Software Engineering and Applications 软件工程与应用, 2025, 14(2), 261-270
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.142024

Hans汉斯

YW-FSVOD：基于YOLO-World的开放词汇小样本目标检测方法

张艺博，张索非

南京邮电大学物联网学院，江苏 南京

收稿日期：2025年2月13日；录用日期：2025年4月3日；发布日期：2025年4月18日

摘　要

小样本目标检测作为计算机视觉的重要分支，致力于模拟人类从极少量样本中学习新目标检测的能力。现有方法依赖大量标注数据，且仍受限于预定义类别表征的刚性约束，因此难以适应开放词汇场景。除此之外，CLIP等视觉－语言预训练模型通过跨模态对齐展现了零样本推理潜力，但其聚焦于图像分类任务，而小样本条件下的目标检测模型性能退化仍面临关键挑战。为此，本文提出了一种基于YOLO-World的开放词汇小样本目标检测方法YW-FSVOD。该方法通过构建语言描述引导的多尺度特征对齐机制，将文本语义嵌入至YOLO-World的视觉编码空间，增强模型对未见类别的泛化能力；并采用预计算文本嵌入替代完整语言模型计算，在保证检测精度的同时实现推理速度的显著提升。实验结果表明，YW-FSVOD在COCO和LVIS数据集上表现优异，精度显著优于传统的小样本目标检测框架。

关键词

小样本目标检测，视觉语言模型，开放词汇，两阶段微调，YOLO-World

YW-FSVOD: Open-Vocabulary Few-Shot Object Detection Method Based on YOLO-World

Yibo Zhang, Suofei Zhang

School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Feb. 13th, 2025; accepted: Apr. 3rd, 2025; published: Apr. 18th, 2025

Abstract

Few-shot object detection, as an important branch of computer vision, aims to simulate the human

文章引用: 张艺博, 张索非. YW-FSVOD: 基于 YOLO-World 的开放词汇小样本目标检测方法[J]. 软件工程与应用, 2025, 14(2): 261-270. DOI: 10.12677/sea.2025.142024

ability to learn new object categories from a minimal number of samples. Existing methods typically rely on large annotated datasets and are constrained by rigid predefined category representations, making them difficult to adapt to open-vocabulary scenarios. Moreover, vision-language pretraining models, such as CLIP, have shown promising zero-shot inference capabilities through cross-modal alignment, but these models are primarily focused on image classification tasks. The performance of object detection models under few-shot conditions still faces significant challenges. To address this issue, we propose a novel open-vocabulary few-shot object detection method, YW-FSVOD, based on YOLO-World. This approach constructs a multi-scale feature alignment mechanism guided by textual descriptions, embedding textual semantics into the visual encoding space of YOLO-World to enhance the model's generalization ability to unseen categories. Additionally, we replace the full language model computation with precomputed text embeddings, significantly improving inference speed while maintaining detection accuracy. The experimental results show that YW-FSVOD performs excellently on both the COCO and LVIS datasets, with accuracy significantly surpassing that of traditional few-shot object detection frameworks.

Keywords

Few-Shot Object Detection, Vision-Language Model, Open-Vocabulary, Two-Stage Fine-Tuning, YOLO-World

1. 引言

传统的深度学习目标检测方法通常依赖于大规模的标注数据集来实现理想的性能，这在许多领域并不现实，因为获取大量标注数据往往需要耗费巨大的时间和资源，特别是在医学影像分析、遥感监测和工业检测等领域。因此，小样本学习(Few-Shot Learning, FSL) [1]成为备受关注的研究课题。

小样本学习的核心理念是通过少量的标注样本训练出具有良好泛化能力的模型，从而在新的任务或类别上实现有效的检测和识别。这一过程不仅要求模型具备从有限样本中学习的能力，还需要具备跨任务迁移知识的能力以提高学习效率。为克服小样本学习带来的挑战，研究者们提出了多种策略，包括迁移学习、元学习、数据增强和生成对抗网络(GAN) [2]等，这些方法为相关领域的应用与发展提供了新的思路与实践途径。

