# 前景增强与语义引导的开放词汇目标检测

刘立波[1, 2+]，赵镛清[1+]，张若楠[3*]
（1. 宁夏大学信息工程学院，宁夏 银川 750021；2. 宁夏“东数西算”人工智能与信息安全重点实验室，宁夏 银川 750021；
3. 宁夏大学 信息与网络空间安全学院，宁夏 中卫 755000）

**摘 要：**开放词汇目标检测中，现有方法的前景监督主要依赖基类标注，导致无人工标注的新类目标在候选生成阶段易被漏检。利用具体类别伪标签补充监督虽可缓解该问题，但伪标签的类别噪声会干扰新类判断。此外，通用模板生成的类别表示难以充分体现候选区域在目标尺度与完整性上的差异，且区域—文本间的语义关联易在多层解码中逐步衰减。针对上述问题，本文提出前景增强与语义引导的开放词汇目标检测方法 FESG-OVD，为学习更通用的物体特征、促进视觉语言模型语义知识向检测任务的迁移提供一种思路。首先，构建类别无关前景增强模块（CAFE），该模块通过目标 Token 聚合多尺度目标结构并引导编码特征增强，利用开放世界伪标签提供的通用物体信息补充前景监督，以减少潜在目标在候选生成阶段的遗漏。其次，设计语义查询增强模块（SQE），该模块通过多模板引导类别表示补充候选并校准语义查询，使类别语义引导候选筛选与查询构建。最后，提出语义保持模块（SP），将 SQE 生成的语义锚点传递至各解码层，根据当前查询状态聚合类别语义并更新查询，使语义引导由解码前延续至多层解码过程。实验结果表明，本文方法在 COCO 数据集上的新类精度达到 45.7%，在 LVIS 数据集上的稀有类精度达到 39.3%，较现有先进方法分别提升 1.7 和 1.5 个百分点，验证了该方法对新类目标的检测能力。

**关键词：**开放词汇目标检测；视觉—语言模型；前景增强；语义查询增强；语义保持
**中图分类号：**TP18
**文献标志码：**A
**doi：**10.11959/j.issn.2096−3750.

# Foreground Enhancement and Semantic Guidance for Open-Vocabulary Object Detection

LIU Libo[1,2+], ZHAO Yongqing[1+], ZHANG Ruonan[3*]
1. School of Information Engineering, Ningxia University, Yinchuan, Ningxia 750021, China
2. Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West, Yinchuan, Ningxia 750021, China
3. School of Information and Cyberspace Security, Ningxia University, Zhongwei, Ningxia 755000, China

**Abstract:** In open-vocabulary object detection, foreground supervision mainly relies on base-class annotations, making novel objects without manual annotations prone to omission during candidate generation. Category errors in category-specific pseudo-labels used as additional supervision may also interfere with novel-class recognition. Meanwhile, category representations derived from generic text templates may not adequately capture differences in object scale and completeness across candidate regions, and region – text associations may weaken during multilayer decoding. To address

**收稿日期：**XXXX−XX−XX；**修回日期：**XXXX−XX−XX
**通信作者：**张若楠，zhangrn@nxu.edu.cn
**基金项目：**宁夏自然科学基金(No. 2024AAC02010)；国家自然科学基金(No. 62262053, No. 62506179)；宁夏科技创新领军人才项目(No. 2022GKLRLX03)；银川市科技项目(No. 2025RC09)
**Foundation Items:** Natural Science Foundation of Ningxia, China(No. 2024AAC02010), The National Natural Science Foundation of China (No. 62262053, No. 62506179), Science and Technology Leading Talent Training Projects of Ningxia (No. 2022GKLRLX03), Yinchuan City Science and Technology Projects(No. 2025RC09)

these issues, we propose Foreground Enhancement and Semantic Guidance for Open-Vocabulary Object Detection (FESG-OVD). The method offers an approach to learning more general object features and transferring semantic knowledge from vision–language models to object detection. First, the Category-Agnostic Foreground Enhancement module (CAFE) uses object tokens to aggregate multiscale object structures and guide the enhancement of encoded features. It supplements foreground supervision with generic object information from open-world pseudo-labels to reduce potential object omissions during candidate generation. Second, the Semantic Query Enhancement module (SQE) uses multi-template guided category representations to supplement candidates and calibrate semantic queries, enabling category semantics to guide candidate selection and query construction. Finally, the Semantic Persistence module (SP) passes the semantic anchors generated by SQE to each decoder layer, where category semantics are aggregated according to the current query state to update the queries, extending semantic guidance from the pre-decoding stage throughout multilayer decoding. Experimental results show that FESG-OVD achieves a novel-class AP50 of 45.7% on COCO and a rare-class AP of 39.3% on LVIS, exceeding the results of the compared state-of-the-art methods by 1.7 and 1.5 percentage points, respectively. These results demonstrate the effectiveness of the proposed method in detecting novel objects.

