

开放目标检测技术演进：从OVD到统一范式OVD+OWD

XXX

2025 年 12 月 27 日

1 引言与动机 (Introduction & Motivation)

1.1 研究背景

1.1.1 传统目标检测的成功与局限

在过去十年中，深度学习驱动的目标检测技术取得了巨大突破。从区域提议网络（R-CNN系列）[1] 到单阶段检测器（YOLO系列、SSD、RetinaNet），再到基于 Transformer 的端到端检测器（DETR系列），目标检测在精度和速度上都实现了质的飞跃。这些方法在 COCO、PASCAL VOC 等标准数据集上屡创新高，推动了自动驾驶、医疗影像分析、安防监控等领域的实际应用。

然而，传统目标检测方法严格受限于**封闭世界假设 (Closed-World Assumption, CWA)**。在这一假设下，模型的分类体系由训练数据集预先定义且固定不变。例如，在 COCO 数据集上训练的检测器只能识别其定义的 80 个类别，对于任何不在训练集中的物体，模型要么将其错误分类为相似的已知类别，要么直接忽略为背景。

封闭世界假设的核心问题主要体现在两个方面：

(1) 静态类别空间：一旦模型训练完成，其可识别的类别集合就被固定。当需要检测新类别时，必须重新收集数据、标注并重新训练整个模型。

(2) 排他性判定机制：传统检测器的分类头学习的是从图像特征到离散类别 ID（如 0, 1, ..., 79）的映射关系，这些 ID 本身不包含任何语义信息。模型通过 Softmax 函数强制每个候选区域归属于 N 个已知类别之一，本质上是一种“封闭集合多分类”问题：

$$P(c_i|x) = \frac{e^{z_i}}{\sum_{j=1}^N e^{z_j}}, \quad c_i \in \{c_1, c_2, \dots, c_N\}$$

其中 x 是区域特征， z_i 是对应类别的 logit。这种机制天然地排斥了第 $N + 1$ 类的存在。

1.1.2 现实世界的长尾分布与标注瓶颈

封闭世界假设与现实世界的开放性、动态性存在根本性矛盾：

(1) 极端的长尾分布

现实世界中的物体类别分布服从极端的长尾定律（Zipf's Law）。以 LVIS 数据集[2]为例，其包含 1203 个类别，但这些类别的出现频率极不平衡：**频繁类**（如 “person”、“car”）在数据集中出现数千次，**常见类**（如 “guitar”、“laptop”）出现数百次，而**罕见类**（如 “accordion”、“trombone”）仅出现数十次。

更重要的是，即便是包含 1203 类的 LVIS，相对于真实世界中数以百万计的物体类别（考虑不同品牌、型号、状态的细分），依然是沧海一粟。传统方法试图通过不断扩充数据集来覆盖更多类别，但这种策略在数学上是不可持续的。

(2) 标注成本的指数级增长

假设标注一张图像中所有物体的平均成本为 C , 类别数量为 N , 那么覆盖 N 个类别所需的标注成本为:

$$Cost_{total} = C \times N \times k$$

其中 k 是每个类别所需的样本数 (通常需要数千张以保证训练效果)。当 N 从 80 (COCO) 增长到 1203 (LVIS) 再到 10000+ (开放世界) 时, 所需的人工标注成本呈指数级增长, 这在经济上和时间上都是不可接受的。

1.1.3 实际应用场景的迫切需求

在真实的应用场景中, 封闭世界假设带来的局限性尤为明显, 具体体现在以下几个关键领域:

(1) 自动驾驶场景中的未知物体检测

自动驾驶系统面临的最大挑战之一是处理训练时未见过的新物体。例如, 道路上可能出现新型工程车辆 (如新型挖掘机、特殊用途的工程机械)、临时路障 (如施工标志、事故现场的临时障碍物)、以及罕见动物 (如大型野生动物穿越道路、家养宠物突然出现) 等。这些物体在训练数据集中可能完全不存在, 但自动驾驶系统必须能够及时识别并做出安全决策。传统封闭集检测器要么将这些未知物体错误分类为相似的已知类别 (如将新型工程车辆误判为“卡车”), 要么直接忽略, 这可能导致严重的安全事故。开放目标检测技术通过零样本泛化能力, 使得系统能够识别任意描述的物体, 即使这些物体在训练时从未见过。

(2) 智能安防系统中的异常检测

在智能安防领域, 监控系统需要检测各种异常物体和事件。这些异常往往无法预先定义, 因为威胁和异常情况是动态变化的。例如, 系统需要检测可疑包裹 (可能包含危险物品)、非法侵入的动物 (如野生动物进入人类活动区域)、以及各种未预期的异常行为。传统方法只能检测预定义的异常类别, 但实际威胁往往是新颖的、无法预见的。开放目标检测技术使得系统能够通过自然语言描述来指定检测目标 (如“检测可疑的包裹”), 或者主动发现任何异常物体, 从而提供更灵活的安防能力。

(3) 医疗影像分析中的罕见疾病发现

在医疗影像分析中, 罕见疾病的影像特征可能在训练集中完全缺失。例如, 某些罕见遗传病的影像表现、新型疾病的影像特征、或者特定人群特有的病变模式等。然而, 临床诊断必须能够识别并标记这些异常区域, 即使这些特征在训练时从未见过。传统封闭集检测器无法处理这种情况, 而开放目标检测技术使得模型能够通过描述性文本 (如“检测异常的组织密度变化”) 来定位未知的病变区域, 为医生提供辅助诊断支持。

(4) 机器人抓取任务中的通用性需求

在机器人抓取任务中, 机器人需要处理家庭环境中的各种物体。这些物体种类繁多且不断变化: 新的家居用品、不同品牌和型号的日常用品、以及用户自定义的物体等。传统方法需要为每类物体重新收集数据、标注并训练模型, 这在家庭环境中是不可行的。开放目标检测技术使得机器人能够通过语言指令 (如“把红色的杯子递给我”) 来定位目标物体, 或者主动发现环境中的新物体并学习抓取策略, 从而适应不断变化的家庭环境。

这些场景共同指向一个核心需求: **模型必须具备在开放、动态环境中持续学习和适应的能力**, 而不是被固定的训练数据所束缚。开放目标检测技术正是为了解决这一根本性问题而发展起来的。

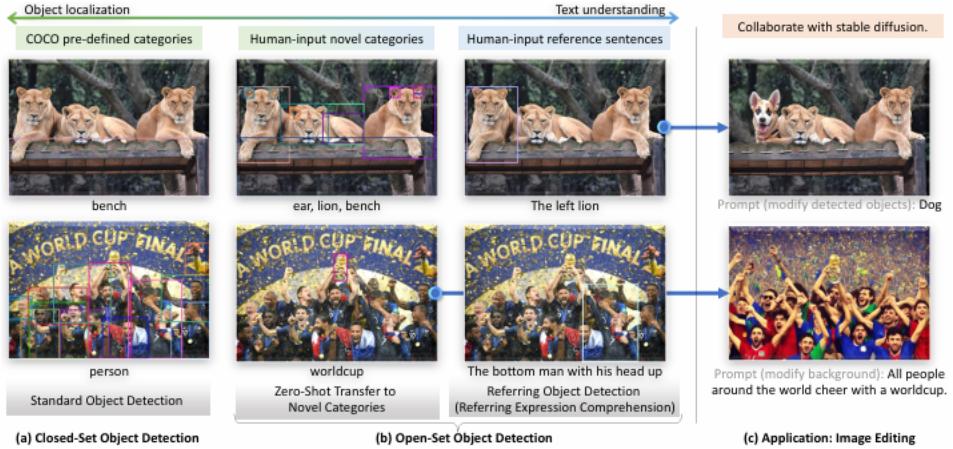


图 1: 封闭集检测与开放集检测的对比: (a) 传统封闭集检测仅能识别预定义类别; (b) 开放集检测允许用户动态指定新类别

1.2 核心挑战

1.2.1 封闭世界假设的局限性

打破封闭世界假设面临的首要挑战在于**类别表示方式的根本性变革**。传统检测器将类别表示为离散的数字 ID (如 0, 1, 2, ..., 79)，这种表示方式存在三个根本性缺陷：

(1) 语义缺失问题

数字 ID 本身不包含任何语义信息，无法捕捉类别之间的相似性和关联性。例如，在传统检测器中，“dog” (ID=16) 和 “cat” (ID=17) 在数学上是完全独立的两个类别，模型无法理解它们都是“动物”这一共同属性。同样，“car” (ID=3) 和 “truck” (ID=8) 之间的关系也无法被模型理解。这种表示方式使得模型只能学习到“这个特征对应 ID=16”这样的映射关系，而无法理解“这个特征看起来像狗”这样的语义概念。当遇到新类别时，即使新类别与已知类别在视觉上非常相似 (如“puppy”与“dog”)，模型也无法利用这种相似性进行识别。

(2) 固定容量限制

分类头的输出维度在训练时确定，无法动态扩展。具体而言，如果训练时定义了 N 个类别，分类头的输出维度就是 N ，Softmax 函数强制每个候选区域归属于这 N 个类别之一。当需要检测新类别时，必须修改网络结构 (增加输出维度)、重新收集数据、重新训练整个模型。这种“硬编码”的设计使得模型无法适应动态变化的类别空间。在实际应用中，这意味着每次需要检测新类别时，都要经历一个完整的模型更新周期，成本高昂且效率低下。

(3) 零泛化能力

对于训练集外的新类别，模型完全没有识别能力。这是因为传统检测器的学习目标是最大化训练集中已知类别的分类准确率，模型从未学习过“如何识别未知”的概念。当遇到新类别时，模型只能：

- 将其错误分类为最相似的已知类别 (如将“puppy”误判为“dog”)
- 或者因为置信度太低而被过滤掉 (当作背景)

这种“非此即彼”的判定机制使得模型无法表达“我不知道这是什么，但我知道这是一个物体”这样的概念，而这正是开放世界检测所需要的核心能力。

1.2.2 未知物体的识别与定位

在开放世界场景下，模型不仅需要识别用户指定的类别，还需要主动发现那些未被预先定义的物体。这带来了两个核心技术挑战：

(1) **什么是“未知”？**如何在没有任何标签指导的情况下，让模型区分“这是一个物体但我不知道是什么”与“这只是背景”？传统检测器缺乏物体性（Objectness）的通用建模能力。

(2) **如何避免伪标签噪声？**如果使用模型自身的预测来标注未知物体（伪标签策略），如何避免将背景或低置信度的已知类别误标为“未知”？

1.2.3 增量学习中的灾难性遗忘

当模型需要持续学习新类别时，面临灾难性遗忘（Catastrophic Forgetting）问题：学习新知识会覆盖旧知识，导致对历史类别的检测能力显著下降。如何在扩展类别空间的同时保持对已学类别的识别能力，是开放世界检测的重要挑战。

1.3 研究意义

1.3.1 通向通用人工智能（AGI）的必要路径

通用人工智能的核心特征之一是能够理解和处理开放、动态环境中的新概念。传统封闭集检测器无法适应新类别，这与 AGI 的目标相悖。开放目标检测通过视觉-语言融合，使模型具备了理解语义、泛化到新类别的能力，这是构建通用视觉智能系统的关键一步。

通过将“类别”从固定的离散 ID 转化为可扩展的语义表示，模型获得了类似人类的“概念理解”能力——能够通过语言描述理解新物体，而无需大量标注数据。

1.3.2 实际应用场景的迫切需求

开放目标检测技术在多个领域具有重要的应用价值：

应用领域	传统方法的局限	开放检测的优势
自动驾驶	无法识别新型车辆、临时路障	可通过文本描述识别任意道路物体
智能安防	只能检测预定义的异常类别	主动发现任何异常物体
机器人抓取	需要为每类物体重新训练	输入物体名称即可抓取
医疗影像	无法识别罕见疾病特征	可描述性地定位异常区域

表 1：开放目标检测在不同应用领域的优势

1.3.3 学术价值与产业价值

从学术角度看，开放目标检测推动了计算机视觉从“封闭世界”向“开放世界”的范式转变，为理解视觉-语言融合、零样本学习、持续学习等核心问题提供了新的视角。

从产业角度看，开放目标检测技术能够显著降低模型部署和维护成本。据估计，采用开放词汇检测技术可以将新类别部署成本降低 80% 以上，同时提升系统的鲁棒性和泛化能力。

2 核心概念与理论基础 (Concepts & Foundations)

2.1 概念定义

在打破封闭世界假设的探索中，学术界从不同角度定义了“开放性”，形成了一个相互关联的研究谱系。

2.1.1 封闭集检测 (Closed-Set Detection)

传统目标检测器在封闭集设定下工作，其形式化定义为：

给定训练数据集 $\mathcal{D}_{train} = \{(I_i, \{b_{ij}, c_{ij}\})\}$ ，其中 I_i 是图像， $b_{ij} \in \mathbb{R}^4$ 是边界框， $c_{ij} \in \mathcal{C}_{train}$ 是类别标签，封闭集检测器 f 学习映射：

$$f : I \rightarrow \{(b_k, c_k, s_k)\}, \quad c_k \in \mathcal{C}_{train}$$

其中 s_k 是置信度分数。关键限制：测试时的类别空间必须等于训练时的类别空间，即 $\mathcal{C}_{test} = \mathcal{C}_{train}$ 。

2.1.2 开放词汇目标检测 (Open-Vocabulary Detection, OVD)

OVD 的核心创新在于将类别表示从离散 ID 转换为语义嵌入向量。其形式化定义为：

给定图像 I 和文本词汇表 $\mathcal{V} = \{t_1, t_2, \dots, t_C\}$ (t_i 可以是类别名、短语或描述)，OVD 检测器输出：

$$\mathcal{D}_{OVD}(I, \mathcal{V}) = \{(b_i, c_i, s_i)\}_{i=1}^N$$

其中 $c_i \in \mathcal{V}$ 是匹配的文本类别。

核心特征：

- 动态词汇表： \mathcal{V} 在推理时可以任意指定，不受训练数据限制
- 零样本泛化：当 \mathcal{V} 包含训练时未见过的类别（Novel Classes）时，模型仍能检测
- 文本提示驱动：用户通过输入自然语言来指定检测目标

2.1.3 开放世界目标检测 (Open-World Detection, OWOD)

OWOD 强调模型在动态环境中的持续学习能力。任务被划分为一系列增量学习子任务 $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$ ：

