# 开放目标检测技术演进：从OVD到统一范式OVD+OWD

xxx

2025 年 12 月 26 日

## 1 引言与背景 (Introduction & Background)

### 1.1 研究背景与动机

#### 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（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。**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。**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：

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

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 $N+1$ 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均衡：**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。

更关键的问题在于，即使投入巨大资源标注了大量类别，模型部署后仍会不断遇到训练时未见过的新物体（如新发布的产品型号、罕见的动植物种类、特定场景下的临时物体等），导致模型性能快速衰减。

### 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。在**自动驾驶**场景中，道路上可能出现工程车辆、临时路障、罕见动物等训练集中不存在的物体，模型必须能够识别这些"未知"目标以保证安全。在**智能安防**领域，监控系统需要检测异常物体（如可疑包裹、非法侵入的动物），这些物体往往无法预先定义。在**医疗影像分析**中，罕见疾病的影像特征可能在训练集中缺失，但临床诊断必须能够识别并标记这些异常区域。

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。

下图直观对比了 (a) 传统封闭集检测与 (b) 开放集检测中人类输入 novel categories 的能力：

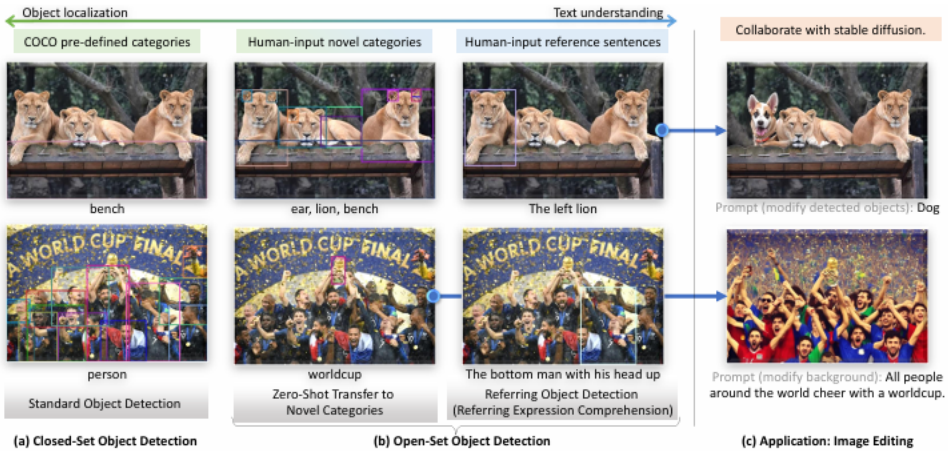


图 1: 封闭集检测与开放集检测的对比

因此，打破 CWA，构建能够适应动态环境、低成本扩展类别的检测系统，成为计算机视觉领域的必然趋势。

## 1.2 范式转移：视觉-语言融合的突破

### 1.2.1 从离散 ID 到语义嵌入

打破封闭世界假设的关键在于**改变类别表示的方式**。传统检测器将类别表示为离散的数字 ID，这种表示缺乏语义内涵，无法泛化到新类别。而视觉-语言模型（VLM）的出现为这一问题提供了革命性的解决方案。

| 对比维度 | 传统分类器 | 开放检测器 |
|---|---|---|
| 类别表示 | 离散 ID（如 0, 1, ..., 79） | 语义嵌入向量（如 512 维） |
| 分类方式 | 学习特征到 ID 的映射 | 计算特征与文本嵌入的相似度 |
| 新类别泛化 | 需要重新训练 | 输入新类别名称即可检测 |
| 数学表达 | $f : \mathbb{R}^d \to \{0, 1, ..., N-1\}$ | $sim(f_{img}(x), f_{text}(t)) \in [0, 1]$ |

表 1: 传统分类器与开放检测器的对比

**传统分类器 vs. 开放检测器的本质区别：**

开放检测器的核心思想是：**将目标检测从"特征到 ID 的映射"转化为"区域-文本匹配"问题**。给定图像区域特征 $v \in \mathbb{R}^d$ 和类别名称的文本嵌入 $t \in \mathbb{R}^d$，通过计算它们在统一语义空间中的相似度来判定类别：

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

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（Zero-shot）情况下识别该类别，无需重新训练。

### 1.2.2 CLIP：视觉-语言对齐的基石

这一范式转移的理论基础源于 OpenAI 的 CLIP（Contrastive Language-Image Pre-training）模型[3]。CLIP 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。

CLIP 的核心机制包括三个方面。**对比学习目标：**最大化匹配图文对的相似度，最小化不匹配对的相似度。给定一个 batch 包含 $N$ 个图文对 $\{(I_i, T_i)\}_{i=1}^N$，InfoNCE 损失为：

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

其中 $\tau$ 是温度参数，$sim(\cdot, \cdot)$ 是余弦相似度。**零样本迁移能力：**通过将类别名称转换为 "a photo of a {class}" 的文本提示，CLIP 能够在未见过的类别上进行分类。这一能力的关键在于：模型在预训练阶段学习到的是**物体的视觉-语义关联**，而非特定的类别 ID。**语义空间的连续性：**在 CLIP 的特征空间中，语义相近的概念（如 "dog" 和 "puppy"）在向量空间中距离较近，这为细粒度识别和泛化提供了基础。

然而，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挑战：CLIP 的预训练是在完整图像与文本之间进行的，缺乏对**局部区域**的精细理解；目标检测需要同时完成**定位**任务，而 CLIP 仅提供分类能力；区域特征与图像级特征的分布存在差异（Region-Image Distribution Gap）。

因此，开放目标检测的核心技术难题是：**如何将 CLIP 的图像级语义对齐能力迁移到区域级检测任务**。

### 1.2.3 区域-文本对齐的技术路径

为了解决上述问题，研究者提出了两条主要技术路径。**知识蒸馏路径（ViLD, RegionCLIP 等）：**利用预训练的 CLIP 作为教师模型，通过蒸馏将其知识迁移到检测器的区域嵌入中。然而，该方法受限于训练数据的类别覆盖，泛化能力有限，且蒸馏过程可能导致语义信息损失。**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，即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 $\{(r_i, t_i)\}$，其中 $r_i$ 是图像区域，$t_i$ 是对应的文本描述（类别名或短语），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。

这种接地预训练的优势在于：直接在区域级进行训练，避免了图像级到区域级的迁移 gap；可以利用更丰富的文本信息（如属性、关系），而不仅仅是类别名；通过大规模数据（如数百万图文对）显著提升泛化能力。

## 1.3 核心概念界定与研究范围

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。根据对"开放性"理解的不同，主要包括以下研究方向：

| 研究方向 | 核心问题 | 对"开放"的定义 | 实用性 | 代表工作 |
|---|---|---|---|---|
| 开放词汇目标检测 (OVD) | 识别任意文本描述的类别 | 类别空间可通过语言无限扩展 | 成熟，零样本能力强 | GLIP, Grounding DINO, YOLO-World |
| 传统开放世界目标检测 (OWOD) | 主动发现未知并持续学习 | 能识别"不知道"并动态扩展知识 | 效果受限，召回率低，噪声大 | ORE, OW-DETR, PROB |
| 基于OVD的开放世界检测（OVD+OWOD） | 结合零样本与主动发现 | 利用OVD能力实现OWOD目标 | 新兴方向，性能显著提升 | OW-OVD, YOLO-UniOW |
| 开放集识别 (OSR) | 拒绝识别不可信样本 | 区分已知和未知，但不学习未知 | 成熟但应用受限 | OpenMax, ARPL |
| 零样本检测 (ZSD) | 基于属性推理未见类别 | 通过语义属性迁移到新类 | 依赖属性标注，泛化性有限 | DELO, SB |
| 少样本检测 (Few-shot) | 用少量样本学习新类 | 快速适应，但需要标注 | 实用但需标注成本 | Meta R-CNN, FSCE |
| 长尾检测 (Long-tail) | 处理类别分布不平衡 | 提升稀有类别性能 | 成熟，针对特定场景 | LVIS-based方法 |

表 2: 开放目标检测相关研究方向对比

### 1.3.1 开放词汇目标检测 (OVD)

给定图像 $I$ 和文本词汇表 $\mathcal{V} = \{t_1, t_2, ..., t_C\}$（$t_i$ 可以是类别名、短语或描述），开放词汇目标检测器 $\mathcal{D}_{OVD}$ 输出检测结果：

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

其中 $b_i \in \mathbb{R}^4$ 是边界框坐标，$c_i \in \mathcal{V}$ 是匹配的文本类别，$s_i \in [0, 1]$ 是置信度分数。

OVD 的核心特征包括：**动态词汇表**，即 $\mathcal{V}$ 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（如 "a blue surfboard"）来指定检测目标。

OVD 的评估范式主要包括：**Zero-shot Transfer**，在 Base 类别上训练，在 Novel 类别上测试（如 COCO 的 48/17 划分，48 个 Base 类 + 17 个 Novel 类）；**Open Vocabulary Evalua-**

**tion**，在大规模数据预训练后，直接在标准数据集（如 LVIS 1203 类）上零样本评估；**Referring Expression Comprehension (REC)**，给定描述性文本（如 "the person wearing a red hat"），定位对应目标。

**典型数据集：**

| 数据集 | 类别数 | 特点 | 用途 |
|---|---|---|---|
| COCO | 80 | 常见物体，均衡分布 | Base/Novel 划分评估 |
| LVIS | 1203 | 长尾分布（rare/common/frequent） | 大词汇量零样本评估 |
| ODinW | 35个数据集 | 真实场景多样性 | 跨域泛化评估 |
| RefCOCO/+/g | - | 包含属性和关系的描述 | 细粒度理解评估 |

