RAPID-OVD：基于区域分类与伪标签生成流程优化的开放词汇目标检测框架

朱锐[1),2)] 许凯瑞[1)] 刘博涵[1)] 卢昭伉[1)] 曹端瑜[1)] 张璇[1),2)]

[1)](云南大学软件学院, 昆明 650500)

[2)](云南省软件工程重点实验室（云南大学软件学院）, 昆明 650500)

摘 要 开放词汇目标检测旨在使模型能够识别训练过程中未见过的新类别,在自动驾驶、智能监控和医疗诊断等领域具有重要应用价值。然而,当前方法面临伪标签质量不佳、区域形变及少样本泛化能力不足三个重要挑战。为此,本文提出了一种基于区域分类与伪标签生成流程优化的开放词汇目标检测框架 RAPID-OVD。该框架采用模块化设计,包含三个核心组件:首先,构建了基于键值缓存机制的 RAPID-CLIP 模块。通过在少样本训练集上构建键值缓存,利用相似度检索和特征融合技术,将少样本领域知识与 CLIP 预训练知识有机结合。引入指数函数进行相似度变换和残差连接融合策略,显著增强了模型在少样本场景下的区域分类能力,仅需 20 轮训练即可达到接近 200 轮的性能水平,训练效率提升 10.2 倍。其次,设计了自适应填充策略(APS)。根据目标区域的形态特征动态选择最优填充方式。对于面积小于 3000 像素且长宽比小于 1.4 的目标采用 1:1 填充比例,对于面积较大且长宽比较大的目标采用 1:1.4 填充比例,在保持目标原始形态的同时确保输入尺寸一致性,有效解决区域提议变形问题。最后,建立了系统性的四阶段伪标签生成与优化流程。通过融合 RPN 定位分数与 RAPID-CLIP 分类分数的多源信息,以及置信度阈值化和非极大值抑制的多阶段筛选技术,显著提升伪标签的准确性和可靠性。实验结果表明,在 COCO-OVD 数据集上,RAPID-OVD 在新类别检测方面达到 40.5%AP_{novel},超越当前最优方法 OV-DQUO 1.3 个百分点,相比直接基线 VL-PLM 提升 8.2 个百分点;在基类检测上保持 56.3%AP_{base} 的竞争力。在 LVIS-OVD 数据集上,稀有类别检测达到 26.5%AP_r,常见类别达到 34.9% AP_c,频繁类别达到 38.1% AP_f,整体性能 34.7% AP。消融实验表明，自适应填充策略使小物体检测能力相对增幅达 29%,RAPID-CLIP 模块在 COCO-OVD 验证集上的区域分类准确率达到 73.75%,相比零样本 CLIP 提升约 9 个百分点，验证了各组件的有效性。

关键词 视觉语言模型；伪标签挖掘；开放词汇目标检测；无标注数据利用；少样本分类；

中图法分类号 TP18

RAPID-OVD：An Open-Vocabulary Object Detection Framework Based on Region Classification and Optimized Pseudo-Label Generation Process

ZHU Rui[1),2)] XU Kai-Rui[1)] LIU Bo-Han[1)] LU Zhao-Kang[1)] CAO Duan-Yu[1)] ZHANG Xuan[1),2)]

[1)](School of Software, Yunnan University, Kunming 650500)

[2)](Yunnan Key Laboratory of Software Engineering（School of Software, Yunnan University), Kunming, 650500)

Abstract Open-vocabulary object detection aims to enable models to recognize novel object categories that are unseen during the training process, demonstrating significant application value in diverse fields including autonomous driving, intelligent monitoring, and medical diagnosis. However, current state-of-the-art approaches

本课题得到国家自然科学基金地区科学项目基金(62362067)、云南省自然科学基金基础研究面上项目基金(202101AT070004,202001BB050031)、云南省院士专家工作站项目基金(202205AF150006)、云南省软件工程重点实验室开放基金项目(2023SE205)资助.朱锐，博士，副教授，中国计算机学会（CCF）专业会员,主要研究领域为计算机视觉等. 许凯瑞，硕士研究生，中国计算机学会（CCF）学生会员,主要研究领域为计算机视觉. 刘博涵，硕士研究生，中国计算机学会（CCF）学生会员,主要研究领域为计算机视觉.卢昭伉，硕士研究生，中国计算机学会（CCF）学生会员,主要研究领域为计算机视觉.曹端瑜，硕士研究生，中国计算机学会（CCF）学生会员,主要研究领域为自然语言处理.张璇（通信作者），博士，教授，中国计算机学会（CCF）专业会员,主要研究领域为自然语言处理等.

face three critical technical challenges that significantly limit their practical deployment: poor pseudo-label quality resulting from inherent limitations of vision-language models, severe region deformation issues caused by naive resizing operations that distort object geometry, and insufficient few-shot generalization capability when adapting to new categories with limited training examples. To comprehensively address these fundamental challenges, we propose RAPID-OVD, an open-vocabulary object detection framework based on region classification enhancement and optimized pseudo-label generation process. The proposed framework adopts a modular design architecture consisting of three core technical components that work synergistically. First, we construct a RAPID-CLIP module based on a key-value caching mechanism. This module builds comprehensive key-value caches on few-shot training sets and leverages sophisticated similarity retrieval and feature fusion techniques to organically combine domain-specific few-shot knowledge with CLIP pre-trained knowledge. By introducing exponential functions for similarity transformation and carefully designed residual connection fusion strategies, the module significantly enhances the model's region classification capability in few-shot scenarios. Remarkably, it achieves performance levels comparable to 200 training epochs with merely 20 epochs, thereby improving training efficiency by an impressive factor of 10.2 times. Second, we design an Adaptive Padding Strategy (APS) that dynamically selects optimal padding approaches based on the morphological characteristics of target regions. Specifically, for targets with areas smaller than 3000 pixels and aspect ratios less than 1.4, a 1:1 padding ratio is strategically applied to maintain visual integrity; for targets with larger areas and higher aspect ratios, a 1:1.4 padding ratio is employed to better preserve geometric structure. This adaptive strategy effectively addresses region proposal deformation issues while simultaneously maintaining original target proportions and ensuring input size consistency required by the vision encoder. Third, we establish a systematic four-stage pseudo-label generation and optimization pipeline. Through intelligent multi-source information fusion of RPN localization scores and RAPID-CLIP classification scores, combined with sophisticated multi-stage filtering techniques including confidence thresholding and non-maximum suppression operations, the pipeline significantly improves both the accuracy and reliability of generated pseudo-labels. Extensive experimental results demonstrate the superior performance of our approach. On the widely-used COCO-OVD benchmark dataset, RAPID-OVD achieves 40.5% AP_{novel} for novel category detection, surpassing the current best-performing method OV-DQUO by 1.3 percentage points and substantially improving upon the direct baseline VL-PLM by 8.2 percentage points, while maintaining highly competitive performance of 56.3% AP_{base} on base categories. On the more challenging LVIS-OVD dataset with long-tail distribution, the framework achieves 26.5% AP_r for rare category detection, 34.9% AP_c for common categories, 38.1% AP_f for frequent categories, and an overall performance of 34.7% AP. Comprehensive ablation studies validate the effectiveness of each component: the adaptive padding strategy achieves a substantial 29% relative improvement in small object detection capability, while the RAPID-CLIP module attains 73.75% region classification accuracy on the COCO-OVD validation set, representing approximately 9 percentage points improvement over zero-shot CLIP baseline. These comprehensive experimental results collectively demonstrate that our method provides a systematic and effective solution for open-vocabulary object detection, achieving a favorable balance between detection performance and computational efficiency.

Key words Vision-Language Model; Pseudo-Label Mining; Open-Vocabulary Object Detection; Unlabeled Data Utilization; Few-shot Classification

1 引言

随着计算机视觉技术的快速发展，目标检测作为计算机视觉领域的核心任务之一，在自动驾驶、智能监控、医疗诊断等众多应用场景中发挥着重要作用[1]。传统的目标检测方法通常采用封闭词汇设定，即模型只能识别训练时预定义的固定类别集合。然而，现实世界中的视觉场景复杂多样，新的目标类别不断涌现，封闭词汇检测方法难以适应这种动态变化的需求，严重限制了其在实际应用中的灵活性和可扩展性。

为了突破传统封闭词汇检测的局限性，开放词汇目标检测（Open-Vocabulary Object Detection, OVD）应运而生。开放词汇目标检测旨在使模型能够识别和定位训练过程中未见过的新类别，从而实现对任意类别的目标检测能力。这一技术的出现具有重要的理论价值和实际意义：从理论角度，它扩展了传统目标检测的边界，使模型具备了更强的泛化能力和适应性；从应用角度，它能够显著降低新场景部署的标注成本，提高系统的实用性和可维护性。

近年来，大规模视觉语言预训练模型，特别是 CLIP[2]（Contrastive Language-Image Pre- training）的出现，为开放词汇目标检测提供了强有力的技术支撑。CLIP 通过在大规模图像-文本对上进行对比学习，学习到了丰富的跨模态语义表示，展现出卓越的零样本识别能力。基于 CLIP 的开放词汇检测方法[3–5]能够利用其强大的语义理解能力，通过文本描述的方式引入新类别的先验知识，从而实现对未见类别的有效检测。这种范式的转变使得目标检测从”为每个任务训练专门模型”向”单一模型适应多种任务”的方向发展[6]。

尽管开放词汇目标检测技术取得了显著进展，但仍面临诸多技术挑战。首先，伪标签质量不佳问题严重影响模型性能[7, 8]。在开放词汇检测中，新类别的监督信号主要来源于利用视觉语言模型生成的伪标签。然而，由于视觉语言模型本身的局限性以及目标检测任务的复杂性，生成的伪标签往往存在分类错误、定位不准确等问题，这些低质量的伪标签会误导模型训练，导致检测性能显著下降。其次，区域提议变形问题对特征表示产生不利影响。现有方法通常将不规则的目标区域直接缩放到固定尺寸以适应预训练模型的输入要求，这种简单的处理方式会导致目标的形状和比例发生严重变形，特别是对于细长物体和小物体，变形现象更为明显，从而影响特征提取的准确性和后续分类的可靠性。最后，少样本场景下的泛化能力不足制约了模型的实际应用[9, 10]。在实际部署中，新类别往往只有少量的标注数据可用，现有方法在这种少样本条件下的性能表现不稳定，难以达到实用化的要求。

针对上述技术挑战，现有研究提出了多种解决方案，但仍存在明显不足。在伪标签优化方面，一些研究尝试通过多模型集成、不确定性估计等方法提升伪标签质量[11, 12]，但这些方法往往计算开销较大，且对于复杂场景的适应性有限。在区域变形处理方面，部分工作采用注意力机制或空间变换网络来缓解变形影响[13, 14]，但这些方法主要关注全局特征调整，对于局部形态保持的效果有限。在少样本学习方面，现有方法多采用元学习或自监督学习策略[15, 16]，但在开放词汇检测的特定场景下，这些通用方法的有效性仍需进一步验证。

基于以上分析，本文提出了一种基于区域分类与伪标签生成流程优化的开放词汇目标检测框架 RAPID-OVD（Region-Adaptive Pseudo-label Improved Detection for Open-Vocabulary Detection）。本文的主要创新与贡献包括以下三个方面：

(1)提出了系统性的伪标签生成流程优化方法。通过四阶段伪标签生成与优化流程，显著提升了伪标签的准确性和可靠性。该方法不仅能够有效过滤错误标签，还能够增强正确标签的置信度，为模型训练提供高质量的监督信号。实验结果表明，优化后的伪标签生成流程使得新类别检测的平均精度达到 40.5%AP_{novel}，超越了现有的主流方法。

(2) 设计了自适应填充策略（Adaptive Padding Strategy, APS），有效解决了区域提议变形问题。该策略根据目标区域的形态特征，动态选择最优的填充方式，在保持目标原始比例的同时，确保输入尺寸的一致性。通过精心设计的形态分析算法，APS 能够自动识别目标的几何特

征，并据此调整填充参数，显著改善了细长物体和小物体的检测效果，相对性能提升达到 29%。

(3)将基于键值缓存的 CLIP 微调方法引入开放词汇目标检测领域，构建了 RAPID-CLIP 模块。该模块通过在少样本训练集上构建键值缓存，利用相似度检索和特征融合技术，显著增强了模型在少样本场景下的分类能力。与传统的全参数微调方法相比，RAPID-CLIP 不仅参数效率更高，而且在区域分类准确率上提升约 6 个百分点，有效提升了开放词汇检测的泛化能力。