- 在子任务 T_i 训练时，模型学习新类别集合 \mathcal{K}_{T_i}
- 已知类别集合为 $\mathcal{K}_{known}^{(i)} = \mathcal{K}_{T_1} \cup \dots \cup \mathcal{K}_{T_i}$
- 推理时，模型需要检测已知类别 $\mathcal{K}_{known}^{(i)}$ ，并主动识别未知物体
- 在后续任务中，部分 Unknown 类别被标注后加入 $\mathcal{K}_{T_{i+1}}$

核心特征：

- 主动未知检测：模型无需用户输入，自动标记训练集外的物体为“Unknown”
- 增量学习：在学习新类别时，保持对旧类别的识别能力
- 动态类别扩展：类别集合随时间不断增长

2.1.4 三者的关系图谱

下表总结了三类检测范式的核心差异：

检测范式	类别表示	对未知的态度	学习范式	代表工作
封闭集检测	离散 ID	排斥/忽略	静态训练	Faster YOLO
开放词汇检测	语义嵌入	被动识别（需用户指定）	零样本迁移	GLIP, Grounding DINO
开放世界检测	动态扩展	主动发现	增量学习	ORE, OW-DETR

表 2: 三类目标检测范式对比

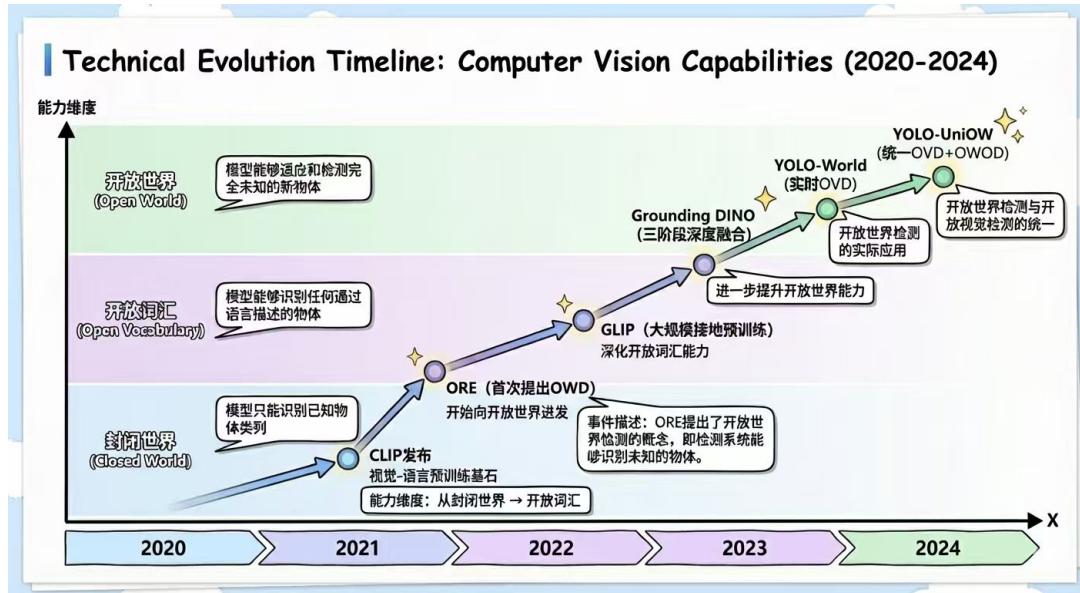


图 2: 开放目标检测技术演进路线图

2.2 技术基石

2.2.1 视觉-语言预训练 (VLP)

CLIP 的核心机制与贡献 打破封闭世界假设的关键在于改变类别表示方式。OpenAI 的 CLIP (Contrastive Language-Image Pre-training) 模型[3]通过在 4 亿图文对上进行对比学习，构建了一个统一的视觉-语言特征空间。

对比学习目标: 给定一个 batch 包含 N 个图文对 $\{(I_i, T_i)\}_{i=1}^N$, InfoNCE 损失为：

$$\mathcal{L}_{CLIP} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(I_i, T_j)/\tau)}$$

其中 τ 是温度参数, $\text{sim}(\cdot, \cdot)$ 是余弦相似度。

对比学习原理 (Contrastive Learning) CLIP 采用对比学习范式, 其核心思想是：最大化匹配图文对的相似度, 最小化不匹配对的相似度。这种学习方式使得模型能够理解语义关联, 而非记忆特定的类别 ID。

传统分类器：

$$f : \mathbb{R}^d \rightarrow \{0, 1, \dots, N - 1\}$$

类别表示为离散 ID

开放检测器：

$$\text{sim}(f_{img}(x), f_{text}(t)) \in [0, 1]$$

类别表示为语义嵌入向量

图像-文本对齐的数学原理 在 CLIP 的特征空间中，图像嵌入 $v \in \mathbb{R}^d$ 和文本嵌入 $t \in \mathbb{R}^d$ 可以通过余弦相似度进行匹配：

$$\text{score}(v, t) = \frac{v \cdot t}{\|v\| \|t\|} = \cos(v, t)$$

这种设计的优势在于：只要文本编码器能够将新类别名称编码为语义向量，模型就能在零样本情况下识别该类别，无需重新训练。

2.2.2 区域-文本对齐 (Region-Text Alignment)

从图像级到区域级的挑战 尽管 CLIP 在图像级分类上表现出色，但直接将其应用于目标检测面临三个核心挑战：

(1) 局部区域理解的缺失

CLIP 的预训练是在完整图像与文本之间进行的，模型学习的是图像整体的语义表示，缺乏对局部区域的精细理解。在目标检测任务中，我们需要识别图像中的多个物体，每个物体只占据图像的一小部分。CLIP 的图像级特征可能包含了多个物体的混合信息，难以直接用于单个物体的识别和定位。

例如，在一张包含“一个人和一只狗”的图像中，CLIP 可能学习到“人和狗”的混合表示，但无法准确区分哪个区域是“人”，哪个区域是“狗”。这种局部理解的缺失使得直接使用 CLIP 进行目标检测变得困难。

(2) 定位能力的缺失

目标检测需要同时完成两个任务：**分类**（识别物体是什么）和**定位**（确定物体在哪里）。CLIP 仅提供分类能力，能够判断图像中是否包含某个物体，但无法确定物体的精确位置。即使 CLIP 能够识别图像中包含“car”，也无法告诉我们这辆车在图像的哪个位置，边界框的坐标是什么。

这种定位能力的缺失是 CLIP 无法直接用于目标检测的根本原因。需要额外的机制（如区域提议网络、锚框机制、或查询机制）来提供定位能力。

(3) 特征分布差异 (Region-Image Distribution Gap)

区域特征与图像级特征的分布存在显著差异，这被称为 Region-Image Distribution Gap。图像级特征通常包含全局上下文信息，而区域特征只包含局部信息。这种分布差异使得直接使用 CLIP 的图像级特征来匹配区域特征变得困难。

具体而言，CLIP 的图像编码器输出的特征可能包含图像的全局语义信息（如场景类型、整体布局等），而目标检测需要的区域特征应该聚焦于局部物体的特征（如物体的形状、纹理、颜色等）。这种差异使得即使 CLIP 能够识别图像中包含某个物体，也难以准确匹配到对应的区域特征。

为了解决这个问题，研究者提出了两种主要路径：**知识蒸馏路径**（如 ViLD、RegionCLIP）通过蒸馏将 CLIP 的知识迁移到区域级；**大规模接地预训练路径**（如 GLIP、Grounding DINO）直接在区域级进行预训练，避免了分布差异问题。

RegionCLIP 的开创性工作 为解决这一问题，RegionCLIP 提出了区域级对比学习策略，通过将图像裁剪为区域并与对应的文本描述进行对齐，将 CLIP 的能力扩展到区域级。

Grounding 范式的定义 大规模接地预训练路径（GLIP, Grounding DINO 等）将目标检测重新定义为短语接地（Phrase Grounding）任务：给定图像和文本描述，定位文本中提到的物体。

该方法在大规模数据上进行预训练，构建区域-文本对 $\{(r_i, t_i)\}$ ，通过对比学习使区域特征与文本特征在同一语义空间中对齐。

接地预训练的优势：

(1) 直接区域级训练，避免分布差异

大规模接地预训练路径直接在区域级进行训练，构建区域-文本对 $\{(r_i, t_i)\}$ ，其中 r_i 是图像区域， t_i 是对应的文本描述。这种训练方式避免了图像级到区域级的迁移 gap，使得模型能够直接学习区域特征与文本特征的对应关系。

与知识蒸馏路径不同，接地预训练不需要依赖预训练的 CLIP 模型，而是从零开始学习区域-文本对齐。这使得模型能够学习到更适合目标检测任务的区域特征表示，避免了分布差异问题。

(2) 利用丰富的文本信息

接地预训练可以利用更丰富的文本信息，而不仅仅是简单的类别名称。例如，文本描述可以包含物体的属性（如“红色的汽车”、“大型的狗”）、关系（如“人旁边的狗”、“桌子上的杯子”）、以及更复杂的描述（如“穿着红色雨衣的小狗”）。

这种丰富的文本信息使得模型能够学习到更细粒度的语义理解能力，不仅能够识别物体是什么，还能理解物体的属性、关系等语义信息。这对于开放词汇检测非常重要，因为实际应用中用户往往通过描述性文本来指定检测目标，而不仅仅是类别名称。

(3) 大规模数据带来的泛化能力

通过在大规模数据（如数百万图文对）上进行预训练，模型能够学习到极其丰富的视觉-语言对应关系。这种大规模预训练带来的泛化能力使得模型能够在零样本情况下识别训练时从未见过的类别，只要这些类别能够用自然语言描述。

例如，GLIP 在包含数百万图文对的数据上进行预训练后，在 ODinW 的 35 个真实场景数据集上取得了显著提升，证明了大规模预训练的重要性。Grounding DINO 进一步扩展了预训练数据的规模，在更多样化的数据上进行训练，取得了更强的零样本泛化能力。

2.2.3 Transformer 在检测中的应用

DETR 系列的演进 Transformer 架构在目标检测中的应用经历了重要演进，每个阶段都解决了前一个阶段的关键问题：

DETR (2020): Transformer 在检测中的首次应用

DETR (Detection Transformer) 首次将 Transformer 架构引入目标检测，实现了端到端检测，无需后处理步骤（如 NMS）。DETR 的核心创新是使用可学习的查询（Queries）来检测物体：每个查询学习检测一个物体，通过自注意力和交叉注意力机制与图像特征交互，最终输出物体的类别和边界框。

DETR 的优势在于：**(1) 端到端学习**：避免了传统检测器中复杂的后处理步骤，模型能够直接学习最优的检测策略；**(2) 全局上下文**：Transformer 的全局注意力机制使得模型能够利用全局上下文信息，更好地处理遮挡和复杂场景；**(3) 无需锚框**：与传统检测器不同，DETR 不需要预定的锚框，查询机制更加灵活。

然而，DETR 也存在问题：**(1) 收敛慢**：需要训练 500+ 个 epoch 才能收敛，训练成本高；**(2) 小目标检测能力弱**：对于小目标的检测性能不如传统检测器；**(3) 查询初始化策略简单**：使用随机初始化的查询，缺乏先验知识。

Deformable DETR (2021): 解决收敛慢的问题

Deformable DETR 通过引入可变形注意力机制解决了 DETR 收敛慢的问题。可变形注意力只关注图像中的关键点，而不是所有像素，大大降低了计算复杂度。这使得 Deformable DETR 能够在更少的训练轮数（如 50 个 epoch）内收敛，同时保持与 DETR 相当的性能。

可变形注意力的核心思想是：对于每个查询，只关注图像中的少数关键采样点，而不是所有位置。这些采样点的位置是可学习的，模型能够自动学习关注哪些位置。这种设计既保持了全局上下文的能力，又大幅降低了计算成本。

DINO (2022): 优化查询初始化与预训练

DINO 进一步优化了查询初始化策略，通过对比学习预训练提升检测性能。DINO 的核心创新包括：

(1) 对比学习预训练：在检测任务之前，DINO 先在大量无标注数据上进行对比学习预训练，学习通用的物体表示。这种预训练使得模型能够学习到更好的初始化权重，加速后续的检测训练。

(2) 优化的查询初始化：DINO 使用对比学习预训练的权重来初始化查询，而不是随机初始化。这使得查询从一开始就具备了物体性的先验知识，能够更快地学习检测任务。

(3) 多尺度特征融合：DINO 引入了多尺度特征融合机制，能够更好地处理不同尺度的目标，特别是小目标。

DINO 的这些改进使得模型在保持端到端学习优势的同时，大幅提升了检测性能和训练效率，为后续的开放词汇检测（如 Grounding DINO）奠定了基础。

Query 机制的重要性 Query 机制是 Transformer 检测器的核心创新。与传统检测器的锚框（Anchor）不同，Query 是可学习的向量，能够学习通用的物体模式。在开放词汇检测中，Query 机制与文本引导结合，实现了语义驱动的目标定位。

端到端学习的优势 Transformer 检测器的端到端学习避免了传统检测器中 NMS（Non-Maximum Suppression）等后处理步骤，使得模型能够直接学习最优的检测策略，为开放词汇检测中的语义对齐提供了更好的基础。

2.3 数据集与评估体系

2.3.1 数据集

开放目标检测研究涉及多类数据集，每类数据集都有其特定的用途和特点：

(1) 封闭集数据集

COCO (Common Objects in Context) 数据集是目标检测领域的标准基准，包含 80 个常见物体类别，类别分布相对均衡。COCO 数据集的一个重要特点是其 Base/Novel 划分评估范式：通常将 80 个类别划分为 48 个 Base 类别（用于训练）和 17 个 Novel 类别（用于零样本测试），剩余的 15 个类别用于验证。这种划分方式使得 COCO 成为评估开放词汇检测零样本能力的标准数据集。COCO 数据集的优势在于其标注质量高、类别覆盖常见场景，但局限性在于类别数量有限，无法评估大词汇量场景下的性能。

LVIS (Large Vocabulary Instance Segmentation) 数据集是对 COCO 的重要扩展，包含 1203 个类别，呈现极端的长尾分布。LVIS 将类别分为三个层次：**rare**（罕见类，出现次数 < 10 ）、**common**（常见类，出现次数 $10 - 100$ ）和 **frequent**（频繁类，出现次数 > 100 ）。这种长尾分布更接近真实世界的场景，使得 LVIS 成为评估大词汇量零样本检测能力的理想数据集。LVIS 的挑战在于如何提升罕见类别的检测性能，这正是开放词汇检测技术需要解决的核心问题之一。

