文章编号:1672-3961(2025)01-0001-14 DOI:10.6040/j.issn.1672-3961.0.2024.162

开放词汇目标检测方法综述

聂秀山,赵润虎,宁阳*,刘新锋
(山东建筑大学计算机科学与技术学院, 山东 济南 250101)

摘要:目标检测方法针对特定场景进行训练,需要识别的物体都要人工标注,检测器只能识别被标注的物体。随着目标检测应用场景逐渐增加,特定场景下训练的目标检测器不能满足多样化场景需求,目标检测方法的泛化性能成为研究者关注热点。不同场景中存在同一物体标签不一致,不同物体特征差异较大等问题,导致在特定场景下训练目标检测器无法泛化到其他场景。针对上述挑战,研究者提出面向开放词汇目标检测方法,利用大量图像-词汇知识将目标检测器从特定场景扩展到开放场景。检测器扩展到开放场景通常有两种方式,即基于大规模图像标题数据方法和基于预训练视觉语言模型方法。基于图像标题数据方法通常需要从大量数据中提取与物体相对应的词汇知识注入检测器,基于视觉语言模型方法则直接利用预训练的知识扩展检测器。开放词汇目标检测模型无需重新训练即可应用在不同场景,更加实用有效。

关键词:开放词汇;开放世界;零样本学习;开放场景目标检测;视觉语言模型
中图分类号:TP391 **文献标志码**:A
引用格式:聂秀山,赵润虎,宁阳,等.开放词汇目标检测方法综述[J]. 山东大学学报(工学版),2025,55(1):1-14.
NIE Xiushan, ZHAO Runhu, NING Yang, et al. Survey of open vocabulary object detection methods [J]. Journal of Shandong University (Engineering Science), 2025, 55(1):1-14.

Survey of open vocabulary object detection methods

NIE Xiushan, ZHAO Runhu, NING Yang*, LIU Xinfeng
(School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, Shandong, China)

Abstract: Object detection methods required training for specific scenes, with objects to be detected manually annotated. The object detector could only recognize the objects that were labeled. As the application scenarios of object detection gradually increased, object detectors trained for specific scenes failed to meet the needs of diverse scenarios. The generalization capability of object detection methods became a hot topic among researchers. In different scenarios, the same object was given inconsistent labels, and significant differences were noted in the features of various objects, causing the object detectors trained in those specific scenes to fail in generalizing to other scenes. Addressing the aforementioned challenges, researchers introduced an open-vocabulary object detection approach. This method utilized extensive image-lexicon knowledge to extend the object detection task from specific to open scenes. Detectors were typically extended to open scenes in two ways: one was through the use of large-scale image-caption data, and the other was via pre-trained visual-language models. The image-caption data method typically required extracting vocabulary knowledge corresponding to objects from a large amount of data and injecting it into the detector. The visual-language model method directly utilized pre-trained knowledge to extend the detector. Open-vocabulary object detection models were able to be applied to different scenes without retraining, which made them more practical and effective.

Keywords: open vocabulary; open world; zero-shot learning; open scene object detection; visual language model

收稿日期:2024-07-12
基金项目:山东省自然科学基金资助项目(ZR202103010201)
第一作者简介:聂秀山(1981—),男,江苏徐州人,教授,博士生导师,博士,主要研究方向为机器学习与数据挖掘、视觉数据智能检索与分析。E-mail:niexiushan@163.com
*通信作者简介**:宁阳(1985—),男,山东济南人,副研究员,硕士生导师,博士,主要研究方向为通用人工智能、机器学习、计算机视觉。E-mail:ningyang20@sdjzu.edu.cn

0 引言

目标检测是计算机视觉领域的核心任务，旨在找出图像中所有感兴趣的目标，确定目标位置和类别，不同目标之间特征差异较大，存在背景干扰，物体遮挡等问题，目标检测一直是一项具有挑战性的问题。随着深度学习技术发展；文献[1]提出了R-CNN网络用于解决目标检测任务，文献[2]提出了经典的两阶段目标检测架构Faster R-CNN，目标检测研究进入高速发展阶段。两阶段算法的思想是在图像上生成大量可能包含对象的候选区域，对候选区域进行筛选后进行分类和回归选出目标对象。这种方法检测精度较高，检测速度相对较慢。一些研究者提出舍弃复杂候选区操作，只需要在图像上预先定义不同大小和比例的锚框，用锚框代替两阶段算法中的候选区域，只需对图像进行一次卷积处理就可以完成对象定位和分类，这类算法被统称为一阶段算法，典型的一阶段算法例如YOLO方法[3]，SSD方法[4]都具有较好性能。近几年Transformer[5]在自然语言处理（natural language processing，NLP）领域大放异彩，具有出色全局建模能力，许多研究者将它引入到计算机视觉领域，提出许多基于transformer的高性能视觉模型[6-7]。最典型的结构DETR[8]直接将目标检测视为集合预测问题，消除了非最大抑制和锚生成过程，简化了检测任务。文献[9]提出了一种基于选择性状态空间模型的序列模型Mamba。Mamba在长序列任务上优异性能与较低的计算复杂度，引起了学术界的广泛关注。视觉领域的研究者受Mamba在语言建模中成就启发，成功将Mamba适应到视觉领域，提出了基于Mamba的各种视觉模型[10-11]，在目标检测任务中取得了不错效果。经过不断发展，目标检测技术已经相当成熟且具有较高扩展性。

传统目标检测方法在不同数据集上能够取得较高性能，取决于一个关键因素，即需要人工对相关数据集进行精致标注，这一工程耗时耗力[12]。模型在封闭场景和小规模数据集上进行训练时，泛化性能相当有限。目标检测领域最常用的数据集COCO[13]包含80个类别，大型数据集LVIS[14]，包含1 203个类别，这些数据集只能涵盖开放场景中少部分类别。在开放场景中，目标类别繁杂，大部分需要检测的目标类别不包含在已经标注过的类别中。想要模型检测这些全新类别对象时，需要增加新类别标注进行重新训练。如何避免昂贵标注实现对大量未知类别目标进行检测是一个极具挑战性的问题。

为了应对开放场景检测任务的挑战，研究者们提出多种类型的任务。文献[15]提出开放集识别（open-set recognition，OSR）的概念，很快扩展到开放集目标检测任务（open-set object detection，OSOD），这种方式允许在测试阶段出现新的、训练阶段未见过的类别，相比传统监督学习，开放集识别更接近现实世界情况。OSOD在训练过程中不涉及任何与新类有关的额外辅助信息，很难准确识别新目标类别。为了能够帮助模型识别新类目标，文献[16]提出少样本目标检测示例（few-shot object detection，FSOD），通过引入极少量样本作为辅助，使模型能够捕获到新类别特征，在未来的任务中能正确检测出这个新类别。为了更进一步实现开放场景检测任务，文献[17]提出了零样本检测（zero-shot object detection，ZSD）任务。旨在训练期间对某些类别完全不提供训练样本来进行未知类别的预测[18]。为了实现这一目标，当前主流ZSD方法用固定的类别语义嵌入替换分类器可学习权重，根据未知类别语义描述从训练样本中提取与未知类别相关的特征实现未知类别推测，例如BERT[19]。语义嵌入缺乏与视觉特征对齐，导致在推理过程中，模型仅根据预定义词嵌入识别新类，限制了模型挖掘未见类视觉信息和关系的能力，导致这些方法在新类检测任务中性能不高。文献[20]提出开放词汇检测（open vocabulary object detection，OVD）任务并设计了检测模型OVR-CNN，通过借助大量与视觉相关的语言词汇数据作为辅助弱监督来覆盖更多的类别。近几年一些OVD改进方法可以直接利用在大规模图像文本对上预训练视觉语言模型（vision-language models，VLM），使检测器不再受限于带标注少数类别，提高检测器泛化能力，识别开放场景中未知物体。OVD相对于ZSD在性能上实现了巨大飞跃，OVD泛化能力具有极大的研究价值，近年来研究者们在开放词汇领域提出了大量检测方法，本研究旨在对OVD方法做分类总结与分析。