随着语言辅助视觉模型的发展，小样本学习取得了进一步的发展。然而，现有的这些语言辅助视觉的小样本研究主要集中在基本的图像分类任务[3]，对目标检测任务的关注相对较少。与图像分类不同，目标检测不仅需要识别对象的类别，还需要在海量候选区域中精确定位目标。这一额外任务显著增加模型复杂度和检测难度。此外，传统的小样本目标检测依赖于手动标注类别来训练模型，难以适应不同领域的数据和任务。每当面对新任务时，都需要进行大量的训练数据，导致模型难以在未知的类别或环境下有效泛化。本文针对这些挑战做出以下创新工作：

1) 开放场景下的轻量化小样本模型微调策略的优化与应用。本文提出并详细探讨了小样本微调对开放模型性能的优化。验证了在少量标注数据的情况下，通过两阶段的微调策略，能够显著提升小样本目标检测性能，特别是在 AP 和 AP50 等指标上均表现出稳定的提升。

2) 在 COCO 和 LVIS 数据集上对不同规模的模型进行性能评估。实验结果表明，所提出的方法在小样本条件下表现出了明显的性能提升。与传统模型相比，该方法在精度和速度均具有较强的优势，并在

泛化能力和任务适应性方面展现出良好的效果。

2. 相关工作

2.1. 小样本学习

小样本目标检测(Few-Shot Object Detection, FSOD)已成为计算机视觉领域的重要研究方向，旨在从有限数量的标记样本中训练出具备良好泛化能力的目标检测模型。在许多场景下，例如医学影像、遥感监测和工业检测等专业领域，获取大规模标注数据集既困难又昂贵，因此这一研究在实际应用中尤为关键。

早期的目标检测方法依赖于人工特征和滑动窗口技术，例如方向梯度直方图(HOG)和尺度不变特征变换(SIFT)，以及支持向量机(SVM)等分类器。这些方法虽在一定程度上提高了目标检测性能，但在复杂场景下仍存在局限性。随着深度学习的兴起，基于卷积神经网络(CNN)的检测方法(如 R-CNN 等) [4]极大地提高了在复杂场景下的检测精度。然而，这些方法在训练时通常需要大量的标注数据，难以适应小样本场景。为了在有限数据下提升检测性能，研究者引入了迁移学习、元学习等小样本学习概念。迁移学习通过在大型数据集上预训练模型并在小数据集上微调，使模型在新任务中取得良好性能；元学习则通过“学习如何学习”的方法使模型快速适应新任务，例如模型无关的元学习(MAML)和原型网络(ProtoNet)等方法表现突出。

目前主流的小样本目标检测方法主要包括：基于度量学习[5]的方法：如孪生网络，通过学习一个良好的特征空间来区分不同类别；基于注意力机制[6]方法：通过关注重要的特征区域来提升目标检测的准确性，如 Attention RPN (区域提议网络)；基于数据增强[7]的方法：通过生成更多训练样本来增强模型的泛化能力。除此之外，小样本目标检测在多模态学习[8]、自监督学习[9]和改进神经网络架构[10]等方面取得了显著进展。多模态学习通过结合图像和文本等多种数据模态来提高模型性能；自监督学习利用无标签数据进行预训练，然后在有限标注数据上微调；而 Vision Transformers (ViT)等架构通过结合 Transformer 技术提升特征提取能力，进一步提高小样本检测效果。

尽管小样本研究领域已经取得了显著进展，但大多数方法仍然依赖预定义和训练好的对象类别，这在开放场景下存在诸多挑战。首先，现有方法通常假设目标类别在训练阶段已经出现过，或至少与训练类别具有相似的特征表示，然而在实际应用中，模型往往需要识别完全未知或不相关的类别，这种局限性导致其检测性能大幅下降；其次，现实场景中的数据分布通常具有长尾特性，即少数主流类别占据了大部分数据，而大量类别的数据量极少，这使得传统小样本目标检测方法难以适应动态变化，并在类别不均衡的情况下难以保持良好的检测效果。最后，在诸如自动驾驶、安防监控等需要高效响应的应用场景下，模型往往需要快速适应新类别，而现有方法大多依赖额外的训练，缺乏足够的灵活性和实时性。因此，如何摆脱对固定类别的依赖，提高模型在开放场景中的泛化能力、适应性和实时性，仍然是当前亟待解决的关键问题。