**Key words:** Open-Vocabulary Object Detection, Vision–Language Models, Foreground Enhancement, Semantic Query Enhancement, Semantic Persistence

## 0　引言

开放词汇目标检测[1-2]（Open-Vocabulary Object Detection，OVD）旨在突破传统封闭集检测器[3-4]对预定义类别范围的限制，使模型在仅使用有限基类标注进行训练的条件下，能够同时识别和定位训练阶段已见的基类与未见的新类。由于新类缺少直接的人工标注，现有方法普遍引入以CLIP[5]为代表的视觉-语言模型（Vision-Language Model，VLM），利用其在大规模图像—文本数据上学习到的跨模态语义知识，将类别文本表示与目标区域的视觉特征进行匹配，从而获得对新类别的识别能力。如何将VLM的开放语义知识有效迁移到检测器，已成为提升OVD模型新类泛化能力的关键。按照研究侧重点，现有OVD方法大致可归纳为弱监督信息扩展和VLM语义知识迁移两类。其中，弱监督信息扩展方法利用图像级标签和类别伪标签，补充基类标注未覆盖的新类训练信息[6]。Detic[7]通过大规模图像级标签扩展检测类别，VL-PLM[8]结合区域候选与CLIP分类结果生成新类伪标签，GLCS-Net[9]进一步利用上下文信息筛选可靠的监督样本。上述方法扩展了新类监督信息的来源，但由VLM生成的类别伪标签通常包含分类与定位误差[10]，其中的类别错误可能干扰新类判断[11-12]。同时，主要依赖基类标注学习的候选生成机制可能低估新类目标的前景得分，使其容易在候选筛选阶段被漏检。

另一类方法侧重VLM语义知识迁移，通过知识蒸馏[13]、区域—文本匹配等方式，将VLM的开放类别语义引入检测器。ViLD[14]通过知识蒸馏对齐区域特征与CLIP视觉特征，RegionCLIP[15]利用图像—文本数据学习区域级语义表示，HD-OVD[16]从实例、类别和图像层面构建分层语义蒸馏，CORA[17]通过区域提示与锚框预匹配使类别文本参与目标定位，OV-DINO[18]进一步利用语言信息进行候选选择与跨模态特征融合。尽管上述方法增强了文本语义与视觉区域之间的联系[19-21]，其语义利用方式仍存在两方面不足。一方面，通用模板生成的类别表示难以充分体现候选区域中目标尺度和完整程度的差异。另一方面，部分方法的区域—文本交互主要集中于解码前或特定阶段，而区域表征会在后续解码中随着视觉信息的融合不断更新，早期建立的区域—文本联系可能逐渐减弱。

综上，如何在减少类别噪声干扰的同时补充新类目标的前景监督，使类别表示更充分地体现候选区域中的目标差异，并在多层解码中保持区域—文本联系，仍是现有OVD方法亟需解决的问题。为解决上述问题，本文提出前景增强与语义引导的开放词汇目标检测方法（Foreground Enhancement and Semantic Guidance for Open-Vocabulary Object Detection，FESG-OVD），为学习更通用的物体特征、促进CLIP语义知识向检测任务的迁移提供一种思路。该方法通过前景增强减少潜在目标在候选生成阶段的遗漏，并将类别语义的引导作用从候选

筛选与查询构建延续至多层解码。本文主要贡献如下：

（1）针对新类目标前景监督不足及具体类别伪标签中的类别噪声干扰问题，构建类别无关前景增强模块（Category-Agnostic Foreground Enhancement，CAFE）。该模块通过可学习的目标Token聚合多尺度目标结构并引导编码特征增强，同时利用开放世界伪标签提供的通用物体信息补充前景监督，以减少基类标注未覆盖的潜在目标在候选生成阶段的遗漏。