从技术架构角度，RAPID-OVD 采用两阶段的设计理念：首先使用类别无关的提议生成器预测候选区域，然后通过 RAPID-CLIP 模块进行精确的区域分类。这种解耦设计不仅简化了训练过程，还提高了系统的模块化程度和可维护性。整个框架在保持计算效率的同时，实现了检测性能的显著提升。

论文的其余部分安排如下：第 2 节回顾相关工作，包括视觉语言模型、CLIP 微调技术和开放词汇目标检测的最新进展； 第 3 节详细阐述 RAPID- OVD 框架的设计理念和技术实现；第 4 节通过大量实验验证所提方法的有效性，并进行深入的分析和讨论；第 5 节总结全文并展望未来的研究方向。

2 相关工作

开放词汇目标检测作为计算机视觉领域的前沿研究方向，融合了视觉语言预训练、迁移学习和目标检测等多个技术领域的最新进展。本章将从视觉和语言模型、CLIP 微调技术以及开放词汇目标检测方法三个方面，系统回顾相关研究的发展脉络，分析现有方法的优势与不足。

2.1 视觉语言模型

视觉语言模型（Vision-Language Models, VLMs） 的发展经历了从任务特定模型到通用表示学习的重要转变。早期的多模态学习工作主要关注特定任务的模型设计，如图像描述生成[17]、视觉问答[18]和指称表达理解[19]等。这些方法通常需要为每个任务单独设计网络架构和训练策略，限制了模型的通用性和可扩展性。

近年来，大规模视觉语言预训练模型的出现标志着该领域的重要突破。CLIP（Contrastive Language-Image Pre-training） 作为里程碑式的工作，通过在 4 亿个图像-文本对上进行对比学习，学习到了丰富的跨模态语义表示。CLIP 采用简洁而有效的对比学习框架：将图像和文本分别编码到共同的表示空间，通过最大化匹配图像-文本对的相似度，同时最小化不匹配对的相似度来进行训练。这种设计使得 CLIP 展现出卓越的零样本迁移能力，在多个视觉任务上无需额外训练即可达到有竞争力的性能。

CLIP 的成功引发了视觉语言预训练模型的研究热潮。ALIGN[20]通过扩大训练数据规模至 12 亿图像-文本对，进一步提升了模型的性能表现。ALBEF[21]引入了动量蒸馏和伪标签技术，在保持训练效率的同时提升了模型的表示能力。BLIP[22]和 BLIP-2[23]系列工作进一步推进了视觉语言理解与生成的统一建模，展现了强大的多模态理解和推理能力。

然而，现有的视觉语言模型在目标检测任务中仍面临诸多挑战。首先，全局-局部表示差异问题：大多数 VLMs 是在图像级别的数据上进行预训练，其学习到的表示更适合全局图像理解，而目标检测需要精确的局部区域表示，这种表示层次的不匹配影响了模型在检测任务上的性能[24]。其次，定位能力有限问题：虽然 VLMs 在图像分类任务上表现优异，但其对目标精确定位的能力相对较弱，这主要源于预训练过程中缺乏显式的空间定位监督信号[25]。最后，计算效率问题：大规模 VLMs 通常具有庞大的参数量和计算需求，直接应用于目标检测任务会带来显著的计算开销，限制了其在实际应用中的部署。

2.2 CLIP微调

为了充分发挥 CLIP 在下游任务中的潜力，研究人员提出了多种微调策略来适应特定任务的需求。这些方法可以大致分为提示学习、适配器微调和缓存机制三大类。

提示学习方法通过优化输入文本提示来改善 CLIP 的性能。CoOp（Context Optimization）[26]是该方向的开创性工作， 通过将文本提示中的上下文词汇设置为

可学习参数，在少样本场景下显著提升了 CLIP 的分类性能。

CoOp 将原始的手工制作提示“a photo of a [CLASS]”扩展为“a photo of a [V1] [V2] ... [Vm] [CLASS]”， 其中[V1]到[Vm]是可学习的连续向量。实验表明，CoOp 在 16 样本设置下相比零样本 CLIP 提升了 2.62 个百分点。

适配器微调方法通过在预训练模型中插入轻量级的适配器模块来实现任务适应。CLIP-Adapter[27]在 CLIP 的视觉和文本编码器中分别引入了轻量级的多层感知机（MLP）适配器，通过残差连接将适配后的特征与原始特征相结合。该方法在 ImageNet 16 样本设置下达到了 63.59%的准确率， 比零样本 CLIP 提升了 3.26 个百分点。同时，CLIP-Adapter 的训练时间显著短于 CoOp，在保持性能的同时提升了训练效率。Side-Tuning[28]提出了并行微调的策略，通过在主干网络旁边添加辅助分支，在不修改预训练权重的情况下实现任务适应。这类方法的优势在于能够保持预训练模型的通用性，同时为特定任务提供必要的适应性调整。

缓存机制方法代表了一种全新的微调范式。Tip-Adapter[29]提出了一种无需训练的 CLIP 适应方法，通过构建键值缓存模型来存储少样本训练集的知识。Tip-Adapter 将少样本训练集的视觉特征作为键（keys），对应的标签 one-hot 编码作为值（values），在推理时通过特征相似度检索相关知识并与 CLIP 的预测结果融合。该方法在无需任何训练的情况下即可达到 62.03%的准确率，超越了零样本 CLIP 的 60.33%。

尽管上述微调方法在图像分类任务上取得了显著成功，但在目标检测场景中的应用仍存在局限性。现有方法主要针对全图分类进行设计，对于区域级的特征表示和空间定位能力的提升有限。此外，大多数方法未充分考虑目标检测中的区域变形问题和伪标签质量优化，同时这也是本文提出方法所解决的问题。

2.3 开放词汇目标检测

开放词汇目标检测旨在使目标检测模型能够识别训练过程中未见过的新类别，是传统封闭词汇检测向更通用人工智能的重要演进。

根据技术路线的不同，现有的开放词汇目标检测方法可以分为四大类[30]：知识蒸馏方法、伪标签方法、基于区域感知的训练方法和迁移学习方法。

知识蒸馏方法通过将大型视觉语言模型的知识迁移到轻量级检测器中实现开放词汇能力。ViLD （ Vision and Language knowledge Distillation）[31]作为该方向的代表性工作， 提出了一种两阶段的知识蒸馏框架： 首先训练一个类别无关的区域提议网络， 然后通过蒸馏损失将 CLIP 的区域表示能力迁移到检测器的分类头。ViLD 在 COCO 开放词汇检测基准上达到了 27.6%AP_{novel} 的性能。这类方法的优势在于能够有效利用大型 VLMs 的丰富语义知识，但存在知识蒸馏效率不高和师生模型容量差异较大的问题。

伪标签方法通过利用视觉语言模型在未标注数据上生成伪标签来扩充训练数据。VL- PLM（ Vision and Language guided Pseudo-Label Mining）[32]提出了一套完整的伪标签生成流水线， 通过融合 CLIP 分数和检测器的目标性分数来优化提议质量， 并通过重复应用检测头来移除冗余提议。该方法在 COCO-OVD 检测基准上取得了 34.4%AP_{novel} 的优异性能，相比 ViLD 提升了 6.8 个百分点。MarvelOVD[25]通过实时集成检测器能力与 VLM 指导来优化学习过程，实现了检测性能的进一步提升。这类方法能够有效利用大规模无标注数据和视觉语言模型的语义理解能力来扩充训练样本，显著提升新类别的检测性能，但其效果十分依赖伪标签的生成质量，且在标签噪声处理、训练稳定性和计算资源消耗方面仍面临挑战。

基于区域感知的训练方法通过设计特殊的训练策略来增强模型的区域理解能力。GLIP（Grounded Language-Image Pre- training）[33]将检测和定位任务统一在一个框架中， 通过生成区域-单词对来提供显式监督。该方法将传统的检测损失重新表述为区域-文本匹配问题， 有效地将视觉定位与语言理解相结合。Detic[34]通过使用图像分类数据来扩展检测器的词汇表， 在 21K 类别上训练的模型展现出强大的开放词汇检测能力。这类方法的优势在于能够直接在目标检测框架内进

行端到端训练，但需要大量的弱监督数据和精心设计的损失函数。

迁移学习方法直接利用预训练的视觉语言模型作为特征提取器。F-VLM（Frozen Vision-Language Model）[35]将预训练的 VLM 参数完全冻结，仅训练新增的检测头，在保持预训练知识的同时实现了开放词汇检测能力。RegionCLIP[36]通过区域特征池化将图像级的 CLIP 特征扩展到区域级，并通过伪标签训练进一步优化性能。这类方法的计算效率较高，但受限于预训练模型的局部特征判别能力。

除去上述提到的这四种方法，新兴的特征空间语义操作方法通过在预训练模型的特征空间直接进行语义推理，减少了对复杂几何变换和显式区域操作的依赖，为开放词汇检测、分割等任务提供了新的解决思路。OVSeg(Open-Vocabulary Semantic Segmentation with Mask-Adapted CLIP)[37]识别了预训练 CLIP 在处理遮罩图像时存在的域偏移问题，提出了 mask-adapted 架构来适配 CLIP 对遮罩区域的处理能力。该方法通过从 image-caption 数据集中挖掘大规模多样化的 mask-category 对进行专门训练，在保持 CLIP 开放词汇泛化能力的同时提升其对遮罩区域的语义理解。CLIP- DIY(CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free)[38]采用完全无需额外训练的策略，通过多尺度 dense inference 直接利用 CLIP 的分类能力进行逐像素预测，并结合无监督目标定位方法提供前景-背景空间指导，实现了 training-free 的开放词汇语义分割。DiffSegmenter(Diffusion Model is Secretly a Training-free Open Vocabulary Semantic Segmenter)[39]创新性地发现文本到图像扩散模型在生成过程中隐含学习了丰富的目标形状和语义对应关系，通过提取扩散模型 U-Net 中的交叉注意力图来定位目标区域、利用自注意力图建立像素间语义关联，避免了传统的 CLIP 特征适配过程。这类方法能够有效利用大规模预训练模型的丰富语义表示实现开放词汇的零样本泛化，但其性能上界受限于预训练特征空间的表示质量，且在细粒度边界精度、小目标检测能力和推理计算效率方面仍存在改进空间。

综合上述相关工作的分析，当前开放词汇目标检测领域仍面临三个重要技术挑战有待解决。首先，伪标签质量不佳问题：现有方法生成的伪标签往往存在分类错误和定位不准确等问题，这主要源于视觉语言模型在复杂空间推理和局部区域理解方面的固有局限性，以及缺乏系统性的质量控制机制。虽然部分工作如 VL-PLM 和 MarvelOVD 尝试通过改进筛选策略来提升伪标签质量，但其最终AP_{novel}徘徊在 32-39%的水平，表明仍有显著改进空间。其次，区域提议变形问题：当前方法普遍采用简单的缩放操作将不规则区域调整为固定尺寸，导致目标形状严重失真，特别是对细长物体和小物体的影响更为显著。尽管一些研究引入了注意力机制或空间变换来缓解此问题，但这些方法主要关注全局调整，对局部形态保持的效果仍然有限。最后，少样本场景下的泛化能力不足：现有的 CLIP 微调方法多针对全图分类任务设计，在区域级特征表示和少样本适应方面存在明显不足，限制了其在实际部署中面对新类别时的性能表现。

3 方法

3.1 RAPID-OVD框架总体架构

RAPID-OVD 框架旨在通过系统性优化伪标签生成流程来提升开放词汇目标检测性能。与现有方法主要关注单一技术环节的改进不同，RAPID- OVD 采用端到端的设计理念，从区域提议生成到最终伪标签优化的完整流程进行协同优化，有效解决了开放词汇目标检测中伪标签质量不佳、区域变形和少样本泛化能力不足三个核心技术挑战。如图 1 所示，RAPID-OVD 框架采用模块化设计，由三个核心组件构成：