1 开放词汇目标检测概念

1.1 开放词汇学习

开放词汇学习主要目的是构建一个能够理解和处理大量词汇模型，不仅仅局限于固定单词[21]。开放词汇学习比传统闭集学习有更强的泛化性，闭

集学习只能识别预定义类别,开放词汇设置允许模型识别预定义类别之外的未知类别。为了实现这种泛化性,开放词汇学习利用语言词汇数据作为辅助监督让模型学习到更多词汇,这些词汇包含未知类别相关信息,可以帮助模型识别未知类别。

开放词汇学习旨在识别超出预定义标签空间的类别。它与早期提出的开放集学习(open-set learning, OSL)、开放世界学习(open-world learning, OWL)、零样本学习(zero-shot learning, ZSL)和少样本学习(few-shot learning, FSL)概念非常相似,都将类别信息分为已知类(基类)和未知类(新类),让模型在已知类标签空间中进行训练,具备识别未知类能力。OSL 目的是在测试过程中对已知类进行明确分类,将训练期间从未见过的对象标记为未知类,不能将未知类直接当作背景,OSL 任务不需要对未知类进行进一步细致分类[22];OWL 是一个动态学习过程,无需对模型进行再训练[23]。这一过程包括对已知对象进行分类,识别未知对象,由人工对未知对象进行标注,将新标注的类别添加到基类,通过这样的方式逐步学习新类别;ZSL 旨在识别训练期间未见过的对象,需要对未知对象进行明确分类[24]。ZSL 和开放词汇关键区别在于开放词汇可以使用语言词汇数据作为辅助监督;FSL 则通过引入一小部分新类的信息进行学习,使模型对新的、未见过的实例进行识别[25]。

开放词汇学习借助大量与视觉相关的语言词汇数据,比 OSL、OWL、ZSL、FSL 更一般、更实用、更有效,如图 1 所示。

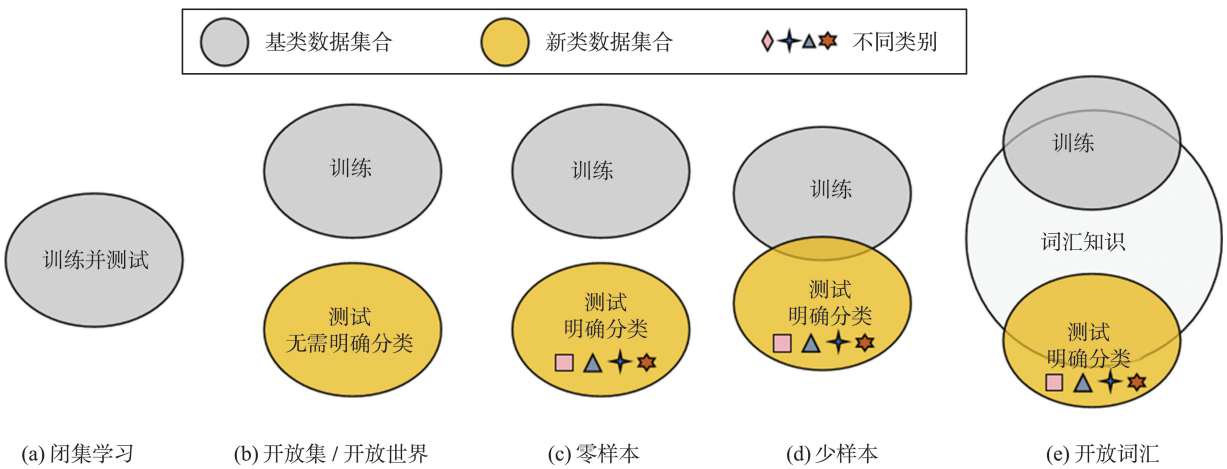

图 1　开放集/开放世界、零样本、少样本和开放词汇之间比较

Fig.1　Comparison between open-set/open-world, zero-shot, few-shot and open vocabulary

在 OSL 和 OWL 中,只需要识别出新类并将它们标记为“未知”;在 ZSL 和 FSL 中,除了要识别出新类,还必须将新类分类为特定的类别;在开放词汇学习中,通过借助语言词汇知识对新类进行分类。语言词汇知识可以是图像标题数据或预训练 VLM 的视觉文本嵌入。语言词汇知识不一定完全包含基类和新类,不会覆盖数据集中的所有类名。相反,语言词汇知识会包含基类和新类之外的类别信息,这一特性可以进一步扩展模型泛化能力。

基于 OSL、ZSL、FSL 的目标检测方法在检测新类时表现不佳,提出了面向开放词汇目标检测方法。这种方法使用额外低成本数据辅助训练或从预训练视觉语言模型中提取新知识,能使模型学习到更大的语言词汇表,帮助模型在更多的类之间进行泛化。

1.2　开放词汇目标检测

开放词汇目标检测(open-vocabulary object detection, OVD)就是利用开放词汇学习的思想,借助图像-文本知识在已知类数据上进行训练,完成未知类目标检测。利用大量额外数据获取足够的知识以覆盖更多的目标检测类别,把学习到的知识迁移到通用目标检测框架进一步训练,使封闭集目标检测器扩展到开放词汇目标检测器。OVD 不再受限于预先标注的类别,极大增强了检测器泛化能力,使检测器能够检测出未知类别目标。开放词汇目标检测技术范式如图 2 所示。

OVD 已成为一个有着巨大潜力的研究方向,针对 OVD 研究工作逐渐增加。尤其在近几年,提出了越来越多的大型视觉语言预训练模型,这些大模型促进了 OVD 快速发展。

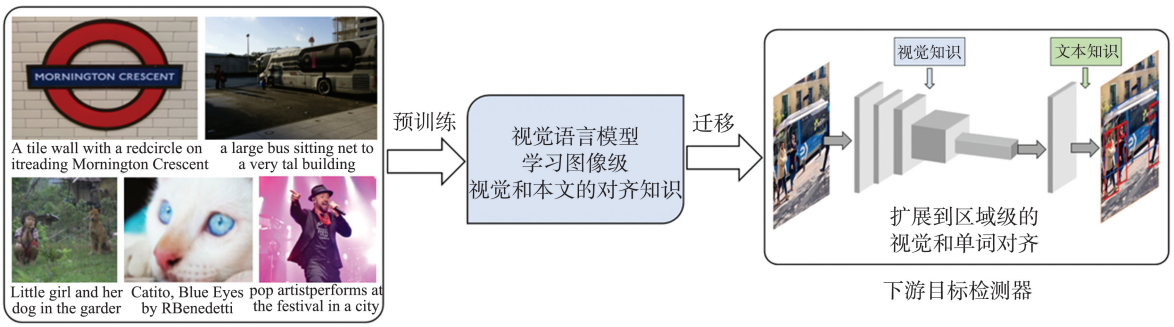

图 2 开放词汇目标检测范式

Fig.2 Open vocabulary object detection paradigm

2 视觉语言建模

人类学习本质上是多模态的,联合利用多种感官有助于更好理解和分析新信息。一些研究者提出了基于大规模视觉语言信息预训练的 VLM,它展现出了强大的零样本泛化能力。可以直接利用 VLM 能力,将闭集检测器扩展到开放词汇检测器。

2.1 大规模视觉语言预训练

大规模视觉语言预训练旨在赋予模型从多模态数据中学习有效信息的能力,指导模型学习视觉语言相关联信息,提高下游任务效果。好的视觉语言预训练可以使模型更好理解给定视觉输入的语义。

以前的一些工作,例如 VisualBERT[26]、ViLBERT[27]等,通过关注视觉和语句之间的密集联系或者使用复杂网络结构来实现视觉和语言联合。最近,在大规模图像-文本对上预训练的视觉语言模型 CLIP[28]在各种视觉任务上表现出显著的零样本性能。CLIP 通过对从互联网抓取的 4 亿个图像-文本对进行对比预训练,预训练只是简单预测哪个图像与哪个文本相对应,这是一种高效且可扩展的学习方法。在推理过程中,CLIP 没有使用分类头进行预测,用提示模板对类名进行填充后输入 CLIP 文本编码器得到对应的文本嵌入,通过图像编码器获得图像嵌入,在文本和图像嵌入两两之间计算余弦相似度,简单选择具有最高相似性得分的类作为预测。CLIP 仅使用简单的 transformer 结构进行视觉语言预训练,通过大规模数据训练证明了针对预测图片与标题配对问题仅靠简单的预训练任务就可以产生更强的可泛化模型。