表 3: OVD 典型数据集

**局限性分析：**

OVD 的本质是"**被动式开放**"，假设所有感兴趣的物体都能被提前命名和描述，模型依赖用户提供准确的类别名称或描述，无法主动发现用户未预期的新物体。

例如，如果用户不知道画面中出现了一只"狐猴"（Lemur），即使模型具备识别能力，也无法将其检测出来，因为用户没有在词汇表中包含该类别。

### 1.3.2 传统开放世界目标检测 (OWOD)

传统开放世界目标检测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 $\mathcal{T} = \{T_1, T_2, ..., T_M\}$：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 $\mathcal{K}^{(i)}_{known} = \mathcal{K}_{T_1} \cup ... \cup \mathcal{K}_{T_i}$；推理时，模型需要检测已知类别 $\mathcal{K}^{(i)}_{known}$，**主动识别未知物体**并标记为"Unknown"，在后续任务中，部分 Unknown 类别被标注后加入 $\mathcal{K}_{T_{i+1}}$。

OWOD 的核心特征包括：**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 $\mathcal{K}_{new}$ 时，保持对旧类别 $\mathcal{K}_{old}$ 的识别能力（避免灾难性遗忘）；**动态类别扩展**，类别集合随时间不断增长：$\mathcal{K}^{(1)} \subset \mathcal{K}^{(2)} \subset ... \subset \mathcal{K}^{(M)}$。

除了标准的 mAP，OWOD 引入专门的未知类别评估指标：**U-Recall**（未知类别的召回率，衡量模型发现未知物体的能力）、**U-mAP**（未知类别的平均精度，更严格地评估未知检测质量）、**Wilderness Impact (WI)**（未知物体对已知类别检测的干扰程度，理想值接近 0）、**Absolute Open-Set Error (A-OSE)**（综合评估未知检测和已知分类的平衡性）。

传统 OWOD 方法面临以下核心困难：**未知类别定义困难**，即什么是"Unknown"？模型如何在没有标签的情况下区分已知和未知？**伪标签噪声严重**，用模型自己的预测作为 Unknown 的监督信号，容易将背景、已知类别误判为未知，导致错误累积。**未知召回率低**，ORE 在 Task 1 仅获得 4.92 的 U-Recall，OW-DETR 也仅有约 7-9 的 U-Recall，远无法满足实际应用需求。**灾难性遗忘**，学习新类别时，旧类别的性能往往显著下降。

典型方法演进包括：**ORE (2021)** 首次提出 OWOD 任务设定，使用能量模型识别未知，但效果有限；**OW-DETR (2022)** 基于 DETR 架构，引入注意力驱动的伪标签生成，仍受制于噪声问题；**PROB (2023)** 利用概率建模区分已知和未知，性能略有提升但未根本解决问题。

**传统OWOD**由于缺乏对未知类别的有效建模方法,这些方法本质上是在"盲目猜测"什么是未知物体,导致性能瓶颈难以突破。

### 1.3.3 基于OVD的开放世界检测:OVD+OWOD新范式

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,形成了一个新的研究范式。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**。

下图展示了基于 OVD 实现开放世界检测的统一框架：

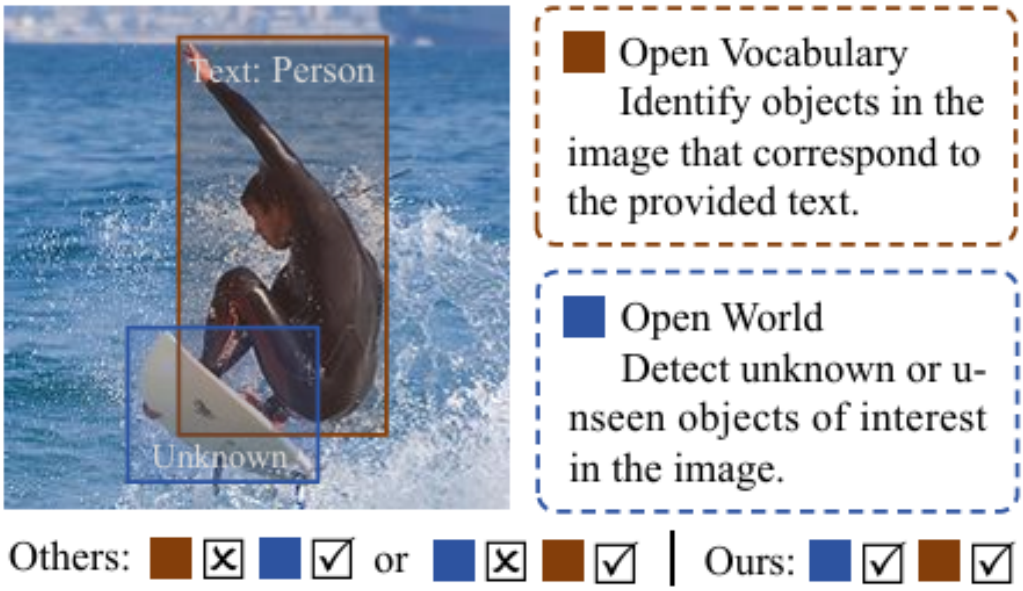


图 2: 基于 OVD 的开放世界检测统一框架

基于 OVD 的开放世界检测关键技术路线包括三个方面。**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 GLIP、Grounding DINO 等 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 (通配符学习)**：YOLO-UniOW 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**：OW-OVD 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。

**性能提升:**

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 (已知类别) | 核心优势 |
|---|---|---|---|
| ORE (传统) | 4.9 | ~50 | 首次提出任务,效果受限 |
| OW-DETR (传统) | 7-9 | ~52 | DETR架构,注意力驱动 |
| 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 4: 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。

### 1.3.4 OVD与开放世界检测的对比与联系

下表总结了三类方法的核心差异:

| 对比维度 | OVD | 传统OWOD | 基于OVD的开放世界检测 |
|---|---|---|---|
| 核心目标 | 识别用户指定的任意类别 | 主动发现未知物体并持续学习 | **统一**：零样本识别+主动发现 |
| 输入需求 | 需要文本提示（类别名/描述） | 无需文本输入 | 训练时无需，推理时可选 |
| 对未知的态度 | **被动识别**：仅识别词汇表中的类别 | **主动感知**：自动标记未训练类别 | **双重能力**：既能被动响应，又能主动发现 |
| 学习范式 | 零样本迁移（预训练 + 冻结） | 增量学习（持续更新） | 预训练OVD + 增量适配 |
| 评估重点 | 新类别的检测精度 | 未知发现 + 持续学习 + 防遗忘 | 两者兼顾 |
| 应用场景 | 用户知道要找什么（如搜索特定物品） | 模型需自主感知环境变化（如安防异常检测） | **通用场景**：同时支持指定检测和主动发现 |

表 5: OVD、传统OWOD与基于OVD的开放世界检测对比

三类方法的核心关系如下：OVD 提供了基础能力，即强大的零样本检测和视觉-语言理解；传统 OWOD 提出了目标愿景，即主动发现未知并持续学习，但技术实现不足；基于 OVD 的开放世界检测实现了能力融合，利用 OVD 的成熟技术来实现 OWOD 的目标，性能远超传统方法。

## 1.4 总结：技术演进路线图

为便于理解，我们将该领域的演进总结为以下4个阶段：

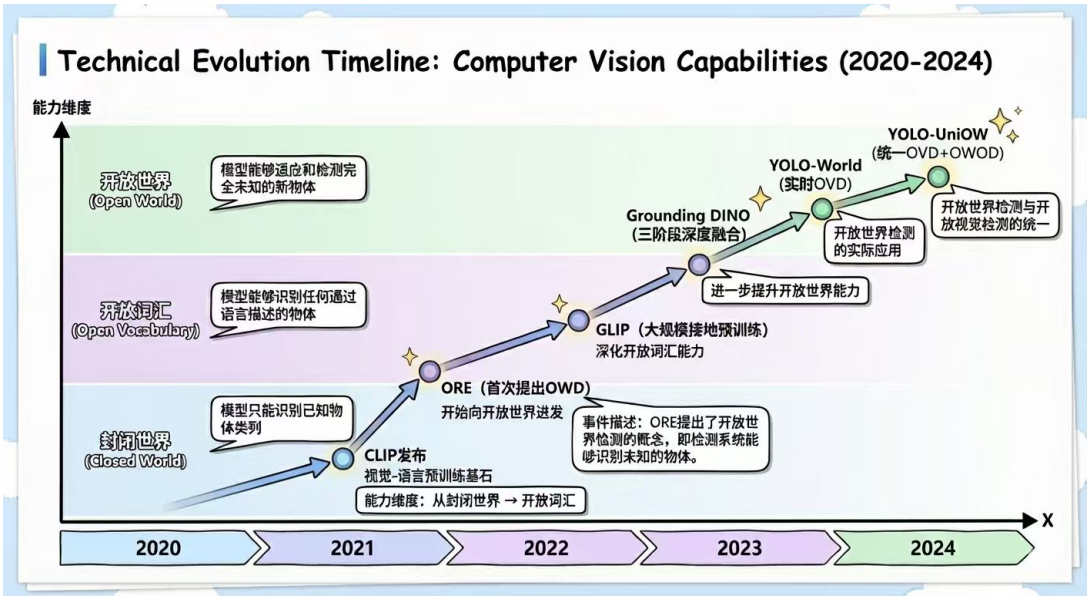


图 3: 开放目标检测技术演进路线图

| 发展阶段 | 核心范式 | 类别表示 | 对未知的态度 | 代表模型 | 关键突破 |
|---|---|---|---|---|---|
| **封闭世界** (2020) | 特征到ID映射 | 离散数字ID | **排斥/忽略**：未知物体被误分类或当作背景 | Faster R-CNN, YOLOv5 | 精度与速度的极致优化 |
| **开放词汇突破** (2021-2023) | 区域-文本匹配 | 文本嵌入向量 | **被动识别**：仅识别用户指定的类别名称 | GLIP, Grounding DINO, YOLO-World | CLIP + 大规模接地预训练 |
| **传统开放世界** (2021-2023) | 主动发现 + 增量学习 | 物体性 + 动态扩展 | **主动感知**：自动标记未知但效果有限 | ORE, OW-DETR, PROB | 任务范式创新 |
| **统一开放世界** (2024-) | **OVD + OWOD 融合** | 文本嵌入 + Wildcard | **双重能力**：零样本识别 + 主动发现 | OW-OVD, YOLO-UniOW | 基于OVD实现OWOD |