(1)两阶段类别无关提议生成器，负责从输入图像中生成可能包含目标对象的候选区域；

(2)区域特征增强模块，通过自适应填充策略保持目标原始形态；

(3)RAPID-CLIP 分类模块，基于键值缓存机制提升少样本场景下的分类能力。

这三个组件协同工作，形成完整的伪标签生成流程：首先从未标注图像中提取候选区域，然

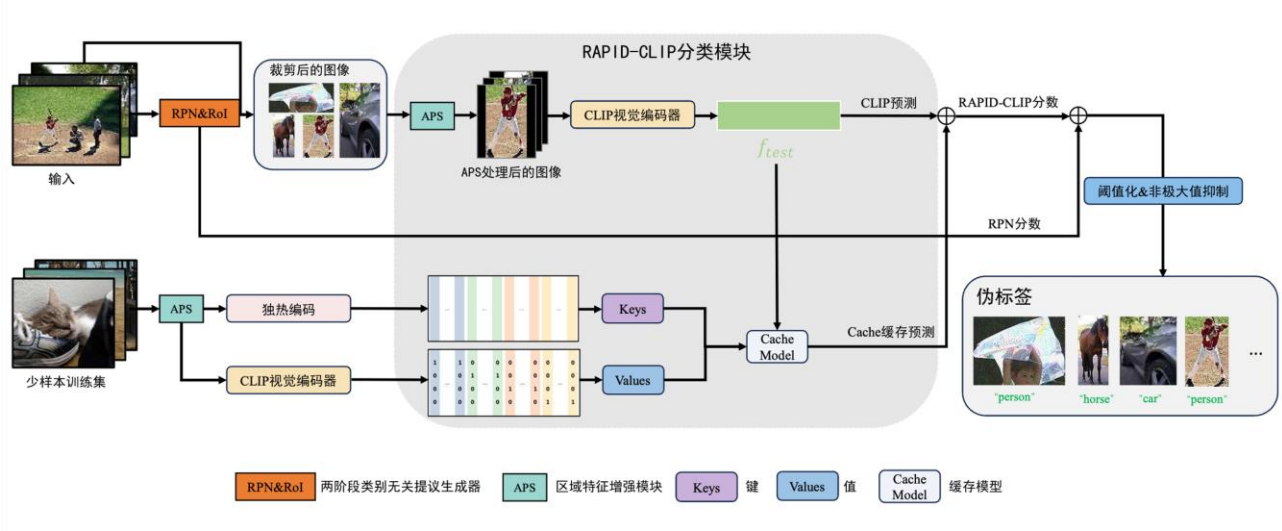


图 1 RAPID-OVD 总体架构

后通过特征增强和优化分类提升伪标签质量，最终通过多阶段筛选生成高质量伪标签用于模型训练。

两阶段类别无关提议生成器基于 Mask R-CNN 架构设计，在训练时忽略具体的类别信息，仅关注目标的存在性和位置信息。这种类别无关的设计使得提议生成器能够有效泛化到训练过程中未见过的新类别，为后续的开放词汇分类提供了高质量的候选区域。

区域特征增强模块通过自适应填充策略解决区域提议在特征提取过程中的变形问题。考虑到CLIP模型要求固定尺寸的输入图像（224×224像素），简单的缩放操作会导致不规则形状目标发生严重变形，特别是对于细长物体和小物体。APS 通过分析目标区域的几何特征，动态选择最优的填充方式，在保持目标原始形态和空间结构信息的同时，确保输入尺寸的一致性要求。

RAPID-CLIP 分类模块构建于 CLIP 框架之上，通过引入基于 Tip-Adapter 的键值缓存机制，显著增强了模型在少样本场景下的区域分类能力。该模块利用少样本训练集构建键值缓存，通过相似度检索和特征融合技术，将少样本知识与 CLIP 的预训练知识有机结合，在无需大规模参数微调的情况下实现了分类性能的显著提升。

3.2 基于键值缓存的CLIP微调

在 RAPID-OVD 框架中，高质量伪标签的生成是提升开放词汇目标检测性能的关键，而生成高质量伪标签的核心在于提升 CLIP 对图像区域分类的准确性。然而，CLIP 作为在大规模图像-文本对上预训练的通用模型，在特定域的少样本场景下往往表现出泛化能力不足的问题。为了解决这一挑战，本文基于 Tip-Adapter 的思想，构建了一个键值缓存模型，通过相似度检索的方式将少样本知识与 CLIP 的预训练知识相结合，同时融入本文提出的自适应填充策略，形成 RAPID-CLIP 模块。基于键值缓存的 CLIP 微调算法如 Algorithm1 所示。

Algorithm 1: 基于键值缓存的CLIP微调算法

Input: 少样本训练集$\{(I_i, y_i)\}_{i=1}^{NK}$，测试区域$I_{\text{test}}$，超参数$\alpha, \beta$

Output: 增强的分类预测logits

// 阶段一：构建键值缓存

for $i = 1$ **to** NK **do**

  $F_{\text{train}}[i] \leftarrow \text{Normalize}(\text{CLIP}_{\text{visual}}(I_i))$// 提取并标准化训练特征

  $L_{\text{train}}[i] \leftarrow \text{OneHot}(y_i)$// 生成独热编码标签

// 阶段二：测试特征提取与相似度计算

$f_{\text{test}} \leftarrow \text{Normalize}(\text{CLIP}_{\text{visual}}(I_{\text{test}}))$// 提取测试特征

$S \leftarrow f_{\text{test}} \cdot F_{\text{train}}^T$// 计算余弦相似度

$A \leftarrow \exp(-\beta(1-S))$// 相似度锐化变换

// 阶段三：缓存预测与特征融合

$P_{\text{adapter}} \leftarrow A \cdot L_{\text{train}}$// 加权聚合缓存预测

$P_{\text{clip}} \leftarrow f_{\text{test}} \cdot W_c^T$// CLIP原始预测

$\text{logits} \leftarrow \alpha \cdot P_{\text{adapter}} + P_{\text{clip}}$// 融合最终预测

return *logits*

3.2.1 键值缓存模型设计

RAPID-CLIP 的核心思想是利用少样本训练集构建键值缓存模型，在无需大规模参数更新的情况下，有效利用少样本数据中的领域特定知识。这种设计将少样本训练集的视觉特征作为键（keys），对应的标签独热编码作为值（values），构成一个可查询的知识库。

在进行分类预测时，首先提取测试图像的视觉特征作为查询（query），然后通过计算查询与缓存中键的相似度来检索相关知识，最后将检索结果与 CLIP 的原始预测进行融合。给定包含 N 个类别的少样本训练集，每个类别有 K 个样本，RAPID-CLIP 首先利用 CLIP 的视觉编码器提取所有训练样本的特征表示。

设训练集为$\{(I_i, y_i)\}_{i=1}^{NK}$，$I_i$表示第$i$个训练图像，$y_i$表示对应的类别标签。键的构建过程如下：对于每个训练样本I_i，通过 CLIP 视觉编码器提取 L2 正则化的特征向量：

$$F_{train}[i] = Normalize\big(CLIP_{visual}(I_i)\big)\#(1)$$

$F_{train} \in \mathbb{R}^{NK\times C}$表示所有训练样本的特征矩阵，$C$为特征维度。值的构建则直接将每个样本的类别标签转换为独热编码：

$$L_{train}[i] = OneHot(y_i)\#(2)$$

$L_{train} \in \mathbb{R}^{NK\times N}$表示所有训练样本的标签矩阵。为了进一步提升缓存模型的表示能力并控制存储开销，RAPID-CLIP 采用了原型聚合策略。具体地，对于拥有较多样本的类别，将同类样本随机分为若干组，计算每组样本特征的均值作为原型表示。这种策略既保持了类别内的多样性，又有效控制了缓存大小。在推理阶段，RAPID-CLIP 首先对输入的区域图像进行特征提取。给定测试区域I_{test}，通过 CLIP 视觉编码器获得其 L2 正则化特征：

$$f_{test} = Normalize\big(CLIP_{visual}(I_{test})\big)\#(3)$$

随后，计算测试特征与缓存中所有键的余弦相似度。由于测试特征和键特征都经过 L2 正则化，其内积等价于余弦相似度：

$$S = f_{test}F_{train}^T\#(4)$$

为了突出最相似的样本并抑制噪声，RAPID-CLIP 引入了指数函数进行相似度变换：

$$A = \exp\big(-\beta(1-S)\big)\#(5)$$

β是控制相似度锐度的超参数。较大的β值会使相似度分布更加尖锐，仅有最相似的样本对预测产生显著影响；较小的β值则会考虑更多样本的贡献，提供更平滑的预测结果。基于计算得到的相似度权重，通过加权聚合缓存中的值来获得适配器的预测：

$$P_{adapter} = A \cdot L_{train}\#(6)$$

最终，RAPID-CLIP 将适配器预测与 CLIP 的原始分类结果进行加权融合：

$$logits = \alpha \cdot P_{adapter} + f_{test}W_c^T\#(7)$$

W_c是 CLIP 分类器的权重矩阵，α是平衡预训练知识与少样本知识的残差比参数。这种融合机制使得模型既能利用 CLIP 强大的预训练知识，又能充分吸收少样本数据中的特定领域信息。

3.2.2 高效微调策略与收敛性分析

为了进一步挖掘少样本数据的潜力并提升 RAPID-CLIP 模块的分类性能，本文设计了高效的微调策略。与传统的全参数微调不同，RAPID-CLIP 仅对缓存中的键进行微调，而保持 CLIP 的编码器权重和缓存中的值不变。这种设计有效避免了过拟合的风险，同时大幅降低了计算开销。

微调过程中的损失函数为：

$$\mathcal{L} = CrossEntropy\big(100 \cdot f_iW_c^T + \alpha \cdot \exp\big(-\beta(1-f_iW^T)\big)L_{train}, y_i\big)\#(8)$$

我们在 7 个基准数据集上进行了详细的收

敛性分析实验,实验结果如表 1 所示。RAPID-CLIP 在 20 轮训练后即可达到接近 200 轮训练的性能水平，平均性能差异为 0.6%，性能保持率为 99.4%，训练时间平均缩短 10.2 倍。

在开放词汇检测的核心基准 COCO- OVD 上，20 轮训练达到 73.75%分类准确率，200 轮训练为 73.89%，性能保持率达到 99.8%。类似的趋势在其他数据集上同样得到验证， 表明 RAPID- CLIP 能够在极少的训练轮次内达到性能收敛,验证了键值缓存机制的高效性。微调过程中，缓存中的键会逐渐向更有利于分类的方向调整，同类样本的特征表示趋于聚集，不同类别间的边界变得更加清晰。这种优化过程可以理解为在少样本约束下对特征空间进行重新组织，使其更适合当前的分类任务。

表 1 不同训练轮次下 RAPID-CLIP 的性能与效率对比。COCO-OVD 数据集上缓存大小设置为 4，其余 6 个数据集上缓存大小均设置为 16。

数据集	方法	训练轮次	训练时间↓(s)	准确率↑(%)	相对增益↑(%)
EuroSAT	Zero-shot CLIP	0	0	45.28	-
	RAPID-CLIP	20	19.21	85.47	88.76
	RAPID-CLIP	200	221.55	86.01	89.95
COCO-OVD	Zero-shot CLIP	0	0	65.17	-
	RAPID-CLIP	20	15.11	73.75	13.17
	RAPID-CLIP	200	168.97	73.89	13.38
Flowers102	Zero-shot CLIP	0	0	65.81	-
	RAPID-CLIP	20	60.52	94.11	43.00
	RAPID-CLIP	200	628.08	94.8	44.05
food101	Zero-shot CLIP	0	0	77.28	-
	RAPID-CLIP	20	54.16	79.36	2.69
	RAPID-CLIP	200	468.36	79.65	3.07
UCF101	Zero-shot CLIP	0	0	61.99	-
	RAPID-CLIP	20	49.83	78.54	26.70
	RAPID-CLIP	200	510.85	78.65	26.88
OxfordPets	Zero-shot CLIP	0	0	85.85	-
	RAPID-CLIP	20	50.31	89.15	3.84
	RAPID-CLIP	200	504.39	89.07	3.75
DTD	Zero-shot CLIP	0	0	42.32	-
	RAPID-CLIP	20	57.28	66.9	58.08
	RAPID-CLIP	200	543.92	67.02	58.36
平均	Zero-shot CLIP	0	0	63.39	-
	RAPID-CLIP	20	43.77	81.04	33.75
	RAPID-CLIP	200	435.10	81.44	34.21