为了进一步提高 CLIP 性能、泛化性以及适应下游任务的能力,研究者们针对 CLIP 提出了各种改进方法。SLIP[29]是一种结合自监督学习和 CLIP 预训练的多任务学习框架,明确证明了自监督学习可以帮助使用语言监督进行视觉表征学习;EVA-CLIP[30]结合表征学习、优化、数据增强等方法提升 CLIP 训练的效率和性能,EVA-CLIP 还将 CLIP 扩展到 180 亿参数,使模型在零样本图像分类中达到了更高精度;FLIP[31]是一种简单有效训练 CLIP 的方法,在训练过程中随机屏蔽和去除大部分图像块,这种机制使 CLIP 能在同样内存占用中学习到更多图像-文本对,提高了精度和速度;CLIPSelf[32]采用自蒸馏方式,不需要添加任何额外区域-文本对,将 CLIP 图像级识别能力调整为局部区域识别,使 CLIP 能够更好适应下游任务。

基于大规模视觉语言信息预训练的 VLM 将图像和语言词汇转化为视觉嵌入和文本嵌入对齐到相同的特征空间中,弥补了视觉和语言数据的差距,这是开放词汇检测任务的基础。模型可以利用对齐特征改进下游任务检测器,使检测器能够在开放场景中识别新类。

2.2 从封闭集检测转向开放词汇检测

受联合视觉语言建模启发,研究者们提出了开放词汇对象检测的概念,它利用图像标题数据连接新类语义和视觉区域,目标就是让模型看到更多的词汇。开放词汇方法能够在未知场景中检测出更多类别,研究者们开始逐渐将注意力转向开放词汇检测。CLIP 提出并开源之后,ViID[33]第一个使用 CLIP 知识来构建开放词汇对象检测器。越来越多的研究工作专注提高开放词汇检测器性能并构建新基准。

目前许多开放词汇方法通过利用 VLM 中学习到的图像文本对齐来构建开放词汇表对象检测器,这样可以根据需求轻松扩展到不同场景,无需收集

相关数据或产生额外的注释代价。如表 1 所示,自 2021 年以来,开放词汇学习相关研究工作数量显著增加[21]。

表 1　开放词汇工作统计
Table 1　Open vocabulary work statistics

年份	开放词汇相关研究工作数量	开放词汇目标检测研究工作数量
2021	2	1
2022	36	17
2023	87	24

3　开放词汇目标检测方法

OVD 的一种方法是直接利用日常生活中获得的大规模图像文本数据,这些数据包含丰富的词汇,十分适合开放词汇检测。另一种常见方法是将固定分类器权重替换为 VLM 模型文本嵌入。如图 3 所示,该方法结合了 VLM 和检测器骨干网学习到的视觉特征知识,使检测器可以通过语义相关的文本嵌入来检测新类。

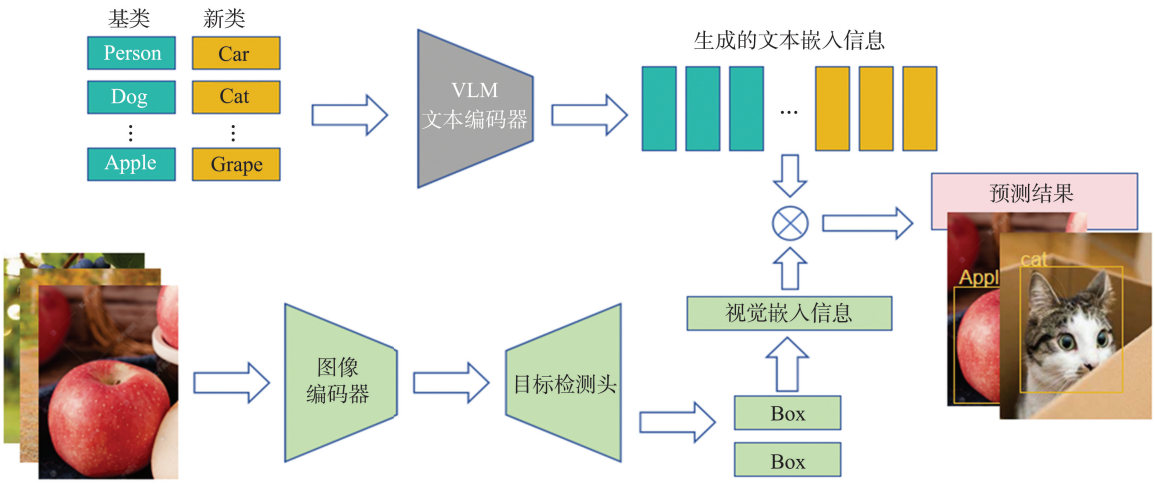

图 3　开放词汇目标检测中常见架构
Fig.3　Common architectures in open vocabulary object detection

3.1　基于大规模外部图像文本数据

3.1.1　区域-文本预训练

大规模图像文本数据包含足够多的知识,足以涵盖大部分数据集的类别。一些方法采用包含数百万个图像文本对的图像标题数据集进行预训练。这种预训练只实现了图像级视觉文本对齐,目标检测中一张图像通常包含多个对象,需要将图像级对齐调整到区域级对齐才能用于检测器检测新类。

针对开放词汇检测任务,OVR-CNN 模型通过图像-标题数据来学习视觉语言映射,利用预训练视觉编码器与学习到的映射关系微调 Faster RCNN 框架,将大量对齐知识注入目标检测模型。OVR-CNN 整体预训练过程为图像级且并未添加局部区域对应关系;为了将图像级对齐调整为局部区域对齐,文献[34]提出了模型 LocOv,通过引入局部匹配技术学习新类标签。这是一个两阶段模型,第一阶段将图像中的对象与文本标题中相应类标签进行匹配,使用预训练文本嵌入模型处理标题以获取部分词特征。使用 Faster R-CNN 对图像进行处理获得目标区域特征,以一种类无关的方式改善图像区域-标题匹配。第二阶段利用已知类标注对模型进行调优,让模型适用于对象检测。模型检测精度相较于 OVR-CNN 取得了很大提升,但两阶段方式会导致整体效率不高;为了更好获取区域-文本的对应关系,文献[35]提出了一个在线提议挖掘方法 MEDet,利用检测数据集和图像标题数据集的小批量数据共同训练目标检测器,在此期间对标题文本进行解析以获取单词文本,这些单词中会包含新类信息。将训练中 RPN 生成的区域提议和单词文本进行匹配,实现在图像-文本数据集上挖掘区域-单词关系,学习丰富的提议级视觉语言知识;为了直接实现区域级预训练,文献[36]提出了一种区域感知图像-文本预训练方法 RO-ViT,该方法使用 ViT 结构,在预训练阶段引入了一个裁剪位置嵌入模块。裁剪位置嵌入模块会随机裁剪和调整位置嵌入区域大小,导致模型将图像视为从更大未知图像中截取的区域,能更好匹配下游检测任务。通过这种方式解决了图像级预训练和开放词汇对象检测之间的差距,但随机裁剪方法会导致丢失大量的上下文信息;文献[37]认为以往方法过于依赖提炼区域级对齐预训练,文章中提出一个简单而有效模型 VLDet,VLDet 不依赖昂贵的基础注释,直接从图像

-文本对训练目标检测器。该模型的关键思想是利用类无关定位头生成类无关区域提议,将区域特征和文本嵌入都视为集合,内积相似度作为对齐分数,利用二分图匹配将每个图像区域在图像-文本对的监督下与词汇对齐;文献[38]提出了一种对比预训练方法DITO,使检测器从有噪声的大规模图像-文本对中学习并适应检测任务。DITO在检测器中使用一种移位窗口学习方法减轻ViT检测器中窗口注意力偏差,学习位置敏感信息。DITO不需要引入伪标记或额外损失。检测器利用预训练获得的图像-文本信息,导致检测器存在原始数据中的缺陷和偏见;为了获得准确对应关系,文献[39]提出CoDet模型,该模型采用了一种全新区域文本挖掘思想。许多图像-文本对的标题中包含相同概念,CoDet每次联合一组图像-文本对,从中发现区域-词对,利用这些区域-词对和候选区域计算相似度,通过文本引导来定位相似对象,发掘新的区域-词对。

