-det：利用未经筛选的图像实现开放词汇检测。见 ECCV，2022。8[20] Mingfei Gao、Chen Xing、Juan Carlos Niebles、Junnan Li、Ran Xu、Wenhao Liu 和 Caiming Xiong。使用伪边界框标签进行开放词汇目标检测。arXiv，2021。8[21] Golnaz Ghiasi、Yin Cui、Aravind Srinivas、Rui Qian、Tsung-Yi Lin、Ekin D. Cubuk、Quoc V. Le 和 Barret Zoph。简单的复制粘贴是一种强大的实例分割数据增强方法。见 CVPR，2021年6月。9[22] Ross Girshick。Fast R-CNN。见 ICCV，2015。1，3[23] Ross Girshick、Jeff Donahue、Trevor Darrell 和 Jitendra Malik。用于精准目标检测和语义分割的丰富特征层级。见 CVPR，2014。1，3[24] Xiuye Gu、Tsung-Yi Lin、Weicheng Kuo 和 Yin Cui。通过视觉与语言知识蒸馏实现开放词汇目标检测。arXiv，2021。3[25] Agrim Gupta、Piotr Dollar 和 Ross Girshick。LVIS：一个大型词汇实例分割数据集。见 CVPR，2019。1，2，3，6，7，8，9，10[26] John A Hartigan 和 Manchek A Wong。算法 AS 136：一种 k-means 聚类算法。《皇家统计学会杂志：C辑（应用统计）》，28(1)：100–108，1979。5[27] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。见 CVPR，2016。6，9[28] Zhaojin Huang、Lichao Huang、Yongchao Gong、Chang Huang 和 Xinggang Wang。Mask Scoring R-CNN。见 CVPR，2019。2[29] Aishwarya Kamath、Mannat Singh、Yann LeCun、Gabriel Synnaeve、Ishan Misra 和 Nicolas Carion。MDETR——用于端到端多模态理解的调制检测。见 ICCV，2021。3，8[30] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv，2014。9[31] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalan- tidis、Li-Jia Li、David A Shamma 等。Visual Genome：通过众包密集图像标注连接语言与视觉。

连接语言与视觉：使用众包密集图像标注。corr abs/1602.07332（2016）。arXiv，2016。8[32] Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、S Popov、M Malloci、T Duerig 等。Open Images 数据集 V4：大规模统一图像分类、目标检测和视觉关系检测。arXiv 2018。arXiv，2018。1，2，3[33] Hei Law 和 Jia Deng。CornerNet：将目标检测为成对关键点。见 ECCV，2018。3[34] Chunyuan Li、Haotian Liu、Liunian Harold Li、Pengchuan Zhang、Jyoti Aneja、Jianwei Yang、Ping Jin、Yong Jae Lee、Houdong Hu、Zicheng Liu 等。ELEVATER：用于评估语言增强视觉模型的基准与工具包。arXiv，2022。2，3[35] Liunian Harold Li、Pengchuan Zhang、Haotian Zhang、Jianwei Yang、Chunyuan Li、Yiwu Zhong、Lijuan Wang、Lu Yuan、Lei Zhang、Jenq-Neng Hwang 等。基于语言的图像预训练。见 CVPR，2022。3，8[36] Tsung-Yi Lin、Piotr Doll´ar、Ross B. Girshick、Kaiming He、Bharath Hariharan 和 Serge J. Belongie。用于目标检测的特征金字塔网络。见 CVPR，2017。9[37] Tsung-Yi Lin、Priya Goyal、Ross B. Girshick、Kaiming He 和 Piotr Doll´ar。密集目标检测的焦点损失。见 ICCV，2017。1，3，6[38] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。见 ECCV，2014。1，2，3，7[39] Wei Liu、Dragomir Anguelov、Dumitru Erhan、Christian Szegedy、Scott E. Reed、Cheng-Yang Fu 和 Alexander C. Berg。SSD：单次多框检测器。见 ECCV，2016。1，3[40] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：使用移位窗口的层级视觉 Transformer。见 ICCV，2021。6，9[41] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。见 ICLR，2018。6，7，9[42]Vicente Ordonez、Girish Kulkarni 和 Tamara Berg。Im2text：使用 100 万张带标题的照片描述图像。NIPS，2011。8[43] Wanli Ouyang、Xiaogang Wang、Cong Zhang 和 Xiaokang Yang。长尾分布下目标检测深度模型微调的影响因素。见 CVPR，2016。6，7[44]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。从自然语言监督中学习可迁移的视觉模型。见 ICML。PMLR，2021。3，5，11[45] Hanoona Rasheed、Muhammad Maaz、Muhammad Uzair Khattak、Salman Khan 和 Fahad Shahbaz Khan。弥合目标级与图像级表征之间的差距，以实现开放词汇检测。见 NIPS，2022。8[46] Joseph Redmon 和 Ali Farhadi。YOLO9000：更好、更快、更强。见 CVPR，2017。3[47] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：借助区域候选网络实现实时目标检测。发表于 NeurIPS，2015。