3.3 自适应填充策略

在开放词汇目标检测任务中， 区域提议经过裁剪后需要输入到 CLIP 模型进行分类， 然而 CLIP 的 ViT-B/32 架构要求固定的 224×224 像素输入尺寸。传统方法通常采用简单的缩放调整将任意尺寸的区域提议统一调整至模型所需的输入规格，但这种处理方式会导致严重的图像变形问题。针对这一关键技术挑战，本文提出了自适应填充策略（Adaptive Padding Strategy, APS），通过动态调整填充方式来保持目标的原始形态和空间结构信息，从而显著提升区域分类的准确性。

3.3.1 区域变形问题分析

在目标检测任务中，待检测目标的边界框在尺寸与宽高比上呈现出显著的多样性，这种多样

图 **2** 区域变形对分类性能的影响对比。三列分别为未经裁剪的原始区域、裁剪后直接缩放导致的变形预测结果及分类置信度和裁剪后经过自适应填充策略(APS)处理预测结果及分类置信度。

性是导致后续处理中区域变形问题的根源。当采

用直接缩放方法将这些区域调整为 224×224 的正方形输入时，原始宽高比与目标正方形输入差异越大的目标，其固有的空间结构信息越容易遭到破坏，从而引入严重的空间失真。

对于长宽比为$r{:}1$（$r > 1$）的细长目标，其在短边方向会被过度拉伸r倍，导致目标的空间几何特征发生根本性改变。这种几何失真不仅破坏了目标的视觉完整性，更重要的是干扰了 CLIP 预训练过程中学习到的视觉-语言对应关系，因为 CLIP 在大规模图像-文本对上的预训练假设输入图像保持相对自然的空间比例关系。

为验证区域提议变形问题对特征表示的不利影响，我们设计了对比实验来评估直接缩放对 CLIP 分类性能的具体影响。实验选择了几个典型的细长物体作为测试案例，通过对比直接缩放和本文提出的自适应填充策略(APS)的效果来直观展示变形问题的严重性。

如图 2 所示，展示了当细长物体被强制缩放为 224×224 时，CLIP 的分类性能出现显著下降甚至出现分类失败的三个典型场景：

(1) 棒球棒案例（长宽比 3.34:1）：直接缩放后被错误分类为网球拍，置信度仅 0.18。这一失败案例直观展示了细长运动器材在几何变形后如何导致类别混淆，CLIP 错误地将变形后的棒球棒识别为形状相似的网球拍。

(2) 红绿灯案例（长宽比 5.10:1）：虽然分类结果正确，但置信度仅为 0.77，相较于 APS 处理后的 0.97 置信度，性能大幅下降。这表明即使预测正确，变形问题仍然对特征表示质量产生不利影响，降低了模型的分类确定性。

(3) 滑雪板案例（长宽比 8.02:1）：虽然分类正确，但从置信度 0.68 提升至 0.83，表明 APS 策略即使在极端长宽比下仍能有效改善特征表示质量。

3.3.2 自适应填充策略

针对区域提议在特征提取过程中的几何失真问题，本文提出了自适应填充策略，该策略基于形态保持原则，在确保输入尺寸一致性的约束条件下，通过几何特征分析实现目标区域空间比例关系的最优保持。APS 根据候选区域的长宽比和面积特征，采用分层决策机制动态确定填充参数，以替代传统的统一缩放变换方法。

自适应填充策略的算法实现包含三个核心步骤：几何特征分析、填充策略选择和自适应填充执行。设原始区域提议的尺寸为(w,h)，其中w和h分别表示宽度和高度。自适应填充策略算法如 Algorithm 2 所示。

首先进行几何特征分析，计算长宽比 r = max(w, h)/min(w, h)和区域面积$A = w \times h$。基于这两个几何特征，算法采用如下分层决策机制确定最优填充策略：

$$Mode = \begin{cases} Ratio1{:}1 \ \ if\ A < T_a\ and\ r < T_r \\ Ratio1{:}1.4 \ \ if\ A \geq T_a\ and\ r \geq T_r \\ Adaptive \ \ otherwise \end{cases} \quad (9)$$

对于面积小于 3000 像素且长宽比相对均衡的小目标，采用 1:1 的填充比例以避免过度拉伸；对于面积较大或长宽比较大的目标，则采用 1:1.4 的填充比例以在保持形状的同时适应模型输入要求。

Algorithm 2: 自适应填充策略算法

Input: 区域尺寸(w,h)，目标尺寸$S = 224$
Output: 填充后的标准化区域I_{padded}

$r \leftarrow \frac{\max(w,h)}{\min(w,h)}$ // 计算长宽比
$A \leftarrow w \times h$ // 计算区域面积
if $(A < 3000) \wedge (r < 1.4)$ **then**
  Mode $\leftarrow$ Ratio$_{1:1}$ // 小目标1:1填充
else if $(A \geq 3000) \wedge (r \geq 1.4)$ **then**
  Mode $\leftarrow$ Ratio$_{1:1.4}$ // 大目标1:1.4填充
else
  Mode $\leftarrow$ Adaptive // 自适应填充
scale $\leftarrow \min\left(\frac{S}{w}, \frac{S}{h}\right)$ // 等比例缩放
$(w_s, h_s) \leftarrow (w \times \text{scale}, h \times \text{scale})$ // 缩放后尺寸
$(\Delta w, \Delta h) \leftarrow \left(\frac{S - w_s}{2}, \frac{S - h_s}{2}\right)$ // 填充像素
$I_{\text{padded}} \leftarrow$ PADDING$(I_{\text{scaled}}, \Delta w, \Delta h, \text{Mode})$ // 执行填充
return I_{padded}

在填充执行阶段，算法首先进行等比例缩放以确保至少一个维度达到输入要求。缩放比例计算如下：

$$scale = \min\left(\frac{S}{w}, \frac{S}{h}\right) \#(10)$$

缩放后的区域尺寸为：

$$(w_{scaled}, h_{scaled}) = (w \times scale, h \times scale) \#(11)$$

随后计算所需的填充像素数量：

$$\Delta w = \frac{S - w_{scaled}}{2},\ \Delta h = \frac{S - h_{scaled}}{2} \#(12)$$

3.4 伪标签生成与优化流程

开放词汇目标检测的核心挑战在于如何从未标注图像中生成高质量的伪标签用于模型训练。传统方法往往依赖单一的分类置信度进行伪标签筛选，忽略了定位质量信息和多源信息的协同作用，导致伪标签质量不佳。本文提出了一套完整的四阶段伪标签生成与优化流程，通过系统性的处理策略和多维度的质量控制机制，显著提升伪标签的准确性和可靠性。

3.4.1 阶段一：区域提议生成

区域提议生成是整个伪标签生成流程的基础环节，其质量直接影响后续分类和筛选的效果。本阶段的核心任务是从未标注的输入图像中提取可能包含目标对象的候选区域，为后续的开放词汇分类提供高质量的空间定位基础。

RAPID-OVD 采用基于 Mask R-CNN 架构的两阶段类别无关提议生成器来完成区域提议任务。与传统的目标检测器不同，该生成器在训练过程中完全忽略具体的类别信息，仅关注目标的存在性和位置信息。在此阶段不使用阈值化和非极大值抑制，目的在于保留尽可能多的候选区域，为后续优化提供充分的候选样本。

阶段一输出每个候选区域的空间坐标信息b_i和对应的 RPN 分数$s_{rpn}^{(i)}$，作为后续区域特征增强的输入。

3.4.2 阶段二：区域特征增强

基于阶段一的候选区域$\left\{b_i, s_{rpn}^{(i)}\right\}$，阶段二首先根据空间坐标信息$b_i$对原始图像进行裁剪获得局部图像块，然后通过自适应填充策略对裁剪图像进行特征增强处理，生成满足 CLIP 输入要求的标准化区域图像。

本阶段采用第 3.3 节详述的自适应填充策略（APS），根据裁剪图像的几何特征动态选择最优填充方案，在满足模型输入尺寸要求的同时最大程度保持目标的原始形态和空间结构信息。

阶段二输出标准化为 224×224 像素的区域图像I_i^{padded}，同时保留 RPN 分数$s_{rpn}^{(i)}$和空间位置信息$\mathrm{b_i}$，作为后续多源信息分类的输入。

3.4.3 阶段三：多源信息分类

基于阶段二的标准化区域图像I_i^{padded}和 RPN 分数$s_{rpn}^{(i)}$，阶段三通过 RAPID-CLIP 模块获得各类别的分类置信度$s_{clip}^{(i,c)}$。然后将其与 RPN 分数融合生成综合评估，确保同时考虑定位质量和分类准确性。

使用本文提出的 RAPID-CLIP 模块对每个标准化区域图像进行分类处理，获得各类别的分类置信信息$s_{final}^{(i,c)}$。融合机制将这两类分数结合，生成综合的置信度评估：

$$s_{final}^{(i,c)} = 0.5 \cdot s_{rpn}^{(i)} + 0.5 \cdot s_{clip}^{(i,c)} \#(13)$$

这种融合策略确保只有在定位质量和分类置信度都较高的情况下，综合分数才会达到较高值，既考虑了定位的准确性，又兼顾了分类的可靠性。

阶段三输出每个候选区域的综合置信度$s_{final}^{(i,c)}$，同时保留空间位置信息b_i供后续 NMS 使用。

3.4.4 阶段四：伪标签质量评估与优化

基于阶段三的综合置信度$s_{final}^{(i,c)}$和空间位置信息b_i，阶段四通过置信度阈值化和非极大值抑制生成最终伪标签。与现有方法仅基于单一分类置信度筛选不同，本阶段采用融合 RPN 与 CLIP 分数的多重筛选机制，显著提升伪标签质量。

首先采用置信度阈值化处理过滤低质量的候选区域。对于每个类别c，保留满足以下条件的候选区域：

$$s_{final}^{(i,c)} \geq \tau \#(14)$$

τ为置信度阈值。置信度阈值τ设置为 0.8 能够在伪标签质量和数量之间取得良好平衡。过低的阈值会引入过多噪声标签，影响模型训练效果

；过高的阈值则会导致有效伪标签的丢失。

经过阈值化处理后，系统应用非极大值抑制（NMS）算法去除重复检测。对于同一类别c的候选区域集合$\mathcal{R}_c = \{r_i \mid s_{final}^{(i,c)} \geq \tau\}$，NMS 过程可表示为：

$$\mathcal{R}_c^{nms} = NMS(\mathcal{R}_c, \theta_{nms}) \#(15)$$

θ_{nms}为 NMS 阈值，本文设置为 0.3，能够有效去除冗余检测框同时保留不同目标的独立检测结果。与传统 NMS 仅基于分类置信度排序不同，本文的 NMS 过程基于融合后的综合置信度$s_{final}^{(i,c)}$进行排序，确保同时考虑定位质量和分类准确性。

为了量化我们多阶段筛选流程所生成伪标签的质量，我们设计了专门的评估指标。设生成的伪标签集合为 $\mathcal{P} = \{(b_j, c_j, s_j)\}_{j=1}^{N_p}$。首先，为了直接评估伪标签与真实目标（Ground Truth）的对齐程度，我们将伪标签集合本身视为一组检测器的预测输出，并在 COCO-OVD 验证集上计算其针对新类别的 Novel AP (AP50)。这种方法可以直观地反映伪标签的定位和分类准确性。此外，我们还统计了伪标签在每张图像上的平均生成数量，即伪标签密度（ρ_{pseudo}）。

表2　不同方法在伪标签生成效率和最终检测性能上的对比

方法	伪标签生成时间↓(s/img)	平均伪标签数/图像	伪标签Novel AP	推理FPS↑	AP_{novel} ↑
VL-PLM [32]	0.54	4.09	20.9	35.80	32.3
MarvelOVD [25]	0.51	3.15	21.4	52.12	38.9
RAPID-OVD(本文方法)	0.67	4.13	21.9	34.18	40.5

定量分析实验如表 2 所示，RAPID-OVD 在伪标签质量上取得了显著提升。尽管多阶段处理流程带来了一定的计算开销增加，却显著提升了伪标签的质量和数量：平均每张图像生成 4.13 个伪标签（ρ_{pseudo} = 4.13），伪标签自身的 Novel AP 达到 21.9%。这些高质量伪标签直接转化为检测性能的显著改善：RAPID-OVD 框架在 COCO-OVD 数据集上最终达到了 40.5% 的 AP_{novel}，超越了现有最先进方法。

4 实验结果与分析

4.1 数据集

本文在开放词汇目标检测领域公认的权威基准数据集上进行了全面实验验证，采用 COCO- OVD 和 LVIS-OVD 两个标准数据集。这两个数据集构成了开放词汇目标检测研究的核心评估基准，几乎所有该领域的主流方法都以此作为性能验证的基准[25, 31, 32]，为不同技术路线的方法提供了公平、全面的比较平台。