（2）针对通用类别表示难以充分体现候选区域中目标尺度和完整程度差异的问题，设计语义查询增强模块（Semantic Query Enhancement，SQE）。该模块在保留原始CLIP类别表示的基础上构建多模板引导类别表示，通过语义辅助的候选补充与区域语义校准，增强文本语义在候选筛选和查询构建中的作用。

（3）针对区域—文本联系可能在多层解码中减弱的问题，提出语义保持模块（Semantic Persistence，SP）。该模块将SQE生成的语义锚点传递至各解码层，根据当前查询状态聚合类别语义，并通过门控残差更新查询，使SQE引入的类别语义在后续解码中持续发挥引导作用。

（4）在COCO和LVIS数据集上开展对比实验。FESG-OVD的新类检测指标$AP_{50}^{Novel}$和稀有类检测指标$AP_r$分别达到45.7%和39.3%，消融实验进一步验证了三个模块及其组合的有效性。

## 1　本文方法

FESG-OVD由类别无关前景增强模块（CAFE）、语义查询增强模块（SQE）和语义保持模块（SP）组成，整体结构如图1所示。

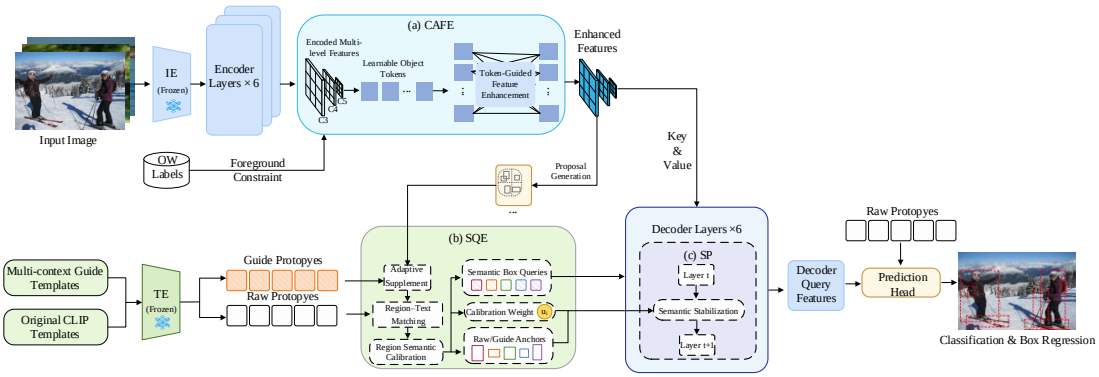



图1　FESG-OVD框架图

CAFE通过目标Token聚合多尺度目标结构并增强编码特征，在候选生成前实施前景增强。这里的“前景增强”，指增强编码特征中潜在目标位置的前景响应，包含两个层面。在特征层面，目标Token聚合的多尺度结构经门控交叉注意力注入编码特征，提高潜在目标位置的响应；在监督层面，前景监督由基类标注扩展为类别无关的通用物体监督，由真实标注与高置信度开放世界伪标签共同提供，训练时据此构建类别无关前景约束，为编码特征增强提供指导。SQE与SP共同实现语义引导。其中SQE利用类别语义补充候选并校准语义查询，构建初始检测查询，同时生成语义锚点和语义校准权重；SP在各解码层中根据当前查询状态，从语义锚点中聚合类别语义，并结合语义校准权重更新查询，在多层解码中保持区域—文本联系。

在候选筛选、查询构建与多层解码三个环节中，语义引导遵循统一机制，均以类别文本表示为语义参照，通过相似度或交叉注意力计算当前表征与类别语义的关联，并以自适应权重将门控后的语义信息经残差更新注入当前表征，可统一表示为$X' = X + \lambda \cdot w(X,S) \cdot Attn(X,S)$，其中$X$为待更新的表征，$S$为类别文本表示，$Attn$为相似度或交叉注意力运算，$w$为依据二者关联构造的自适应权重，$\lambda$为残差系数。三个环节的差异在于权重$w$的构造，候选筛选依据引入类别语义前后的前景概率变化构造语义增益权重，查询构建依据区域类别匹配的区分度构造类别区分度权重，多层解码依据当前查询状态构造查询状态权重。