额外的图像-文本对中潜在包含新类信息,学习更多图像-文本对可以提高模型识别新类能力,增强检测器泛化能力。模型需要在图像-文本对数据上训练学习图像和文本对应关系,将学习到的对应关系迁移到检测器上进行微调适应下游检测任务,预训练过程需要更多计算成本对额外数据集进行训练,对资源有限情况不友好。

3.1.2 伪标签文本对齐

通用对象检测器在常用数据集上进行训练学习到的词汇量有限,想要构建鲁棒的通用对象检测器就需要扩展到更大的标签空间和更大的训练数据集。获取数千个类别的注释成本非常高,一些方法考虑引入额外图像-文本数据明确构建伪区域-文本对解决这个问题。

文献[40]针对上述问题提出了Detic模型,该模型使用ImageNet21K分类图像数据集和目标检测数据集一起对检测模型进行联合训练,对于带有框标签的图像,使用Faster R-CNN进行训练,对于图像级标注,只对固定区域提议的特征进行分类训练。Detic利用了图像分类数据丰富性,以更大的词汇表进行推理,这意味着Detic能够检测出多样化的目标类别,不仅仅局限于有限的类别;为了获取更细粒度的伪区域-文本对,文献[41]提出了RegionCLIP模型,利用CLIP将CC3M数据集图像区域和模板文本进行匹配获取伪区域-文本对,将伪文本对输入模型预训练图像编码器学习区域信息,使用人工注释的数据集对视觉编码器进行微调以适应不同检测任务。RegionCLIP可以获取大量的伪区域-文本对,其中会包含大量的噪声数据,对检测器产生严重干扰;文献[42]为了使用VLM从未标记图像中挖掘语义丰富的伪标签并以边界框的形式生成,提出了VL-PLM模型。该模型将一个未标记图像输入到一个两阶段类别不可知检测器中获得区域建议,对应区域图输入到CLIP图像编码器中获得视觉嵌入,使用CLIP文本编码器和模板提示生成类名文本嵌入。对于每个区域,通过点积计算区域嵌入和文本嵌入之间相似度分数,根据类别不可知检测器分数和VLM模型分数生成最终伪标签,进一步减少了噪声。VL-PLM在获取视觉嵌入时,输入裁剪图像区域,会导致对象丢失重要的上下文信息;文献[43]提出了一种从大规模图像标题对中自动生成不同对象伪边界框注释的方法PB-OVD,该方法利用图像编码器和文本编码器从图像-标题对中提取视觉和文本嵌入。不同的是,这些视觉和文本嵌入会经过交叉注意力转化为多模态特征,为图像中每个感兴趣对象输出激活图,选择与激活图有最大重叠的对象提议作为伪边界框标签;以往伪标签方法都需要额外的图像-文本数据来生成为标签,增加了额外训练成本。针对这个问题,文献[44]提出DST-Det模型,该模型引入了一种动态自训练策略为新类生成伪标签,不需要引入额外数据。作者认为检测器只在基类中训练并把基类识别为前景,背景中潜在包含大量新类对象,在PRN生成提议之后,将负面提议区域特征输入到CLIP编码器计算与文本的匹配分数,当分数大于设定阈值就作为新类伪标签加入基类一起进行下一轮训练。DST-Det是一个动态生成伪标签的迭代优化模型,可以直接应用在CLIPSelf上获得更强性能。

伪标签文本对齐方法利用除真实标注之外的图像-文本对构建伪区域-文本对,这是一种硬对齐方式,即一个区域只能对应一个单词。伪标签能够有效提高模型对新类检测能力,大部分伪标签生成方法整体流程是两阶段的,即利用图像文本数据生成伪标签,用伪标签和目标检测数据集训练检测器。

3.2 基于预训练视觉语言模型

3.2.1 知识蒸馏

知识蒸馏目的是将视觉知识直接蒸馏到闭集检测器中。视觉语言模型在预训练阶段获得的知识量远大于闭集检测器的知识量,一个简单方法就是将VLM中的新类别知识提炼出来加入到基类中训练检测器。

一种早期解决方法是文献[33]提出的 ViLD 模型,通过知识蒸馏方式,将图像和文本知识从预训练开放词汇图像分类模型中传递到两阶段检测器中,解决 OVD 中训练数据有限问题。ViLD 使用 VLM 图像编码器计算裁剪区域图像嵌入,使用文本编码器获取类别文本嵌入,将文本嵌入作为区域分类器。ViLD 利用 RoI Align 将预先计算好的提议送到检测器中获取区域嵌入并最小化区域嵌入和图像嵌入之间距离,使检测框嵌入和 VLM 推断的图像文本嵌入对齐,区域裁剪方式能够使图像编码器适应区域嵌入;文献[45]认为大量研究只从 VLM 中提取对象级知识转移到检测器,忽略了全局场景理解。为了解决这个问题,提出 OADP 金字塔架构,使用全局、块和对象 3 个蒸馏模块构成了 1 个分层蒸馏金字塔,这种机制可以将更多样化的知识从 CLIP 转移到检测器,弥补了对象级蒸馏中缺失的关系信息;相比较 ViLD 两阶段检测器,文献[46]提出使用更高效的单阶段检测器模型 HierKD。单阶段检测器中缺少与类别无关的对象建议,阻碍了对未见对象的知识蒸馏,导致性能严重下降。HierKD 采用了分层知识蒸馏方法,引入全局级语言-视觉知识蒸馏模块。通过将全局级知识蒸馏与实例级知识蒸馏相结合,同时学习可见类和未见类知识;文献[47]对经典的 DETR 模型进行改进提出了 OV-DETR,为了将 DETR 转变为开放词汇检测器,在训练时,将从 CLIP 中获得的嵌入信息送入 transformer 解码器,将学习目标制定为输入查询与相应对象之间的二进制匹配,学习精确对应关系以便在测试时推广到未见过的查询;文献[48]指出除了单个区域嵌入的蒸馏之外,应该明确学习共存的视觉概念以鼓励模型理解场景,提出 BARON 方法。对于每个提议附近区域进行采样,形成多组区域包,将包中区域嵌入投影到句子形成的词嵌入空间,输入 CLIP 文本编码器获得区域包嵌入。裁剪每组区域包图像输 CLIP 图像编码器获得裁剪区域嵌入。通过训练对齐这些裁剪区域嵌入和区域包嵌入,除了使用简单新类类名进行文本提炼之外,一些模型还利用了更细粒度的信息,包括属性、标题和对象关系等;文献[49]提出 PCL 模型指出模型需要细粒度标签来提取关于新对象更丰富的知识。PCL 利用图像文本模型生成了许多从不同角度描述对象实例的标题,这些伪标题标签为知识蒸馏提供了更加密集的样本。

知识蒸馏是一种有效设计方式,模型能够直接利用 VLM 中的知识,不需要进行繁琐预训练。模型使用知识蒸馏结合各种策略将提取的知识转移到闭集检测器中,这种设计方式能够让检测器获取识别新类的能力。VLM 自身识别能力受预训练数据规模限制,利用 VLM 进行知识蒸馏的目标检测模型识别能力受 VLM 限制。

3.2.2　迁移学习

迁移学习方法与知识蒸馏方法不同。迁移学习主要利用 VLM 图像编码器直接对检测数据进行微调,或通过冻结 VLM 图像编码器提取视觉特征用于下游检测任务。

文献[50]提出 F-VLM 模型是一种基于迁移学习的方法,它仅使用冻结的 VLM,无需知识蒸馏和定制预训练。F-VLM 利用冻结的 CLIP 图像编码器作为图像特征提取骨干,仅训练检测头。一方面将检测头结果与 CLIP 文本编码器进行对比计算得出检测器分数,另一方面将 RPN 生成的提议作用在图像编码器的特征图上获取区域嵌入,将区域嵌入与文本嵌入做对比计算得出 VLM 分数,两个分数进行加权计算得出最终的检测分数;文献[51]提出 OWL-ViT 模型可以直接利用 CLIP,在 CLIP 的图像编码器后增加检测头,对中等大小的检测数据进行微调迁移到检测任务。OWL-ViT 网络采用类似 DETR 结构,为了简化网络删除了解码器,对于图像编码器的输出,直接用线性层作为轻量级对象分类器,用 MLP 头部作为定位头,将预训练编码器转移到开放词汇对象检测。