2.2. 目标检测器 YOLO

YOLO (You Only Look Once) [11]-[13]系列的发展显著推动了目标检测领域的进步，提供了一种在速度与精度之间取得平衡的实时检测框架。与 Faster R-CNN [14]和 Mask R-CNN [15]等双阶段检测器不同，YOLO 模型采用单阶段架构，这不仅简化了检测流程、减少了计算开销，也增强对多尺度物体的检测能力。YOLO 团队持续优化网络架构(YOLOv1-YOLOv8 版本)，实现了其在各种场景下的强适应性和高检测性能。

随着视觉语言模型在开放词汇目标检测(Open-Vocabulary Object Detection, OVD)中的应用，检测器迫

切需要具备识别未见类别的能力。YOLO-World [16]作为 YOLO 系列的一个创新扩展，结合视觉语言建模，实现了开放词汇检测，并在零样本场景中表现出色。YOLO-World 通过引入可重参数化的视觉－语言路径聚合网络(RepVL-PAN)和区域－文本对比损失来支持高效的开放词汇检测。这种方法使 YOLO-World 能够在传统固定词汇模型与现代需求之间架起桥梁，为现实世界应用中的上下文感知检测提供了有前途的解决方案。通过集成视觉－语言预训练，YOLO-World 在多样化场景下的检测更加精准和鲁棒。此外，与其他开放视觉小样本模型如 GroundingDINO [17]相比，YOLO-World 由于采用单阶段检测，参数量更小，计算成本更低，更便于部署，为开放场景下的目标检测提供了一个轻量化、高效的解决方案。虽然其在零样本目标检测有很好的效果，但是小样本目标检测领域依然缺失。本文在 YOLO-World 的基础上进行改进，使其更加适应小样本目标检测任务，并进一步提升其检测性能与泛化能力。

3. 基于 YOLO-World 的二阶段微调方法

本文采用二阶段微调方法对 YOLO-World 模型进行改进，将其强大的零样本场景检测能力拓展至小样本目标检测任务，从而更好地解决目标检测中样本不足问题。

3.1. 第一阶段：基础模型

基础模型主要由三个核心组件组成：文本编码器、YOLO 视觉检测器以及可重新参数化的视觉语言路径聚合网络(RepVL-PAN)，模型整体架构如图 1 所示。其中，文本编码器负责将输入文本转换为文本嵌入，供后续模型处理；YOLO 视觉检测器中的图像编码器负责从输入图像中提取多尺度特征，以捕捉丰富的视觉信息；RepVL-PAN 通过跨模态融合机制将图像特征和文本嵌入结合，增强两者的联合表示能力。这种跨模态融合机制不仅提升了视觉和语言信息的表达效果，也增强了 YOLO-World 在多模态任务中的鲁棒性和高效性，使其在小样本目标检测等任务中表现更优。