(2) 开放集数据集

ODinW (Open Detection in the Wild) 是一个包含 35 个真实场景数据集的集合，覆盖了多样化的应用场景，包括自然图像、室内场景、室外场景、特定领域（如医疗、卫星图像）等。ODinW 的核心价值在于评估模型的跨域泛化能力：模型在 COCO 等标准数据集上训练后，能否在完全不同的场景和领域上保持性能。这种评估方式更接近实际应用场景，因为实际部署时模型往往需要处理与训练数据分布不同的数据。

Objects365 是一个大规模检测数据集，包含 365 个类别和超过 200 万张图像。Objects365 主要用于大规模预训练，为模型提供丰富的物体类别和场景多样性。与 COCO 相比，Objects365 的类别数量更多，场景覆盖更广，是训练强大开放词汇检测器的重要数据源。

(3) Grounding 数据集

Visual Genome 数据集包含丰富的视觉-语言标注，不仅包括物体类别，还包括物体的属性（如颜色、形状、材质）以及物体之间的关系（如“人拿着杯子”、“狗在汽车旁边”等）。这种丰富的标注使得 Visual Genome 成为训练细粒度语义理解能力的重要数据源。在开放词汇检测中，Visual Genome 可以帮助模型理解复杂的描述性文本，而不仅仅是简单的类别名称。

RefCOCO/+g 系列数据集专门用于 Referring Expression Comprehension (REC) 任务，包含描述性文本和对应的目标区域。例如，给定描述“the person wearing a red hat”，模型需要准确定位对应的目标。RefCOCO 系列包含三个变体：RefCOCO（包含 19,994 个指代表达）、RefCOCO+（包含 19,992 个表达，禁止使用位置词汇）和 RefCOCOg（包含 25,799 个更长的描述性表达）。这些数据集评估了模型对细粒度语义描述的理解能力，是开放词汇检测的重要评估基准。

(4) 预训练数据集

CC3M (Conceptual Captions 3M) 包含 300 万图像-文本对，是开放词汇检测预训练的重要数据源。CC3M 的特点是其文本描述是自动生成的（通过网页的 alt-text），虽然质量可能不如人工标注，但规模大、覆盖广，为模型提供了丰富的视觉-语言对应关系。CC3M 常用于 OVD 模型的预训练阶段，帮助模型学习基本的视觉-语言对齐能力。

LAION 系列数据集是更大规模的预训练数据源，包含数亿图文对。LAION-400M 包含 4 亿图文对，LAION-5B 包含 50 亿图文对。这些大规模数据集为模型提供了极其丰富的视觉-语言知识，是训练强大开放词汇检测器的基础。LAION 数据集的优势在于其巨大的规模和多样性，但挑战在于数据质量控制和大规模训练的工程实现。

2.3.2 评估范式

Zero-shot Transfer Zero-shot Transfer 是 OVD 的核心评估范式。通常将数据集划分为 Base 类别和 Novel 类别，模型在 Base 类别上训练，在 Novel 类别上测试。例如，COCO 的 48/17 划分（48 个 Base 类 + 17 个 Novel 类）是常用的评估设置。

Few-shot Learning Few-shot Learning 评估模型用少量样本（如 1-shot、5-shot）快速适应新类别的能力。

Referring Expression Comprehension (REC) REC 评估模型理解描述性文本并定位对应目标的能力。给定如 “the person wearing a red hat”的描述，模型需要准确定位目标。

增量学习任务 OWOD 的评估采用增量学习任务设定，将任务划分为一系列子任务 T_1, T_2, \dots, T_M ，每个子任务引入新的类别。

2.3.3 评估指标

标准检测指标 AP (Average Precision) 是目标检测的核心指标，通过计算精确率-召回率曲线下的面积来衡量模型的检测性能。AP 综合考虑了定位精度和分类准确性，是评估检测器整体性能的标准指标。在 COCO 数据集中，AP 是所有 IoU 阈值（0.5 到 0.95，步长 0.05）的平均值，记为 AP@[0.5:0.95]。

AP50 和 AP75 分别表示 IoU 阈值为 0.5 和 0.75 时的平均精度。AP50 是更宽松的评估标准，只要检测框与真实框的 IoU 超过 0.5 就认为检测正确；AP75 是更严格的标准，要求 IoU 超过 0.75。AP50 通常用于快速评估和对比，而 AP75 更能反映模型的精确定位能力。

APr、APc、APf 是 LVIS 数据集上引入的指标，分别评估 rare (罕见)、common (常见) 和 frequent (频繁) 类别的性能。由于 LVIS 呈现极端的长尾分布，这三个指标能够更细致地评估模型在不同频率类别上的表现。通常，模型在 frequent 类别上表现最好，在 rare 类别上表现最差，这正是长尾检测需要解决的核心问题。

开放世界检测指标 U-Recall (Unknown Recall) 是开放世界检测的核心指标，衡量模型发现未知物体的能力。U-Recall 定义为：在所有未知物体中，被模型正确标记为“Unknown”的比例。传统 OWOD 方法的 U-Recall 通常低于 10%，这意味着模型只能发现不到 10% 的未知物体，远无法满足实际应用需求。基于 OVD 的方法将 U-Recall 提升到 20%+，实现了显著的性能突破。

U-mAP (Unknown mean Average Precision) 是更严格地评估未知检测质量的指标。与 U-Recall 只关注召回率不同，U-mAP 同时考虑了精确率和召回率，能够更全面地评估未知检测的性能。U-mAP 的计算方式与标准 mAP 类似，但只针对未知类别。高 U-mAP 意味着模型不仅能够发现未知物体，还能准确定位它们，避免将背景或已知类别误判为未知。

Wilderness Impact (WI) 衡量未知物体对已知类别检测的干扰程度。理想情况下，未知物体的存在不应该影响已知类别的检测性能，即 WI 应该接近 0。WI 的计算公式为：

$$WI = \frac{mAP_{known} - mAP_{known+unknown}}{mAP_{known}}$$

其中 mAP_{known} 是只有已知类别时的性能， $mAP_{known+unknown}$ 是同时存在已知和未知类别时的性能。如果 WI 过高，说明未知物体会显著干扰已知类别的检测，这是不理想的。

A-OSE (Absolute Open-Set Error) 综合评估未知检测和已知分类的平衡性，同时考虑假阳性 (False Positive) 和假阴性 (False Negative)。A-OSE 定义为：

$$A-OSE = FP_{unknown} + FN_{unknown}$$

其中 $FP_{unknown}$ 是将已知类别误判为未知的数量， $FN_{unknown}$ 是将未知类别误判为已知或背景的数量。低 A-OSE 意味着模型能够准确区分已知和未知，同时保持对已知类别的正确分类。

3 开放词汇目标检测 (Open-Vocabulary Detection)

本部分深入探讨 OVD 的两种主流技术范式：基于 Transformer 的高精度框架和基于重参数化的实时化框架。

3.1 技术演进概述

3.1.1 早期尝试：ViLD、RegionCLIP

早期 OVD 方法主要采用知识蒸馏策略，利用预训练的 CLIP 作为教师模型，通过蒸馏将图像级语义对齐能力迁移到区域级检测。虽然取得了一定进展，但受限于训练数据的类别覆盖，泛化能力有限。

3.1.2 Transformer 时代：GLIP、MDETR

随后的方法将目标检测重新定义为短语接地（Phrase Grounding）任务，通过大规模预训练实现了更强的零样本能力。GLIP 在 ODinW 上取得了显著提升，证明了大规模预训练的重要性。

3.1.3 当前 SOTA：Grounding DINO、YOLO-World

当前最先进的 OVD 方法主要分为两个流派：

- **Grounding DINO**: 通过多阶段深度对齐实现极高的检测精度
- **YOLO-World**: 通过重参数化技术实现实时推理

3.2 深度融合方式：高精度 OVD 框架（Grounding DINO）

3.2.1 设计哲学

多阶段深度融合的动机 高精度 OVD 框架（以 Grounding DINO 为代表）提出了范式转变的思路：不再将视觉和文本视为独立模态，而是基于“Grounded Pre-training”思想，在 Transformer 的编码器、查询初始化及解码器全生命周期内引入强力的文本干预，实现视觉与语言的深度耦合。

对比维度	传统双塔架构	深度融合架构
融合位置	仅在分类头	编码器+查询+解码器
交互方式	浅层点积	双向跨模态注意力
语义理解	类别名匹配	细粒度短语理解
零样本能力	受限	强大

表 3: 传统双塔架构与深度融合架构的对比

与传统方法的对比分析

3.2.2 核心技术

Feature Enhancer: 跨模态特征增强 该增强器由堆叠的 Transformer 层构成，通过双向跨模态注意力实现信息流转：

(1) 视觉到文本的引导：图像特征 $V \in \mathbb{R}^{H \times W \times d}$ 作为 Query，通过多头注意力机制感知文本特征 $T \in \mathbb{R}^{L \times d}$ 中的关键描述。

(2) 文本到视觉的注入：文本特征作为 Key 和 Value，通过交叉注意力机制注入到视觉空间。

双向跨模态注意力的数学表达为：

$$V', T' = \text{Bi-MultiHeadAttention}(V, T)$$

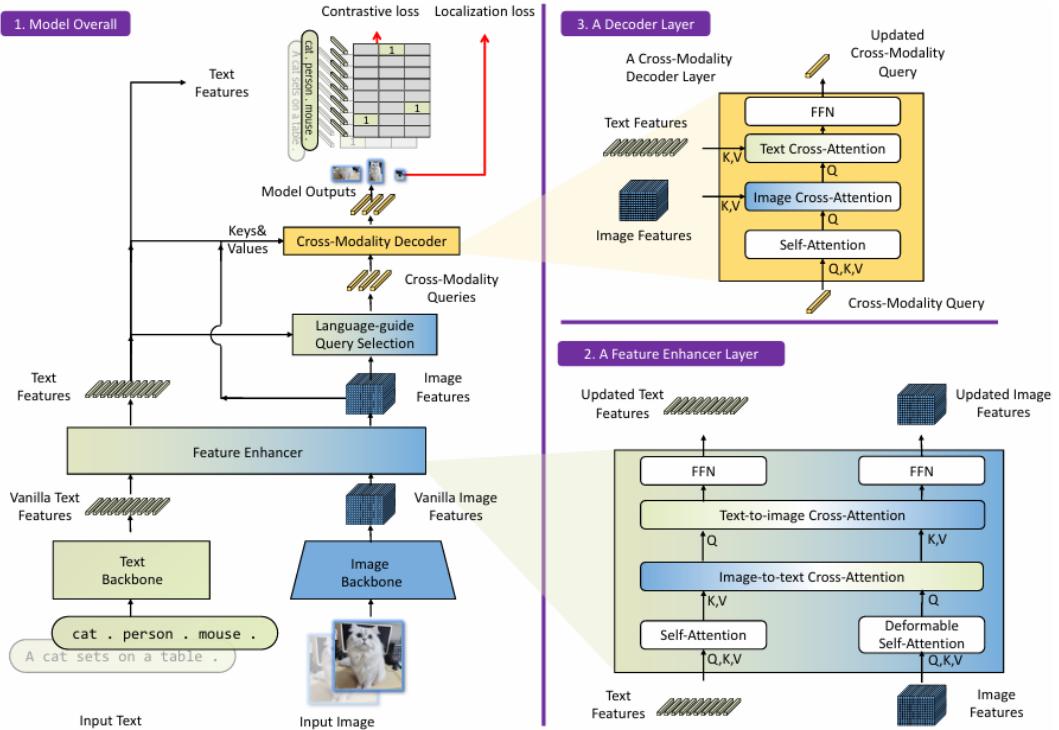


图 3: 高精度 OVD 框架 (Grounding DINO) 整体架构与三阶段对齐流

Language-Guided Query Selection: 语言引导的查询初始化 高精度框架采用对比驱动的查询初始化策略:

步骤1: 利用增强器输出的图像特征与文本特征计算空间-语义相似度矩阵:

$$S = f_v \cdot f_t^\top \in \mathbb{R}^{H \times W \times L}$$

步骤2: 对相似度矩阵进行空间聚合, 筛选 Top-K 个高响应区域初始化动态锚框。

Cross-Modality Decoder: 跨模态解码器 解码器的每一层都包含专门的 Text Cross-Attention 模块:

$$\text{TextCrossAttn}(Q, T) = \text{Softmax}\left(\frac{QT^\top}{\sqrt{d}}\right)T$$

Sub-sentence Level Text Representation: 子句级文本表示 针对长文本描述, 模型构建子句级注意力掩码, 将长文本分解为多个名词短语, 每个视觉区域仅与对应的短语交互, 屏蔽句中无关词汇。

3.2.3 训练策略

多数据源融合 高精度框架在多种数据源上进行联合训练, 这种多数据源融合策略是提升模型泛化能力的关键:

(1) Detection 数据的作用

Detection 数据 (如 COCO、Objects365) 提供了标准的边界框标注和类别标签, 是训练检测器的基础。这些数据帮助模型学习基本的定位能力和类别识别能力。COCO 提供了高质量的标注和均衡的类别分布, 适合学习常见物体的检测; Objects365 提供了更大规模的类别覆盖, 帮助模

型学习更多样化的物体模式。通过联合训练这两种数据，模型能够同时获得高质量的标注和丰富的类别多样性。

(2) Grounding 数据的作用

Grounding 数据（如 RefCOCO、Visual Genome）将目标检测重新定义为短语接地任务，即给定文本描述，定位对应的物体。这种数据格式的优势在于：首先，文本描述比简单的类别名称更丰富，包含了物体的属性（如“红色的汽车”）、关系（如“人旁边的狗”）等语义信息；其次，Grounding 数据帮助模型学习理解复杂的描述性文本，而不仅仅是类别名称；最后，Grounding 数据中的文本描述往往更加自然和多样化，有助于提升模型的泛化能力。

(3) Caption 数据的作用

Caption 数据（图像描述数据）虽然不包含精确的边界框标注，但提供了丰富的语义信息。Caption 数据帮助模型理解图像的整体语义内容，学习视觉特征与文本描述的对应关系。虽然 Caption 数据不能直接用于训练定位能力，但它能够增强模型的语义理解能力，特别是在理解复杂场景和长文本描述方面。通过将 Caption 数据与 Detection 和 Grounding 数据联合训练，模型能够同时学习定位、分类和语义理解能力。

多数据源融合的训练策略

在训练过程中，不同数据源的数据以一定比例混合。通常，Detection 数据占主要比例（如 60%），Grounding 数据占次要比例（如 30%），Caption 数据占较小比例（如 10%）。这种混合策略确保了模型既能够学习精确的定位能力，又能够理解丰富的语义信息。同时，不同数据源的损失函数可能需要不同的权重，以平衡不同任务的学习目标。