表 6: 开放目标检测技术演进阶段总结

本文聚焦于**开放词汇目标检测(OVD)**及其在**开放世界场景下的扩展应用(OVD+OWOD)**。首先，从**技术成熟度**来看，OVD 通过 CLIP 等大规模视觉-语言预训练取得了突破性进展，在零样本检测任务上已展现出强大的泛化能力和实用价值。其次，**传统 OWOD 的局限**在于，虽然传统 OWOD（如 ORE、OW-DETR）首次提出了主动发现未知物体的概念，但由于缺乏有效的未知物体建模方法，实际效果不佳——未知召回率普遍低于 10%，且伪标签噪声严重。第三，**（OVD+OWOD）新范式的出现**：近期研究（OW-OVD、YOLO-UniOW 等）发现，可以基于成熟的 OVD 检测器，通过引入"未知"类别的特殊建模（如 Wildcard Learning、属性选择等），实现 OWOD 的目标，性能远超传统方法。最后，**统一框架的价值**在于，理想的开放感知系统应同时具备**被动响应**（用户指定类别）和**主动发现**（自动标记未知）的能力，这正是基于 OVD 实现统一开放世界检测的核心价值。

因此,本文将深入探讨OVD的核心技术,以及如何在OVD基础上实现开放世界能力，即OVD+OWOD,而对传统OWOD方法仅作简要介绍。

# 2 OVD 核心技术范式——从深度融合到实时推理

在目标检测领域，从"封闭世界"向"开放世界"转化的核心技术枢纽在于开放词汇目标检测（OVD）。本部分将深入探讨 OVD 的两种主流技术范式：一种是追求极致语义对齐与定位精度的"高精度流派"，基于 Transformer 架构实现多阶段深度对齐；另一种是致力于工业级端到端部署的"实时化流派"，通过重参数化技术实现高效推理。这两者共同构成了当前开放感知领域的技术基石。

## 2.1 高精度 OVD 框架：基于 Transformer 的多阶段深度对齐

在开放词汇检测（OVD）任务中，核心挑战在于如何将离散的文本标签与连续的视觉空间表示进行对齐。高精度 OVD 框架提出了一种范式转变的思路：它不再将视觉和文本视为独立模态，而是基于"Grounded Pre-training"思想，摒弃了传统检测器仅在逻辑层进行类别映射的局限。该框架通过在 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，实现了视觉与语言的深度耦合。

下图展示了该框架的整体架构及其三阶段对齐流： 注：该架构清晰呈现了文本与图像特征从早期特征增强，到语言引导的查询选择，再到最终跨模态解码的闭环过程。

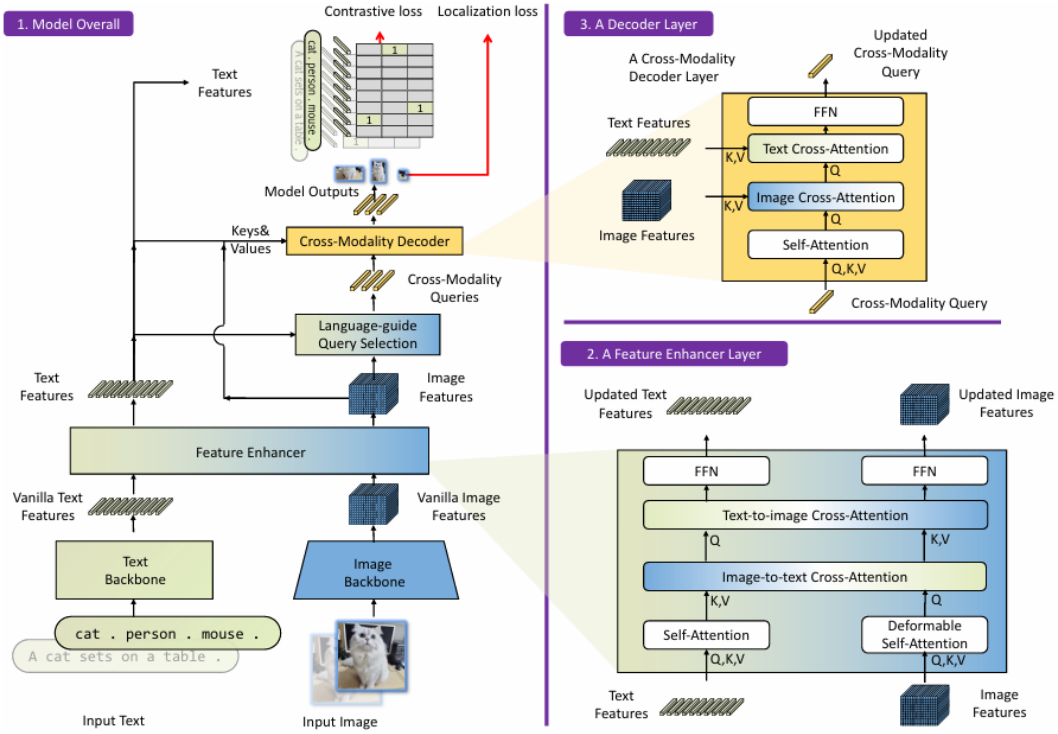


图 4: 基于 Transformer 的高精度 OVD 框架整体架构与三阶段对齐流

### 2.1.1 跨模态特征增强器

传统开放集检测器通常采用"双塔架构"，仅在最后的分类头进行浅层的点积对齐，这导致视觉特征缺乏语义指导。高精度 OVD 框架引入了**跨模态特征增强器（Cross-Modal Feature Enhancer）**，在特征提取的早期阶段实现视觉与语言的深度耦合。

**架构设计：**该增强器由堆叠的 Transformer 层构成，采用**Deformable Self-Attention**机制高效处理多尺度视觉特征。与传统双塔架构不同，增强器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（$H, W$ 为特征图尺寸，$L$ 为文本序列长度，$d$ 为特征维度）。这一过程使得视觉特征能够根据文本语义动态调整空间权值分布，聚焦于与文本描述相关的区域。

**（2）文本到视觉的注入：**文本特征 $T$ 作为 Key 和 Value，通过交叉注意力机制注入到视觉空间。这种注入不是简单的特征拼接，而是利用全局上下文消除语义歧义，使得视觉特征在进入后续模块之前就已经具备语义敏感性。

双向跨模态注意力的数学表达为：

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

其中：

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

**优势分析：**这种**紧密融合**策略确保了视觉特征在进入预测头之前，就已经过语义过滤和增强。相比传统方法仅在最后进行浅层对齐，早期融合能够：

- **提升长尾类别识别能力：**通过文本指导，模型能够更好地理解罕见类别的视觉特征
- **增强复杂指代性表达的理解：**对于"穿着红色雨衣的小狗"这类复合描述，早期融合使得模型能够同时考虑多个语义属性

- **减少语义歧义**：文本信息的早期注入有助于消除视觉特征中的歧义，提升检测精度

### 2.1.2 语言引导的查询选择

为了在无限的语义空间中精确定位，高精度 OVD 框架改进了标准 DETR 的查询机制，利用语言先验来初始化对象查询。这是该框架的核心创新之一。

**传统查询机制的局限**：标准 DETR 使用可学习的查询向量（Learnable Queries），这些查询在训练过程中学习通用的物体模式，但缺乏对特定文本描述的针对性。在开放词汇场景下，当用户输入新的类别描述时，通用查询可能无法有效定位目标。

**对比驱动的查询初始化**：高精度框架采用**对比驱动的查询初始化（Contrastive-Driven Query Initialization）**策略。具体流程如下：

**步骤1：相似度计算**。利用增强器输出的图像特征 $f_v \in \mathbb{R}^{H\times W\times d}$ 与文本特征 $f_t \in \mathbb{R}^{L\times d}$ 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：区域响应聚合**。对相似度矩阵进行空间聚合，得到每个空间位置的文本响应强度：

$$R = \text{MaxPool}(\text{Softmax}(S, \text{dim} = -1)) \in \mathbb{R}^{H\times W}$$

**步骤3：Top-K 区域选择**。根据响应强度筛选出 Top-K 个高响应区域的索引：

$$\text{Indices} = \text{Top-K}(R, k = K)$$

**步骤4：动态锚框初始化**。这些索引被用于初始化**动态锚框（Dynamic Anchor Boxes）**的位置坐标。具体而言，对于选中的区域索引 $(h_i, w_i)$，对应的锚框中心位置初始化为：

$$\text{Anchor}_i = \left(\frac{w_i}{W}, \frac{h_i}{H}\right)$$

而内容查询（Content Queries）则保持可学习状态，在训练过程中进一步优化。

**优势分析**：这种设计使得解码器从一开始就具备了”语义先验”，能够：

- **直接跳过无关区域**：查询初始化阶段就排除了与文本描述无关的背景区域
- **聚焦高相关候选目标**：解码器只需在少量高响应区域进行精细化定位，大幅提升效率
- **增强零样本泛化能力**：对于训练时未见过的类别描述，模型仍能通过语义相似度找到相关区域