COCO-OVD 数据集作为开放词汇检测的基础标准，包含超过 12 万张涵盖 80 个类别的图像，展现了复杂的自然场景、多种尺度物体、各种遮挡情况和丰富的环境变化。遵循开放词汇检测领域的标准设置[31, 32]，本文将 COCO 类别分为两组：48 个已知基础类别（base categories）和 17 个未知新类别（novel categories）。在开放词汇检测设置中，训练过程仅使用基础类别的标注信息，而新类别的图像作为未标注数据参与训练，测试时评估模型对新类别的检测能力。这种设置有效模拟了开放词汇场景，测试模型对训练过程中未见过类别的泛化能力。

48 个基础类别包括常见的日常物体，如人物、汽车、椅子等，这些类别在训练阶段提供完整的边界框标注信息。17 个新类别则包括飞机、火车、船只等物体，这些类别的图像在训练集中存在但不提供标注信息，仅在测试阶段用于评估检测性能。COCO-OVD 的类别设置和数据分布为开放词汇检测方法提供了标准化的评估环境，确保了实验结果的可重现性和可比性。

LVIS-OVD 数据集作为大规模长尾分布的挑战性基准，为本文方法提供了更严格的泛化能力验证。LVIS-OVD 数据集包含 1203 个类别，具有更大的词汇规模和更丰富的长尾分布特性，是目前开放词汇检测领域词汇规模最大的评估

基准。在开放词汇设置中，将 337 个稀有类别（rare categories）设置为新类别，其余 866 个常见类别（frequent and common categories） 作为基础类别。LVIS-OVD 数据集的长尾分布特性为评估模型在稀有类别上的检测能力提供了更具挑战性的测试环境，能够更好地反映方法在实际应用中面对罕见目标的鲁棒性。

COCO-OVD 和 LVIS-OVD 两个数据集在评估维度上形成了有效互补，共同构成了全面的性能验证体系。首先，从评估覆盖度角度，COCO-OVD 提供了标准化的基础评估环境，其类别划分和数据规模已成为该领域的事实标准，几乎所有主流方法[25, 31, 32]都在此基准上报告性能；LVIS-OVD 则从大规模词汇和长尾分布的角度提供了更严格的挑战性测试，验证方法的极限泛化能力。其次，从技术挑战维度，两个数据集分别侧重不同的技术难点：COCO-OVD 主要考验模型在标准开放词汇设置下的基础性能，而 LVIS-OVD 重点测试模型处理大词汇量和稀有类别的能力。最后，从实验验证的充分性角度，当前开放词汇检测领域的顶级会议论文普遍采用这两个数据集作为主要评估基准[25, 31, 32]，表明学术界已形成共识：这两个数据集能够全面、充分地验证开放词汇检测方法的有效性。

4.2 评价指标

开放词汇目标检测的性能评估需要从检测精度和泛化能力两个维度进行综合考量，以全面衡量模型在处理已知类别和未知类别时的实际应用效果。

本研究采用目标检测领域的标准指标平均精度（Average Precision，AP）来评估开放词汇目标检测的精度表现。AP 指标通过计算不同 IoU 阈值下的精确率-召回率曲线下面积，能够有效反映检测系统在定位和分类方面的综合性能。本文主要使用以下几个关键指标：

AP_{novel} (新类)：新类别平均精度，专门用于评估模型对训练过程中未见过类别的检测能力。这是开放词汇目标检测最核心的评价指标，直接反映了模型的开放词汇泛化性能。

AP_{base} (基类)：基础类别平均精度，用于评估模型对训练集中已标注类别的检测效果。该指标反映了引入开放词汇能力后模型在原有任务上的性能保持情况。

AP_{all} (全部)：所有类别的综合平均精度，通过加权平均基础类别和新类别的检测精度得到，为模型整体性能提供全局评估。

为了验证模型的泛化能力，本研究在多个不同规模和特点的数据集上进行评估。在 LVIS 数据集上，根据类别出现频率将 1203 个类别划分为三个子集：稀有类别（rare）、常见类别（common）和频繁类别（frequent）。相应地使用AP_r、AP_c和AP_f分别评估模型在不同频率类别上的表现。

考虑到 RAPID-CLIP 微调模块的重要作用，本文同时采用分类准确率（Classification Accuracy）来评估区域级别的分类性能。分类准确率定义为正确分类的区域提议数量与总区域提议数量的比值：

$$分类准确率 = \frac{正确分类区域数量}{区域总数量} \times 100\% \#(16)$$

该指标主要用于消融实验中分析不同填充策略和微调方法对分类性能的影响，有助于理解检测性能提升的内在机制。

4.3 基线系统

为了全面评估提出方法的有效性，本文选择了开放词汇目标检测领域具有代表性的方法进行对比实验。根据不同的技术路线和性能表现，本文主要与以下基线系统进行对比：

(1) VLDet[40]

VLDet 提出了一种端到端的视觉语言框架来解决开放词汇目标检测问题。该方法将图像区域特征和词嵌入视为两个集合，通过双分匹配算法寻找区域-词对应关系的最优匹配。VLDet 直接从图像-文本对中训练目标检测器，无需依赖昂贵的定位标注或从分类导向的视觉模型中进行知识蒸馏。

(2) BIND[41]

BIND 引入了 Built-in Detector 的设计理念，采用编码器-解码器结构和两阶段训练框架。第一阶段通过图像-文本双编码器学习区域-词对齐，第二阶段使用 DETR 风格的解码器进行检测训练。与传统的迁移学习方法不同，BIND 消除了从预训练视觉语言模型到现成检测器的知

识迁移需求，并设计了自适应锚点提议网络来显著提升检测效率。

(3)OADP[42]

OADP 是一种基于知识蒸馏的改进方法，它在标准的区域-文本特征对齐蒸馏管线中引入了一个辅助任务。该辅助任务要求模型预测整张图像中存在的所有类别词汇，以此来正则化特征学习过程，有效地缓解了学生检测器在学习新知识时对基础类别知识的灾难性遗忘问题。

(4) CAKE[43]

CAKE 提出了类别感知知识提取框架，包含类别特定知识蒸馏分支(CSKD)和类别泛化区域提议网络(CG-RPN)。CSKD 通过对象级和聚类级双重蒸馏构建类别特定特征集，CG-RPN 利用该特征集指导提议生成并缓解基础类别过拟合。CAKE 作为即插即用框架，可显著提升现有方法在新类别上的检测性能。

(5) BARON[44]

BARON 是知识蒸馏类开放词汇检测的改进方法，将上下文相关的区域分组为包（bag），通过视觉语言模型文本编码器将其处理为类似句子的嵌入。通过上下文信息增强和包级别的特征对齐，在知识提取与学生检测器学习潜力之间实现了更好的平衡。

表3 在COCO-OVD数据集上与最先进（SOTA）方法的比较。遵循以往的研究，开放视觉检测（OVD）重点关注新类别上的性能表现。RAPID-OVD(本文方法)在AP50新类测试中达到了最佳性能40.5%，OV-DQUO方法达到次佳性能39.2%。

方法	期刊会议	检测器	基准测试	AP50(泛化)		
				新类↑	基类↑	全部↑
VLDet [40]	ICLR 23	Faster R-CNN	WS-OVD	32	50.6	45.8
BIND[41]	CVPR 24	Built-in Detector	WS-OVD	36.3	54.7	50.2
OADP[42]	CVPR 23	Faster R-CNN	G-OVD	35.6	55.8	50.5
CAKE[43]	AAAI 25	Faster R-CNN	G-OVD	39.1	58.1	53.1
BARON[44]	CVPR 23	Faster R-CNN	KD-OVD	35.8	58.2	52.3
LBP[45]	CVPR 24	Faster R-CNN	KD-OVD	37.8	58.7	53.2
CORA[46]	CVPR 23	DAB-DETR	C-OVD	35.1	35.5	35.4
SIA-OVD[47]	MM 24	DAB-DETR	C-OVD	35.5	40.3	39.3
SHiNe[48]	CVPR 24	CenterNet2	C-OVD	24.3	54.4	44.9
OV-DQUO[49]	AAAI 25	DINO	C-OVD	39.2	-	-
VL-PLM[32]	ECCV 22	Mask R-CNN	PL-OVD	32.3	54.0	48.3
SAS-Det[7]	CVPR 24	Faster R-CNN	PL-OVD	37.4	58.5	53.0
MarvelOVD[25]	ECCV 24	Faster R-CNN	PL-OVD	38.9	56.4	51.8
RAPID-OVD(本文方法)	-	Mask R-CNN	PL-OVD	**40.5**	56.3	52.2

表4 在LVIS-OVD上与SOTA方法的比较。RAPID-OVD(本文方法)在AP_r测试中达到了最佳性能26.5%，MarvelOVD方法达到次佳性能26%。

方法	骨干网络	AP_r ↑	AP_c ↑	AP_f ↑	AP ↑
VLDet[40]	RN50	21.7	29.8	34.3	30.1
BARON[44]	RN50	23.2	29.3	32.5	29.5
OADP[42]	ViT-B/32	21.9	28.4	32.0	28.7
MarvelOVD[25]	RN50	26.0	34.2	36.9	34.2
RAPID-OVD(本文方法)	RN50	**26.5**	34.9	38.1	34.7

(6)LBP[45]

LBP (Language-guided Box Propagation) 是一种半监督的开放词汇检测方法。它旨在通过利用语言模型，从少量标注样本（exemplars）出发，将边界框标注自动传播到整个数据集中未标注的图像。该方法通过视觉和文本的相似性度量，在大量数据上生成高质量的伪标签，显著降低了人工标注成本。

(7) CORA[46]

CORA 是基于分类器的开放词汇检测方法，采用 DAB-DETR 检测器架构。该方法冻结视觉语言模型编码器参数，同时使用基础类别标签对 ResNet-50 骨干网络中的位置嵌入进行微调，以适应开放词汇检测任务。CORA 通过参数高效的微调策略，在计算资源受限的情况下实现了具有竞争力的检测性能。

(8) SIA-OVD[47]

SIA-OVD 引入形状不变适配器（Shape-Invariant Adaptor）来弥合全局图像和局部区域之间的表示差距。该方法通过专门设计的适配器模块增强了视觉语言模型在局部特征判别方面的能力。SIA-OVD 在处理不同形状和尺度的目标时表现出较好的鲁棒性。

(9) SHiNe[48]

SHiNe 是一种无需训练的开放词汇目标检测增强方法，通过特征偏移和邻域评分策略提升预训练检测器的性能。该方法可以无缝集成到任何现有的开放词汇检测器中，通过动态调整区域特征在嵌入空间中的位置并利用邻域信息进行类别判别，显著改善了模型在处理不同粒度词汇时的鲁棒性和检测精度。

(10) OV-DQUO[49]

OV-DQUO 提出了一种基于去噪文本查询训练和开放世界未知对象监督的开放词汇 DETR 框架来解决置信度偏差问题。该方法通过通配符匹配策略，使检测器学习未知对象与通用语义文本嵌入的配对关系，缓解基础类别和新类别间的置信度偏差；采用去噪文本查询训练，从开放世界未知对象合成前景背景查询-框对进行对比学习，增强新类别与背景的区分能力；引入感兴趣区域查询选择模块，结合区域-文本相似性与置信度分数实现平衡的提议选择。OV-DQUO 直接解决了现有方法忽视的置信度偏差根本问题，在开放词汇检测基准上取得显著性能提升。

(11) VL-PLM[32]

VL-PLM 是伪标签开放词汇检测的开创性工作，也是本文方法的直接基线。该方法通过建立完整的伪标签生成流水线，利用视觉语言模型为未见过的类别生成高质量的边界框伪标签。VL-PLM 采用类别无关的区域提议生成器和 CLIP 模型，通过 RoI 特征提取和分类头重用技术改进边界框提议质量。

(12) SAS-Det[7]

SAS-Det 通过引入语义感知稀疏化（Semantic-Aware Sparsification）模块来优化开放词汇检测流程。该模块能够在区域提议网络之后，有效过滤掉大量与前景物体无关的、语义信息贫乏的背景候选框，使得后续的分类器可以将计算资源集中在更有可能是目标的区域上，从而提升了检测的效率与准确性。

(13) MarvelOVD[25]