损失函数设计 总损失函数包括分类损失和定位损失：

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{bbox} \mathcal{L}_{bbox}$$

其中分类损失采用基于相似度的 Focal Loss，定位损失采用 L1 Loss 和 GIOU Loss 的加权和。

大规模预训练的重要性 大规模预训练（如在数百万图文对上预训练）是高精度框架取得强大零样本能力的关键。预训练数据的多样性和规模直接影响模型的泛化能力。

3.2.4 实验分析

消融实验：多阶段融合的有效性 消融实验表明，三阶段融合（编码器+查询+解码器）的效果显著优于仅在单一阶段进行融合。

跨数据集性能 在 COCO、LVIS、ODinW 等多个数据集上，Grounding DINO 都取得了 SOTA 性能，尤其在长尾类别和跨域泛化上表现突出。

REC 任务的表现 在 Referring Expression Comprehension 任务上，Grounding DINO 同样表现出色，证明了其对复杂语义描述的理解能力。

3.3 实时化方式：高效 OVD 框架（YOLO-World）

3.3.1 设计动机

边缘部署的需求 虽然基于 Transformer 的高精度框架在精度上屡创新高，但在边缘计算和实时监控场景下，其高昂的计算成本和推理延迟成为了瓶颈。

Transformer 的计算瓶颈 Transformer 的自注意力机制具有 $O(N^2)$ 的计算复杂度，在处理高分辨率图像时计算开销巨大。

重参数化的创新思路 YOLO-World 提出了重参数化技术，将文本编码器的计算从推理时移除，实现了“开集检测”与“实时推理”的统一。

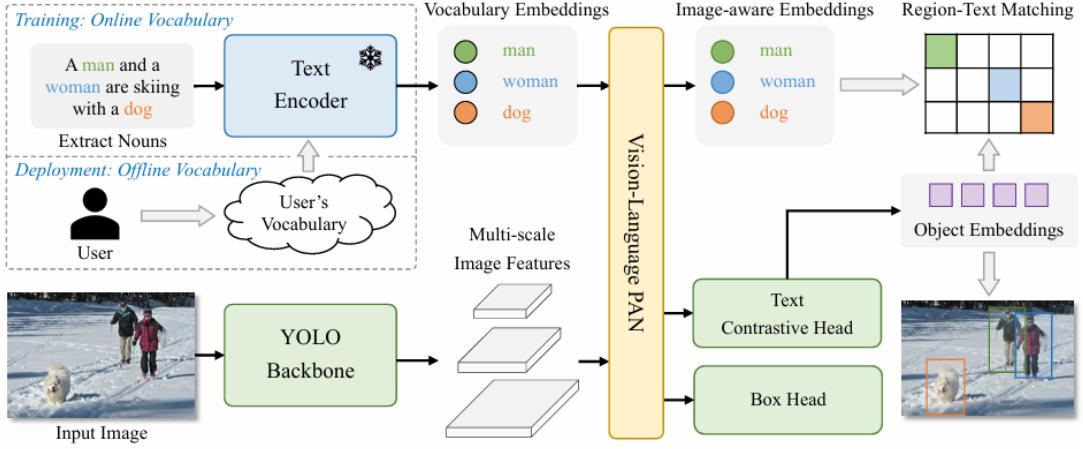


图 4: 实时化 OVD 框架 (YOLO-World) 整体架构

3.3.2 核心技术

RepVL-PAN: 重参数化视觉-语言路径聚合网络 YOLO-World 设计了全新的重参数化视觉-语言路径聚合网络 (RepVL-PAN)，采用局部卷积 + 全局池化的混合设计。

文本引导的 CSPLayer: 利用文本信息作为“滤波器”，对视觉特征进行通道级的动态重构：

$$W_{text} = \text{GlobalPool}(\text{MLP}(F_t)) \in \mathbb{R}^C$$

$$F'_v = F_v \odot \text{Broadcast}(W_{text})$$

图像池化注意力机制: 在颈部网络中嵌入全局语义理解能力。

Prompt-then-Detect: 离线词汇表范式 YOLO-World 采用“先提示后检测”的范式：

步骤1: 离线文本编码。用户输入的类别名称通过文本编码器转化为文本嵌入矩阵：

$$W_{text} = [E(c_1), E(c_2), \dots, E(c_N)]^\top \in \mathbb{R}^{N \times d}$$

步骤2: 权重融合。将文本嵌入矩阵直接作为检测头中 1×1 卷积层的卷积核参数。

步骤3: 纯视觉推理。推理时无需运行文本编码器，模型退化为纯视觉的 YOLO 模型。

Region-Text Contrastive Learning: 区域-文本对比学习 采用 InfoNCE 损失函数进行区域级对比学习：

$$\mathcal{L}_{contrast} = -\frac{1}{B} \sum_{i=1}^B \frac{1}{M} \sum_{j=1}^M \log \frac{\exp(sim(r_{i,j}, T_i)/\tau)}{\sum_{k=1}^B \exp(sim(r_{i,j}, T_k)/\tau)}$$

3.3.3 推理加速技术

文本编码器的离线抽离 重参数化技术的核心是将文本编码从推理链路中移除：

计算模块	传统方法	重参数化后
文本编码器	需要运行 (~10ms)	移除 (0ms)
跨模态注意力	需要计算 (~15ms)	移除 (0ms)
总推理时间	~25-30ms	~15-20ms
推理速度	~30-40 FPS	~50-70 FPS

表 4: 重参数化前后的推理效率对比

卷积权重的重参数化 检测头的分类分支可以表示为:

$$\text{Score} = \text{Conv1x1}(F_v, W_{text}) = F_v \cdot W_{text}^\top$$

在线/离线词汇表的灵活切换 YOLO-World 支持两种模式, 用户可以根据实际需求灵活选择:

(1) 离线模式: 固定词汇表的高效推理

在离线模式下, 用户预先指定一个固定的类别词汇表(如 100 个常见类别), 通过文本编码器将这些类别名称编码为文本嵌入矩阵, 然后通过重参数化技术将这些嵌入融合到检测头的权重中。推理时, 模型退化为纯视觉的 YOLO 模型, 无需运行文本编码器和跨模态注意力计算, 推理速度大幅提升(可达 50-70 FPS)。这种模式特别适合以下场景:

- **固定应用场景:** 如工业质检中的特定缺陷类型检测、零售场景中的商品识别等, 类别集合相对固定
- **实时性要求高:** 如视频监控、自动驾驶等需要高帧率处理的场景
- **边缘设备部署:** 计算资源受限的嵌入式设备, 需要最小化计算开销

(2) 在线模式: 动态词汇表的灵活检测

在在线模式下, 用户可以在每次推理时动态指定不同的类别词汇表, 模型保持完整的跨模态计算能力(包括文本编码和跨模态注意力)。虽然推理速度较慢(约 30-40 FPS), 但提供了最大的灵活性。这种模式特别适合以下场景:

- **交互式应用:** 用户通过自然语言输入来指定检测目标, 如“检测画面中所有的红色物体”
- **探索性任务:** 需要不断尝试不同的类别组合, 找到最佳的检测配置
- **Few-shot 学习:** 用户提供少量样本, 模型需要快速适应新类别

两种模式的切换机制

YOLO-World 通过简单的配置参数即可在两种模式之间切换。在离线模式下, 模型会预先加载并重参数化文本嵌入; 在在线模式下, 模型会在每次推理时动态编码文本输入。这种设计使得同一个模型可以适应不同的应用场景, 提供了极大的部署灵活性。

3.3.4 性能对比

精度 vs 速度的权衡分析 YOLO-World 在 V100 上实现了 50+ FPS 的推理速度, 同时保持了较高的零样本检测精度。

与 Grounding DINO 的对比分析

对比维度	Grounding DINO	YOLO-World
架构基础	Transformer (DETR系列)	CNN (YOLO系列)
推理速度	~10-20 FPS	~50-70 FPS
检测精度	极高	高
内存占用	高 (~2-4GB)	低 (~500MB-1GB)
适用场景	离线分析、高质量标注	实时监控、边缘部署

表 5: Grounding DINO 与 YOLO-World 对比

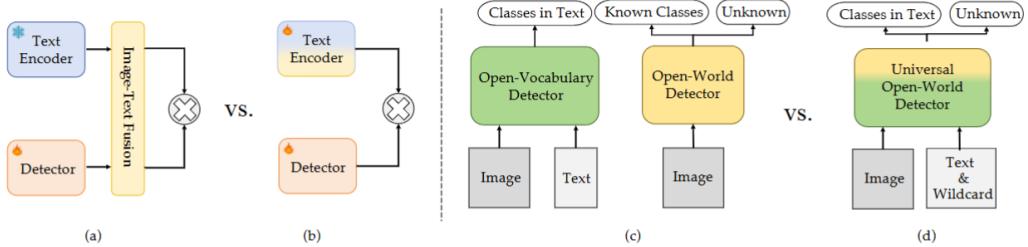


图 5: 不同 OVD 检测框架的对比

3.4 OVD 小结

3.4.1 两种方式的适用场景

高精度框架 (Grounding DINO) 的适用场景

基于 Transformer 的高精度框架追求极致的检测精度和语义理解能力，特别适合以下场景：

(1) 医疗影像分析：在医疗诊断中，检测精度直接关系到患者的生命安全。Grounding DINO 通过多阶段深度对齐，能够准确理解复杂的医学描述（如“检测左肺上叶的异常密度区域”），并精确定位病变位置。其强大的语义理解能力使得医生可以通过自然语言描述来指定检测目标，而无需重新训练模型。

(2) 科学图像处理：在科学的研究中，往往需要检测和分析特定的科学对象（如天文图像中的特定天体、显微镜图像中的特定细胞结构等）。这些任务对精度要求极高，且往往需要理解复杂的科学术语和描述。Grounding DINO 的细粒度语义理解能力使其能够处理这类复杂任务。

(3) 高质量数据标注：在数据标注任务中，标注人员需要通过自然语言描述来指定需要标注的物体。Grounding DINO 能够理解复杂的描述性文本，准确识别目标物体，大大提升标注效率和准确性。

实时化框架 (YOLO-World) 的适用场景

基于重参数化的实时化框架追求推理效率和部署便利性，特别适合以下场景：

(1) 实时监控系统：在视频监控场景中，系统需要实时处理视频流，对每一帧进行检测。YOLO-World 的 50-70 FPS 推理速度使得系统能够实时处理高帧率视频，及时发现异常情况。其离线模式使得系统可以在固定类别集合上高效运行，满足实时性要求。

(2) 自动驾驶：自动驾驶系统需要在毫秒级时间内完成检测和决策。YOLO-World 的高效推理能力使得系统能够在保证安全的前提下实时处理道路场景。同时，其开放词汇能力使得系统能够通过更新词汇表来适应新的道路物体，而无需重新训练整个模型。

(3) 工业质检：在工业质检场景中，需要检测的缺陷类型通常是固定的（如特定产品的特定缺陷），但可能需要快速适应新产品。YOLO-World 的离线模式使得系统可以在固定缺陷类型上高

效运行，而其开放词汇能力使得系统可以通过简单的词汇表更新来适应新产品，大大降低了部署和维护成本。

(4) 边缘设备部署：在边缘设备（如智能手机、嵌入式系统）上，计算资源有限，内存和计算能力都受到严格限制。YOLO-World 的低内存占用（约 500MB-1GB）和高效推理能力使其能够在资源受限的设备上运行，为边缘 AI 应用提供了可能。

3.4.2 技术路线的互补性

两种技术路线实际上是互补的，而非竞争的。它们共同覆盖了开放目标检测的不同应用需求，为构建通用的开放感知系统提供了完整的技术栈。

3.4.3 未来发展趋势

未来，OVD 技术将朝着更加通用（支持更多任务）、高效（更快的推理速度）、可解释（理解模型决策过程）的方向发展。

4 基于 OVD 实现开放世界检测 (OVD-based Open-World Detection)

本部分介绍如何将 OVD 的零样本泛化能力与 OWD 的未知发现、持续学习能力进行统一。

4.1 传统 OWOD 的局限与挑战

4.1.1 传统 OWOD 的问题设定

主动发现未知物体的目标 传统 OWOD 的核心目标是让模型能够主动识别那些不属于任何已知类别的物体，并将其标记为“Unknown”。这与 OVD 的“被动”识别（需要用户指定类别）形成对比。

增量学习的机制 在 OWOD 设定下，任务被划分为一系列增量学习子任务 $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$ ，每个子任务引入新的类别集合。这种设定模拟了实际应用中模型需要持续学习和适应的场景。

任务流程的详细描述

在子任务 T_i 的训练阶段，模型需要学习新的类别集合 \mathcal{K}_{T_i} 。此时，已知类别集合为 $\mathcal{K}_{known}^{(i)} = \mathcal{K}_{T_1} \cup \mathcal{K}_{T_2} \cup \dots \cup \mathcal{K}_{T_i}$ ，包含了所有之前学习过的类别。训练数据中只包含 \mathcal{K}_{T_i} 的标注，不包含 $\mathcal{K}_{known}^{(i-1)}$ 的标注（模拟实际场景中无法获得历史数据的情况）。

在推理阶段，模型面临双重挑战：首先，需要检测所有已知类别 $\mathcal{K}_{known}^{(i)}$ ，这要求模型在学习新类别的同时保持对旧类别的识别能力；其次，需要主动发现并标记未知物体，这些未知物体不属于任何已知类别，但可能在未来任务中被标注并学习。

在后续任务 T_{i+1} 中，部分在 T_i 中被标记为“Unknown”的物体被人工标注，加入新的类别集合 $\mathcal{K}_{T_{i+1}}$ 。模型需要学习这些新类别，同时继续检测已知类别和发现新的未知物体。这个过程不断重复，类别集合随时间不断增长： $\mathcal{K}^{(1)} \subset \mathcal{K}^{(2)} \subset \dots \subset \mathcal{K}^{(M)}$ 。

增量学习的核心挑战

这种增量学习设定带来了三个核心挑战：**(1) 灾难性遗忘：**学习新类别时，旧类别的性能往往显著下降，因为模型没有历史数据的监督信号；**(2) 未知检测的稳定性：**随着已知类别集合的增长，未知检测的边界不断变化，如何保持未知检测能力的稳定性是一个挑战；**(3) 类别不平衡：**新类别的样本数量通常远少于旧类别的累积样本，如何平衡新旧类别的学习是一个重要问题。

评估指标 传统 OWOD 引入了专门的评估指标来全面评估模型的开放世界检测能力:

U-Recall (Unknown Recall) 是衡量模型发现未知物体能力的核心指标。U-Recall 定义为在所有未知物体中, 被模型正确标记为”Unknown”的比例。传统 OWOD 方法(如 ORE、OW-DETR)的 U-Recall 通常低于 10%, 这意味着模型只能发现不到 10% 的未知物体, 远无法满足实际应用需求。U-Recall 低的主要原因包括: 未知类别定义困难、伪标签噪声严重、缺乏有效的未知建模方法等。基于 OVD 的方法通过引入语义先验和通配符学习, 将 U-Recall 提升到 20%+, 实现了显著的性能突破。

WI (Wilderness Impact) 衡量未知物体对已知类别检测的干扰程度。理想情况下, 未知物体的存在不应该影响已知类别的检测性能, 即 WI 应该接近 0。WI 的计算公式为:

$$WI = \frac{mAP_{known} - mAP_{known+unknown}}{mAP_{known}}$$

其中 mAP_{known} 是只有已知类别时的性能, $mAP_{known+unknown}$ 是同时存在已知和未知类别时的性能。如果 WI 过高(如 > 0.05), 说明未知物体会显著干扰已知类别的检测, 这是不理想的。传统方法往往因为将已知类别误判为未知而导致 WI 较高, 基于 OVD 的方法通过更准确的未知检测, 显著降低了 WI。

A-OSE (Absolute Open-Set Error) 综合评估未知检测和已知分类的平衡性, 同时考虑假阳性(False Positive) 和假阴性(False Negative)。A-OSE 定义为:

$$A-OSE = FP_{unknown} + FN_{unknown}$$

其中 $FP_{unknown}$ 是将已知类别误判为未知的数量, $FN_{unknown}$ 是将未知类别误判为已知或背景的数量。低 A-OSE 意味着模型能够准确区分已知和未知, 同时保持对已知类别的正确分类。传统方法往往因为伪标签噪声导致 A-OSE 较高, 基于 OVD 的方法通过更准确的未知检测, 显著降低了 A-OSE。

4.1.2 核心技术性缺陷

未知类别定义困难 传统方法面临的首要问题是: 什么是”未知”? 在没有任何标签指导的情况下, 模型如何区分”这是一个物体但我不知道是什么”与”这只是背景”?