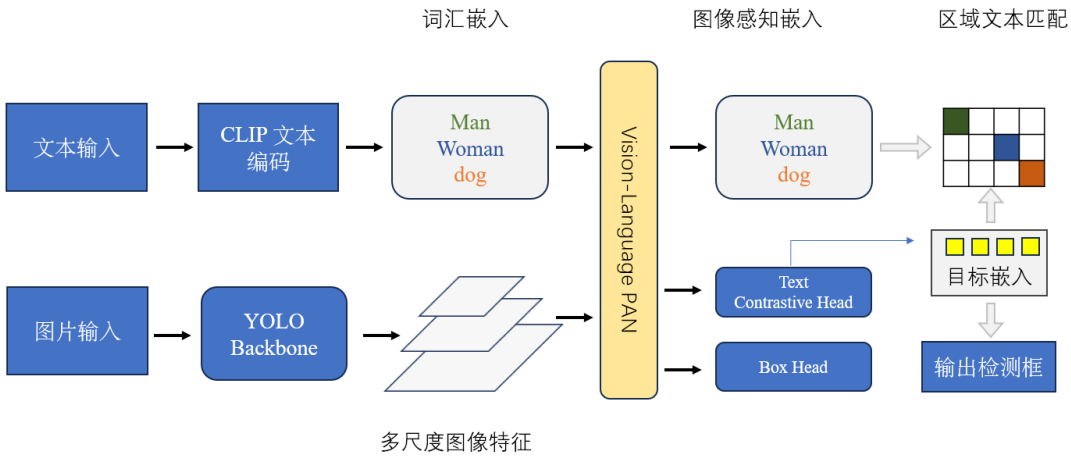

Figure 1. YOLO-World architecture
图 1. YOLO-World 结构图

在基本模型训练过程中，YOLO-World 通过区域－文本对比损失进行学习，以提升目标检测的跨模态理解能力。对于给定的马赛克样本 I 及其对应的文本 T，YOLO-World 输出 K 个对象预测 $\{B_k, s_k\}_{k=1}^{K}$，并生成注释集合 $\Omega = \{B_i, t_i\}_{i=1}^{N}$。为确保预测结果与真实标签的准确对齐，该方法采用了任务对齐的标签分配策略，将预测结果与真实注释进行匹配，并将每个正样本的预测分配给相应的文本索引作为分类标签。区域－文本对比损失 L_{con} 通过优化对象与文本(区域－文本)相似性和对象－文本分配之间的交叉熵来构

建。此外，为了提高边界框预测的准确性，该方法结合了 IoU (Intersection over Union)损失和分布式焦点损失(DFL)进行优化，使模型在目标定位上更加精确，同时增强检测的稳定性和泛化能力。

因此，总的训练损失函数定义为：

$$L(I) = L_{con} + \lambda_I \cdot (L_{iou} + L_{dlf}) \tag{1}$$

其中，λ_I 是一个指示因子。当输入图像 I 来自检测或图像－文本关联数据时，$\lambda_I = 1$；当图像来自图像－文本数据时，$\lambda_I = 0$。由于图像－文本数据集中存在噪声边界框，回归损失仅对具有精确边界框的样本进行计算。

3.2. 第二阶段：视觉文本多尺度微调

为了使预训练的 YOLO-World 模型适配小样本目标检测，本文提出了一种高效的微调策略。该策略在充分利用预训练的多模态特征的同时，聚焦于模型的特定组件进行优化，而无需重新训练整个模型，从而在保证性能的同时提升效率和任务适应性。具体而言，主干网络采用混合结构，用于同时处理视觉和文本输入，如图 2 所示。该结构在使用提取图像特征的基础上，进一步结合文本信息，使模型能够更好地理解小样本场景下的目标，并提高跨模态的泛化能力。为适配小样本目标检测任务，本文在视觉特征提取及文本嵌入机制上进行了针对性调整，以提升模型的泛化能力和推理效率。

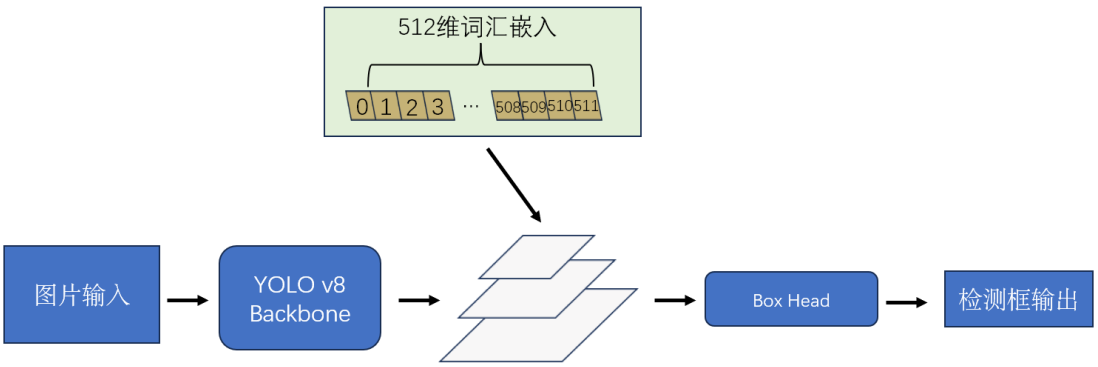

Figure 2. Multi scale fine-tuning of visual text model structure
图 2. 视觉文本多尺度微调后模型结构