**实例说明**：假设用户输入的文本提示是”穿着红色雨衣的小狗”。在这一阶段，模型会：

1. 计算图像特征与”红色”、”雨衣”、”狗”等词汇的语义相似度
2. 识别出同时满足多个语义属性的区域（如既包含”红色”又包含”雨衣”特征，且具有”狗”的形状）
3. 在这些高响应区域初始化查询向量，使得后续解码过程能够直接聚焦于目标物体

查询向量（Queries）从诞生的那一刻起，就携带了目标的几何位置先验和语义先验，这是该框架在零样本检测任务中表现出极强鲁棒性的关键原因。

### 2.1.3 跨模态解码器与子句级去噪

在解码阶段，高精度 OVD 框架引入了定制化的**文本交叉注意力（Text Cross-Attention）**层，并配合子句级处理机制，进一步细化对齐粒度。

**文本交叉注意力机制：**与标准 DETR 解码器不同，该框架的每一层解码器都包含专门的 Text Cross-Attention 模块。给定查询特征 $Q \in \mathbb{R}^{K\times d}$（$K$ 为查询数量）和文本特征 $T \in \mathbb{R}^{L\times d}$，文本交叉注意力计算为：

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

这一机制使得每个查询向量能够动态地从文本描述中提取相关信息，实现查询级别的语义对齐。

**子句级掩码机制：**针对长文本描述，模型构建了特殊的**子句级注意力掩码（Phrase-Level Attention Mask）**。该机制的核心思想是：将长文本分解为多个名词短语（Noun Phrases），每个视觉区域仅与对应的短语交互，屏蔽句中无关词汇。

具体实现中，对于文本描述 "A man holding a blue umbrella"，模型首先进行短语分割：

- 短语1："man"
- 短语2："holding"
- 短语3："blue umbrella"

然后构建掩码矩阵 $M \in \{0,-\infty\}^{K\times L}$，其中 $M_{i,j} = 0$ 表示查询 $i$ 可以与文本位置 $j$ 交互，$M_{i,j} = -\infty$ 表示禁止交互。修改后的注意力计算为：

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

**优势分析：**子句级掩码机制带来了以下优势：

- **防止语义干扰：**不会因为出现了"blue"就让模型去寻找蓝色的衣服，而是严格将其限制在"umbrella"的属性上
- **确保细粒度属性的精确归属：**每个属性（如颜色、形状）都与其所属的物体正确关联
- **降低假阳性率：**通过精确的语义匹配，显著减少了 OVD 场景中的误检

**循环交互解码：**解码器通过多层迭代，不断利用文本特征作为辅助信息来更新边界框坐标。与传统方法不同，该框架中文本特征不仅是分类的依据，更作为空间偏移的偏置项直接参与了边界框（BBox）的回归计算。

具体而言，在第 $l$ 层解码器中，边界框回归不仅依赖于视觉特征，还融合了文本语义信息：

$$\Delta bbox^{(l)} = \text{MLP}([f^{(l)}_{visual}, f^{(l)}_{text}])$$

其中 $f^{(l)}_{text}$ 是经过 Text Cross-Attention 后的文本特征，$[\cdot,\cdot]$ 表示特征拼接。这种设计使得模型能够：

- **引导遮挡目标的精确重建：**当目标被部分遮挡时，文本描述提供了额外的语义约束
- **处理边缘模糊的目标：**文本信息帮助模型在边界不清晰的情况下做出更准确的定位
- **实现像素级坐标微调：**通过多层迭代，模型能够对遮挡或尺度极小的目标（如远景中的车辆）进行精细化定位

通过这种**视觉-文本协同解码**机制，高精度 OVD 框架在复杂场景下仍能保持极高的检测精度。

#### 2.1.4 分类与定位的对齐损失

针对开放词汇检测的特殊性，高精度 OVD 框架重新设计了损失函数，以更好地适应动态类别空间和语义对齐需求。

**开放集分类损失**：传统检测器使用固定类别的交叉熵损失，这在开放词汇场景下不再适用。该框架采用**基于相似度的分类损失**，通过计算查询特征与文本嵌入的点积相似度来判定类别。

对于每个查询 $q_i$ 和文本词汇表 $\mathcal{V} = \{t_1, t_2, ..., t_C\}$，分类分数计算为：

$$s_{i,j} = \frac{q_i \cdot t_j}{\|q_i\| \|t_j\|} = \cos(q_i, t_j)$$

然后采用 **Focal Loss** 来处理正负样本不平衡问题：

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \alpha_j (1 - p_{i,j})^{\gamma} \log(p_{i,j}) \cdot y_{i,j}$$

其中 $p_{i,j} = \text{Softmax}(s_{i,j})$ 是归一化的相似度分数，$y_{i,j}$ 是真实标签，$\alpha_j$ 是类别权重，$\gamma$ 是聚焦参数。

**损失设计的优势**：

- **适应动态类别空间**：无需预先定义固定类别数，可以处理任意大小的词汇表
- **语义相似度驱动**：通过余弦相似度，模型能够识别语义相近的类别（如"dog"和"puppy"）
- **解决样本不平衡**：Focal Loss 能够自动关注难样本，提升训练效果
- **零样本泛化**：在 Zero-shot 迁移时，模型能够根据语义相似度进行准确判别

**定位损失**：在定位分支，框架保留了标准的 **L1 Loss** 和 **GIOU Loss** 来确保边界框回归的几何精度：

$$\mathcal{L}_{bbox} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{GIOU}$$

其中 L1 Loss 确保坐标的精确性：

$$\mathcal{L}_{L1} = \sum_{i=1}^{N} |bbox_i^{pred} - bbox_i^{gt}|$$

GIOU Loss 考虑边界框的重叠和尺度：

$$\mathcal{L}_{GIOU} = 1 - \text{GIOU}(bbox^{pred}, bbox^{gt})$$

**总损失函数**：最终的训练损失为分类损失和定位损失的加权和：

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

通过这种**语义对齐与几何精度的联合优化**，高精度 OVD 框架能够在保持高定位精度的同时，实现强大的零样本泛化能力。

### 2.2 实时化 OVD 框架：重参数化带来的感知革命

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（FLOPs）和推理延迟成为了瓶颈。实时化 OVD 框架的出现填补了这一空白，它

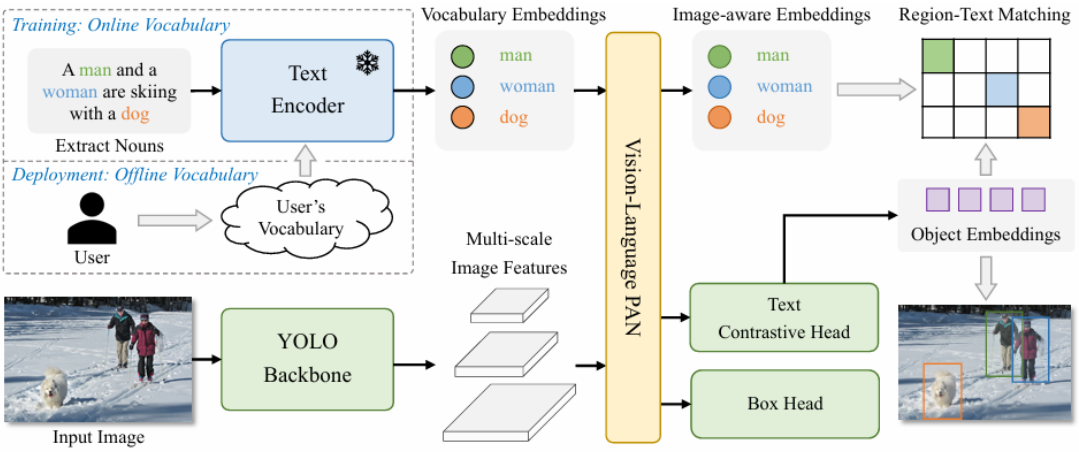


图 5: 实时化 OVD 框架整体架构

实现了"开集检测"与"实时推理"的完美统一。该框架在 V100 上能以 50+ FPS 的速度实现高精度的零样本检测，证明了轻量级检测器通过架构创新，同样可以驾驭复杂的开放词汇场景。

下图展示了实时化 OVD 框架的整体架构，清晰呈现了其如何将文本编码器、YOLO 骨干网以及特征融合模块有机结合： 注：如图所示，该框架包含图像编码器、文本编码器以及核心的 *RepVL-PAN* 模块，最后通过对比头输出目标位置与类别嵌入。

下图进一步对比了不同检测框架的设计理念：

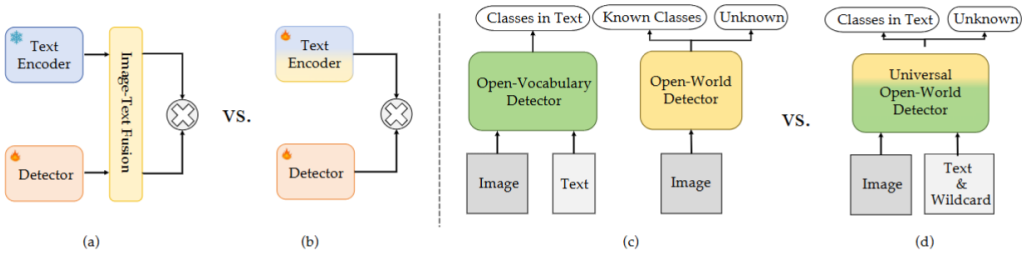


图 6: 不同 OVD 检测框架的对比

### 2.2.1 RepVL-PAN 与视觉-语义交互

实时化 OVD 框架并没有简单地将文本特征拼接到 YOLO 的检测头上，而是设计了一种全新的**重参数化视觉-语言路径聚合网络（Reparameterized Vision-Language Path Aggregation Network, RepVL-PAN）**。这是该框架的核心创新，实现了高效的多模态特征融合。