### 1.1　类别无关前景增强（CAFE）

现有方法通常依据编码特征的前景得分筛选候

选位置。当训练中的前景监督主要来自基类标注时，基类标注未覆盖的潜在目标可能获得较低的前景得分，从而在候选筛选时被忽略。为此，CAFE利用开放世界伪标签补充前景监督，通过可学习的目标Token聚合多尺度目标结构并引导编码特征增强。该模块包括多尺度目标Token建模、Token指导的编码特征增强和类别无关前景约束三个部分，细节如图2所示。

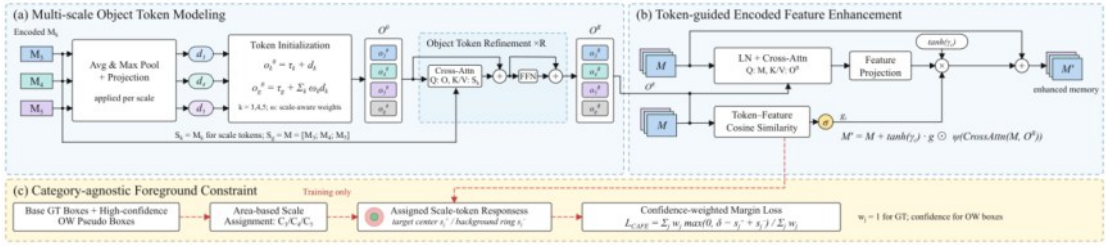



图2　CAFE模块细节图

### 1.1.1　多尺度目标Token建模

给定输入图像$I$，提取在空间分辨率和语义层次上互补的C3、C4和C5三个尺度的视觉特征，经特征投影和编码器处理后得到对应的编码特征$M_3$、$M_4$和$M_5$。将三个尺度的编码特征展平并拼接，得到完整编码特征$M=[M_3;M_4;M_5]$。若所有目标Token均从完整编码特征中自由聚合信息，不同Token可能关注相似的目标区域。为保留不同尺度的目标结构，CAFE设置三个尺度Token和一个全局Token。分别对M3、M4和M5进行平均池化与最大池化，以概括整体特征分布并保留显著目标特征。将两种池化结果拼接并经过特征映射，得到三个尺度描述向量。每个尺度描述向量与对应的可学习Token相加，形成当前图像的尺度Token初始表示。全局Token根据三个尺度描述向量预测融合权重，将加权得到的多尺度描述与可学习的全局Token相加，形成其初始表示。三个尺度Token分别聚合对应尺度的目标结构，全局Token则补充跨尺度信息。

在特征交互过程中，三个尺度Token分别与对应尺度的编码特征执行交叉注意力，全局Token则与完整编码特征$M$进行交互。记第$k$个Token在第$r-1$个Token更新块后的表示为$o_k^{r-1}$，其读取的编码特征为$S_k$。对于三个尺度Token，$S_k$为对应尺度的编码特征；对于全局Token，$S_k=M$。第$r$次更新过程表示为：

$$\tilde{o}_k^r = o_k^{r-1} + MHA\left(LN\left(o_k^{r-1}\right),LN\left(S_k\right),LN\left(S_k\right)\right),\#(1)$$

$$o_k^r = \tilde{o}_k^r + FFN\left(LN\left(\tilde{o}_k^r\right)\right),\#(2)$$

其中，MHA、LN和FFN分别表示多头注意力、层归一化和前馈网络。经过$R$个Token更新块后，四个Token共同组成最终目标Token集合$O^R$。该集合聚合了当前图像中不同尺度的目标结构，用于引导后续的编码特征增强。

### 1.1.2　Token指导的编码特征增强

为将目标Token聚合的目标结构用于候选生成与后续解码，CAFE利用$O^R$对完整编码特征$M$进行逐位置增强。对于第$i$个空间特征$m_i$，分别计算其与各目标Token的余弦相似度，取其中的最大值并转换为空间门控权重：

$$g_i = \sigma\left(\frac{\max_k cos\left(LN\left(m_i\right),LN\left(o_k^R\right)\right) - b}{\tau_a}\right),\#(3)$$