知识蒸馏方法需要将 ROI 重复送入 VLM 图像编码器中,这一过程会消耗大量内存;迁移学习方法通常直接利用 VLM 图像编码器,只需要少量的额外计算资源就能够取得较好效果,这取决于 VLM 强大的泛化能力。

3.2.3　提示学习

提示(Prompt)是一段文本或语句。在图像识别领域中,Prompt 则可以是一个图片描述、标签或分类信息。利用提示建模是一种有效技术,通过将学习到的提示合并到基础模型中,可以使基础模型适应各种下游任务。

获取类名的文本嵌入需要向预训练 VLM 文本编码器输入提示来生成,用它作为区域分类器监督检测器训练。这种模式成功的关键因素是制定合理提示,需要对文字描述进行仔细调整和巧妙设计。为了避免费力的提示工程,文献[52]提出 DetPro 模型将图像中负面建议(一般为背景信息)纳入到训练中,提出了一种上下文分级方案,将图像前景中正面建议分离出来,进行针对性提示训

练;CLIP是以场景为中心进行训练,目标检测是以对象为中心,为了使CLIP文本空间适应以对象为中心的图像,文献[53]提出了一种区域提示学习模型PromptDet,它将一系列可学习的向量添加到文本输入中,这些提示向量不对应于任何实际的具体单词,就好像是一个虚拟标记,帮助文本嵌入空间更好地对齐以对象为中心的视觉表示。提示模板中加入详细描述,有助于减轻词汇歧义;文献[54]提出了一个高效开放词汇对象检测模型Prompt-OVD,它利用CLIP中类嵌入作为提示,引导transformer解码器检测基类和新类对象。Prompt-OVD还提出了RoI掩码注意和RoI修剪技术有助于充分利用CLIP的零样本分类能力;预训练视觉语言模型是在整个图像上训练,将其应用于区域识别任务时难免会发生偏差,为了缓解这个问题,文献[55]提出CORA模型,利用区域提示将CLIP调整为区域分类器。区域提示添加在CLIP图像编码器骨干网中,对图像编码器进行微调,减轻全图像特征和区域特征之间分布差距。锚预匹配将调整后CLIP作为分类器,在锚框匹配前先对锚框进行预分类以获取泛化的对象定位;背景框中潜在包含大量开放词汇类别物体,已有方法忽视背景类别多样性,导致背景中未标注的潜在类别特征具有模糊性。为了解决这个问题,文献[56]提出LBP方法,从背景中挖掘潜在新类别特征,学习背景提示。为了建立背景物体与潜在类别对应关系,对所有背景候选区域视觉嵌入特征执行聚类,每个聚簇中心代表一种潜在背景类别,训练时跟据聚簇中心与背景区域的特征计算相似度,通过设定阈值为训练产生的背景框赋予潜在类别标签,帮助模型更好对背景中的潜在类别进行区分。

Prompt能起到一个提示的作用,帮助模型学习到与提示相关的知识,不需要太多计算资源且效果较好。提示学习没有使用额外的训练数据,模型泛化性依靠从VLM中提取知识的能力,增加提示对模型性能的提高是有限的。

4 数据集、评估指标与模型性能比较

4.1 数据集与评估指标

OVD方法主要关注COCO、LVIS和V3Det[57]数据集。V3Det是最近提出的超大规模数据集,它包含的类别超过了13 000个,非常适合用于开放词汇检测。除了3个常用数据集,一些方法在验证模型的零样本泛化能力时会直接使用其他数据集进行预测,例如Pascal VOC[58]数据集和Objects365[59]数据集。需要将数据集按照类别分为基类与新类,只有标记为基类的注释才能参与训练,新类只能用作预测。3个常用数据集的OVD设置如表2所示。

表2 开放词汇对象检测常用数据集

Table 2 Common datasets for open vocabulary object detection 单位:个

数据集	类别		训练集		验证集		
	基类数量	新类数量	图象数量	基类对象数量	图象数量	基类数量	新类对象数量
COCO	48	17	107 761	665 387	4 836	28 538	4 614
LVIS	866	337	100 170	1 264 884	19 809	243 507	1 200
V3Det	6 709	6 495	132 437	836 203	29 821	136 479	83 950

COCO数据集共包含80个类别,其中48个作为基类,17个作为新类。对于使用COCO数据集的OVD设置,COCO的评估指标为边框A_P,用A_{Pn}表示新类A_P,A_{Pb}表示基类A_P。

LVIS数据集是为长尾目标检测任务设计的,共有1 203个类别,其中866个frequent类和common类作为基类,而377个rare类充当新类。LVIS v1的OVD评估指标为掩码A_P。其中,A_{Pr}表示新类A_P。

V3Det数据集是一个庞大的视觉检测数据集,包含13 204个类别、243×10^3个图像和$1\ 753\times10^3$个框注释,在开放词汇设置中,V3Det数据集评估指标与COCO数据集相同。

4.2 开放词汇检测方法性能对比

OVD重点关注模型检测新类的能力。在性能对比时,重点需要关注各方法在新类上的精度。本节中所对比的数据均来自已经公开的论文。

4.2.1 在COCO数据集上的表现

在COCO数据集上执行开放词汇设置评估指标分别是新类和基类的框A_P。开放词汇主要关注新类A_P,其他A_P对于结果评估并没有太大影响。如表3所示,在基于ResNet骨干网的方法中,LBP方法在新类上取得了最高值,$A_P=37.8$。这说明LBP提出在背景中挖掘新类的方法非常有效,之前的方法忽视了背景类中潜在的大量新类信息,如何有效从背景中提取新类是进一步要研究的问题。进一步使用增强的ResNet骨干网,在训练中引入额外数据,CORA方法取得了最高值,$A_P=43.1$。

CORA 利用视觉提示将 CLIP 微调为区域级分类器,这足以说明区域级图像-文本匹配对开放词汇检测任务的重要性。在基于 ViT 方法中,使用 ViT-B/16 作为骨干网训练时,DST-Det 方法取得了最高值,A_P = 41.3。DST-Det 中部分基本思想与 LBP 类似,需要从背景中挖掘新类信息,DST-Det 还可以直接应用在 CLIPSelf 上进行训练,强大的 VLM 能使 DST-Det 获得更强泛化性能。

表 3　COCO 数据集上的检测表现

Table 3　Detection performance on the COCO dataset

骨干网	方法/模型	文本模型	检测器	额外数据集	A_{Pn}	A_{Pb}	A_{Pall}
RN50	PB-OVD[43]	CLIP ViT-B/32	Mask R-CNN	COCO Cap、VG、SBU	30.8	46.1	42.1
	CoDet[39]	CLIP	CenterNet2	COCO Cap	30.6	52.3	46.6
	CORA[55]	CLIP	DAB-DETR	Ø	**35.1**	35.5	35.4
RN50-C4	Region CLIP[41]	CLIP RN50	Faster R-CNN	CC3M	**31.4**	57.1	50.4
	OADP[45]	CLIP	Faster R-CNN	COCO Cap	30.0	53.3	47.2
	OV-DETR[47]	CLIP	Def-DETR	Ø	29.4	61.0	52.7
	Detic[40]	CLIP	CenterNet2	COCO Cap	27.8	47.1	45.0
RN50-FPN	ViLD[33]	CLIP ViT-B/32	Mask R-CNN	Ø	27.6	59.5	51.3
	LocOv[34]	CLIP	Faster R-CNN	COCO Cap	28.6	51.3	45.7
	VLDet[37]	CLIP	Faster R-CNN	COCO Cap	32.0	50.6	45.8
	VL-PLM[42]	CLIP	Mask R-CNN	COCO Cap	34.4	60.2	53.5
	HierKD[46]	CLIP	ATSS	COCO Cap	20.3	51.3	43.2
	BARON[48]	CLIP	Faster R-CNN	Ø	34.0	60.4	53.5
	LBP[56]	CLIP	Faster R-CNN	Ø	**37.8**	58.7	53.2
	F-VLM[50]	CLIP RN50	Mask R-CNN	Ø	28.0	—	39.6
	PromptDet[53]	CLIP	Mask R-CNN	LAION	26.6	—	50.6
RN50×4	Region CLIP[41]	CLIP RN50x4	Faster R-CNN	CC3M	39.3	61.6	55.7
	CORA[55]	CLIP RN50x4	DAB-DETR	COCO Cap	**43.1**	60.9	56.2
ViT-B/16	DST-Det[44]	CLIPSelf	Mask R-CNN	Ø	**41.3**	—	—
	RO-ViT[36]	CLIP	Mask R-CNN	Ø	30.2	—	41.5
	CLIPSelf[32]	CLIP	Mask R-CNN	Ø	37.6	54.9	50.4
	DITO[38]	DITO Pretrain	Faster R-CNN	Ø	38.6	—	48.5
	Prompt-OVD[54]	CLIP ViT-L/14	Def-DETR	Ø	30.6	63.5	54.9