**架构设计理念：**与 Transformer 架构的全局注意力机制不同，RepVL-PAN 采用**局部卷积 + 全局池化**的混合设计，在保持 YOLO 实时性的同时注入语义敏感度。

**文本引导的 CSPLayer：**在训练阶段，RepVL-PAN 引入了文本引导的 CSPLayer（Cross Stage Partial Layer）。该模块的核心思想是：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。

具体而言，给定视觉特征 $F_v \in \mathbb{R}^{H\times W\times C}$ 和文本特征 $F_t \in \mathbb{R}^{L\times d}$（经过文本编码器编码后的类别嵌入），文本引导的权重计算为：

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

然后通过通道级乘法调整视觉特征：

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

其中 $\odot$ 表示逐元素乘法，Broadcast 将权重广播到空间维度。

**优势分析**：这种设计的优势在于：

- **计算效率高**：相比 Transformer 的 $O(N^2)$ 复杂度，卷积操作仅为 $O(N)$，大幅降低计算成本
- **保留空间结构**：卷积操作保留了视觉特征的空间结构信息，有利于精确定位
- **动态语义过滤**：文本信息动态调整不同通道的权重，使得模型能够聚焦于与文本描述相关的视觉特征

**图像池化注意力机制**：为了增强全局语义理解，RepVL-PAN 在颈部网络中嵌入了**图像池化注意力（Image Pooling Attention）**模块。该模块首先对视觉特征进行全局池化：

$$F_{global} = \text{GlobalAvgPool}(F_v) \in \mathbb{R}^C$$

然后计算空间注意力权重：

$$A = \text{Sigmoid}(\text{Conv}([F_v, \text{Broadcast}(F_{global})])) \in \mathbb{R}^{H \times W}$$

最后应用注意力权重：

$$F''_v = F_v \odot A$$

**设计意义**：图像池化注意力机制使得 YOLO 这种局部感知见长的模型也能获得大尺度语义视野，这对于理解复杂场景和长距离依赖关系至关重要。同时，该机制的计算开销极小，不会影响模型的实时性能。

**多尺度特征融合**：RepVL-PAN 继承了 YOLO 的多尺度特征金字塔设计，在 P3、P4、P5 三个尺度上进行特征融合。每个尺度都应用了文本引导的 CSPLayer 和图像池化注意力，确保不同尺度的特征都具备语义敏感性。

#### 2.2.2 "先提示，后检测"：推理效率的质变

与高精度框架的实时交互不同，实时化 OVD 框架采用了**"先提示后检测（Prompt-then-detect）"**的范式。这是该框架实现高效推理的关键创新。

**重参数化技术原理**：在训练阶段，模型学习视觉特征与文本嵌入的对齐关系。在推理时，如果用户预设了固定的类别词汇表（Offline Vocabulary），可以将文本编码器移除，将文本嵌入重新参数化为网络权重。

**具体实现流程**：

**步骤1：离线文本编码**。用户输入的类别名称集合 $\mathcal{V} = \{c_1, c_2, ..., c_N\}$ 通过文本编码器（如 CLIP Text Encoder）转化为文本嵌入矩阵：

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

其中 $E(\cdot)$ 是文本编码函数，$N$ 是类别数，$d$ 是嵌入维度。

**步骤2：权重融合**。将文本嵌入矩阵直接作为检测头中 $1 \times 1$ 卷积层的卷积核参数。具体而言，检测头的分类分支可以表示为：

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

其中 $F_v \in \mathbb{R}^{H\times W\times d}$ 是视觉特征，Score $\in \mathbb{R}^{H\times W\times N}$ 是分类分数。

**步骤3：纯视觉推理**。经过重参数化后，模型在推理时无需运行文本编码器，也无需进行在线的跨模态注意力计算。从计算链路上看，它退化为了一个纯视觉的 YOLO 模型。

**效率提升分析**：重参数化技术带来的效率提升主要体现在：

| 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） |
| 检测头 | 标准计算 | 标准计算 |
| 总推理时间 | ∼25-30ms | ∼15-20ms |
| 推理速度 | ∼30-40 FPS | ∼50-70 FPS |

表 7: 重参数化前后的推理效率对比

**优势总结**：

- **大幅降低推理延迟**：移除文本编码和跨模态计算，推理速度提升 50-100%
- **降低内存占用**：无需存储文本编码器的中间结果，内存占用减少约 30%
- **简化部署流程**：推理时只需加载视觉模型，部署更加简单
- **支持动态词汇表更新**：当需要添加新类别时，只需重新计算文本嵌入并更新卷积权重，无需重新训练模型

**应用场景**：这种"离线编码、在线匹配"的逻辑，彻底解决了 OVD 落地难的痛点。特别适合以下场景：

- **固定类别检测**：如工业质检中的特定缺陷类型检测
- **实时监控系统**：需要高帧率处理的视频监控场景
- **边缘设备部署**：计算资源受限的嵌入式设备
- **大规模批量处理**：需要处理大量图像的离线任务

通过重参数化技术，实时化 OVD 框架在保持强大零样本能力的同时，实现了工业级的推理效率，为开放词汇检测的广泛应用奠定了基础。

#### 2.2.3 区域-文本对比学习策略

为了在大规模无标注数据中学习泛化性，实时化 OVD 框架引入了**区域-文本对比学习（Region-Text Contrastive Learning）**策略。这是该框架实现强大零样本能力的关键训练机制。

**对比学习目标**：给定一个 batch 包含 $B$ 个图像-文本对 $\{(I_i, T_i)\}_{i=1}^{B}$，对于每个图像 $I_i$，模型提取 $M$ 个候选区域特征 $\{r_{i,j}\}_{j=1}^{M}$。对比学习的目标是：最大化匹配的（区域，文本）对的相似度，最小化不匹配对的相似度。

**损失函数设计**：采用 InfoNCE 损失函数：

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

其中 $sim(\cdot,\cdot)$ 是余弦相似度，$\tau$ 是温度参数（通常设为 0.07）。

**大规模预训练数据：**框架在 CC3M（Conceptual Captions 3M）等大规模图文对数据集上进行预训练。这些数据集包含数百万个图像-文本对，覆盖了丰富的物体类别和场景。

**训练策略：**

- **多尺度区域采样：**从不同尺度的特征图中采样候选区域，确保覆盖不同大小的物体
- **困难负样本挖掘：**选择与正样本相似度较高的负样本进行对比，提升模型的判别能力
- **温度参数调节：**通过调整温度参数 $\tau$，控制相似度分布的尖锐程度，影响模型的泛化能力

**优势分析：**

- **利用大规模无标注数据：**无需精细标注，仅需图像-文本对即可训练
- **学习通用语义表示：**通过对比学习，模型学到了物体与文本描述之间的通用对应关系
- **强大的零样本泛化：**即使在没有经过精细标注的类别（如某些特定品牌的商品、罕见动物等）上，模型也能凭借大规模预训练带来的常识进行准确预测
- **跨域适应能力：**通过大规模数据的预训练，模型具备了跨域泛化的能力

**与高精度框架的对比：**实时化框架的对比学习策略与高精度框架的主要区别在于：

- **数据规模：**实时化框架更依赖大规模预训练数据来学习泛化能力
- **对齐粒度：**高精度框架通过多阶段对齐实现细粒度匹配，实时化框架通过对比学习实现粗粒度但高效的对齐
- **计算效率：**实时化框架的对比学习主要在训练阶段进行，推理时通过重参数化移除，不影响效率

通过区域-文本对比学习，实时化 OVD 框架在保持高效推理的同时，实现了强大的零样本检测能力，为开放词汇检测的广泛应用提供了技术基础。

## 2.3 两种技术范式的对比分析

总结来看，基于 Transformer 的高精度框架代表了开放目标检测的”精度上限”，它通过复杂的多阶段对齐实现了对复杂指令的精准解析，是离线分析和高质量标注任务的首选。而基于重参数化的实时化框架则代表了”效率广度”，它利用重参数化技术将开放能力平民化，让实时嵌入式设备具备了识别万物的可能。

**从架构设计角度看：**

**（1）视觉特征提取的”保真度”：**高精度框架的深度融合模型虽然强大，但视觉特征被语言高度”污染”了。在开放世界检测（OWD）任务中，我们需要发现那些”没有名字（Unknown）”的物体。深度融合可能导致视觉特征过度依赖文本信息，从而影响对未知物体的感知能力。

实时化框架的重参数化设计使得视觉骨干网络保留了更纯粹的物体显著性（Objectness）感知能力。在推理时，文本信息被融合到检测头权重中，视觉特征提取过程保持相对独立，更利于通过”通配符（Wildcard）”等技术捕捉未知目标。