3.2.1. 视觉特征提取方法

为了高效提取图像中的丰富视觉信息，本文采用了 YOLO-World 框架中的主干网络——YOLOv8 作为基础视觉特征提取模型。YOLOv8 融合了 EfficientNet 与 CSPNet 两大架构优势，通过解耦头部设计和先进的数据增强策略，可以实现高精度目标定位和实时检测。而为了减轻模型的计算开销、提升推理速度及细粒度特征捕捉不足的问题，本文在第二阶段对原始模型结构进行了针对性优化。具体改进措施为：

结构简化与模块替换：第二阶段转向对较小的数据集的优化，如本文中 COCO 数据集，以更好适应具体的目标类别，文本特征的引导不再是最关键的任务，无需再依赖第一阶段复杂的视觉语言融合。因此本文移除原模型中的 T-CSPLayers 模块和 Image-Pooling Attention 模块(即移除了 RepVL-PAN 结构)，简化模型结构，以降低模型复杂度，减少参数冗余，提升推理速度。

多尺度卷积特征提取：采用改进的卷积网络架构实现多尺度特征提取，通过生成分层特征图，使模

型在保持计算高效性的同时，能够捕获更丰富的局部细粒度信息与全局上下文语义。

这种改进策略不仅可以简化模型结构，还能降低对大规模数据的依赖，而且在小样本情景下可以显著提升特征表达能力和模型鲁棒性。

3.2.2. 文本嵌入模块

针对目标检测任务中类别语义信息的有效利用，本文提出了一种基于预先计算提示嵌入的文本嵌入模块。与传统的 YOLO-World 系列方法不同，我们在第二阶段训练过程中不再引入额外的语言编码器，而是直接将预先生成的文本嵌入作为语义输入，为检测头提供与类别紧密关联的特征信息。

预计算的类语义嵌入：文本嵌入采用基于 CLIP 的文本编码器或图文融合模型预先生成，将每个类别的名称映射为固定维度(D = 512)的特征向量。在微调阶段，为进一步降低语义噪声，我们仅使用单词作为文本输入，而非完整语句，从而使预计算的类语义嵌入更加精炼。以 COCO 数据集为例，我们为其中 80 个类别分别生成了对应的 512 维文本特征向量，并在模型加载时直接注入检测框架中。

多模态特征融合：在整个检测网络中，这些预计算且冻结的文本嵌入在 Neck 阶段与多尺度视觉特征进行深度融合，使得生成的特征图中蕴含了显式的类别语义信息。借助这种融合策略，检测头(Head)无需在推理阶段在线调用文本编码器，即可利用融合后的多模态特征直接进行分类与定位预测。该设计大大降低了计算复杂度和推理时延，提升了检测效率，尤其在小样本场景下展现出更为显著的优势。

通过将预先计算的文本嵌入引入到目标检测流程中，本文方法可以有效解决传统方法中仅依赖纯视觉特征进行类别判别时可能遇到的多类分布不均和高阶语义信息缺乏的问题。

4. 实验

4.1. 数据集

第一阶段训练基于 YOLO-World 所提供的预训练模型，该模型已在基础类别数据上进行了充分训练，具备良好的通用视觉特征提取能力。其常用数据集组合及其特点如下：O365 + GoldG + CC3M-Lite：此组合的预训练模型涵盖更广泛的视觉概念，但其精度相对略低；O365 + GoldG：该组合包含较少的检测类别，但模型的精度更高。为了进一步提升模型对不同类别的适应能力，本文采用 O365 + GoldG 预训练模型，以在精度和泛化能力之间取得平衡。

在第二阶段微调过程中，我们使用 COCO 数据集[18] (Common Objects in Context)对模型进行进一步训练和算法评估，以验证其在具体任务上的检测性能。COCO 数据集提供了丰富的图像和标注信息，是目标检测领域中最为广泛使用的标准数据集之一。为了实现小样本学习，我们从原始的 COCO 数据集中提取了一个子集，目的是为每个类别提供少量的图像用于训练。在此过程中，我们限制了每个类别的最大图像数量，并根据不同的实验需求生成了多个子集。例如，对于每个类别，我们创建了 10-shot 和 30-shot 子集。这些子集保证了每个类别的样本数量在有限的范围内，使得模型能够在有限的样本条件下进行有效的训练和评估。在 COCO 数据集的标准注释文件基础上，我们通过随机选择每个类别的图像，生成了多个不同规模的子集，这些子集包含了训练所需的图像及其相关标注。每个子集的构建都以类别为单位，确保了各个类别在不同实验条件下的图像数量可控。通过这样的处理，我们能够对小样本目标检测进行更加精细的实验和模型评估。