注:黑体数字为新类最佳性能指标;Ø 表示没有引入额外数据集;—表示没有明确值。

4.2.2　在 LVIS 数据集上的表现

在 LVIS 数据集上使用 A_{Pr} 表示方法检测新类别的能力。如表 4 所示,在使用标准 ResNet-50 骨干网时,LBP 方法依然优于其他方法,取得了最高值,A_{Pr} = 24.1。LBP 在两个数据集上优于同结构的方法,足以说明该方法的有效性。当使用更强的 ResNet 骨干网时,DST-Det 方法取得了最高值,A_{Pr} = 34.5。DST-Det 在训练过程中不需要引入额外数据。DST-Det 和 LBP 本质上都是从检测出的背景区域中挖掘新类信息,这说明图像背景中潜在包含大量新类,许多方法忽略了这个关键问题,如何从背景中挖掘新类信息变得尤为重要。在基于 ViT 方法中,同样使用 ViT-B/16 作为骨干,DITO 取得了最高值,A_{Pr} = 32.5。在表格后半部分,对比了一些方法在更强骨干网下的性能,其中 DITO 方法使用了 ViT-L/16 作为骨干网,引入额外公共数据集 DataComp-1B 进行训练,取得了最高值,A_{Pr} = 40.4,远超其他方法。

4.2.3　在 V3Det 数据集上的表现

V3Det 数据集 V3Det 是最近提出的超大规模数据集,包含的类别超过了 13 000 个,基本类超过 6 000 个。在 V3Det 数据集上训练代价过大,很少有方法在这个数据集上进行训练,对比数据不够充足。如表 5 所示,就现有数据对比,DST-Det 方法在该数据集上取得了最优的结果,使用标准 ResNet-50 骨干网训练,A_{Pn} = 7.2。替换更强的骨干网之后 DST-Det 检测性能得到显著提高,A_{Pn} = 13.5。

表 4 LVIS 数据集上的检测表现

Table 4 Detection performance on the LVIS dataset

骨干网	方法/模型	文本模型	检测器	额外数据集	A_{Pr}	A_{Pc}	A_{Pf}	A_{Pall}
RN50	VLDet[37]	CLIP	CenterNet2	CC3M	21.7	29.8	34.3	30.1
	CoDet[39]	CLIP	CenterNet2	CC3M	**23.4**	30.0	34.6	30.7
	CORA[55]	CLIP	DAB-DETR	Ø	22.2	—	—	—
RN50-C4	MEDet[35]	CLIP	Faster R-CNN	CC3M、CC	**22.4**	—	—	34.4
	Region CLIP[41]	CLIP RN50	Faster R-CNN	CC3M	17.1	27.4	34.0	28.2
	OADP[45]	CLIP	Faster R-CNN	Ø	21.9	28.4	32.0	28.7
	OV-DETR[47]	CLIP	Def-DETR	Ø	17.4	25.0	32.5	26.6
RN50-FPN	ViLD[33]	CLIP ViT-B/32	Mask R-CNN	Ø	16.6	24.6	30.3	25.5
	Detic[40]	CLIP	Mask R-CNN	IN-21K	17.8	26.3	31.6	26.8
	BARON[48]	CLIP	Faster R-CNN	Ø	22.6	27.6	29.8	27.6
	LBP[56]	CLIP	Faster R-CNN	Ø	**24.1**	29.5	32.8	29.9
	F-VLM[50]	CLIP RN50	Mask R-CNN	Ø	18.6	—	—	24.2
	DetPro[52]	CLIP ViT-B/32	Mask R-CNN	Ø	19.8	25.6	28.9	25.9
	PromptDet[53]	CLIP	Mask R-CNN	LAION	21.4	23.3	29.3	25.3
RN50×64	DST-Det[44]	CLIP RN50x64	Mask R-CNN	Ø	**34.5**	—	—	—
	F-VLM[50]	CLIP RN50x64	Mask R-CNN	Ø	32.8	—	—	34.9
ViT-B/16	RO-ViT[36]	CLIP	Mask R-CNN	Ø	28.0	—	—	30.2
	CLIPSelf[32]	CLIP	Mask R-CNN	Ø	25.3	—	—	—
	DITO[38]	DITO Pretrain	Faster R-CNN	ALIGN	32.5	—	—	34.0
	Prompt-OVD[54]	CLIP ViT-L/14	Def-DETR	Ø	23.1	—	—	24.2
ViT-H/14	OWL-ViT[51]	CLIP	DETR	Ø	**25.6**	—	—	34.7
ViT-L/16	RO-ViT[36]	CLIP	Mask R-CNN	LAION-2B	32.4			32.9
	CLIPSelf[32]	CLIP	Mask R-CNN	Ø	34.9	—	—	—
	DITO[38]	DITO Pretrain	Faster R-CNN	DataComp-1B	**40.4**	—	—	37.7
Swin-B	VLDet[37]	CLIP	CenterNet2	CC3M	26.3	39.4	41.9	38.1
	CoDet[39]	CLIP	CenterNet2	CC3M	**29.4**	39.5	43.0	39.2
Swin-L	PCL[49]	CLIP ViT-L/14	Def-DETR	Ø	**24.7**	—	—	38.7

注:黑体数字为新类最佳性能指标;Ø 表示没有引入额外数据集;—表示没有明确值。

表 5 V3Det 数据集上的检测表现

Table 5 Detection performance on the V3Det dataset

骨干网	方法/模型	文本模型	检测器	额外数据集	A_{Pn}	A_{Pb}	A_{Pall}
RN50	Detic[40]	CLIP	CenterNet2	IN-21K	6.7	30.2	17.1
RN50	Region CLIP[41]	CLIP	Faster R-CNN	CC3M	3.1	22.1	12.6
RN50	DST-Det[44]	CLIP	Mask R-CNN	Ø	**7.2**	—	—
RN50×64	DST-Det[44]	CLIP	Mask R-CNN	Ø	**13.5**	—	—

注:黑体数字为新类最佳性能指标;Ø 表示没有引入额外数据集;—表示没有明确值。

4.2.4 迁移到其他数据集上的表现

为了验证开放词汇方法的有效性和泛化能力,一些方法将在 COCO 或 LVIS 数据集上训练好的模型,不经过任何微调,直接使用其他数据集进行检测,测试模型泛化能力。如表 6 所示,VL-PLM 在 COCO 数据集上进行训练,在另外 3 个数据集上泛化表现最佳。OV-DETR 和 CoDet 在 LVIS 数据集上进行训练,OV-DETR 在 VOC 数据集中表现最佳,CoDet 在 COCO 和 Object365 上整体表现最佳。从现有试验数据可以看出,开放词汇目标检测各种方法的确能够获取一定的零样本泛化能力,足以证明这些开放词汇方法的有效性。需要针对开放词汇检测的性能进行提高,使模型具有更强的泛化能力。表中 VOC 表示 PASCAL VOC,是一个用于图像识别的数据集,包括 20 个对象类别。Objects365 是一个大规模对象检测数据集,它包含超过 60 万个训练图像,365 个对象类别,以及超过 1 000 万个高质量边界框。