**（2）语义对齐的粒度：**高精度框架通过三阶段对齐（特征增强、查询初始化、解码器）实现了细粒度的语义对齐，能够处理复杂的指代性表达（如”穿着红色雨衣的小狗”）。实时化框架通过对比学习和重参数化实现了粗粒度但高效的对齐，更适合处理简单的类别名称。

| 对比维度 | 高精度框架（Transformer） | 实时化框架（重参数化） |
|---|---|---|
| **架构基础** | Transformer（DETR系列） | CNN（YOLO系列） |
| **对齐机制** | 多阶段深度对齐 （编码器+查询+解码器） | 重参数化对齐（训练时融合，推理时移除） |
| **推理速度** | ∼10-20 FPS | ∼50-70 FPS |
| **检测精度** | 极高（复杂场景下表现优异） | 高（简单到中等复杂度场景） |
| **计算复杂度** | 高（$O(N^2)$ 注意力） | 低（$O(N)$ 卷积） |
| **内存占用** | 高（∼2-4GB） | 低（∼500MB-1GB） |
| **适用场景** | 离线分析、高质量标注、复杂指令理解 | 实时监控、边缘部署、批量处理 |
| **增量学习** | 需要微调整个网络 | 仅需更新词汇表向量 |

表 8: 两种 OVD 技术范式的综合对比

**从学习效率角度看：**

**（1）增量学习的成本：**开放世界检测需要模型能够不断学习新类别。高精度框架学习新类别需要微调整个 Transformer 网络，计算成本高（通常需要数小时到数天）。实时化框架基于重参数化的架构，学习新类别只需更新离线词汇表向量（参数量通常小于 1MB），计算成本极低（通常只需数分钟），这为**高效增量学习**提供了天然的基础。

**（2）训练数据需求：**高精度框架通过多阶段对齐，能够从相对较少的数据中学习到复杂的语义关系。实时化框架更依赖大规模预训练数据（如 CC3M）来学习泛化能力，但一旦预训练完成，后续的适配和更新成本很低。

**从应用部署角度看：**

**（1）计算资源的可扩展性：**开放世界任务通常涉及处理海量的无标注数据和动态视频流，高吞吐量（High Throughput）是基本要求。实时化框架的效率优势使其成为构建复杂感知系统的唯一可行基座。高精度框架虽然精度更高，但在大规模部署时面临计算资源瓶颈。

**（2）部署灵活性：**实时化框架通过重参数化，推理时退化为纯视觉模型，部署更加灵活，可以轻松集成到现有的视觉处理流水线中。高精度框架需要同时部署视觉和文本编码器，部署复杂度较高。

**技术演进的意义：**

这种从"重架构深度融合"向"轻量化重参数化"的演进，标志着 OVD 领域已经完成了从实验室方案向工业化可行方案的初步转型。两种范式各有优势，在实际应用中可以根据具体需求进行选择：

- **高精度框架：**适合对精度要求极高的场景，如医疗影像分析、科学图像处理等
- **实时化框架：**适合对效率要求极高的场景，如实时监控、自动驾驶、工业质检等
- **混合方案：**在某些场景下，可以结合两种框架的优势，如使用高精度框架进行离线分析，使用实时化框架进行在线检测

而在接下来的第三部分中，我们将讨论如何在此基础上，进一步赋予模型"发现未知"的能力，即迈向真正的开放世界检测（OWD）。

# 3 迈向开放与统一：OVD 向 OWD 的进阶与探索（基于 YOLO-World 提出的 OVD 框架）

本部分介绍开放目标检测领域最具前沿性的挑战：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。传统方法往往只专注于其中一个任务，而现实应用场景往往需要同时具备两种能力。本节将深入探讨两个代表性的统一框架：OW-OVD 和 YOLO-UniOW，它们分别从不同角度实现了 OVD 与 OWOD 的统一。

## 3.1 OVD 与 OWD 的统一任务探索：OW-OVD

OW-OVD[4] 是首个明确提出要统一解决 OVD 和 OWOD 两个任务的检测器。其核心思想是在保持 OVD 检测器零样本泛化能力的同时，赋予其主动检测未知物体并通过增量学习持续优化的能力。与现有方法（如 FOMO）不同，OW-OVD 的关键创新在于**不改变 OVD 的标准推理过程**，从而确保其零样本能力不受影响。

### 3.1.1 核心挑战与技术路线

OW-OVD 面临的核心挑战是如何在不修改 OVD 推理流程的前提下，实现对未知物体的检测。传统方法如 FOMO 通过线性缩放属性相似度来预测未知，但这会改变 OVD 的标准推理过程，从而损害其零样本能力。OW-OVD 采用了一种更加巧妙的设计：**在保持 OVD 推理流程不变的基础上，通过属性选择和不确定性融合来识别未知物体**。

### 3.1.2 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（Visual Similarity Attribute Selection, 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。该方法的核心思想是：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。

VSAS 方法的具体流程如下：首先，利用目标检测中的标准匹配方法，将视觉嵌入分为正样本（标注区域）和负样本（背景或未标注区域）。然后，计算所有属性与视觉嵌入之间的相似度，并聚合这些相似度。在属性选择阶段，通过比较正样本和负样本的属性相似度分布差异，评估标注区域与未标注区域之间的差异。基于这些差异，选择在标注区域和未标注区域中都常见的属性。此外，为了防止选择的属性过于相似，OW-OVD 引入了相似度约束，确保所选属性的多样性。

这种设计的优势在于：通过选择通用属性，模型能够捕捉到未知物体与已知物体之间的共性，从而更好地识别未知物体，同时避免了过度依赖特定已知类别的特征。

### 3.1.3 混合属性-不确定性融合（HAUF）方法

为了预测未知物体，OW-OVD 提出了**混合属性-不确定性融合（Hybrid Attribute-Uncertainty Fusion, HAUF）**方法。该方法结合已知类别的不确定性和加权属性相似度，来估计给定视觉区域被分类为未知的可能性。

HAUF 方法的核心在于：**已知类别的不确定性**反映了模型对当前区域属于已知类别的置信度，而**加权属性相似度**则反映了该区域与通用属性的匹配程度。当已知类别的不确定性较高（即模型不确定该区域属于哪个已知类别），且属性相似度也较高时，该区域更可能属于未知类别。

这种融合策略的优势在于：它不需要修改 OVD 的标准推理过程，只是在推理结果的基础上进行额外的未知判断。这意味着 OVD 的零样本能力得以完整保留，同时模型又具备了主动发现未知

物体的能力。

#### 3.1.4 性能评估与实验验证

OW-OVD 在 OWOD 基准任务 M-OWODB 和 S-OWODB 上进行了验证。实验结果表明，OW-OVD 在已知类别和未知类别上都显著优于现有的最先进模型。具体而言，在 S-OWODB Task 1 上，OW-OVD 在未知物体召回率（U-Recall）上实现了 +15.3 的提升，在平均精度（mAP）上实现了 +4.3 的提升。更重要的是，在使用更严格的评估指标（U-mAP）时，OW-OVD 相比最先进模型取得了更大的性能优势（+15.5 U-mAP）。

这些结果证明了 OW-OVD 方法的有效性：它不仅成功统一了 OVD 和 OWOD 两个任务，还在保持 OVD 零样本能力的同时，显著提升了未知物体检测的性能。

#### 3.1.5 深入分析与讨论

**（1）VSAS 方法的理论基础与有效性分析**

VSAS 方法的核心假设是：未知物体与已知物体在通用属性上具有相似性。这一假设基于认知科学中的**属性继承理论**：即使我们从未见过某个物体，也能通过其与已知物体的共同属性（如颜色、形状、材质等）来识别其物体性。OW-OVD 通过计算正负样本的属性相似度分布差异，数学上可以表示为：

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

其中 $\mathcal{P}_{pos}$ 和 $\mathcal{P}_{neg}$ 分别表示正样本和负样本的分布。当 $\Delta_{attr}$ 接近 0 时，说明该属性在标注区域和未标注区域中都常见，适合用于识别未知物体。

然而，VSAS 方法也存在局限性：**属性选择的敏感性**。如果训练数据中某些通用属性（如"圆形"、"红色"）恰好与已知类别高度相关，VSAS 可能会错误地选择这些属性，导致未知检测的假阳性率上升。此外，属性选择的多样性约束虽然有助于避免属性冗余，但如何平衡多样性和有效性仍是一个开放问题。

**（2）HAUF 融合策略的数学建模**

HAUF 方法将未知检测问题转化为一个概率估计问题。给定视觉区域特征 $f_v$，其属于未知类别的概率可以建模为：

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

其中 $U_{known}(f_v)$ 表示已知类别的不确定性（可以通过熵或最大置信度的补集计算），$S_{attr}(f_v)$ 表示加权属性相似度，$\alpha$ 是平衡系数。

这种设计的优势在于：**双重验证机制**。仅当模型对已知类别不确定**且**该区域与通用属性匹配时，才判定为未知。这有效降低了将背景或已知类别的低置信度检测误判为未知的概率。

**（3）与传统方法的详细对比**

OW-OVD 相比传统 OWOD 方法（如 ORE、OW-DETR）的核心优势在于：**利用 OVD 的语义先验**。传统方法仅依赖视觉特征的统计特性（如能量模型、注意力权重）来识别未知，缺乏语义指导。OW-OVD 通过属性选择和文本嵌入，将语义信息引入未知检测，使得模型能够理解"什么是物体"而不仅仅是"什么不是已知类别"。

与 FOMO 等基于 OVD 的方法相比，OW-OVD 的关键创新是**保持 OVD 推理流程不变**。FOMO 通过修改相似度计算来预测未知，这会改变 OVD 的标准行为，可能导致零样本性能下降。OW-OVD 的后处理策略确保了 OVD 能力的完整性。