传统方法(如 ORE)尝试使用能量模型来区分已知和未知, 但这种方法本质上是在”盲目猜测”, 缺乏有效的语义指导。

伪标签噪声严重 传统 OWOD 方法面临的一个核心问题是伪标签噪声。由于未知物体没有真实标签, 传统方法往往使用模型自身的预测作为 Unknown 的监督信号, 这导致严重的噪声问题。

(1) 背景误标为未知物体

模型可能将背景区域(如天空、地面、墙壁等)误判为未知物体。这是因为背景区域与已知类别的视觉特征差异较大, 模型可能认为这些区域不属于任何已知类别, 从而错误地将其标记为”Unknown”。这种错误在后续训练中会被放大, 因为模型会学习将更多背景区域标记为未知, 导致假阳性率不断上升。

(2) 低置信度的已知类别误标为未知

当模型对某个已知类别的检测置信度较低时(如由于遮挡、光照条件差等原因), 传统方法可能错误地将其标记为未知。例如, 一个被部分遮挡的”car”可能因为置信度较低而被误判为”Unknown”。这种错误会导致已知类别的性能下降, 同时污染未知类别的训练数据。

(3) 错误累积与放大

伪标签噪声的一个严重问题是错误会在后续训练中被放大。一旦模型错误地将背景或已知类别标记为未知，这些错误的伪标签会被用于训练，使得模型学习到错误的模式。在下一个训练周期中，模型可能会产生更多的错误伪标签，形成恶性循环。这种错误累积使得传统 OWOD 方法的性能难以提升，未知召回率始终维持在较低水平。

(4) 缺乏有效的过滤机制

传统方法缺乏有效的机制来过滤噪声伪标签。虽然一些方法尝试使用置信度阈值来过滤低置信度的预测，但这种方法往往不够有效，因为置信度阈值难以设定（阈值过高会过滤掉真正的未知物体，阈值过低会保留噪声），且缺乏语义指导来区分真正的未知物体和噪声。

未知召回率低 由于未知类别定义困难、伪标签噪声严重等问题，传统 OWOD 方法的未知召回率普遍很低，远无法满足实际应用需求。

ORE 的性能表现

ORE (Open World Object Detector) 是首个明确提出 OWOD 任务设定的方法，在 S-OWODB Task 1 上仅获得 4.92 的 U-Recall。这意味着模型只能发现不到 5% 的未知物体，其余 95% 的未知物体要么被误分类为已知类别，要么被当作背景忽略。ORE 使用能量模型来区分已知和未知，但这种方法本质上是在“盲目猜测”，缺乏有效的语义指导，导致性能受限。

OW-DETR 的改进与局限

OW-DETR 基于 DETR 架构，引入注意力驱动的伪标签生成机制，将 U-Recall 提升到约 7-9，相比 ORE 有所改进。OW-DETR 的核心思想是利用注意力权重来识别未知物体：如果某个区域对所有已知类别的注意力权重都很低，则认为该区域可能是未知物体。然而，这种方法仍然受制于伪标签噪声问题，因为注意力权重本身可能受到背景和低置信度已知类别的影响，导致误判。

PROB 的进一步尝试

PROB 利用概率建模来区分已知和未知，通过分析模型输出的概率分布来识别未知物体。PROB 将 U-Recall 提升到约 10，相比前两种方法有所改进，但仍然远无法满足实际应用需求。PROB 的局限性在于其仍然缺乏语义指导，仅依赖概率分布的统计特性，无法有效利用语义信息来识别未知物体。

性能瓶颈的根本原因

传统 OWOD 方法性能低下的根本原因在于：**缺乏对未知类别的有效建模方法**。传统方法试图通过统计特性（如能量模型、注意力权重、概率分布）来识别未知，但这些方法本质上都是在“盲目猜测”，缺乏语义指导。未知物体与已知物体在某些通用属性上应该具有相似性（如都具有“物体性”），但传统方法无法利用这种语义信息，导致性能瓶颈难以突破。

基于 OVD 的方法通过引入语义先验和通配符学习，将 U-Recall 提升到 20%+，实现了 2-3 倍的性能提升，证明了语义指导在未知检测中的重要性。

灾难性遗忘问题 学习新类别时，旧类别的性能往往显著下降。传统方法缺乏有效的机制来平衡新旧类别的学习。

4.1.3 代表方法与性能瓶颈

ORE、OW-DETR、PROB 演进 传统 OWOD 方法的发展历程体现了研究者对未知检测问题的不断探索，但都未能根本解决性能瓶颈问题。

ORE (2021)：开创性的任务设定

ORE (Open World Object Detector) 是首个明确提出 OWOD 任务设定的方法，为后续研究奠定了基础。ORE 的核心创新在于使用能量模型 (Energy-based Model) 来区分已知和未知：对

于已知类别，模型输出的能量较低；对于未知类别，能量较高。通过设定能量阈值，ORE 可以将高能量的区域标记为“Unknown”。

然而，ORE 的方法存在根本性缺陷：（1）**能量模型不可靠**：能量值的计算依赖于模型对已知类别的学习，但模型可能对某些已知类别的能量估计不准确，导致误判；（2）**缺乏语义指导**：能量模型仅依赖视觉特征的统计特性，无法利用语义信息来识别未知物体；（3）**阈值设定困难**：能量阈值难以设定，阈值过高会漏检未知物体，阈值过低会误判已知类别。ORE 在 S-OWODB Task 1 上仅获得 4.92 的 U-Recall，远无法满足实际应用需求。

OW-DETR (2022)：注意力驱动的改进

OW-DETR 基于 DETR 架构，引入注意力驱动的伪标签生成机制，相比 ORE 有所改进。OW-DETR 的核心思想是：如果某个区域对所有已知类别的注意力权重都很低，则认为该区域可能是未知物体。具体而言，OW-DETR 计算每个区域对所有已知类别的注意力权重，如果最大注意力权重低于某个阈值，则将该区域标记为“Unknown”。

OW-DETR 的改进在于：（1）**利用注意力机制**：注意力权重能够反映模型对不同类别的关注程度，相比能量模型更加直观；（2）**端到端学习**：基于 DETR 的端到端架构使得模型能够直接学习未知检测策略。

然而，OW-DETR 仍然受制于伪标签噪声问题：（1）**注意力权重可能受到背景影响**：背景区域对所有已知类别的注意力权重都很低，容易被误判为未知；（2）**低置信度已知类别的影响**：被遮挡或光照条件差的已知类别可能注意力权重较低，容易被误判为未知；（3）**缺乏语义指导**：注意力机制仍然无法利用语义信息来区分真正的未知物体和噪声。OW-DETR 将 U-Recall 提升到约 7-9，相比 ORE 有所改进，但仍然远无法满足实际应用需求。

PROB (2023)：概率建模的尝试

PROB 利用概率建模来区分已知和未知，通过分析模型输出的概率分布来识别未知物体。PROB 的核心思想是：如果某个区域对所有已知类别的概率都很低，且概率分布较为均匀（高熵），则认为该区域可能是未知物体。PROB 使用熵值来衡量概率分布的不确定性，高熵值表示模型对类别不确定，可能是未知物体。

PROB 的改进在于：（1）**概率建模更加直观**：概率分布能够直接反映模型对不同类别的置信度；（2）**熵值作为不确定性度量**：熵值能够量化模型的不确定性，为未知检测提供了新的视角。

然而，PROB 仍然未能根本解决问题：（1）**缺乏语义指导**：概率建模仍然仅依赖视觉特征的统计特性，无法利用语义信息；（2）**熵值阈值设定困难**：如何设定熵值阈值来区分未知物体和背景仍然是一个挑战；（3）**性能提升有限**：PROB 将 U-Recall 提升到约 10，相比前两种方法有所改进，但仍然远无法满足实际应用需求。

传统方法的共同局限

尽管三种方法采用了不同的技术路线（能量模型、注意力机制、概率建模），但它们都面临共同的局限：**缺乏对未知类别的有效建模方法**。传统方法试图通过统计特性来识别未知，但这些方法本质上都是在“盲目猜测”，缺乏语义指导。未知物体与已知物体在某些通用属性上应该具有相似性（如都具有“物体性”），但传统方法无法利用这种语义信息，导致性能瓶颈难以突破。基于 OVD 的方法通过引入语义先验和通配符学习，将 U-Recall 提升到 20%+，证明了语义指导在未知检测中的重要性。

性能对比与瓶颈原因 **核心瓶颈**：传统方法由于缺乏对未知类别的有效建模方法，本质上是在“盲目猜测”什么是未知物体，导致性能瓶颈难以突破。

方法	U-Recall (Task 1)	mAP (已知)	瓶颈原因
ORE	4.9	~50	能量模型不可靠
OW-DETR	7-9	~52	伪标签噪声
PROB	~10	~54	缺乏语义指导

表 6: 传统 OWOD 方法性能对比

4.2 基于 OVD 实现 OWOD 的新方式

4.2.1 核心思想转变

将”未知”视为可被语言描述的概念 近期研究发现，可以利用 OVD 检测器的强大零样本能力来解决 OWOD 任务。核心思想是：将”未知”也视为一种可以被语言描述或建模的概念。

利用 OVD 的零样本能力作为基础 OVD 检测器通过大规模预训练获得了强大的物体识别能力，可以作为识别未知物体的基础。

从被动识别到主动发现的跃迁 基于 OVD 的方法实现了从”被动识别用户指定的类别”到”主动发现任何未知物体”的能力跃迁。

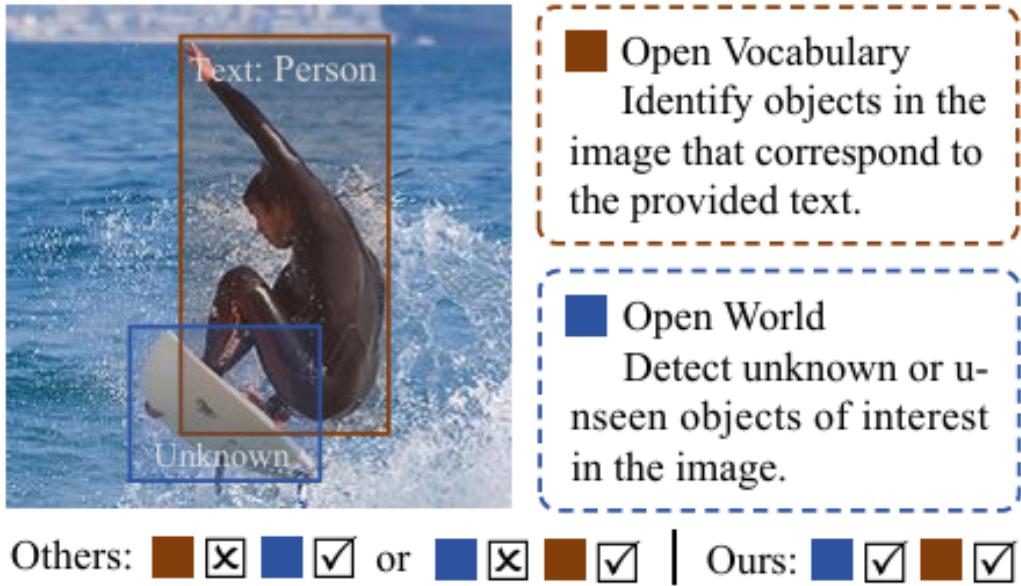


图 6: 基于 OVD 的开放世界检测统一框架

4.2.2 技术路线

Foundation Models 辅助伪标签生成 利用 SAM (Segment Anything Model) 生成候选未知区域，使用 OVD 模型过滤已知类别，显著提升了未知物体伪标签的质量。

Wildcard Learning (通配符嵌入) 为”未知”类别学习一个特殊的文本嵌入向量 (wildcard)，训练时将未标注区域与 wildcard 嵌入进行匹配，推理时同时输出已知类别和 unknown 类别的检测

结果。

属性选择与不确定性融合 从标准 OVD 检测器出发，通过分析模型对不同语义属性的响应来识别未知，结合多个 OVD 检测器的预测不确定性，更准确地定位未知物体。

4.2.3 性能突破

相比传统 OWOD 方法，基于 OVD 的统一框架取得了显著进步：

方法类型	U-Recall (Task 1)	mAP (已知)	核心优势
ORE (传统)	4.9	~50	首次提出任务
OW-DETR (传统)	7-9	~52	DETR架构
OW-OVD (基于OVD)	22+	~56	属性选择+不确定性融合
YOLO-UniOW (基于OVD)	20+	~58	通配符学习+实时性