MarvelOVD 提出了一种多视角词汇蒸馏（multi-view vocabulary distillation）的策略。它不仅使用单一的类别名称作为监督信号，还通过大型语言模型为每个类别生成了包含属性、关系、上下文等信息的丰富文本描述（即"多视角"词汇）。这种更具信息量的监督信号能够显著提升学生检测器对于新类别的理解和泛化能力。

4.4 实验结果

为了全面验证 RAPID-OVD 框架的有效性，本文在 COCO-OVD 和 LVIS-OVD 数据集上进行了详细的实验评估。实验主要从开放词汇检测性能、消融实验分析、不同尺寸物体的检测性能分析和可视化结果分析四个方面展开，以全面衡量提出方法在不同维度上的性能表现。

4.4.1 开放词汇检测性能对比

表 3 和表 4 分别展示了 RAPID-OVD 与代表性基线方法在 COCO-OVD 和 LVIS-OVD 数据集上的开放词汇检测性能对比。实验结果表明，本文提出的 RAPID-OVD 框架在新类别检测方面取得了显著的性能提升。

从表 3 可以看出，RAPID-OVD 在新类别检测（AP_{novel}）上达到了 40.5%的平均精度，相比

于所有基线方法都取得了显著提升。相比于直接基线 VL-PLM，RAPID-OVD 提升了 8.2 个百分点（从 32.3%提升至 40.5%）；相比于当前性能最优的 OV-DQUO 方法，RAPID-OVD 提升了 1.3 个百分点（从 39.2%提升至 40.5%）。

在更具挑战性的 LVIS-OVD 数据集上，RAPID-OVD 继续展现出优异的泛化能力。如表 4 所示，RAPID-OVD 在稀有类别（AP_r）检测上达到了 26.5%的性能，相比次佳方法 MarvelOVD 的 26.0%提升了 0.5 个百分点。同时，在常见类别（AP_c）和频繁类别（AP_f）上分别达到了 34.9%和 38.1%，均超越了所有基线方法。

表 5 进一步展示了本文方法与当前伪标签生成最优基线 MarvelOVD 在 COCO-OVD 验证集上 17 个新类别的详细检测性能对比及伪标签数量分析。

RAPID-OVD 在 17 个新类别中的 11 个类别上实现了性能提升，在 6 个类别上出现性能下降。整体而言，方法在新类别检测上取得了 1.6 个百分点的提升。

表5　各目标新类别在COCO-OVD验证集上的检测性能(AP50 个) 及伪标签数量分析

检测性能(AP50 个)

方法	Airplane	Bus	Cat	Dog	Cow	Elephant	Umbrella	Tie	Snowboard
MarvelOVD[25]	68.66	75.16	63.37	72.19	54.45	78.51	23.81	18.03	8.65
RAPID-OVD(本文方法)	69.31	75.46	51.83	71.26	46.33	77.56	21.74	20.65	15.43
Δ性能变化	+0.65	+0.30	-11.54	-0.93	-8.12	-0.95	-2.07	+2.62	+6.78

方法	Skateboard	Cup	Knife	Cake	Couch	Keyboard	Sink	Scissors
MarvelOVD[25]	18.35	32.17	0.96	31.15	38.00	46.28	12.04	25.80
RAPID-OVD(本文方法)	22.53	27.53	6.93	34.36	44.72	47.16	18.65	33.82
Δ性能变化	+4.18	-4.64	+5.97	+3.21	+6.72	+0.88	+6.61	+8.02

伪标签数量

方法	Airplane	Bus	Cat	Dog	Cow	Elephant	Umbrella	Tie	Snowboard
MarvelOVD[25]	36213	51768	31613	28420	51776	45305	47827	90675	92395
RAPID-OVD(本文方法)	30443	37401	60577	38054	20534	30832	95138	111869	190102
Δ数量变化	-5770	-14367	+28944	+9634	-31242	-14473	+47311	+21194	+97707

方法	Skateboard	Cup	Knife	Cake	Couch	Keyboard	Sink	Scissors
MarvelOVD[25]	90472	73946	396710	83772	41081	40940	39130	20929
RAPID-OVD(本文方法)	181705	40002	23547	43420	24728	35304	39918	44598
Δ数量变化	+91233	-33944	-373163	-40352	-16353	-5636	+788	+23669

结合伪标签数量与性能变化的关联分析，可将 17 个类别的表现归纳为四种典型情况：

(1)理想协同效应：伪标签数量增加且性能提升，如 Snowboard(+97707 个伪标签，+6.78%)、Skateboard(+91233 个，+4.18%)等 5 个类别。

(2)质量优于数量：伪标签数量减少但性能提升，如 Knife(-373163 个伪标签，+5.97%)、Couch(-16353 个，+6.72%)等 6 个类别。特别是 Knife 类别的大幅改善证明了高质量伪标签筛选的有效性。

(3)伪标签质量与数量的矛盾：Cat、Dog、Umbrella 等 3 个类别虽然伪标签数量增加，但性能仍下降。以 Cat 类别为例，伪标签从 31613 增加到 60577(+28944)，但性能下降 11.54%，表明单纯增加伪标签数量并不能保证性能提升。

(4)过度筛选导致的样本不足：Elephant、Cow、Cup 等 3 个类别的伪标签数量减少且性能下降。如 Cow 和 Cup 类别的伪标签分别减少

31242 和 33944 个，性能相应下降 8.12%和 4.64%，表明 RAPID-OVD 的多阶段筛选策略对某些类别可能过于严格。

对于部分类别(如 Cat、Cow、Umbrella 等)的性能下降，我们认为这主要源于当前方法采用全局统一的自适应填充策略(T_a=3000, T_r=1.4)，未能充分考虑不同类别的视觉特性差异，导致 RAPID-CLIP 在区域分类时的准确性下降，进而影响整体检测性能。

4.4.2 消融实验分析

为了深入理解 RAPID-OVD 框架中各个组件的贡献，本文进行了详细的消融实验分析。主要分析包括微调对 CLIP 分类性能的增强、不同填充策略对分类准确率的影响和微调和填充对检测性能与伪标签质量的影响。

(1) 微调对 CLIP 分类性能的增强

COCO-OVD 数据集中一张图片可能有多个对应的标签，为了验证微调方法对图像分类准确性的性能提升，先按照标注文件对其进行了裁剪处理，使得裁剪后的每一张图片只对应一个标签，并按照标签类别进行归类。在未特殊说明的情况下，缓存大小默认设置为 16。 表 6 展示了微调在 COCO-OVD 验证集上的四项消融研究。

表 **6** COCO-OVD 验证集上 RAPID-CLIP 的消融实验（%）。从上到下分别为：残差比 α、锐度比 β、缓存大小以及扩展缓存实验。

微调的消融研究						
残差比α	0.0	**0.5**	1.0	2.0	3.0	4.0
	65.17	**66.30**	65.31	64.35	64.67	63.21
锐度比β	0.5	1.5	**2.5**	3.5	4.5	5.5
	66.23	66.13	**66.30**	66.25	65.39	65.18
缓存大小	0	1	2	4	**8**	16
	65.17	65.65	66.17	66.8	**67.40**	66.3
更多的缓存大小	**缓存大小设置**	16	32	64	**128**	
	fine-tuning	66.3	66.97	68.05	**68.13**	
	Zero shot	65.17				

超参数 α：控制将缓存模型中新调整的预测与预训练的 CLIP 相结合的程度。 较大的 α 表示使用更多来自少量样本学习训练集的知识，而较少使用其他知识。本文将 α 从 0.0 变化到 4.0，并将超参数 β 设置为 2.5。 当 α 为 0.0 时，模型在不使用少量样本学习知识的情况下等效于零样本学习 CLIP。

锐度比 β： 当 β 较大时，只有在嵌入空间中与测试图像最相似的训练样本对预测的影响较大，反之亦然。在表 6 的第二部分中， α 为 0.5 时，观察到 β 的变化具有有限的影响， β 的值为 2.5 时达到最佳性能。

缓存大小：本文探究了缓存大小对微调性能的影响。在给定 16-shot 训练集的情况下，并未简单地缓存每个类别的全部 16 个样本，而是构建了大小介于 0 到 16 之间的缓存。以大小为 8 的情况为例，本文将每类的 16 个样本随机划分为 8 个均匀组，通过计算每组中 2 个样本特征的平均值获得 8 个原型表示。考虑到这种随机划分可能影响性能表现，本文进行了 5 次实验并报告了平均得分。表 6 部分的结果表明，缓存的样本数量越多，达到的准确率就越高。

(2) 自适应填充策略参数敏感性分析

自适应填充策略的有效性依赖于两个关键阈值参数的合理设置：面积阈值T_a用于区分小目标与常规目标，长宽比阈值T_r用于区分规整目标与细长目标。为系统评估参数设置对模型性能的影响并验证参数的跨数据集适应性，我们首先在 COCO-OVD 数据集上进行全面的参数搜索分析，随后在 LVIS-OVD 数据集上验证参数的泛化能力。

在 COCO-OVD 数据集上，我们通过三组实验系统评估参数设置的最优性。第一组实验中，我们保持T_r=1.4 不变，测试不同T_a取值对检测性能的影响。实验结果如表 7 所示，T_a=3000 时AP_{novel}达到 40.5%的峰值性能，性能变化呈现明显的倒 U 型曲线。

表**7** 面积阈值T_a对检测性能的影响（固定T_r = 1.4）

T_a	**AP**$_{\text{novel}}$ ↑	**AP**$_{\text{base}}$	**AP**$_{\text{all}}$
2000	39.9	55.9	51.7
2500	40.2	56.1	51.9
3000	**40.5**	**56.3**	**52.2**
3500	40.0	56.0	51.8
4000	39.7	55.6	51.4

第二组实验中，我们保持T_a=3000 不变，测试不同长宽比阈值T_r对检测性能的影响。实验结果如表 8 所示，同样呈现出明显的性能峰值特征，T_r=1.4 时AP_{novel}达到 40.5%的最优性能。

为进一步验证参数组合的协同效应，第三组实验选取了代表性的边界参数组合进行综合性能测试，结果如表 9 所示。我们有意选择了T_a和T_r的较小值组合（2000, 1.2）、较大值组合（4000, 1.6）以及混合边界组合，以全面评估参数设置的鲁棒性。实验结果表明，偏离最优参数的组合均导致性能明显下降，验证了(T_a=3000, T_r=1.4)的最优性。

表 **8** 长宽比阈值T_r对检测性能的影响（固定T_a = 3000）

T_r	**AP**$_{\text{novel}}$ ↑	**AP**$_{\text{base}}$	**AP**$_{\text{all}}$
1.2	38.4	53.4	49.5
1.3	39.7	55.7	51.5
1.4	**40.5**	**56.3**	**52.2**
1.5	39.6	55.9	51.6
1.6	39.0	55.4	51.1

表 **9** COCO-OVD 数据集上的参数组合性能对比

T_a	T_r	**AP**$_{\text{novel}}$ ↑	**AP**$_{\text{base}}$	**AP**$_{\text{all}}$
3000	1.4	**40.5**	**56.3**	**52.2**
2000	1.2	37.8	53.0	48.7
2000	1.6	38.4	55.0	50.7
3000	1.2	38.4	53.4	49.5
3000	1.6	39.0	55.4	51.1
4000	1.2	37.6	52.7	48.5
4000	1.6	38.2	54.7	50.4

针对上述参数设置的跨数据集适应性问题，我们在 LVIS-OVD 数据集上进行了泛化性验证。LVIS-OVD 具有更大的类别规模（1203 类）和更显著的长尾分布特征，构成了严格的参数泛化性测试环境。通过统计相同阈值下的目标分布情况（表 10 和表 11），我们量化了两个数据集的分布差异。在最优参数设置下，COCO-OVD 与 LVIS-OVD 的小目标比例分别为 48.4%和 69.6%，规整目标比例分别为 34.2%和 40.0%。两个数据集在目标几何分布上的显著差异使其成为评估参数泛化能力的有效测试基准。

我们在 LVIS-OVD 上测试了与 COCO-OVD 相同的参数组合， 结果如表 12 所示。 实验结果表明，(T_a=3000, T_r=1.4)在 LVIS-OVD 上同样达到最优性能，AP_r达到 26.5%。