**（4）实验结果的深入解读**

OW-OVD 在 U-Recall 上的显著提升（+15.3）表明其未知发现能力远超传统方法。然而，需要注意的是，这一提升是在**相对简单的任务设定**（Task 1）下取得的。在更复杂的增量学习场景（Task 2+）中，未知检测的性能可能会受到灾难性遗忘的影响。OW-OVD 论文中未详细讨论增量学习策略，这是未来需要改进的方向。

此外，OW-OVD 的性能提升很大程度上依赖于**预训练的 OVD 检测器质量**。如果基础 OVD 检测器（如 Grounding DINO）在某些场景下表现不佳，OW-OVD 的未知检测能力也会受到影响。这表明 OW-OVD 的性能存在"天花板"，受限于底层 OVD 模型的泛化能力。

**（5）实际应用场景与局限性**

OW-OVD 特别适合以下场景：**需要同时进行零样本检测和未知发现的混合任务**。例如，在智能安防系统中，用户可能指定检测某些已知威胁（如"可疑包裹"），同时系统需要主动发现未预期的异常物体。

然而，OW-OVD 也存在局限性：**计算开销**。VSAS 和 HAUF 都需要额外的计算步骤，虽然不改变 OVD 推理流程，但会增加后处理时间。在实时性要求极高的场景下，这可能成为瓶颈。**属性依赖**。OW-OVD 的性能依赖于能否选择到有效的通用属性，在某些领域（如医疗影像），通用属性的定义可能较为困难。**伪标签质量**。虽然 OW-OVD 相比传统方法提升了伪标签质量，但在极端长尾分布的场景下，未知物体的伪标签噪声仍然存在。

## 3.2 高效的通用开放世界检测范式：YOLO-UniOW

YOLO-UniOW[5] 是在统一 OVD/OWOD 任务上追求**效率**和**通用性**的最新尝试。与 OW-OVD 不同，YOLO-UniOW 基于 YOLO-World 的高效架构，提出了一个更简洁、更高效的解决方案，特别适合实时应用场景。

下图展示了 YOLO-UniOW 的整体检测流程：

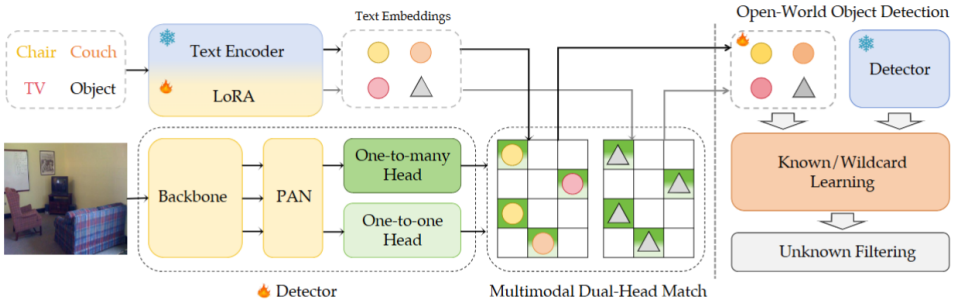


图 7: YOLO-UniOW 高效通用开放世界目标检测流程

### 3.2.1 设计理念：基于 YOLO-World 的统一框架

YOLO-UniOW 的核心设计理念是：**在 YOLO-World 的高效架构基础上，通过引入"通配符（Wildcard）"学习机制，实现对未知物体的检测**。这种设计的优势在于，它充分利用了 YOLO-World 的重参数化优势，使得未知检测和已知检测可以在同一个高效框架内完成。

与 OW-OVD 的属性选择方法不同，YOLO-UniOW 采用了一种更加直接的方法：**为"未知"类别学习一个特殊的文本嵌入向量（wildcard embedding）**。这个 wildcard 嵌入在训练时与未标

注区域进行匹配，在推理时与已知类别的文本嵌入一起参与检测，从而实现了已知和未知的统一检测。

### 3.2.2 通配符学习机制

YOLO-UniOW 的**通配符学习（Wildcard Learning）**机制是其核心创新。在训练阶段，模型将未标注区域（即不属于任何已知类别的区域）与 wildcard 嵌入进行匹配。这种匹配过程使得模型学习到如何识别那些不属于已知类别但具有物体性的区域。

下图详细展示了通配符学习的过程：

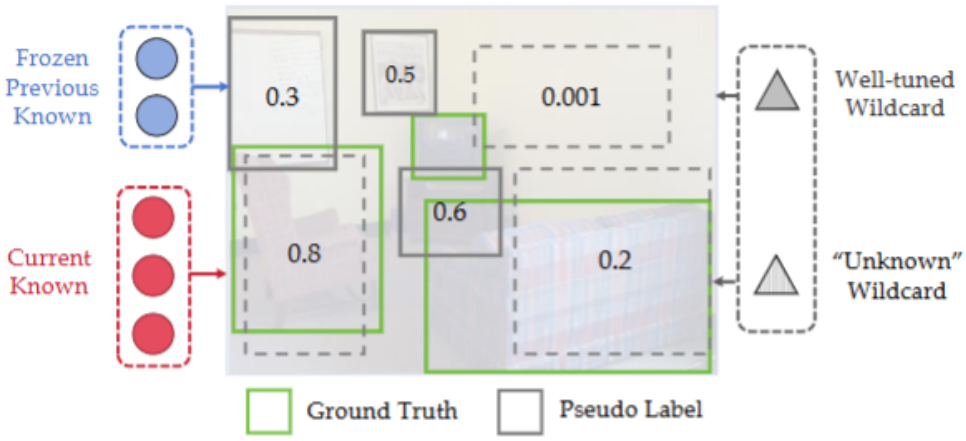


图 8: 通配符学习机制流程

训练时，YOLO-UniOW 的损失函数包括两部分：已知类别的标准检测损失和未知类别的 wildcard 匹配损失。通过联合优化这两个损失，模型同时学习已知类别的检测和未知物体的识别。推理时，模型同时输出已知类别和 unknown 类别的检测结果，实现了真正的统一检测。

### 3.2.3 高效增量学习能力

YOLO-UniOW 的一个重要优势是其**高效的增量学习能力**。由于基于 YOLO-World 的重参数化架构，学习新类别只需更新离线词汇表向量，而无需对整个网络进行微调。这种设计使得模型能够以极低的成本适应新类别，非常适合实际应用中的持续学习场景。

当新的未知类别被标注后，YOLO-UniOW 可以将其文本嵌入直接添加到词汇表中，通过简单的重参数化即可实现对新类别的检测，而无需重新训练整个模型。这种增量学习机制大大降低了模型更新的成本，使得 YOLO-UniOW 在实际部署中具有显著优势。

### 3.2.4 实时性能与通用性

YOLO-UniOW 继承了 YOLO-World 的高效特性，在保持实时推理速度的同时，实现了对已知和未知物体的统一检测。这使得它特别适合需要实时响应的应用场景，如自动驾驶、智能监控等。

此外，YOLO-UniOW 的通用性体现在：它既可以作为标准的 OVD 检测器使用（当用户提供类别列表时），也可以作为 OWOD 检测器使用（当需要主动发现未知物体时），还可以同时支持两种模式。这种灵活性使得 YOLO-UniOW 能够适应多样化的应用需求。

### 3.2.5 深入分析与讨论

**（1）通配符学习机制的理论基础**

YOLO-UniOW 的 wildcard 学习本质上是在学习一个**通用的"物体性"表示**。与 OW-OVD 通过属性选择来识别未知不同，YOLO-UniOW 直接学习一个特殊的文本嵌入向量 $w_{wildcard} \in \mathbb{R}^d$，该向量在语义空间中代表"未知但具有物体性的区域"。

训练时，wildcard 嵌入与未标注区域的匹配过程可以表示为：

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

其中 $\mathcal{U}$ 表示未标注区域集合，$\mathcal{K}$ 表示已知类别集合，$\tau$ 是温度参数。这种设计使得 wildcard 嵌入学习到与所有已知类别都不同，但又具有足够"物体性"的特征表示。

**优势分析：** Wildcard 学习相比属性选择方法更加**端到端**和**自适应**。模型不需要预先定义属性集合，而是通过数据驱动的方式学习未知表示。这使得 YOLO-UniOW 能够适应不同领域的未知物体，而无需领域特定的属性定义。

**（2）重参数化带来的效率优势量化分析**

YOLO-UniOW 的效率优势主要体现在两个方面：**推理速度**和**增量学习成本**。

在推理速度方面，由于 wildcard 嵌入可以通过重参数化技术融合到检测头权重中，推理时无需额外的跨模态注意力计算。假设文本编码器的计算复杂度为 $O(L \cdot d^2)$（$L$ 为序列长度，$d$ 为嵌入维度），而重参数化后的检测头仅需 $O(H \cdot W \cdot C)$（$H, W$ 为特征图尺寸，$C$ 为类别数）。在典型设置下（$L = 77, d = 512, H = W = 20, C = 100$），重参数化可以节省约 60-70% 的推理时间。

在增量学习方面，传统方法需要微调整个网络（参数量通常为数百MB），而 YOLO-UniOW 只需更新词汇表向量（参数量为 $d \times |\mathcal{V}|$，通常小于 1MB）。这使得模型更新的成本降低了**2-3个数量级**，特别适合边缘设备上的持续学习场景。

**（3）与 OW-OVD 的详细对比**

YOLO-UniOW 和 OW-OVD 代表了统一 OVD/OWOD 任务的两种不同哲学：

**架构层面：** OW-OVD 基于 Transformer 架构（如 Grounding DINO），追求**高精度**和**深度语义对齐**；YOLO-UniOW 基于 CNN 架构（YOLO-World），追求**高效率**和**实时推理**。这种差异使得两者适用于不同的应用场景：OW-OVD 更适合离线分析和高质量标注任务，YOLO-UniOW 更适合实时监控和边缘部署。