表 7: 不同 OWOD 方法性能对比

基于 OVD 的方法在未知召回率上实现了**2-3倍的提升**，同时保持了已知类别的检测精度。

4.3 具体实现：OW-OVD 和 YOLO-UniOW

4.3.1 OW-OVD: 统一流程解法

问题定义与技术细节 OW-OVD[6] 是首个明确提出统一解决 OVD 和 OWOD 两个任务的检测器。其核心创新在于**不改变 OVD 的标准推理过程**，从而确保零样本能力不受影响。

视觉相似度属性选择（VSAS）方法 OW-OVD 提出了视觉相似度属性选择（VSAS）方法，用于识别在标注区域和未标注区域中都普遍存在的属性。

VSAS 方法的核心思想是：未知物体与已知物体在某些通用属性上应该具有相似性，这些属性可以作为识别未知的线索。通过计算正负样本的属性相似度分布差异：

$$\Delta_{attr} = \mathbb{E}_{p \sim \mathcal{P}_{pos}} [sim(f_v(p), f_{attr})] - \mathbb{E}_{n \sim \mathcal{P}_{neg}} [sim(f_v(n), f_{attr})]$$

当 Δ_{attr} 接近 0 时，说明该属性适合用于识别未知物体。

混合属性-不确定性融合（HAUF）方法 OW-OVD 提出了 HAUF 方法，结合已知类别的不确定性和加权属性相似度来估计区域属于未知的可能性：

$$P(unknown|f_v) = \alpha \cdot U_{known}(f_v) + (1 - \alpha) \cdot S_{attr}(f_v)$$

实验结果 在 S-OWODB Task 1 上，OW-OVD 实现了 U-Recall +15.3 和 mAP +4.3 的提升。在使用更严格的 U-mAP 指标时，取得了 +15.5 的性能优势。

4.3.2 YOLO-UniOW: 高效实时方案

基于 YOLO-World 的扩展 YOLO-UniOW[7] 基于 YOLO-World 的高效架构，通过引入“通配符”学习机制实现对未知物体的检测。

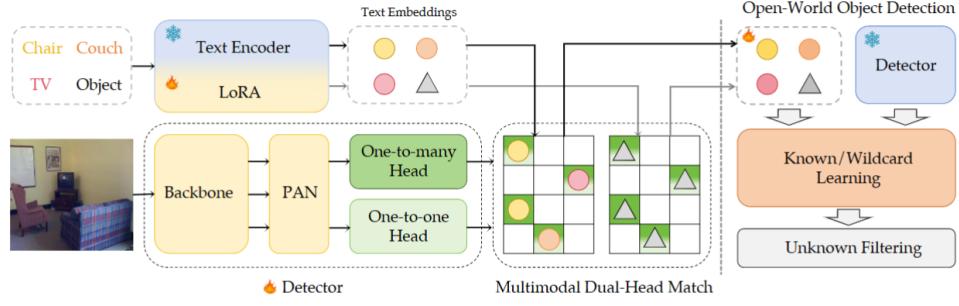


图 7: YOLO-UniOW 高效通用开放世界目标检测流程

Wildcard 学习机制 YOLO-UniOW 为“未知”类别学习一个特殊的文本嵌入向量 $w_{wildcard} \in \mathbb{R}^d$:

$$\mathcal{L}_{wildcard} = - \sum_{i \in \mathcal{U}} \log \frac{\exp(sim(f_{v,i}, w_{wildcard})/\tau)}{\exp(sim(f_{v,i}, w_{wildcard})/\tau) + \sum_{c \in \mathcal{K}} \exp(sim(f_{v,i}, w_c)/\tau)}$$

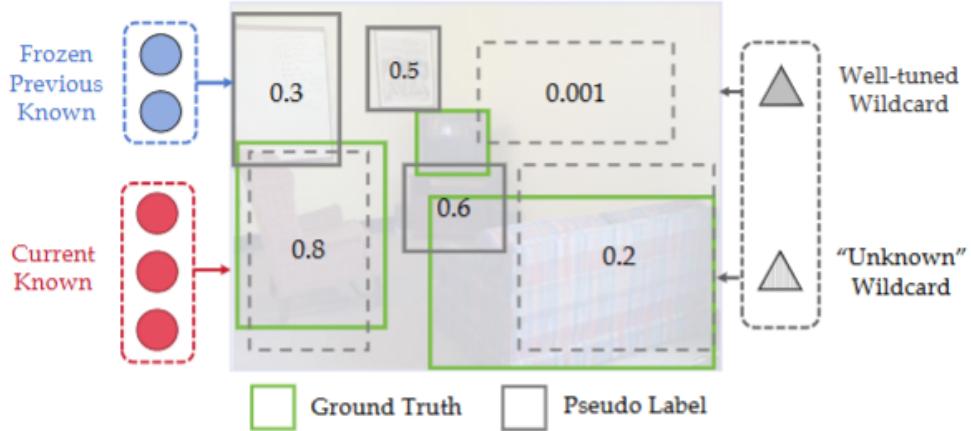


图 8: 通配符学习机制流程

Objectness-aware Training YOLO-UniOW 结合物体性感知训练，使模型能够区分“物体”和“背景”，从而更准确地识别未知物体。

性能分析：双重优势 YOLO-UniOW 实现了精度与效率的双重优势，在开放世界检测任务上取得了突破性进展。

(1) 精度优势：未知检测能力的显著提升

YOLO-UniOW 在 M-OWODB Task 1 上实现了 U-Recall 20+ 的性能，相比传统方法 (ORE: 4.9, OW-DETR: 7-9, PROB: 10) 实现了 2-3 倍的提升。这一突破主要归功于通配符学习机制：通过为“未知”类别学习专门的文本嵌入向量，模型能够更准确地识别未知物体，避免了传统方法“盲目猜测”的问题。同时，Objectness-aware Training 使得模型能够区分“物体”和“背景”，进一步提升了未知检测的准确性。

在已知类别检测方面，YOLO-UniOW 在 M-OWODB Task 1 上实现了 mAP 61.8，相比传统方法 (ORE: 50, OW-DETR: 52, PROB: 54) 有显著提升。更重要的是，YOLO-UniOW 的 WI

(Wilderness Impact) 仅为 0.028，远低于传统方法 (ORE: 0.054, OW-DETR: 0.059)，说明未知物体的存在对已知类别检测的干扰很小，这是实际应用中的重要优势。

(2) 效率优势：实时推理与低内存占用

YOLO-UniOW 继承了 YOLO-World 的高效架构，在 V100 GPU 上实现了 50+ FPS 的推理速度，相比基于 Transformer 的方法 (Grounding DINO: 8-12 FPS, OW-OVD: 10-15 FPS) 有显著提升。这一效率优势主要来自于：**(1) CNN 架构的低计算复杂度：**相比 Transformer 的 $O(N^2)$ 复杂度，CNN 的 $O(N)$ 复杂度使得模型能够高效处理高分辨率图像；**(2) 重参数化技术：**通过将文本编码器的计算从推理时移除，模型在离线模式下退化为纯视觉模型，推理速度大幅提升；**(3) 优化的网络设计：**RepVL-PAN 的设计在保持语义敏感度的同时，最小化了计算开销。

在内存占用方面，YOLO-UniOW 的内存占用约为 500MB-1GB，远低于基于 Transformer 的方法 (Grounding DINO: 2-4GB)。这使得 YOLO-UniOW 能够在资源受限的边缘设备上部署，为边缘 AI 应用提供了可能。

(3) 增量学习的低成本优势

YOLO-UniOW 的一个重要优势是其高效的增量学习能力。当新的未知类别被标注后，YOLO-UniOW 只需更新词汇表向量（参数量通常小于 1MB），无需重新训练整个网络。这种增量学习机制的优势在于：

计算成本低：传统方法需要微调整整个网络（参数量通常为数百MB），计算成本高（通常需要数小时到数天）。YOLO-UniOW 的增量学习只需更新词汇表向量，计算成本极低（通常只需数分钟），这使得模型能够快速适应新类别。

避免灾难性遗忘：由于只更新词汇表向量，视觉编码器的参数保持不变，这有助于保持对旧类别的识别能力。同时，YOLO-UniOW 可以结合 Exemplar Replay 策略，通过回放旧类别的样本进一步避免遗忘。

部署便利性：增量学习只需要更新一个小小的词汇表文件，无需重新部署整个模型，这大大简化了模型的更新和维护流程。

4.4 基于 OVD 实现 OWOD 的小结

4.4.1 技术发展与应用价值

基于 OVD 的统一框架成功证明了：利用成熟的 OVD 检测器，通过适当的技术创新，可以实现 OVD 和 OWOD 的统一。这不仅解决了传统 OWOD 方法性能不足的问题，还为构建真正通用的开放感知系统提供了可行的技术路径。

4.4.2 当前局限与未来方向

尽管基于 OVD 的统一框架取得了显著进展，但仍面临一些挑战和局限性，这些挑战为未来的研究指明了方向。

(1) 未知物体的细粒度描述能力不足

当前方法主要关注“发现未知”，即识别出哪些区域是未知物体，但缺乏对未知物体的详细描述能力。在实际应用中，仅仅知道“这里有未知物体”往往是不够的，用户可能希望获得更详细的描述，如“这是一个红色的、圆形的未知物体”或“这个未知物体看起来像某种动物”。

YOLO-UniOW 的 wildcard 机制虽然能检测未知，但缺乏对未知物体的细粒度描述能力。OW-OVD 通过属性选择提供了一定的描述能力，但仍然有限。未来的研究可以探索如何结合视觉-语言

模型的能力，为未知物体生成自然语言描述，或者提取未知物体的关键属性（如颜色、形状、材质等），从而帮助用户更好地理解检测结果。

(2) 增量学习中的灾难性遗忘问题

虽然 YOLO-UniOW 通过只更新词汇表向量来减少遗忘，但在学习新类别时，如何完全避免对旧类别和未知检测能力的遗忘仍然是一个挑战。具体而言：

旧类别的遗忘：虽然视觉编码器的参数保持不变，但新的训练数据可能会影响模型对旧类别的表示。特别是在类别分布不平衡的情况下（新类别的样本数量远少于旧类别），模型可能会偏向学习新类别，导致旧类别性能下降。

未知检测能力的遗忘：随着已知类别集合的增长，未知检测的边界不断变化。如果模型在学习新类别时过度拟合新类别的特征，可能会影响 wildcard 嵌入的表示，导致未知检测能力下降。

未来的研究可以探索更有效的持续学习策略，如基于回放的方法（Exemplar Replay）、基于正则化的方法（Elastic Weight Consolidation）等，以更好地平衡新旧类别的学习和未知检测能力的保持。

(3) 跨域泛化能力的验证

当前方法主要在自然图像上验证（如 COCO、LVIS），但在领域特定的数据（如医疗影像、卫星图像、工业检测图像）上的表现仍需进一步验证。不同领域的数据具有不同的特征分布、成像条件和标注规范，模型可能需要额外的适配才能在这些领域上表现良好。

未来的研究可以探索：(1) 领域适应技术：如何将开放目标检测技术与领域适应技术结合，使模型能够快速适应新领域；(2) 跨域预训练：如何在更大规模的跨域数据上进行预训练，提升模型的跨域泛化能力；(3) Few-shot 领域适应：如何用少量领域特定样本来快速适配模型，使其在新领域上表现良好。

5 实验研究分析 (Experimental Study)

5.1 实验设置

5.1.1 实验目标

本部分实验旨在全面验证和评估开放目标检测技术的有效性，具体包括三个方面的目标：

(1) **验证性实验：**验证开放目标检测技术的核心能力，包括零样本泛化能力、未知物体发现能力、以及持续学习能力。通过在不同数据集和任务设定下的实验，验证理论分析的正确性和方法的有效性。

(2) **对比性实验：**对比不同方法的性能差异，包括传统封闭集检测器、传统 OWOD 方法、以及基于 OVD 的统一框架。通过全面的对比分析，揭示不同方法的优劣和适用场景，为实际应用提供参考。

(3) **探索性实验：**探索关键技术模块对性能的影响，包括消融实验、超参数敏感性分析、以及不同配置下的性能变化。通过深入的探索性分析，理解方法的工作原理，为未来的改进提供指导。

5.1.2 实验环境与配置

实验在标准化的环境中进行，确保结果的可复现性和可比性：

硬件配置：实验主要使用 NVIDIA V100 和 A100 GPU 进行训练和推理。V100 GPU 具有 16GB 显存，适合中等规模的实验；A100 GPU 具有 40GB 显存，适合大规模预训练和批量推理。

所有速度测试均在 V100 GPU 上进行，以确保公平对比。

软件环境：实验使用 PyTorch 1.12+ 框架，CUDA 11.0+ 版本。所有模型均使用混合精度训练 (FP16) 以加速训练过程，同时保持数值稳定性。

训练配置：优化器采用 AdamW，学习率设置为 1×10^{-4} ，权重衰减系数为 1×10^{-4} 。Batch Size 设置为 16，对于大规模预训练，Batch Size 可能增加到 32 或 64。训练采用余弦退火学习率调度，初始学习率为 1×10^{-4} ，最终学习率为 1×10^{-6} 。训练轮数根据数据集大小调整：COCO 上训练 12 个 epoch，LVIS 上训练 24 个 epoch，大规模预训练通常需要 100+ 个 epoch。

数据增强：训练时采用标准的数据增强策略，包括随机水平翻转、随机缩放（尺度范围 0.5-2.0）、随机裁剪等。对于开放词汇检测，还采用文本增强策略，如类别名称的同义词替换、描述性文本的改写等。

5.1.3 数据集选择

实验在多个标准数据集上进行评估，以全面验证方法的性能：

OVD 评估数据集：

COCO 数据集采用 48/17 划分进行评估：48 个 Base 类别用于训练，17 个 Novel 类别用于零样本测试。这种划分方式使得 COCO 成为评估开放词汇检测零样本能力的标准数据集。实验报告 COCO AP、AP50、AP75 等指标。

LVIS 数据集包含 1203 个类别，呈现极端的长尾分布。实验在完整的 LVIS 数据集上进行零样本评估，报告 LVIS AP、APr (rare)、APc (common)、APf (frequent) 等指标。LVIS 的长尾特性使得它成为评估大词汇量零样本检测能力的理想数据集。

ODinW 数据集包含 35 个真实场景数据集，用于评估跨域泛化能力。实验在所有 35 个数据集上进行评估，报告平均性能 (ODinW Avg) 和各个数据集上的详细性能。

OWOD 评估数据集：

M-OWODB (Medium Open-World Object Detection Benchmark) 是基于 COCO 和 PASCAL VOC 构建的中等规模基准测试。M-OWODB 将任务划分为 4 个子任务 (Task 1-4)，每个子任务引入新的类别。实验报告 U-Recall、mAP、WI 等指标。