4.2. 实验环境和训练策略

4.2.1. 实验环境

实验使用的计算机的 GPU 是 Tesla K80 (11 GB)*12。CPU 是 40 vCPU Intel(R) Xeon(R) Silver 4210

CPU @ 2.20GHz。PyTorch 版本为 1.11.0，Python 版本为 3.9.0，CUDA 版本为 11.2。

4.2.2. 训练策略

在训练过程中，所有输入图像的分辨率统一为 640×640，以确保实验结果的一致性；学习率从预训练时的 2×10^{-3} 缩小 20 倍至 1×10^{-4}；训练的最大轮次为 80 轮，并设置了每个 GPU 的训练批次大小为 4，这样即使在 Tesla K80 偏旧的设备上训练，也有很快的微调模型的速度。另外，我们采用参数冻结策略来稳定模型学习。具体而言，冻结主干网络(backbone)的前四层，同时对 neck 和 head 部分的所有层进行冻结，使其在训练过程中不更新参数。此举确保视觉特征提取保持稳定、特征融合中引入的预计算提示嵌入信号保持一致，并避免检测头对嵌入特征的破坏性改动，从而在整体训练流程中有效维护小样本检测能力与多模态语义特征对齐的稳定性。此外，训练过程中每隔 5 个周期进行一次验证。训练过程中还使用了 EMA 平滑技术，以提高模型的稳定性和准确性。

4.3. 实验结果

4.3.1. 多尺度模型适应性实验

针对 YW-FSVOD 模型，本小节通过配置 L (Large)、M (Medium)、S (Small)三种参数规模的模型来评估该方法的性能表现，并将实验结果整理如下表 1。

Table 1. Accuracy performance under different shot settings
表 1. 不同 shot 下精度表现

模型	Config	AP	AP50	AP75
YOLO-World-v2-S	Zero-shot	37.6	52.3	40.7
YW-FSVOD-S (Ours)	10-shot	37.8	52.9	41.4
YW-FSVOD-S (Ours)	30-shot	38.5	53.4	41.8
YOLO-World-M	Zero-shot	42.8	58.3	46.4
YW-FSVOD-M (Ours)	10-shot	43.0	58.6	47.1
YW-FSVOD-M (Ours)	30-shot	43.5	59.1	47.5
YOLO-World-v2-L	Zero-shot	45.7	61.6	49.8
YW-FSVOD-L (Ours)	10-shot	45.9	61.8	50.3
YW-FSVOD-L (Ours)	30-shot	**46.3**	**62.3**	**50.7**

实验结果表明，本实验对于不同参数大小的 YW-FSVOD (L、S、M)均带来了显著的性能提升。在 YW-FSVOD 版本中，通过 10-shot 微调，模型的 AP 从零样本阶段的 45.76%提升至 45.96%，而 30-shot 微调将其进一步提升至 46.36%。类似地，AP50 也从零样本的 49.8%提升至 50.3%和 50.7%。由此可见，即使只有少量的标注样本，也能够显著增强模型的视觉特征学习能力，并提高检测精度。

对于 YOLO-World-v2-S 和 YOLO-World-v2-M 版本，小样本微调同样展现了较为显著的提升效果。YOLO-World-v2-S 通过 30-shot 微调，其 AP 从 37.65%提升至 38.55%，AP50 则从 40.7%提升至 41.8%；而 YOLO-World-v2-M 在 30-shot 微调下，AP 从 42.85%提升至 43.55%，AP50 从 46.4%提升至 47.5%。这些结果表明，小样本微调在模型性能方面表现出重要作用，特别是在泛化能力和任务适应性方面。