表 **10** COCO-OVD 与 LVIS-OVD 数据集在不同面积阈值T_a下的目标分布对比

T_a	COCO-OVD		LVIS-OVD	
	小目标数量	比例(%)	小目标数量	比例(%)
2000	375,448	41.9	960,752	63.4
2500	408,260	45.5	1,013,320	66.9
3000	434,473	48.4	1,054,090	69.6
3500	456,962	51.0	1,086,828	71.7
4000	476,131	53.1	1,114,364	73.6

表 **11** COCO-OVD 与 LVIS-OVD 数据集在不同长宽比阈值T_r下的目标分布对比

T_r	COCO-OVD		LVIS-OVD	
	规整目标数量	比例(%)	规整目标数量	比例(%)
1.2	172,399	19.2	361,721	23.9
1.3	242,008	27.0	492,467	32.5
1.4	306,461	34.2	605,310	40.0
1.5	361,766	40.3	701,270	46.3
1.6	411,879	45.9	783,972	51.8

表 **12** LVIS-OVD 数据集上的参数组合性能对比

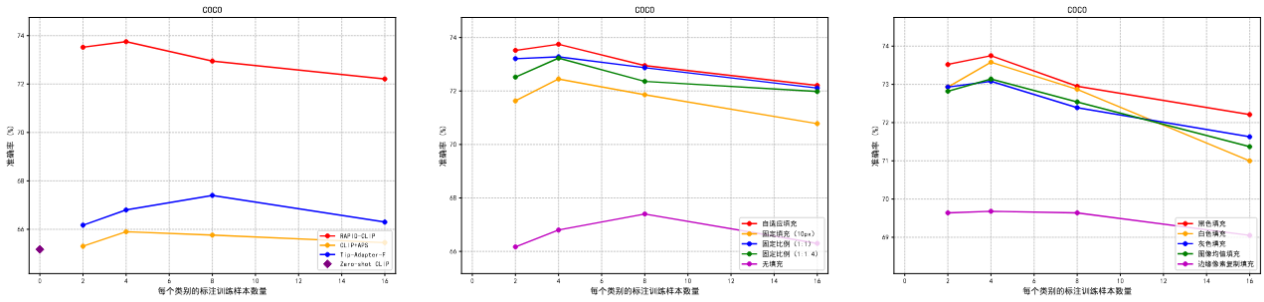
T_a	T_r	**AP**$_r$ ↑	**AP**$_c$	**AP**$_f$	**AP**
3000	1.4	**26.5**	**34.9**	**38.1**	**34.7**
2000	1.2	24.8	32.7	35.7	32.5
2000	1.6	25.3	33.5	36.6	33.3
3000	1.2	25.1	33.1	36.1	33.0
3000	1.6	25.6	33.8	36.9	33.6

4000	1.2	24.6	32.4	35.4	32.2
4000	1.6	25.0	33.0	36.0	32.8

跨数据集验证实验表明，本文提出的参数设置(T_a=3000, T_r=1.4)不仅在 COCO-OVD 上达到最优性能，在目标分布特征显著不同的 LVIS-OVD 上同样保持最优，验证了参数的跨数据集适应性和泛化能力。

(3) 不同策略对分类准确率的影响分析

本小节通过系列实验探究了不同策略对 CLIP 模型分类准确率的影响。为全面评估不同方式对模型性能的影响，本文设计了三组对比实验，分别从方法对比、填充策略和填充颜色三个维度进行分析。图 3 展示了不同填充策略对分类准确率的影响。



(a) 方法对比实验　(b) 填充策略对比实验　(c) 填充颜色对比实验

图 3 不同填充策略对分类准确率的影响

方法对比实验：首先将本文提出的 RAPID-CLIP 方法与现有的几种方法进行对比。如图 3（a）所示，RAPID-CLIP 在各个样本量下均显著优于其他方法。在 4 个样本/类的条件下，RAPID-CLIP 达到了 73.75%的准确率，比基线方法 Tip-Adapter 高出约 6 个百分点。值得注意的是，即使在低样本量（2 个样本/类）的情况下，RAPID-CLIP 也表现出色，准确率达到 73.52%。随着样本量增加到 16 个/类，各方法性能差距仍然明显，RAPID-CLIP 保持领先地位。

填充策略对比实验：为了验证填充策略的有效性，本文对比了五种填充策略：自适应填充、固定填充（10px）、固定比例（1:1）、固定比例（1:1.4）和无填充。 如图 3（b）所示，所有填充策略均显著优于无填充方法，证实了填充操作对提升分类性能的重要性。 自适应填充策略在所有样本量条件下表现最佳，特别是在 4 个样本/类时达到峰值 73.75%。固定比例（1:1）策略次之，与自适应填充性能接近。无填充策略明显落后，准确率仅为 66.67%，这表明未经处理的裁剪区域会导致严重的图像变形，从而影响 CLIP 的识别能力。

填充颜色对比实验： 本文进一步研究了不同填充颜色对分类准确率的影响. 如图 3（c）所示，五种填充颜色中，黑色填充表现最佳，在各个样本量条件下始终保持领先；白色填充和灰色填充次之，它们在 4 个样本/类时表现接近，但随着样本量增加，性能差异逐渐显现。图像均值填充也表现良好，而边缘像素复制填充的效果显著低于其他策略，准确率仅为 69.56%左右。这可能是因为边缘复制方法引入了不自然的视觉模式，干扰了 CLIP 对目标区域的理解。

(4) 微调和填充对检测性能与伪标签质量的影响

在开放词汇目标检测任务中，分类准确率的提升并不总是能直接等比例地转化为检测性能的提升。

这主要是因为伪标签的质量不仅依赖于类别预测的准确性，还受到定位精度等多种因素的综合影响。为了分析了不同填充策略对检测性能和伪标签质量的影响。在已经完成键值缓存微调 CLIP 的基础上，本文采取了不同的填充策略。

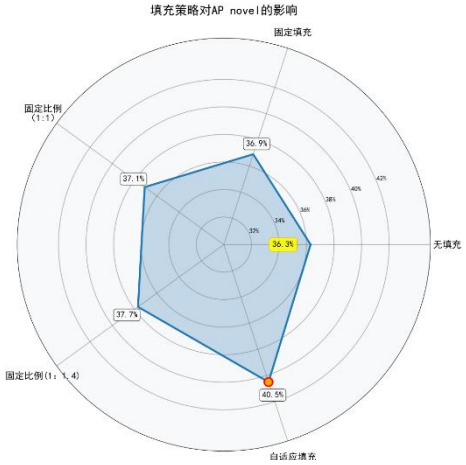


图 4 填充策略对AP_{novel}的影响

图 4 展示了不同填充策略对AP_{novel}的影响。本文的实验结果表明，改善区域提议的视觉表征不仅提升了类别预测的准确性，同时也增强了定位的精确度，进而提高了 AP 指标。 如图所示，不同填充策略对检测性能产生了明显差异. 以无填充方法作为基线，引入填充操作后检测性能得到了普遍提升。其中，自适应填充策略表现最佳，AP_{novel}达到 40.5%，相对基线提升了 4.2 个百分点。 固定比例（1:1）策略次之，提升了 0.8 个百分点. 值得注意的是，即使简单的固定填充（10px）也带来了 0.6 个百分点的提升，证实了填充操作在保持区域提议原始形态方面的普遍有效性。

(5) 分数融合权重分析

在 RAPID-OVD 框架的伪标签生成阶段，本文采用 RPN 分数与 CLIP 分数的加权融合策略来确定候选区域的最终置信度。RPN 擅长空间定位，能准确评估候选区域包含前景目标的可能性，但缺乏语义判别能力；CLIP 具备强大的开放词汇语义理解能力，能准确匹配区域内容与类别描述，但作为图像级预训练模型，其对区域定位质量的评估能力有限。两者在信息维度上互补，融合权重的选择直接影响伪标签的筛选质量。

为系统评估不同融合权重的影响，本文在 COCO-OVD 数据集上进行了权重消融实验。实验以 0.2 为步长覆盖权重空间[0.0, 1.0]，并在等权重点 0.5 附近加密采样（0.4, 0.5, 0.6），共测试 7 个配置。实验结果如表 13 所示。

表 **13** RPN 分数与 CLIP 分数融合权重对检测性能的影响

RPN权重	CLIP权重	**AP**$_{\text{novel}}$ ↑	**AP**$_{\text{base}}$	**AP**$_{\text{all}}$
0.0	1.0	36.8	54.8	50.1
0.2	0.8	38.6	55.4	51.0
0.4	0.6	39.8	56.0	51.8
0.5	0.5	**40.5**	**56.3**	**52.2**
0.6	0.4	40.1	56.1	51.9
0.8	0.2	37.3	55.6	50.8
1.0	0.0	35.2	55.1	49.9

实验结果表明，等权重配置（0.5:0.5）取得最优性能，AP_{all}达到 52.2%。当权重偏离 0.5 时性能显著下降，纯 CLIP 和纯 RPN 配置分别下降 2.1 和 2.3 个百分点。新类检测性能对权重变化更为敏感（波动范围 5.3 个百分点），而基类性能相对稳定（波动 1.5 个百分点），验证了 RPN 定位能力与 CLIP 语义理解能力互补的重要性。

(6) 组件累积贡献分析

为深入理解 RAPID-OVD 框架中各技术组件的贡献，本文进行了系统的累积效应分析。如表 14 所示，各技术组件呈现出明显的递进式性能提升效应：

自适应填充策略通过保持目标原始形态贡献 2.5 个百分点的改进，RAPID-CLIP 微调借助键值缓存机制进一步提升 2.4%，完整的四阶段伪标签优化流程通过多源信息融合最终实现 3.3%的增益。三个核心组件的协同作用使AP_{novel}从 VL-PLM 基线的 32.3%显著提升至 40.5%，总计 8.2%的性能改进，验证了框架设计的有效性和各组件间的协同效应。

4.4.3 不同尺寸物体的检测性能分析

表 15 全面展示了本自适应填充策略对检测性能的提升效果。为了深入剖析其内在机制，我们对其伪标签质量进行了定量评估。数据显示，新策略带来了全面且一致的性能优化。在所有评估维度中，对小物体的改进最为显著：不仅平均精度（AP, IoU=0.50:0.95）从 0.107 大幅提升至 0.138（相对增幅 29.0%），其检出率（AR, maxDets=100）也从 0.146 提升至 0.183（相对增幅 25.3%）。

表 **14** RAPID-OVD 各技术组件的累积效应分析

方法	AP_{novel}	Δ相对改进	AP_{base}	AP_{all}
VL-PLM[32]	32.3	-	54.0	48.3
+ 自适应填充策略(APS)	34.8	+2.5	54.8	49.6
+ RAPID-CLIP微调	37.2	+4.9	55.1	50.4
+ 完整的四阶段优化流程	40.5	+8.2	56.3	52.2

4.4.4 可视化结果分析

图 5 展示了 RAPID-OVD 与直接基线方法 VL- PLM 和当前最优的 OV-DQUO 方法在 COCO-OVD 验证集上的检测结果对比。RAPID-OVD 在检测覆盖率和准确性上都取得了显著提升。推理阶段阈值设置为τ =0.7，θ_{nms} =0.3。

场景一：户外运动场景

第一组户外运动场景包含 person 和 skis 两个目标。VL-PLM 未检测出任何目标；OV- DQUO 和 RAPID-OVD 都成功检测出两个目标，但 RAPID-OVD 在 skis 上置信度提升至 83%。

场景二：办公室场景

第二组办公室场景包含了 keyboard、mouse、laptop、tv 等多种办公设备。VL-PLM 仅能检测到部分目标，存在较多遗漏；OV-DQUO 在检测覆盖率上有所提升

表 **15** 不同尺寸物体的检测性能

评估指标	小物体				中等物体				大物体				所有物体			
	基线	自适应	绝对↑	相对↑(%)	基线	自适应	绝对↑	相对↑(%)	基线	自适应	绝对↑	相对↑(%)	基线	自适应	绝对↑	相对↑(%)
AP (IoU=0.50:0.95)	0.107	0.138	0.031	+29.0	0.191	0.209	0.018	+9.4	0.212	0.219	0.007	+3.3	0.154	0.165	0.011	+7.1
AP (IoU=0.50)	0.251	0.272	0.021	+8.4	0.323	0.334	0.011	+3.4	0.341	0.350	0.009	+2.6	0.288	0.299	0.011	+3.8
AP (IoU=0.75)	0.163	0.166	0.003	+1.8	0.228	0.245	0.017	+7.5	0.246	0.250	0.004	+1.6	0.193	0.202	0.009	+4.7
AR (maxDets=100)	0.146	0.183	0.037	+25.3	0.327	0.357	0.030	+9.2	0.428	0.436	0.008	+1.9	0.297	0.317	0.020	+6.7