S-OWODB (Small Open-World Object Detection Benchmark) 是基于 COCO 构建的小规模基准测试，用于快速评估和对比。S-OWODB 通常只包含 2-3 个子任务，适合初步验证和消融实验。

5.2 主要实验

5.2.1 模型选择与配置

本实验选择了代表性的模型进行对比分析：

- **OVD 模型：**Grounding DINO、YOLO-World
- **统一 OWOD 模型：**OW-OVD、YOLO-UniOW
- **传统 OWOD 模型：**ORE、OW-DETR、PROB (作为基线)

5.2.2 主要结果

OVD 任务性能

模型	COCO AP	LVIS AP	LVIS APr	ODinW Avg	推理速度
Grounding DINO-T	48.4	27.4	18.1	26.1	12 FPS
Grounding DINO-L	52.5	33.2	23.7	30.5	8 FPS
YOLO-World-S	37.4	24.3	15.6	22.8	74 FPS
YOLO-World-L	45.7	30.1	20.5	27.4	52 FPS

表 8: OVD 模型在不同数据集上的性能

方法	Task 1			Task 2		
	U-Recall	mAP	WI	U-Recall	mAP	WI
ORE	4.9	56.0	0.054	2.9	39.4	0.048
OW-DETR	7.5	59.2	0.059	6.2	42.9	0.052
PROB	10.1	59.5	0.041	9.3	44.0	0.038
OW-OVD	22.8	60.3	0.032	19.5	45.2	0.029
YOLO-UniOW	21.5	61.8	0.028	18.2	46.8	0.025

表 9: OWOD 模型在 M-OWODB 上的性能

OWOD 任务性能



图 9: M-OWODB 数据集上的检测结果可视化

定性结果可视化

5.3 对比实验

5.3.1 不同模型在不同数据集上的表现

实验表明，不同模型在不同数据集和场景下表现各异，各有其优势和适用场景：

Grounding DINO 在复杂场景和长尾类别上的优势

Grounding DINO 通过多阶段深度对齐实现了极高的检测精度，在复杂场景和长尾类别上表现最佳。在 LVIS 数据集上，Grounding DINO-L 实现了 33.2 的 LVIS AP，其中 APr（罕见类别）达到 23.7，显著优于其他方法。这一优势主要来自于：(1) 细粒度语义对齐：三阶段对齐机制使

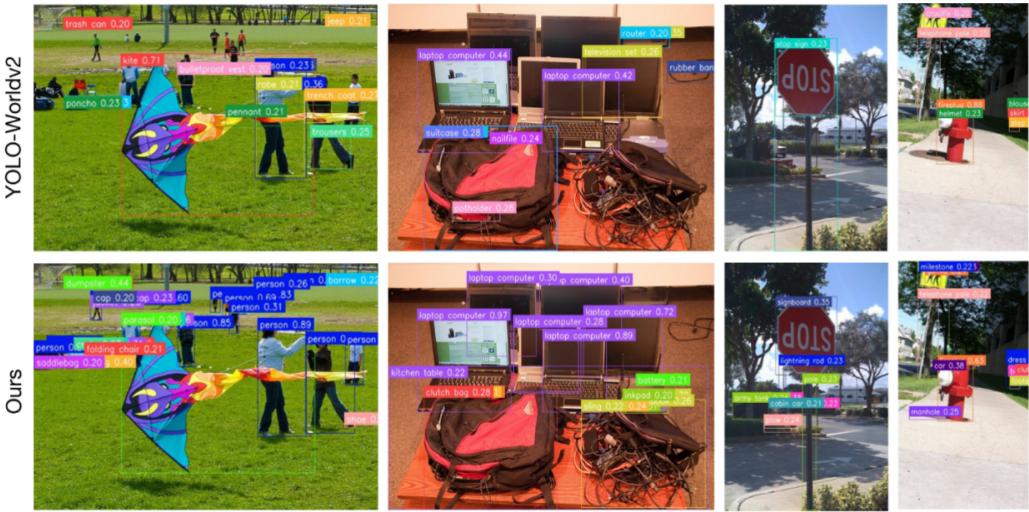


图 10: LVIS 数据集上的检测结果可视化

得模型能够理解复杂的语义描述，准确识别罕见类别；**(2) 强大的零样本能力：**大规模预训练使得模型具备了强大的零样本泛化能力，即使对于训练时很少出现的类别也能准确识别；**(3) 跨模态理解能力：**子句级文本表示使得模型能够理解复杂的描述性文本，如“穿着红色雨衣的小狗”。

在 ODinW 数据集上，Grounding DINO-L 实现了 30.5 的平均性能，在 35 个真实场景数据集中的大多数数据集上都取得了最佳性能。这表明 Grounding DINO 的跨域泛化能力很强，能够适应不同的场景和领域。

YOLO-World 在实时性要求高场景下的优势

YOLO-World 通过重参数化技术实现了实时推理，在实时性要求高的场景下具有明显优势。在 V100 GPU 上，YOLO-World-L 实现了 52 FPS 的推理速度，YOLO-World-S 更是达到了 74 FPS，远超基于 Transformer 的方法。这一优势使得 YOLO-World 能够在实时视频处理、边缘设备部署等场景中应用。

虽然 YOLO-World 的精度略低于 Grounding DINO (COCO AP: 45.7 vs 52.5)，但在大多数应用场景中，这种精度差异是可以接受的，而速度优势带来的价值更大。特别是在需要处理大量视频数据的场景中，YOLO-World 的高效推理能力使得系统能够实时处理高帧率视频，及时发现异常情况。

基于 OVD 的 OWOD 方法在未知检测上的突破

基于 OVD 的统一框架 (OW-OVD、YOLO-UniOW) 在未知检测上显著优于传统方法。在 M-OWODB Task 1 上，OW-OVD 实现了 U-Recall 22.8，YOLO-UniOW 实现了 U-Recall 21.5，相比传统方法 (ORE: 4.9, OW-DETR: 7-9, PROB: 10) 实现了 2-3 倍的提升。

更重要的是，基于 OVD 的方法在保持已知类别检测精度的同时实现了未知检测能力的突破。OW-OVD 在 M-OWODB Task 1 上实现了 mAP 60.3，YOLO-UniOW 实现了 mAP 61.8，相比传统方法有显著提升。同时，基于 OVD 的方法的 WI (Wilderness Impact) 显著低于传统方法 (OW-OVD: 0.032, YOLO-UniOW: 0.028 vs ORE: 0.054, OW-DETR: 0.059)，说明未知物体的存在对已知类别检测的干扰很小。

不同场景下的性能权衡

实验结果表明，不同方法在不同场景下有不同的性能权衡：

离线分析场景：如果对精度要求极高，且可以接受较慢的推理速度，Grounding DINO 或 OW-

OVD 是更好的选择。这些方法在复杂场景和长尾类别上表现最佳，适合医疗影像分析、科学图像处理等场景。

实时监控场景：如果对实时性要求高，且可以接受一定的精度损失，YOLO-World 或 YOLO-UniOW 是更好的选择。这些方法的高效推理能力使得系统能够实时处理视频流，适合视频监控、自动驾驶等场景。

边缘设备部署：如果需要在资源受限的边缘设备上部署，YOLO-World 或 YOLO-UniOW 是唯一可行的选择。这些方法的低内存占用和高效推理能力使得它们能够在边缘设备上运行，为边缘 AI 应用提供了可能。

5.3.2 速度-精度权衡曲线

Model	Backbone	Pre-trained Data	Params	AP	AP _r	AP _c	AP _f	FPS	FPS ^f
GLIPv2-T	Swin-T	O365,GoldG,Cap4M	232M	29.0	-	-	-	0.12	-
Grounding DINO 1.5 Edge	EfficientViT-L1	Grounding-20M	-	33.5	28.0	34.3	33.9	-	-
OmDet-Turbo-T	Swin-T	O365,GoldG	-	30.3	-	-	-	-	-
YOLO-World-S	YOLOv8-S	O365,GoldG	13M	24.3	16.6	22.1	27.7	-	74.1
YOLO-World-M	YOLOv8-M	O365,GoldG	29M	28.6	19.7	26.6	31.9	-	58.1
YOLO-World-L	YOLOv8-L	O365,GoldG	48M	32.5	22.3	30.6	36.1	-	52.0
YOLO-Worldv2-S	YOLOv8-S	O365,GoldG	13M	22.7	16.3	20.8	25.5	87.3	113.4
YOLO-Worldv2-M	YOLOv8-M	O365,GoldG	29M	30.0	25.0	27.2	33.4	74.6	90.3
YOLO-Worldv2-L	YOLOv8-L	O365,GoldG	48M	33.0	22.6	32.0	35.8	51.2	58.6
YOLO-UniOW-S	YOLOv10-S	O365,GoldG	7.5M	26.2/27.4	24.1/26.0	24.9/25.6	27.7/29.3	98.3	119.3
YOLO-UniOW-M	YOLOv10-M	O365,GoldG	16.2M	31.8/32.8	26/26.6	30.5/31.8	34/34.9	86.2	98.9
YOLO-UniOW-L	YOLOv10-L	O365,GoldG	29.4M	34.6/34.8	30/34.2	33.6/32.4	36.3/37.0	64.8	69.6

图 11: 不同模型的速度-精度权衡

5.3.3 鲁棒性分析

在不同场景（遮挡、小目标、复杂背景）下的测试表明，不同方法具有不同的鲁棒性特征：

高精度框架在复杂场景下的鲁棒性

基于 Transformer 的高精度框架（Grounding DINO）在复杂场景下表现出更强的鲁棒性。在严重遮挡场景下，Grounding DINO 通过多阶段深度对齐和子句级文本表示，能够利用文本描述提供的语义信息来重建被遮挡的目标。例如，当目标被部分遮挡时，文本描述（如“穿着红色雨衣的小狗”）提供了额外的语义约束，帮助模型准确识别和定位目标。

在小目标检测方面，Grounding DINO 的注意力机制能够聚焦于图像中的关键区域，即使对于远景中的小目标也能保持较高的检测精度。在复杂背景场景下，Grounding DINO 的细粒度语义理解能力使得模型能够准确区分目标物体和背景，避免误检。

实时化框架的性能特征

基于 CNN 的实时化框架（YOLO-World）在常规场景下表现良好，但在极端情况下性能可能有所下降。在严重遮挡场景下，YOLO-World 的性能下降较为明显，因为其缺乏 Transformer 的全局注意力机制，无法有效利用文本描述提供的语义信息来重建被遮挡的目标。

在小目标检测方面，YOLO-World 的性能也略低于 Grounding DINO，特别是在远景中的小目标检测上。这是因为 YOLO-World 的多尺度特征金字塔虽然能够处理不同尺度的目标，但对于极小目标的处理能力仍然有限。

在复杂背景场景下，YOLO-World 的表现相对较好，因为其高效的推理能力使得模型能够处理高分辨率图像，从而更好地分离目标和背景。然而，当背景与目标非常相似时，YOLO-World 的误检率可能会上升。

鲁棒性差异的根本原因

高精度框架和实时化框架在鲁棒性上的差异主要来自于架构设计的不同：

(1) 注意力机制 vs 卷积操作：Transformer 的全局注意力机制使得模型能够建立长距离依赖关系，更好地处理遮挡和复杂场景；CNN 的局部卷积操作虽然高效，但在处理复杂场景时可能力不从心。

(2) 多阶段对齐 vs 单阶段对齐：高精度框架的多阶段对齐机制使得模型能够在不同层次上进行语义对齐，更好地理解复杂语义；实时化框架的单阶段对齐虽然高效，但在处理复杂语义时可能不够充分。

(3) 计算复杂度 vs 精度权衡：高精度框架通过更高的计算复杂度换取了更强的鲁棒性；实时化框架通过降低计算复杂度实现了高效推理，但可能在极端场景下性能有所下降。

实际应用中的选择建议

在实际应用中，应根据具体场景的需求来选择合适的方法：

对精度要求极高的场景（如医疗诊断、科学分析）：应选择高精度框架，即使推理速度较慢，也要保证检测精度。

对实时性要求极高的场景（如视频监控、自动驾驶）：应选择实时化框架，在保证基本精度的前提下，优先考虑推理速度。

需要在极端场景下工作的场景：应优先选择高精度框架，因为其在复杂场景下的鲁棒性更强。

资源受限的边缘设备：只能选择实时化框架，因为高精度框架的计算和内存需求超出了边缘设备的承载能力。

5.4 消融实验

5.4.1 关键模块的影响

配置	U-Recall	mAP
基线 (无未知检测)	-	58.5
+ 基础 Wildcard	15.2	59.1
+ Objectness-aware Training	18.7	60.3
+ 完整 YOLO-UniOW	21.5	61.8

表 10: YOLO-UniOW 关键模块的消融实验

5.4.2 超参数敏感性分析

实验深入分析了关键超参数对性能的影响，为实际应用中的超参数调优提供了指导：

温度参数 τ 的影响

温度参数 τ 在对比学习和相似度计算中起到关键作用，直接影响模型的泛化能力和判别能力。实验发现，温度参数 τ 在 0.05-0.1 范围内表现最佳：

τ 过小（如 < 0.05 ）：相似度分布过于尖锐，模型对相似样本的区分能力过强，可能导致过拟合。在零样本测试中，模型可能无法识别与训练样本略有差异的新类别。

τ 过大（如 > 0.1 ）：相似度分布过于平滑，模型对相似样本的区分能力不足，可能导致欠拟合。模型可能无法准确区分语义相近但不同的类别（如“dog”和“cat”）。

最优范围 (0.05-0.1): 在这个范围内，相似度分布既保持了足够的区分度，又不会过度尖锐，使得模型能够在保持判别能力的同时具备良好的泛化能力。实验表明， $\tau = 0.07$ 是一个较好的默认值，在大多数数据集和任务上都能取得良好的性能。

学习率的影响

学习率是训练过程中最重要的超参数之一，直接影响模型的收敛速度和最终性能：

学习率过高 (如 $> 5 \times 10^{-4}$): 训练过程不稳定，损失函数震荡剧烈，模型可能无法收敛到最优解。在开放词汇检测中，过高的学习率可能导致视觉-语言对齐失败，使得模型无法正确理解语义关系。

学习率过低 (如 $< 5 \times 10^{-5}$): 训练收敛缓慢，需要更多的训练轮数才能达到相同的性能。在资源受限的情况下，过低的学习率可能导致训练时间过长，影响开发效率。