表 6　模型迁移到其他数据集上的性能

Table 6　The performance of the model when transferred to other datasets

训练数据集	方法/模型	COCO			Object365			LVIS	VOC	
		A_P	A_{P50}	A_{P75}	A_P	A_{P50}	A_{P75}	AP50	A_{P50}	A_{P75}
COCO	VLDet[37]	—	—	—	—	—	—	10.0	61.7	—
	PB-OVD[43]	—	—	—	—	6.9	—	8.0	59.2	—
	VL-PLM[42]	—	—	—	—	**10.9**	—	**22.2**	**67.4**	—
LVIS	RO-ViT[36]	—	—	—	14.0	22.3	14.9	—	—	—
	Detic[40]	—	—	—	—	21.5	—	—	—	—
	CoDet[39]	**39.1**	57.0	**42.3**	**14.2**	20.5	15.3	—	—	—
	ViLD[33]	36.6	55.6	39.8	11.8	18.2	12.6	—	72.2	56.7
	BARON[48]	36.2	55.7	39.1	13.6	**21.0**	14.5	—	76.0	58.2
	OV-DETR[47]	38.1	**58.4**	41.1	—	—	—	—	**76.1**	**59.3**
	F-VLM[50]	32.5	53.1	34.6	11.9	19.2	12.6	—	—	—
	DetPro[52]	34.9	53.8	37.4	12.1	18.8	12.9	—	—	—

注:黑体数字为最佳性能指标;—表示没有明确值。

4.2.5　主要方法所需计算资源对比

OVD 方法从大量图像-文本对数据和 VLM 中提取知识训练下游目标检测器,导致 OVD 方法参数量和计算量较大,训练过程需要消耗大量计算资源,如表 7 所示。不利于 OVD 在开放场景中应用。

表 7　模型计算资源对比

Table 7　Model computing resource comparison

方法/模型	计算资源	训练时间/h
VL-PLM[42]	GPU A100×8	—
CoDet[39]	GPU A100×8	20
CORA[55]	GPU A100×8	27
OV-DETR[47]	GPU V100×8	—
VLDet[37]	GPU V100×8	17

注:—表示没有明确值。

4.2.6　轻量化开放词汇方法参数量对比

为了在资源有限环境下应用 OVD 方法,一些研究开始关注轻量化开放词汇设置,例如 OVLW-DETR[60]。现有轻量化方法直接使用冻结 VLM 作为整个模型骨干网,添加可训练检测头适应目标检测任务。当前针对轻量化 OVD 研究较少,是一个有潜力的研究方向。如表 8 所示。

表 8　轻量化模型参数量对比

Table 8　Comparison of the parameter size of lightweight models

方法/模型	参数/10^7	推理延迟/s
DST-Det[44]	22.9	—
OVLW-DETR[60]	12.0	6.09

注:—表示没有明确值。

5　总结与展望

本研究围绕目标检测领域新兴的方法-开放词汇目标检测,系统阐述了该方法的研究现状。对关键技术进行了综述,总结了各方法的优缺点,在多个数据集上对各个方法的泛化性能进行了公平详细地分析和比较。

当前开放词汇工作已经取得显著性进展,模型泛化能力显著增强,针对开放词汇对象检测的研究仍在初级阶段。在基于大规模外部数据训练方法中,区域-文本预训练方式需要在大量数据上训练提取图像和文本对应关系,预训练过程缓慢,需要耗费大量计算资源,对资源有限用户不友好;伪标签文本对齐方式,直接裁剪图像学习区域级伪标签会导致丢失大量上下文信息,生成伪标签质量不高,泛化能力有限。在基于 VLM 的方法中,知识蒸馏和迁移学习都利用 VLM 中的知识微调下游检测器,知识蒸馏为了适应区域对齐需要反复将区域提议送入 VLM 编码器,大大增加了模型计算量;迁移学习直接利用 VLM 编码器获取图像级特征,与区域级检测不匹配,性能提升有限;提示学习在大模型微调中是一种有效方法,高效提示微调需要复杂的提示设计步骤,模型泛化能力仍受 VLM 限制。

现阶段无论是基于大规模外部数据训练方法,还是基于预训练 VLM 方法,大多在解决图像级别预训练与实例级别检测任务之间的粒度差异问题。OVD 方法提取区域-单词之间的对应关系较弱并且嘈杂,zero-shot 能力仍然受数据规模限制。设计高效方法消除虚假区域-单词对负面影响,提高关联质量,是一个重要挑战。增加数据规模一定程度上可以提升模型效果,开放词汇方法涉及大量数据和 VLM 模型,本身需要大量计算资源,盲目增加数据规模会降低模型效率,针对开放词汇任务设计轻量化检测模型是一个重点问题。

未来开放词汇检测逐步完善,在纯图像检测之

外应用开放词汇或许将成为主流趋势,关于开放词汇未来发展,有以下几个方向:

(1)更强的零样本泛化能力:VLM 使用大量数据进行预训练,它们的泛化能力仍然存在上限。与 VLM 相比,大语言模型包含更多的文本概念,涵盖更多的对象类别。如何利用好大语言模型知识实现更强的泛化性能是未来需要探索的方向。

(2)轻量化检测模型:现有开放词汇检测模型计算量和参数量较大,模型训练需要大量计算资源,推理速度不佳。利用迁移学习思想构建轻量化开放场景检测模型是一个重点研究趋势。

(3)开放词汇视频分析[61]:在实际应用中,视频数据使用更频繁。视频存在额外时间轴,这使得视频中对象比图像中对象具有更加丰富多样的关系。为视频中所有类别收集足够的注释是不切实际的,需要利用开放词汇的零样本泛化能力来解决视频分析问题。

(4)探索开放词汇 3D 场景理解:与图像和视频相比,3D 数据注释成本更高,尤其是对于密集场景。将 VLM 中的知识从 2D 场景中对齐到 3D 场景,实现 3D 场景泛化是未来的研究方向。

(5)与人-物交互[62]相结合:将开放词汇扩展到现实场景,结合多模态语言大模型,实现用户意图推理和场景互动交互式检测。

(6)统一开放词汇检测与分割:建立一个通用模型,实现开放词汇任务统一,同时完成 2D 和 3D 开放词汇感知[63]。当前适用于所有任务和数据集的通用基础模型几乎没有被触及,可以保持期待。

参考文献:

[1] GIRSHICK R, DONAHUE J, DARRELl T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(1): 142-158.

[2] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[3] REDMON J, FARHADI A. Yolov3: An incrementalimprovement [EB/OL]. (2018-04-08)[2024-05-28]. https://arxiv.org/abs/1804.02767.

[4] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: single shot multibox detector[C]//Proceedings of the Computer Vision-ECCV 2016 Workshops. Berlin, Germany: Springer, 2016: 21-37.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2024-05-28]. https://arxiv.org/abs/1706.03762.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transforme-rs for image recognition at scale[EB/OL]. (2021-06-03)[2024-05-28]. https://arxiv.org/abs/2010.11929.

[7] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 10012-10022.

[8] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proc-eedings of the Computer Vision-ECCV 2020 Workshops. Berlin, Germany: Springer, 2020: 213-229.

[9] GU A, DAO T. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. (2024-05-31)[2024-06-17]. https://arxiv.org/abs/2312.00752.

[10] ZHU L, LIAO B, ZHANG Q, et al. Vision mamba: efficient visual representation learning with bidirect-ional state space model[EB/OL]. (2024-02-10)[2024-06-17]. https://arxiv.org/abs/2401.09417.

[11] HUANG T, PEI X, YOU S, et al. Localmamba: visual state space model with windowed selective scan[EB/OL]. (2024-03-14)[2024-06-17]. https://arxiv.org/abs/2403.09338.

[12] ZOU Z, CHEN K, SHI Z, et al. Object detection in 20 years: a survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[13] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]//Proceedings of the Computer Vision-ECCV 2014 Workshops. Berlin, Germany: Springer, 2014: 740-755.

[14] GUPTA A, DOLLAR P, GIRSHICK R. Lvis: a da-taset for large vocabulary instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 5356-5364.

[15] SCHEIRER W J, DE REZENDE ROCHA A, SAPKOTA A, et al. Toward open set recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(7): 1757-1772.

[16] KANG B, LIU Z, WANG X, et al. Few-shot object detection via feature reweighting[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 8420-8429.

[17] BANSAL A, SIKKA K, SHARMA G, et al. Zero-shot object detection[C]//Proceedings of the Computer Vision- ECCV 2018 Workshops. Berlin, Germany: Springer, 2018: 384-400.