图 5 三种不同方法的可视化结果

，但在某些小目标（如 keyboard、mouse）的检测上仍显不足；RAPID-OVD 实现了最全面的目标覆盖，成功检测出 keyboard (88%)、mouse (98%)、laptop (99%)、tv (97%)等多个目标，置信度普遍较高。

场景三：室内餐桌场景

第三组室内餐桌场景进一步验证了各方法在日常生活环境中的适应性。该场景包含了 knife、cup、spoon、bowl、person 等多种日常物品的检测挑战。VL-PLM 检测到 5 个目标，包括 knife (71%)、cup (83%, 79%)、spoon (77%)、person (72%)；OV-DQUO 检测到 7 个目标，在 person (91%)和 knife (89%)的检测上有所提升，并新增了 bowl (77%)的检测；RAPID-OVD 表现最为出色，检测到 8 个目标实例，不仅在关键目标上实现了置信度的显著提升——knife (96%)、spoon (92%)、person (93%)、bowl (80%)，还保持了对多个 cup 目标的稳定检测。

场景四：工作站场景

第四组工作站场景展示了更为复杂的办公环境检测挑战。该场景包含了 person、tv、mouse、keyboard、laptop 等多种办公设备和工具的混合检测任务。VL-PLM 检测到 4 个目标，包括 tv (88%)、person (87%, 77%)、mouse (84%)，但遗漏了重要的 keyboard 和 laptop 目标；OV-DQUO 检测到 6 个目标，成功识别出 person (97%, 93%)、tv (94%)、mouse (92%)、keyboard (73%)、laptop (82%)，在检测覆盖率上有显著提升；RAPID-OVD 在此场景中继续保持了优异的性能表现，同样检测到 6 个目标，但在置信度方面实现了全面提升——person (100%, 96%)、tv (97%)、mouse (94%)、keyboard (75%)、laptop (88%)，相比前两种方法在检测精度和置信度方面都有明显改善。

5 总结与展望

在本研究中，首次探讨并解决了开放词汇目标检测中伪标签质量不佳、区域提议变形以及少样本场景泛化能力不足的关键技术问题。为了克服现有方法在复杂视觉场景下对未见类别检测能力的性能瓶颈，本文提出了 RAPID-OVD 框架。

本文首先提出了一个基于键值缓存机制的 RAPID-CLIP 模块，通过构建可查询的知识库实现少样本知识与预训练知识的有效融合，并进一步引入相似度检索和特征融合技术，从而更精确地学习区域级特征表示。此外，本文还设计了自适应填充策略，根据目标区域形态动态调整填充方式，有效保持物体原始比例和空间结构信息，并采用多阶段筛选技术优化伪标签生成流程，以提供更高质量的监督信号，从而优化模型并提升其在新类别检测上的泛化能力。通过在 COCO-OVD 和 LVIS-OVD 等权威数据集上的大量实验验证，相较于现有的最先进算法，RAPID-OVD 能够更准确地识别和定位未见类别目标，在新类别检测上达到 40.5%的AP_{novel}，显著提升了开放词汇目标检测任务的有效性。

尽管本文提出的 RAPID-OVD 模型有效提升了开放词汇目标检测任务的性能，但仍面临以下挑战：

(1) RAPID-OVD 在每个候选区域上执行 CLIP 特征提取和键值缓存检索，涉及大量的相似度计算和特征融合操作，导致推理时间和内存消耗较高。未来工作可探索知识蒸馏技术，将 RAPID-CLIP 的能力迁移至轻量级网络，或通过特征缓存和批处理优化降低计算复杂度，以满足实时检测应用的需求；

(2) RAPID-OVD 主要针对二维图像中的目标检测任务，但三维场景理解、视频序列分析和实例分割同样是计算机视觉领域的重要研究课题。因此，如何将 RAPID-OVD 的核心思想拓展至三维点云检测、视频时序建模和像素级分割任务，例如设计时空特征融合机制以捕捉视频中的动态目标特征，或引入几何先验知识实现三维空间中的开放词汇检测；

(3) RAPID-OVD 通过键值缓存机制实现少样本知识的检索和融合，以增强模型对新类别的理解能力。然而，近期研究表明，基于大语言模型的多模态推理和上下文学习在零样本理解方面可能具有显著优势。因此，如何将这些先进的多模态大模型技术融入开放词汇检测框架，利用

其强大的语言理解和推理能力进一步提升对复杂场景和长尾类别的检测性能，并探索端到端的多模态学习范式。

(4) RAPID-OVD 在追求新类别检测能力突破的过程中，基类性能出现了下降，这反映了当前开放词汇检测领域普遍存在的性能权衡挑战。这种权衡主要源于有限模型容量下新类伪标签学习与基类监督学习的资源竞争，以及方法组件针对新类分布特征的优化偏向。未来工作将重点探索多任务均衡学习策略，设计动态权重分配机制以平衡新类学习与基类知识保持；研发通用化自适应填充策略，特别是针对具有复杂几何形态特征的类别（如细长物体、不规则形状目标等）设计更加鲁棒的区域处理机制，以减少方法对特定类别分布的偏向性依赖；引入渐进式知识蒸馏机制，在提升新类检测能力的同时有效维护基类特征表示的稳定性，致力于在保持AP_{novel}领先优势的前提下显著提升AP_{base}性能，实现开放词汇目标检测中新类扩展能力与基类稳定性的双重优化。

(5) 自适应填充策略的阈值参数（T_a=3000,T_r=1.4）经由在 COCO-OVD 数据集上的网格搜索法确定。尽管跨数据集验证实验表明该参数配置具有良好的泛化性能，但当前工作尚未对参数选择的合理性建立系统化的理论解释框架。实验观察表明，参数配置对模型性能存在显著影响，这一现象揭示了开放词汇检测系统中超参数与任务性能之间复杂的耦合关系，值得开展更深入的机理性研究。未来工作将从理论层面深入分析模型性能对关键超参数的敏感性机制，探究参数配置与数据集特性、模型架构之间的本质关联，为超参数的合理选择提供理论指导；同时，探索自动化超参数优化方法，通过数据驱动或元学习等技术手段，实现超参数的自适应调整，以提升方法在多样化应用场景中的鲁棒性与通用性。

参 考 文 献

[1] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection//Advances in Neural Information Processing Systems. Vancouver, Canada, 2024: 107984-108011.

[2] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision// Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2021: 8748-8763.

[3] Sun Z, Fang Y, Wu T, et al. Alpha-clip: A clip model focusing on wherever you want//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 13019-13029.

[4] Yu C, Liu X, Wang Y, et al. Tf-clip: Learning text-free clip for video-based person re-identification//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 6764-6772.

[5] Tan C, Tao R, Liu H, et al. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025: 7184-7192.

[6] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[7] Zhao S, Schulter S, Zhao L, et al. Taming self-training for open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 13938-13947.

[8] Kim J, Cho E, Kim S, et al. Retrieval-augmented open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 17427-17436.

[9] Han G, Lim S N. Few-shot object detection with foundation models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 28608-28618.

[10] Su B, Zhang H, Li J, et al. Toward generalized few-shot open-set object detection. IEEE Transactions on Image Processing, 2024, 33: 1389-1402.

[11] Kage P, Rothenberger J C, Andreadis P, et al. A review of pseudo-labeling for computer vision. arXiv preprint arXiv:2408.07221, 2024.

[12] Li H, Wu Z, Shrivastava A, et al. Rethinking pseudo labels for semi-supervised object detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2022: 1314-1322.

[13] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 4794-4803.

[14] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks//Advances in Neural Information Processing Systems. Montreal, Canada, 2015.

[15] Xin Z, Chen S, Wu T, et al. Few-shot object detection: Research advances and challenges. Information Fusion, 2024, 107: 102307.

[16] Madan A, Peri N, Kong S, et al. Revisiting few-shot object detection with vision-language models//Advances in Neural Information Processing Systems. Vancouver, Canada, 2024: 19547-19560.

[17] Park S, Paik J. RefCap: image captioning with referent objects attributes. Scientific Reports, 2023, 13(1): 21577.

[18] Kim J J, Lee D G, Wu J, et al. Visual question answering based on local-scene-aware referring expression generation. Neural Networks, 2021, 139: 158-167.

[19] Wang Y, Ji Z, Wang D, et al. Towards unsupervised referring expression comprehension with visual semantic parsing. Knowledge-Based Systems, 2024, 285: 111318.

[20] Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision// Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2021: 4904-4916.

[21] Li J, Selvaraju R, Gotmare A, et al. Align before fuse: Vision and language representation learning with momentum distillation//Advances in Neural Information Processing Systems. Vancouver, Canada, 2021: 9694-9705.

[22] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 12888-12900.

[23] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 19730-19742.

[24] Lee J, Chun S, Yun S. Toward Interactive Regional Understanding in Vision-Large Language Models. arXiv preprint arXiv:2403.18260, 2024.

[25] Wang K, Cheng L, Chen W, et al. Marvelovd: Marrying object recognition and vision-language models for robust open-vocabulary object detection// Proceedings of the European Conference on Computer Vision. Milan, Italy, 2024: 106-122.

[26] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[27] Gao P, Geng S, Zhang R, et al. Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 2024, 132(2): 581-595.

[28] Zhang J O, Sax A, Zamir A, et al. Side-tuning: a baseline for network adaptation via additive side networks// Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 698-714.

[29] Zhang R, Zhang W, Fang R, et al. Tip-adapter: Training-free adaption of clip for few-shot classification// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 493-510.

[30] Zhu C, Chen L. A survey on open-vocabulary detection and segmentation: Past, present, and future. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8954-8975.

[31] Gu X, Lin T Y, Kuo W, et al. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921, 2021.

[32] Zhao S, Zhang Z, Schulter S, et al. Exploiting unlabeled data with vision and language models for object detection// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 159-175.

[33] Li L H, Zhang P, Zhang H, et al. Grounded language-image pre-training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 10965-10975.

[34] Zhou X, Girdhar R, Joulin A, et al. Detecting twenty-thousand classes using image-level supervision// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 350-368.

[35] Kuo W, Cui Y, Gu X, et al. F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639, 2022.

[36] Zhong Y, Yang J, Zhang P, et al. Regionclip: Region-based language-image pretraining//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16793-16803.

[37] Liang F, Wu B, Dai X, et al. Open-vocabulary semantic segmentation with mask-adapted clip//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7061-7070.

[38] Wysoczańska M, Ramamonjisoa M, Trzciński T, et al. Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2024: 1403-1413.

[39] Wang J, Li X, Zhang J, et al. Diffusion model is secretly a training-free open vocabulary semantic segmenter. IEEE Transactions on Image Processing, 2025.

[40] Wu S, Zhang W, Jin S, et al. Aligning bag of regions for open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 15254-15264.

[41] Lin C, Sun P, Jiang Y, et al. Learning object-language alignments for open-vocabulary object detection. arXiv preprint arXiv:2211.14843, 2022.

[42] Zhang H, Zhao Q, Zheng L, et al. Exploring region-word alignment in built-in detector for open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 16975-16984.

[43] Wang L, Liu Y, Du P, et al. Object-aware distillation pyramid for open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 11186-11196.

[44] Ma S, Qian D, Ye K, et al. Cake: Category aware knowledge extraction for open-vocabulary object detection//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025: 5982-5990.

[45] Li J, Zhang J, Li J, et al. Learning background prompts to discover implicit knowledge for open vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 16678-16687.

[46] Wu X, Zhu F, Zhao R, et al. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7031-7040.

[47] Wang Z, Zhou W, Xu J, et al. SIA-OVD: Shape-Invariant Adapter for Bridging the Image-Region Gap in Open-Vocabulary Detection//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 4986-4994.

[48] Liu M, Hayes T L, Ricci E, et al. Shine: Semantic hierarchy nexus for open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 16634-16644.

[49] Wang J, Chen B, Kang B, et al. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025: 7762-7770.

ZHU Rui, Ph. D., professor. His main research is computer vision.

XU Kai-Rui, Master student. His research interest is computer vision.

LIU Bo-Han, Master student. His research interest is computer vision.

LU Zhao-Kang, Master student. His research interest is computer vision.

Cao Duan-Yu, Master student. His research interest is natural language processing.

ZHANG Xuan, Ph. D., professor. Her main research is natural language processing.