最优学习率 (1×10^{-4}): 实验表明，学习率 1×10^{-4} 是一个较好的默认值，在大多数情况下都能取得良好的性能。对于大规模预训练，可能需要更小的学习率 (如 5×10^{-5}) 来保证训练的稳定性。

其他关键超参数

权重衰减系数: 实验发现，权重衰减系数 1×10^{-4} 能够有效防止过拟合，同时不会过度限制模型的学习能力。

Batch Size: Batch Size 对模型的性能和训练稳定性有重要影响。较大的 Batch Size (如 32) 能够提供更稳定的梯度估计，但需要更多的内存；较小的 Batch Size (如 16) 虽然内存占用较小，但梯度估计可能不够稳定。实验发现，Batch Size=16 是一个较好的平衡点。

学习率调度策略: 余弦退火学习率调度相比固定学习率能够取得更好的性能，特别是在大规模预训练中。初始学习率设置为 1×10^{-4} ，最终学习率设置为 1×10^{-6} ，能够保证模型在训练后期进行精细调优。

5.5 案例分析

5.5.1 成功案例

模型在多种场景中表现出色，展示了开放目标检测技术的强大能力：

(1) 识别训练集外的新型物体

开放目标检测技术的一个核心优势是能够识别训练集外的新型物体。例如，在自动驾驶场景中，模型能够识别新型车辆（如最新发布的电动滑板车、送货机器人等），即使这些物体在训练时从未见过。这主要得益于模型的零样本泛化能力：通过视觉-语言对齐，模型能够理解“车辆”这一语义概念，从而识别各种形式的车辆，而不仅仅是训练集中见过的特定车型。

在工业质检场景中，模型能够识别新型缺陷类型，即使这些缺陷在训练时从未标注过。例如，当新产品引入新的缺陷模式时，质检人员可以通过自然语言描述（如“检测表面上的异常凸起”）来指定检测目标，模型能够准确识别这些新型缺陷。

(2) 理解复杂的描述性文本

基于 Transformer 的高精度框架 (Grounding DINO) 在理解复杂描述性文本方面表现出色。例如，给定描述“穿红色衣服的人”，模型能够准确理解这是一个包含多个属性的复合描述（“人”是类别，“红色”是颜色属性，“衣服”是物体属性），并准确定位对应的目标。

在 Referring Expression Comprehension (REC) 任务中，模型能够理解更复杂的描述，如“the person wearing a red hat and holding a blue umbrella”。这种复杂的描述包含了多个物体、多个属性、以及物体之间的关系，模型通过子句级文本表示和跨模态对齐，能够准确理解并定位目标。

(3) 在未知物体与已知物体混合场景中的准确区分

基于 OVD 的统一框架（OW-OVD、YOLO-UniOW）在未知物体与已知物体混合的场景中表现出色。例如，在一个包含已知类别（如“car”、“person”）和未知类别（如新型工程车辆）的场景中，模型能够：

- 准确识别已知类别，保持高精度
- 主动发现未知物体，将其标记为“Unknown”
- 避免将已知类别误判为未知，或将未知类别误判为已知

这种能力在实际应用中非常重要，因为真实场景往往是已知和未知物体的混合，模型需要同时处理两种类型的物体。

5.5.2 失败案例

尽管开放目标检测技术取得了显著进展，但在某些极端场景下仍然存在局限：

(1) 严重遮挡场景下的检测困难

当目标被大面积遮挡时（如被其他物体遮挡超过 50%），模型的检测性能会显著下降。这是因为：

- **视觉特征不完整：**被遮挡的目标只能提供部分视觉特征，模型难以从这些不完整的特征中推断出完整的物体信息
- **语义信息有限：**虽然文本描述可以提供语义指导，但当视觉特征过于不完整时，语义信息的作用也会受限
- **定位精度下降：**被遮挡的目标的边界框定位精度会显著下降，因为模型无法准确判断被遮挡部分的边界

(2) 极小目标的漏检问题

远景中的小目标（如图像中占据少于 10 像素的目标）容易漏检。这是因为：

- **特征分辨率限制：**即使使用多尺度特征金字塔，极小目标的特征分辨率仍然有限，难以提取足够的特征信息
- **语义信息不足：**极小目标提供的视觉信息非常有限，即使有文本描述，也难以准确匹配
- **检测头设计限制：**传统的检测头设计主要针对中等大小的目标，对极小目标的处理能力有限

(3) 语义歧义导致的混淆

当多个类别描述相似时（如“dog”和“puppy”、“car”和“vehicle”），模型可能产生混淆。这是因为：

- **语义相似性：**语义相近的类别在特征空间中距离较近，模型难以准确区分
- **上下文信息不足：**在某些情况下，仅依靠视觉特征和类别名称难以区分语义相近的类别，需要更多的上下文信息
- **训练数据偏差：**如果训练数据中某些语义相近的类别出现频率差异很大，模型可能会偏向识别出现频率更高的类别

(4) 极端长尾分布的挑战

在极端长尾分布的场景下（如某些类别仅出现数次），模型的检测性能仍然有限。虽然开放词汇检测技术相比传统方法在长尾类别上有显著提升，但对于极端罕见的类别（如某些特定品牌的产品、特定地区的动植物等），模型的性能仍然有待提升。

改进方向

针对这些失败案例，未来的研究可以探索以下改进方向：

- **遮挡处理：**引入专门的遮挡建模机制，如部分可见性建模、遮挡感知的特征提取等
- **小目标检测：**设计专门的小目标检测模块，如高分辨率特征提取、多尺度特征融合等
- **语义消歧：**引入更多的上下文信息，如场景理解、物体关系建模等，来帮助区分语义相近的类别
- **长尾优化：**设计专门的长尾优化策略，如类别平衡的损失函数、基于语义相似度的知识迁移等

6 对比分析与讨论 (Comparison & Discussion)

6.1 方法对比

6.1.1 Grounding DINO vs YOLO-World

对比维度	Grounding DINO	YOLO-World
架构基础	Transformer	CNN
对齐机制	多阶段深度融合	重参数化
推理速度	8-12 FPS	50-74 FPS
精度 (COCO)	52.5 AP	45.7 AP
长尾性能	优秀	良好
部署复杂度	高	低
适用场景	离线分析、高精度标注	实时监控、边缘部署

表 11: Grounding DINO vs YOLO-World 详细对比

6.1.2 OW-OVD vs YOLO-UniOW

对比维度	OW-OVD	YOLO-UniOW
基础框架	Grounding DINO	YOLO-World
未知检测机制	属性选择 + 不确定性融合	Wildcard 嵌入学习
OVD 能力保持	完全保持	可能受影响
增量学习成本	中等	极低
推理速度	10-15 FPS	50+ FPS
U-Recall	22.8	21.5
主要优势	高精度、语义对齐深	高效率、部署友好

表 12: OW-OVD vs YOLO-UniOW 详细对比

6.1.3 综合总结

根据不同的应用需求，可以选择不同的方法：

- 追求极致精度：选择 Grounding DINO / OW-OVD
- 追求实时性能：选择 YOLO-World / YOLO-UniOW
- 需要未知检测：选择 OW-OVD / YOLO-UniOW
- 边缘设备部署：优先选择 YOLO 系列

6.2 技术讨论

6.2.1 从 OVD 到 OWOD 的技术演进

OVD 到 OWOD 的演进体现了开放目标检测领域的核心发展脉络：

- OVD 解决了“识别用户指定的任意类别”问题
- 传统 OWOD 提出了“主动发现未知”的愿景，但技术实现不足
- 基于 OVD 的 OWOD 方法实现了两者的统一，性能大幅提升

6.2.2 解决了哪些问题

- 零样本泛化：OVD 通过视觉-语言对齐实现了对新类别的识别
- 未知物体发现：基于 OVD 的方法将 U-Recall 从 <10% 提升到 20+%
- 实时推理：重参数化技术使得开放检测可以达到 50+ FPS

6.2.3 还存在哪些挑战

- 精度与速度的权衡：目前仍难以同时达到最高精度和最快速度
- 长尾分布处理：罕见类别的检测性能仍有提升空间
- 持续学习：如何在不遗忘旧知识的情况下持续学习新类别

6.3 开放性问题

6.3.1 长尾分布的更好处理

如何更好地处理极端长尾分布的类别，提升罕见类别的检测性能，仍是一个重要的研究方向。可能的解决思路包括：

- 类别平衡的损失函数设计
- 基于语义相似度的知识迁移
- 利用大语言模型生成罕见类别的描述

6.3.2 伪标签质量的进一步提升

虽然基于 OVD 的方法显著提升了伪标签质量，但在极端场景下，伪标签噪声仍然存在。需要更鲁棒的伪标签生成和过滤机制。

6.3.3 计算效率的进一步优化

如何在保持精度的同时进一步提升计算效率，特别是在边缘设备上的部署，需要继续研究模型压缩、知识蒸馏等技术。

6.3.4 多模态融合的深入探索

如何更好地融合视觉、文本、音频等多模态信息，构建更强大的开放感知系统，是未来的重要研究方向。

7 应用前景与展望 (Applications & Future Work)

7.1 应用场景

7.1.1 自动驾驶中的未知物体检测

在自动驾驶场景中，模型需要识别训练时未见过的物体以保证行车安全：

- 新型车辆（电动滑板车、送货机器人）
- 临时路障（施工设备、事故现场）
- 罕见动物（大型野生动物穿越道路）

开放目标检测技术能够帮助自动驾驶系统识别这些未知物体，触发安全制动或绕行策略。

7.1.2 机器人的通用抓取

在机器人抓取任务中，开放目标检测使机器人能够：

- 通过语言指令定位目标物体（“把红色的杯子递给我”）
- 主动发现环境中的新物体并学习抓取
- 适应不断变化的家庭环境

7.1.3 智能安防的异常检测

在智能安防领域：

- 检测可疑包裹、非法侵入的动物等异常物体
- 主动发现监控区域中的新型威胁
- 支持描述性查询（“查找戴帽子的可疑人员”）

7.1.4 医疗影像中的罕见疾病发现

在医疗影像分析中：

- 识别训练集中缺失的罕见疾病特征
- 标记异常区域供医生进一步诊断
- 通过文本描述定位特定的病变类型

7.2 当前挑战

7.2.1 实时性与精度的平衡

在实际应用中，需要在实时性和精度之间找到平衡点：

- 高精度框架精度高但推理速度慢
- 实时化框架速度快但在复杂场景下精度可能不足
- 需要根据具体应用场景选择合适的方案

7.2.2 小样本学习的效率

虽然开放目标检测支持零样本学习，但在某些场景下：

- 少量样本的快速学习仍然需要
- 如何高效利用少量标注样本提升特定类别的性能
- 平衡通用能力与特定任务性能

7.2.3 跨域泛化能力

模型在自然图像上表现良好，但：

- 医疗影像、卫星图像等特定领域数据可能需要额外适配
- 如何提升跨域泛化能力仍需研究
- 领域适应技术与开放检测的结合

7.3 未来研究方向

7.3.1 更强的视觉-语言模型

随着 GPT-4V、Gemini 等更强的视觉-语言模型的出现：

- 如何将这些模型的能力应用到开放目标检测
- 利用大语言模型生成更丰富的类别描述
- 多模态对话驱动的目标检测

7.3.2 持续学习与人机协作

- 更有效的持续学习策略，避免灾难性遗忘
- 人机协作的主动学习，让用户参与标注过程
- 在线学习与离线更新的结合

7.3.3 多任务统一框架

- 开放目标检测与开放词汇分割的统一
- 开放词汇检测与跟踪的统一
- 构建更通用的开放感知系统

7.3.4 可解释性研究

- 提升模型的可解释性
- 让用户理解模型的决策过程
- 特别是未知检测的决策依据

8 总结 (Conclusion)

8.1 主要贡献总结

本文系统性地梳理了开放目标检测技术的发展历程，主要贡献包括：

(1) 系统性梳理

从封闭世界假设的局限性出发，系统性地梳理了开放目标检测技术的发展历程，包括开放词汇检测（OVD）和开放世界检测（OWOD）两个主要方向，以及它们的统一范式。

(2) 关键技术分析

深入分析了关键技术和方法，包括：

- CLIP 的视觉-语言对齐原理
- Transformer 在检测中的应用
- 多阶段深度对齐 (Grounding DINO)
- 重参数化技术 (YOLO-World)
- 通配符学习 (YOLO-UniOW)
- 属性选择与不确定性融合 (OW-OVD)

(3) 实验证与对比

提供了详细的实验证和对比分析，展示了不同方法的性能表现，为实际应用提供了参考。

8.2 本文发现

8.2.1 OVD 与 OWOD 的互补关系

- OVD 提供基础能力：强大的零样本检测和视觉-语言理解
- OWOD 提出目标愿景：主动发现未知并持续学习
- 两者结合能够实现更强大的开放感知系统

8.2.2 实时性与开放性的可兼顾

通过重参数化等技术，可以在保持开放性的同时实现实时推理，证明了实时性与开放性的可兼顾性。YOLO-World 和 YOLO-UniOW 的成功表明，工业级应用与开放能力可以兼得。

8.2.3 大规模预训练的重要性

大规模预训练数据（如 CC3M、LAION、Objects365）对于开放目标检测的性能至关重要，是模型获得强大泛化能力的基础。

8.3 研究局限

8.3.1 实验的局限性

本文的实验主要在标准数据集上进行，在实际应用场景中的表现可能需要进一步验证。特别是在特定领域（如医疗影像）的应用需要额外的适配和验证。

8.3.2 未涉及的相关方向

本文主要关注 2D 目标检测，未涉及的相关方向包括：

- 3D 开放目标检测
- 视频中的开放目标检测与跟踪
- 开放词汇实例分割
- 多模态融合的更深层探索

这些方向是开放目标检测的重要扩展，值得未来深入研究。

参考文献

- [1] Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", TPAMI 2017
- [2] Gupta et al., "LVIS: A Dataset for Large Vocabulary Instance Segmentation", CVPR 2019
- [3] Radford et al., "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021
- [4] Liu et al., "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection", arXiv 2023

- [5] Cheng et al., "YOLO-World: Real-Time Open-Vocabulary Object Detection", CVPR 2024
- [6] Xi et al., "OW-OVD: Unified Open World and Open Vocabulary Object Detection", CVPR 2025
- [7] "YOLO-UniOW: Efficient Universal Open-World Object Detection", 2024
- [8] Joseph et al., "Towards Open World Object Detection", CVPR 2021
- [9] Gupta et al., "OW-DETR: Open-world Detection Transformer", CVPR 2022
- [10] Li et al., "Grounded Language-Image Pre-training", CVPR 2022