4.3.2. 与传统小样本模型对比

在本实验中，我们提出的 YW-FSVOD 模型在多个标准小样本物体检测任务上表现优异，并与传统的方法进行了对比。根据表 2 中的结果，我们可以看到，YW-FSVOD 在视觉语言模型下，10-shot 和 30-

shot 设置的配置在多个评价指标(AP、AP50 和 AP75)都显著的高于传统小样本模型。

具体而言，YW-FSVOD-S、YW-FSVOD-M 和 YW-FSVOD-L 在 10-shot 的 AP 分数上分别为 37.8、43.0 和 45.9，相较于其他方法(如 TFA w/ fc、FSCE 和 MPSR)提升了约 27.8 至 35.9 个点。在 30-shot 设置中，YW-FSVOD-L 更是表现出色，AP 分数高达 46.3，领先于现有最好的方法 Meta Faster R-CNN，其 AP50 和 AP75 也分别超过了 59.1 和 50.7，充分展示了我们的模型在小样本检测任务中的强大能力。

Table 2. Comparison of object detection models on the COCO dataset
表 2. COCO 数据集上的目标检测模型对比

Model/Method	10-shot AP	10-shot AP50	10-shot AP75	30-shot AP	30-shot AP50	30-shot AP75
TFA w/ fc [10]	10.0	19.2	9.2	13.4	24.7	13.2
TFA w/ cos [10]	10.0	19.1	9.3	13.7	24.9	13.4
MPSR [20]	9.8	17.9	9.7	14.1	25.4	14.2
FSCE [9]	11.9	-	10.5	16.4	-	16.2
Meta Faster R-CNN [21]	12.7	25.7	10.8	16.6	31.8	15.8
YW-FSVOD-S (Ours)	37.8	52.9	41.4	38.5	53.4	41.8
YW-FSVOD-M (Ours)	43.0	58.6	47.1	43.5	59.1	47.5
YW-FSVOD-L (Ours)	**45.9**	**61.8**	**50.3**	**46.3**	**62.3**	**50.7**

4.3.3. 参数效率及资源受限场景的适用性

本研究将新提出的 YW-FSVOD 与 YOLO-World 和 Grounding DINO-T 在 LVIS [19]进行了对比，后两者代表了零样本检测领域的先进方法。需要注意的是，虽然 YOLO-World 和 Grounding DINO-T 能够在没有特定类别样本的情况下直接进行推理，但本文的 YW-FSVOD 系列模型是基于小样本检测进行微调的，需要依赖于少量标注样本，因此无法与零样本检测方法直接等量对比。然而，即便在这种情况下，我们的新方法在 LVIS 数据集上的表现仍展示了显著的优势，尤其是在参数量和推理速度(FPS)方面，如表 3 所示。具体而言，YW-FSVOD-S 仅具有 12.8M 参数，其推理速度(76.2 FPS)优于 YOLO-World-S (13M 参数，74.1 FPS)，并且在多个指标(包括 AP、APr、APc、APf)上均取得了更高的表现。此外，YW-FSVOD-M 仅包含 28M 参数，相较于 Grounding DINO-T (172M 参数)，在保持极小模型规模的同时，YW-FSVOD-M 以 59.0 FPS 的推理速度远超越了 Grounding DINO-T，且在检测精度上也具有优势。这表明 YW-FSVOD 在计算效率和检测精度之间实现了良好的平衡，具有实际应用潜力，特别是在需要快速推理和低计算成本的场景中具有独特优势。

Table 3. The evaluation results of different models on LVIS
表 3. 不同模型在 LVIS 上的评估结果

Method	Backbone	Params	Pre-trained Data	FPS	AP	APr	APc	APf
Grounding DINO-T	Swin-T	172M	O365, GoldG	1.5	25.6	14.4	19.6	32.2
YOLO-World-S	YOLOv8-S	13M	O365, GoldG	74.1	26.2	19.1	23.6	29.8
YW-FSVOD-S(Ours)	YOLOv8-S	12.8M	O365, GoldG	76.2	26.8	20.5	24.2	30.6
YOLO-World-M	YOLOv8-M	29M	O365, GoldG	58.1	31.0	23.8	29.2	33.9
YW-FSVOD-M (Ours)	YOLOv8-M	28.5M	O365, GoldG	**59.0**	**32.4**	**24.6**	**30.1**	**35.0**