1、3、6、7、9、10[48] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein、Alexander C. Berg 和 Li Fei-Fei。ImageNet 大规模视觉识别挑战赛。IJCV，2015。3、8[49] Shuai Shao、Zeming Li、Tianyuan Zhang、Chao Peng、Gang Yu、Xiangyu Zhang、Jing Li 和 Jian Sun。Objects365：用于目标检测的大规模高质量数据集。发表于 ICCV，2019。1、2、3、5、7、8[50] Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。Conceptual Captions：一个经过清理、采用上位词标注的图像替代文本数据集，用于自动图像描述。发表于第 56 届计算语言学协会年会论文集（第 1 卷：长篇论文），2018。3、8[51] Peize Sun、Rufeng Zhang、Yi Jiang、Tao Kong、Chenfeng Xu、Wei Zhan、Masayoshi Tomizuka、Lei Li、Zehuan Yuan、Changhu Wang 和 Ping Luo。Sparse R-CNN：通过可学习的候选区域实现端到端目标检测。发表于 CVPR，2020。1[52] Zhi Tian、Chunhua Shen 和 Hao Chen。用于实例分割的条件卷积。发表于 ECCV，2020 2。Zhi Tian、[53] Chunhua Shen、Hao Chen 和 Tong He。FCOS：全卷积单阶段目标检测。CoRR，abs/1904.01355，2019。3、6、7、9[54] Jiaqi Wang、Kai Chen、Rui Xu、Ziwei Liu、Chen Change Loy 和 Dahua Lin。CARAFE：特征的内容感知重组。发表于 ICCV，2019。1[55] Jiaqi Wang、Kai Chen、Rui Xu、Ziwei Liu、Chen Change Loy 和 Dahua Lin。Carafe++：统一的特征内容感知重组。TPAMI，2021。1[56] Jiaqi Wang、Kai Chen、Shuo Yang、Chen Change Loy 和 Dahua Lin。通过引导锚定生成区域候选。发表于 CVPR，2019。1[57] Jiaqi Wang、Wenwei Zhang、曹宇航、Kai Chen、Jiangmiao Pang、Tao Gong、Jianping Shi、Chen Change Loy 和 Dahua Lin。侧向感知边界定位，实现更精确的目标检测。发表于 ECCV，2020。1[58]Jiaqi Wang、Wenwei Zhang、Yuhang Zang、曹宇航、Jiangmiao Pang、Tao Gong、Kai Chen、Ziwei Liu、Chen Change Loy 和 Dahua Lin。用于长尾实例分割的 Seesaw 损失。发表于 CVPR，2021。6、7、10[59] Xinlong Wang、Tao Kong、Chunhua Shen、Yuning Jiang 和 Lei Li。SOLO：根据位置分割目标。发表于 ECCV，2020。2[60] Yuxin Wu、Alexander Kirillov、Francisco Massa、Wan-Yen Lo 和 Ross Girshick。Detectron2。https://github. com/facebookresearch/detectron2，2019。6、9[61] Alireza Zareian、Kevin Dela Rosa、Derek Hao Hu 和 Shih-Fu Chang。利用图像说明进行开放词汇目标检测。发表于 CVPR，2021。3[62] Hao Zhang、Feng Li、Shilong Liu、Lei Zhang、Hang Su、Jun Zhu、Lionel Ni 和 Harry Shum。DINO：采用改进去噪锚框的 DETR，实现端到端目标检测。发表于 ICLR，2022。3 、 6 、 7 、 9 、 10

[63] Pan Zhang、Bo Zhang、Ting Zhang、Dong Chen、Yong Wang 和 Fang Wen。原型伪标签去噪与目标结构学习，用于领域自适应语义分割。发表于 CVPR，2021。2[64] Pan Zhang、Bo Zhang、Ting Zhang、Dong Chen 和 Fang Wen。用于半监督语义分割的鲁棒互学习。arXiv，2021。2[65] Shifeng Zhang、Cheng Chi、Yongqiang Yao、Zhen Lei 和 Stan Z Li。通过自适应训练样本选择弥合基于锚框与无锚框检测之间的差距。发表于 CVPR，2020。6[66]Shilong Zhang、Xinjiang Wang、Jiaqi Wang、Jiangmiao Pang、Chengqi Lyu、Wenwei Zhang、Ping Luo 和 Kai Chen。用于端到端目标检测的密集区分查询。发表于 CVPR，2023 年 6 月。1[67] Yuanhan Zhang、Qinghong Sun、Yichun Zhou、Zexin He、Zhenfei Yin、Kun Wang、Lu Sheng、Yu Qiao、Jing Shao 和 Ziwei Liu。Bamboo：通过人机协同持续构建超大规模视觉数据集。arXiv，2022。3、9[68] Shiyu Zhao、Zhixing Zhang、Samuel Schulter、Long Zhao、BG Vijay Kumar、Anastasis Stathopoulos、Manmohan Chandraker 和 Dimitris N Metaxas。利用无标签数据和视觉语言模型进行目标检测。发表于 ECCV，2022。8[69] Yiwu Zhong、Jianwei Yang、Pengchuan Zhang、Chunyuan Li、Noel Codella、Liunian Harold Li、Luowei Zhou、Xiyang Dai、Lu Yuan、Yin Li 等。RegionCLIP：基于区域的语言图像预训练。发表于 CVPR，2022。3、8、9、11[70] Xingyi Zhou、Rohit Girdhar、Armand Joulin、Philipp Kr¨ahenb¨uhl 和 Ishan Misra。利用图像级监督检测两万类目标。发表于 ECCV，2022。3 、 8 、 9 、 11[71] Xingyi Zhou 、 Vladlen Koltun 和 Philipp Kr¨ahenb¨uhl。概率两阶段检测。arXiv，2021。3、6、7、9[72] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。Deformable DETR：用于端到端目标检测的可变形Transformer。arXiv，2020。3、6、7、9