[18] ZHU P, WANG H, SALIGRAMA V. Zero shot detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(4): 998-1010.

[19] DEVLIN J, CHANG M, LEE K, et al. Bert: pretraining of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2024-06-17]. https://arxiv.org/abs/1810.04805.

[20] ZAREIAN A, ROSA K D, HU D H, et al. Open vocabulary object detection using captions[EB/OL]. (2021-05-14)[2024-06-17]. https://arxiv. org/abs/2011.10678.

[21] WU J, LI X, XU S, et al. Towards open vocabulary learning: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 5092-5113.

[22] GENG C, HUANG S, CHEN S. Recent advances in open set recognition: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3614-3631.

[23] JOSEPH K J, KHAN S, KHAN F S, et al. Towards open world object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2021: 5830-5840.

[24] ROMERA-PAREDES B, TORR P. An embarrassingly simple approach to zero-shot learning[C]//Proceedings of the 32nd International Conference on Machine Learning. New York, USA: ACM, 2015: 2152-2161.

[25] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: a survey on few-shot learning[J]. ACM Computing Surveys, 2020, 53(3): 1-34.

[26] LI L H, YATSKAR M, YIN D, et al. Visualbert: a simple and performant baseline for vision and language[EB/OL]. (2019-08-09)[2024-06-17]. https://arxiv.org/abs/1908.03557.

[27] LU J, BATRA D, PARIKH D, et al. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[EB/OL]. (2019-08-06)[2024-06-17]. https://arxiv.org/abs/1908.02265.

[28] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. New York, USA: ACM, 2021: 8748-8763.

[29] MU N, KIRILLOV A, WAGNER D, et al. Slip: self-supervision meets language-image pre-training[C]//Proceedings of the Computer Vision-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 529-544.

[30] SUN Q, FANG Y, WU L, et al. Evaclip: improved training techniques for clip at scale[EB/OL]. (2023-03-27)[2024-06-17]. https://arxiv.org/abs/2011.10678.

[31] LI Y, FAN H, HU R, et al. Scaling language-image pretraining via masking[C]//Proceedings of the I-EEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 23390-23400.

[32] WU S, ZHANG W, XU L, et al. Clipself: vision-transformer distills itself for open-vocabulary dense prediction[EB/OL]. (2024-01-24)[2024-06-17]. https://arxiv.org/abs/2310.01403.

[33] GU X, LIN T Y, KUO W, et al. Open-vocabulary object detection via vision and language knowledgedistillation[EB/OL]. (2022-05-12)[2024-06-17]. https://arxiv.org/abs/2104.13921.

[34] BRAVO M A, MITTAL S, BROX T. Localized vision-language matching for open-vocabulary object detection[C]//Proceedings of the Pattern Recognition: 44th DAGM German Conference. Berlin, Germany: Springer, 2022: 393-408.

[35] CHEN P, SHENG K, ZHANG M, et al. Open vocabulary object detection with proposal mining and prediction equalization[EB/OL]. (2022-11-24)[2024-06-17]. https://arxiv.org/abs/2206.11134.

[36] KIM D, ANGELOVA A, KUO W. Region-aware pretraining for open-vocabulary object detection with vision transformers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 11144-11154.

[37] LIN C, SUN P, JIANG Y, et al. Learning object-language alignments for open-vocabulary object detection[EB/OL]. (2022-11-27)[2024-06-17]. https://arxiv.org/abs/2211.14843.

[38] KIM D, ANGELOVA A, KUO W. Detection-oriented image-text pretraining for open-vocabulary detec-tion[EB/OL]. (2023-09-29)[2024-06-17]. https://arxiv.org/abs/2310.00161v1.

[39] MA C, JIANG Y, WEN X, et al. Codet: co-occurrence guided region-word alignment for open-vocabulary object detection[J]. Advances in Neural Information Processing Systems, 2024, 36: 71078-71094.

[40] ZHOU X, GIRDHAR R, JOULIN A, et al. Detecting twenty-thousand classes using image-level supe-rvision[C]//Proceedings of the Computer Vision-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 350-368.

[41] ZHONG Y, YANG J, ZHANG P, et al. Regionclip: region-based language-image pretraining[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 16793-16803.

[42] ZHAO S, ZHANG Z, SCHULTER S, et al. Exploiting unlabeled data with vision and language models for object detection[C]//Proceedings of the Computer Vi-

sion-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 159-175.

[43] GAO M, XING C, NIEBLES J C, et al. Open vocabulary object detection with pseudo bounding-box labels [C]//Proceedings of the Computer Vision-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 266-282.

[44] XU S, LI X, WU S, et al. Dst-det: simple dynamic self-training for open-vocabulary object detection [EB/OL]. (2024-04-01) [2024-06-17]. https://arxiv.org/abs/2310.01393.

[45] WANG L, LIU Y, DU P, et al. Object-aware distillation pyramid for open-vocabulary object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 11186-11196.

[46] MA Z, LUO G, GAO J, et al. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 14074-14083.

[47] ZANG Y, LI W, ZHOU K, et al. Open-vocabulary detr with conditional matching [C]//Proceedings of the Computer Vision-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 106-122.

[48] WU S, ZHANG W, JIN S, et al. Aligning bag of regions for open-vocabulary object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 15254-15264.

[49] CHO H C, JHOO W Y, KANG W, et al. Open-vocabulary object detection using pseudo caption la-bels [EB/OL]. (2023-03-23) [2024-06-17]. https://arxiv.org/abs/2303.13040.

[50] KUO W, CUI Y, GU X, et al. Fvlm: open-vocabulary object detection upon frozen vision and lang-uage models [EB/OL]. (2023-02-23) [2024-06-17]. https://arxiv.org/abs/2209.15639.

[51] MINDERER M, GRITSENKO A, STONE A, et al. Simple open-vocabulary object detection [C]//Proceedings of the Computer Vision-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 728-755.

[52] DU Y, WEI F, ZHANG Z, et al. Learning to prompt for open-vocabulary object detection with vision-language model [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 14084-14093.

[53] FENG C, ZHONG Y, JIE Z, et al. Promptdet: tow-ards open-vocabulary detection using uncurated images[C]//Proceedings of the Computer Vision-ECCV 2022 Workshops. Berlin, Germany: Springer, 2022: 701-717.

[54] SONG H, BANG J. Prompt-guided transformers forend-to-end open-vocabulary object detection [EB/OL]. (2023-03-25) [2024-06-17]. https://arxiv.org/abs/2303.14386.

[55] WU X, ZHU F, ZHAO R, et al. CORA: adapting clip for open-vocabulary detection with region prompting and anchor prematching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 7031-7040.

[56] LI J, ZHANG J, LI J, et al. Learning background prompts to discover implicit knowledge for open vocabulary object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2024: 16678-16687.

[57] WANG J, ZHANG P, CHU T, et al. V3det: vast vocabulary visual detection dataset [EB/OL]. (2023-10-05) [2024-06-17]. https://arxiv.org/abs/2304.03752.

[58] EVERINGHAM M, ESLAMI S M, GOOL L, et al. The pascal visual object classes challenge: a retrospective [J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[59] SHAO S, LI Z, ZHANG T, et al. Objects365: a large-scale, high-quality dataset for object detection [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 8430-8439.

[60] WANG Y, SU X, CHEN Q, et al. OVLW-DETR: open-vocabulary light-weighted detection transformer [EB/OL]. (2024-07-15) [2024-07-26]. https://arxiv.org/abs/2407.10655.

[61] GAO K, CHEN L, ZHANG H, et al. Composition-al prompt tuning with motion cues for open-vocabulary video relation detection[EB/OL]. (2023-02-01) [2024-06-17]. https://arxiv.org/abs/2302.00268.

[62] LI L, XIAO J, CHEN G, et al. Zero-shot visual relation detection via composite visual cues from large language models[J]. Advances in Neural Information Processing Systems, 2024, 36: 50105-50116.

[63] ZHU C, CHEN L. A survey on open-vocabulary detection and segmentation: past, present, and future [EB/OL]. (2024-04-15) [2024-06-17]. https://arxiv.org/abs/2307.09220.

（编辑：陈燕）