5. 结论

本文提出了一种基于 YOLO-World 的开放词汇小样本目标检测方法 YW-FSVOD，旨在解决传统小样本目标检测方法在开放场景中的应用限制。本文引入 YOLO-World 视觉－语言模型的多模态能力，并在视觉模块移除了 RepVL-PAN 结构，降低模型复杂度，减少参数冗余；而对语言模块不再引入额外的语言编码器，将预先生成的文本嵌入作为语义输入，为检测头提供与类别紧密关联的特征信息。最后在训练时对 backbone 的前四层、neck 和 head 部分冻结。

实验表明，在 10-shot 和 30-shot 下，YW-FSVOD 模型的精度对比传统的小样本检测模型有显著优势。而与现有的视觉语言模型 Grounding DINO-T 对比，其参数量和速度都有显著的优势，证明了本方法高效的轻量设计。

参考文献

[1] Wang, Y.Q., Yao, Q.M., Kwok, J.T. and Ni, L.M. (2020) Generalizing from a Few Examples. *ACM Computing Surveys*, **53**, 1-34. https://doi.org/10.1145/3386252

[2] Song, Y.S., Wang, T., Cai, P.Y., Mondal, S.K. and Sahoo, J.P. (2023) A Comprehensive Survey of Few-Shot Learning: Evolution, Applications, Challenges, and Opportunities. *ACM Computing Surveys*, **55**, 1-40. https://doi.org/10.1145/3582688

[3] Zhang, R.r., Hu, X.f., Li, B.h., Huang, S.y., Deng, H.q., Qiao, Y., *et al.* (2023) Prompt, Generate, Then Cache: Cascade of Foundation Models Makes Strong Few-Shot Learners. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 15211-15222. https://doi.org/10.1109/cvpr52729.2023.01460

[4] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/cvpr.2014.81

[5] Yan, X.P., Chen, Z.L., Xu, A.N., Wang, X.X., Liang, X.D. and Lin, L. (2019) Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 9576-9585. https://doi.org/10.1109/iccv.2019.00967

[6] Kang, B.Y., Liu, Z., Wang, X., Yu, F.H., Feng, J.S. and Darrell, T. (2019) Few-Shot Object Detection via Feature Reweighting. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 8419-8428. https://doi.org/10.1109/iccv.2019.00851

[7] Hariharan, B. and Girshick, R. (2017) Low-Shot Visual Recognition by Shrinking and Hallucinating Features. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 3037-3046. https://doi.org/10.1109/iccv.2017.328

[8] Gu, X.Y., Lin, T.-Y., Kuo, W.C. and Cui, Y. (2021) Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. arXiv: 2104.13921.

[9] Sun, B., Li, B.H., Cai, S.C., Yuan, Y. and Zhang, C. (2021) FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 7348-7358. https://doi.org/10.1109/cvpr46437.2021.00727

[10] Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E. and Yu, F. (2020) Frustratingly Simple Few-Shot Object Detection. arXiv: 2003.06957.

[11] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/cvpr.2016.91

[12] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/cvpr.2017.690

[13] Wang, C., Bochkovskiy, A. and Liao, H.M. (2023) Yolov7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 7464-7475. https://doi.org/10.1109/cvpr52729.2023.00721

[14] Ren, S., He, K.M., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv: 1506.01497.

[15] He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/iccv.2017.322

[16] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X. and Shan, Y. (2024) Yolo-World: Real-Time Open-Vocabulary Object Detection. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 16901-16911. https://doi.org/10.1109/cvpr52733.2024.01599

[17] Madan, A., Peri, N., Kong, S. and Ramanan, D. (2023) Revisiting Few-Shot Object Detection with Vision-Language Models. arXiv: 2312.14494.

[18] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al.* (2014) Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[19] Gupta, A., Dollar, P. and Girshick, R. (2019) LVIS: A Dataset for Large Vocabulary Instance Segmentation. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5351-5359. https://doi.org/10.1109/cvpr.2019.00550

[20] Wu, J.X., Liu, S.T., Huang, D. and Wang, Y.H. (2020) Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 456-472. https://doi.org/10.1007/978-3-030-58517-4_27

[21] Han, G.X., Huang, S.Y., Ma, J.W., He, Y.C. and Chang, S.-F. (2022) Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 780-789. https://doi.org/10.1609/aaai.v36i1.19959