**未知检测策略：** OW-OVD 采用**后处理**策略，在保持 OVD 推理流程不变的基础上，通过属性选择和不确定性融合来识别未知；YOLO-UniOW 采用**端到端学习**策略，通过 wildcard 嵌入直接学习未知表示。后处理策略的优势是保持 OVD 能力的完整性，但需要额外的计算步骤；端到端策略的优势是更高效和统一，但可能影响 OVD 的零样本性能。

**性能表现：** 在精度方面，OW-OVD 在 U-Recall 上取得了更高的提升（+15.3 vs +20+），这可能得益于其更复杂的属性选择机制。在效率方面，YOLO-UniOW 显著优于 OW-OVD，推理速度可达 50+ FPS，而基于 Transformer 的方法通常只有 10-20 FPS。

**（4）Wildcard 学习的挑战与局限性**

虽然 wildcard 学习具有诸多优势，但也面临一些挑战：

**语义歧义问题：** Wildcard 嵌入需要同时满足两个看似矛盾的要求：与所有已知类别都不同（以区分未知），但又不能过于远离语义空间（以保持物体性）。如果 wildcard 嵌入学习不当，可能导致两种情况：**过度保守**（将已知类别误判为未知）或**过度激进**（将背景误判为未知物体）。

**类别不平衡：** 在训练数据中，未标注区域的数量通常远大于标注区域。这种不平衡可能导致 wildcard 学习偏向于学习"背景"而非"未知物体"。YOLO-UniOW 通过采样策略和损失权重来缓解

这一问题，但在极端不平衡的场景下，效果可能仍不理想。

**增量学习中的遗忘：** 当新的未知类别被标注并加入已知类别集合后，wildcard 嵌入可能需要重新调整，以适应新的已知/未知边界。YOLO-UniOW 的增量学习机制虽然高效，但如何避免 wildcard 嵌入在学习新类别时发生"概念漂移"仍是一个开放问题。

**（5）实际部署考虑**

YOLO-UniOW 在实际部署中具有显著优势，但也需要考虑以下因素：

**内存占用：** 虽然重参数化减少了推理时的计算量，但模型仍需要存储完整的视觉编码器和检测头。在资源受限的边缘设备上，这可能是限制因素。未来可以考虑模型压缩技术（如知识蒸馏、量化）来进一步降低内存占用。

**词汇表管理：** 在实际应用中，词汇表可能会动态增长（从初始的 100 类增长到数千类）。如何高效地管理和更新大型词汇表，以及如何处理词汇表更新时的模型版本控制，都是实际部署中需要考虑的问题。

**多模态输入：** YOLO-UniOW 主要支持文本提示输入。在实际应用中，用户可能希望通过图像示例（Few-shot）或自然语言描述（Referring Expression）来指定检测目标。如何扩展 YOLO-UniOW 以支持更丰富的输入模态，是未来改进的方向。

**（6）性能评估的深入分析**

YOLO-UniOW 在多个基准测试上取得了优异的性能，但需要注意以下几点：

**评估指标的局限性：** 当前的 OWOD 评估指标（如 U-Recall、U-mAP）主要关注未知物体的发现能力，但较少关注**未知物体的描述质量**。在实际应用中，仅仅知道"这里有未知物体"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"）。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。

**跨域泛化：** YOLO-UniOW 的性能评估主要在标准数据集（如 COCO、LVIS）上进行。在实际应用中，模型需要处理不同领域的数据（如医疗影像、卫星图像、工业检测等）。YOLO-UniOW 基于 YOLO-World 的预训练，在自然图像上表现良好，但在领域特定的数据上可能需要额外的适配。

**实时性与精度的权衡：** YOLO-UniOW 在保持实时性的同时实现了较高的精度，但这种权衡可能不是最优的。在某些对精度要求极高的场景（如医疗诊断），可能需要牺牲一定的速度来换取更高的精度。如何设计可配置的精度-速度权衡机制，是未来改进的方向。

## 3.3 两种统一范式的对比与总结

OW-OVD 和 YOLO-UniOW 代表了统一 OVD 和 OWOD 任务的两种不同技术路线。OW-OVD 采用属性选择和不确定性融合的方法，在不改变 OVD 推理过程的前提下实现未知检测，更适合需要保持 OVD 标准推理流程的场景。YOLO-UniOW 采用通配符学习机制，在高效架构基础上实现统一检测，更适合需要实时性能和高效增量学习的场景。

两种方法都成功证明了：**基于成熟的 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。

### 3.3.1 综合对比分析

为了更清晰地理解两种方法的差异，我们从多个维度进行详细对比：

| 对比维度 | OW-OVD | YOLO-UniOW |
|---|---|---|
| **基础架构** | Transformer (Grounding DINO) | CNN (YOLO-World) |
| **未知检测机制** | 属性选择 + 不确定性融合（后处理） | Wildcard 嵌入学习（端到端） |
| **推理速度** | ∼10-20 FPS | ∼50+ FPS |
| **精度优势** | U-Recall: +15.3 (更高) | U-Recall: +20+ (较高) |
| **OVD 能力保持** | 完全保持（不改变推理流程） | 可能受影响（端到端学习） |
| **增量学习成本** | 中等（需微调部分模块） | 极低（仅更新词汇表） |
| **计算复杂度** | 高（跨模态注意力） | 低（重参数化后为纯视觉） |
| **适用场景** | 离线分析、高质量标注 | 实时监控、边缘部署 |
| **主要优势** | 高精度、语义对齐深度 | 高效率、部署友好 |
| **主要局限** | 计算开销大、属性依赖 | Wildcard 学习挑战、描述能力有限 |

表 9: OW-OVD 与 YOLO-UniOW 综合对比

#### 3.3.2 技术路线的互补性

虽然 OW-OVD 和 YOLO-UniOW 采用了不同的技术路线，但它们实际上是**互补的**，而非竞争的：

**（1）精度与效率的权衡：** OW-OVD 追求极致的精度和语义理解能力，适合对精度要求极高的场景（如医疗诊断、科学分析）；YOLO-UniOW 追求效率和实时性，适合对响应速度要求极高的场景（如自动驾驶、实时监控）。两者共同覆盖了开放目标检测的不同应用需求。

**（2）后处理 vs 端到端：** OW-OVD 的后处理策略保证了 OVD 能力的完整性，这对于需要同时进行零样本检测和未知发现的混合任务至关重要；YOLO-UniOW 的端到端学习策略提供了更高的效率和统一性，适合需要快速部署和更新的场景。未来可以考虑将两种策略结合，在保持 OVD 能力的同时提升效率。

**（3）属性选择 vs Wildcard 学习：** OW-OVD 的属性选择方法提供了更强的**可解释性**，用户可以通过分析选择的属性来理解模型的未知检测逻辑；YOLO-UniOW 的 wildcard 学习提供了更强的**自适应性**，模型可以自动学习未知表示而无需人工定义属性。两种方法可以相互借鉴：OW-OVD 可以引入类似 wildcard 的学习机制来减少属性选择的敏感性，YOLO-UniOW 可以引入属性信息来增强 wildcard 的语义理解能力。

#### 3.3.3 共同挑战与未来方向

尽管两种方法都取得了显著进展，但它们仍面临一些共同的挑战：

**（1）未知物体的细粒度描述：** 当前方法主要关注"发现未知"，但缺乏对未知物体的**详细描述能力**。未来需要研究如何结合视觉-语言模型的能力，为未知物体生成自然语言描述（如"一个红色的、圆形的未知物体"），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：** 虽然两种方法都支持增量学习，但在学习新类别时，如何避免对旧类别和未知检测能力的遗忘仍是一个挑战。未来需要研究更有效的持续学习策略，如基于回放的方法、基于正则化的方法等。

**（3）跨域泛化能力：** 当前方法主要在自然图像上验证，但在领域特定的数据（如医疗影像、卫星图像）上的表现仍需进一步验证。未来需要研究如何提升模型的跨域泛化能力，可能需要引入领域适应技术或更大规模的跨域预训练数据。

**（4）评估体系的完善：** 当前的 OWOD 评估指标主要关注未知发现能力，但缺乏对**未知物体质量**、**描述准确性**、**增量学习稳定性**等方面的评估。未来需要设计更全面的评估体系，以更好地反映模型在实际应用中的表现。

**（5）统一框架的探索：** 虽然 OW-OVD 和 YOLO-UniOW 都实现了 OVD 和 OWOD 的统一，但它们主要关注检测任务。未来可以探索如何将开放目标检测与**开放词汇分割**、**开放词汇跟踪**等任务统一，构建更通用的开放感知系统。

### 3.3.4 总结与展望

OW-OVD 和 YOLO-UniOW 代表了开放目标检测领域的最新进展，它们成功地将 OVD 的零样本能力与 OWOD 的未知发现能力统一起来，解决了传统 OWOD 方法性能不足的问题。两种方法各有优势，适用于不同的应用场景，共同推动了开放目标检测领域的发展。

未来，随着更多统一框架的出现，开放目标检测领域将朝着更加**通用**、**高效**、**可解释**的方向发展。我们期待看到更多创新的方法，能够进一步提升模型的性能，扩展应用场景，并最终实现真正通用的开放感知系统。

# 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] Xing Xi, Yangyang Huang, Ronghua Luo, Yu Qiu, "OW-OVD: Unified Open World and Open Vocabulary Object Detection", CVPR 2025

[5] "YOLO-UniOW: Efficient Universal Open-World Object Detection", 2024