其中，$b$表示相似度偏置，$\tau_a$为温度系数。$g_i$越大，表示当前位置的特征与目标Token所聚合的目标结构越匹配。取最大值可保留当前位置与目标Token集合的最强匹配，避免平均操作削弱较强的匹配结果。随后，以$m_i$作为Query、$O^R$作为Key和Value执行交叉注意力，对输出进行特征映射和空间门控加权，再以残差方式更新原始编码特征：

$$m_i' = m_i + \tanh\left(\gamma_c\right) g_i \psi\left(MHA\left(LN\left(m_i\right),LN\left(O^R\right),LN\left(O^R\right)\right)\right),\#(4)$$

其中，$\psi$表示特征映射，$\gamma_c$为可学习的残差系数。各位置增强后的特征共同组成增强编码特征$M'$，用于候选生成，并作为后续解码器读取的图像特征。

### 1.1.3　类别无关前景约束

前述编码特征增强由目标Token引导。为使目标Token能够关注基类标注之外的潜在目标，CAFE在训练阶段利用真实标注和高置信度开放世

界伪标签构建类别无关前景约束，对目标Token进行监督。根据监督框面积，将其分配至C3、C4或C5对应的尺度，使监督目标的大小与特征图的空间分辨率相匹配，再在相应尺度Token的响应图上选取目标中心区域与邻近背景区域。将目标中心区域作为前景位置，将监督框周围且不属于其他监督框的区域作为背景位置。这里的Token响应反映相应尺度Token与空间特征的匹配程度。记第$j$个监督框中心区域的Token响应为$s_j^+$，邻近背景区域的Token响应为$s_j^-$，类别无关前景约束定义为：

$$L_{\mathrm{CAFE}}=\frac{1}{\sum_j w_j}\sum_j w_j \max\left(0,\delta - s_j^+ + s_j^-\right),\#(5)$$

其中，$\delta$表示前景与背景之间的响应间隔。对于真实标注，$w_j=1$；对于开放世界伪标签，$w_j$取其置信度。该约束促使目标中心区域的Token响应高于邻近背景，引导目标Token区分潜在目标与背景，为前述编码特征增强提供训练指导。开放世界伪标签提供的是类别无关的物体监督。利用其位置与置信度，可以补充基类标注之外的前景监督，同时减少错误具体类别信息对训练的直接干扰。

### 1.2　语义查询增强模块（SQE）

CAFE增强了用于候选生成的编码特征，SQE进一步将类别语义引入候选筛选与查询构建。考虑到候选区域中的目标尺度和完整程度存在差异，SQE保留原始CLIP类别表示，同时利用描述这些差异的文本模板构建引导类别表示，用于补充候选并校准语义查询。该模块从候选筛选和查询构建两个环节增强文本语义的作用，包括多模板引导表示、语义引导的候选补充和区域语义校准三个部分，细节如图3所示。

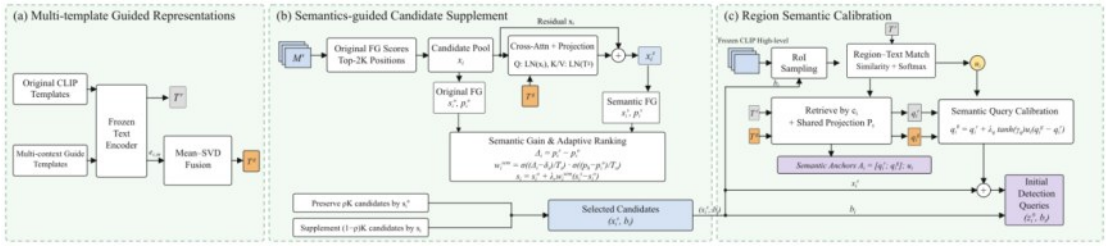



图3　SQE模块细节图

#### 1.2.1　多模板引导表示

候选区域中目标的占比和完整程度会随目标尺度及定位结果发生变化，同一类别的目标可能呈现常规视图、局部放大、小尺度、部分可见或物体统称等不同状态，这些状态可由目标在候选框中的占比与可见比例两个变量刻画。通用模板生成的类别表示难以充分体现上述差异，为此，SQE在保留原始类别表示的同时，利用描述不同目标尺度和完整程度的文本模板构建引导类别表示。从语义信息的角度看，语义信息的本质在于含义的同一性，同一语义可通过同义映射由不同符号形式表达[22-23]。为同一类别设置多个文本模板，即为该类别构建在不同描述下的同义映射集合，使类别表示不依赖于单一句式；而每个模板对应上述一种目标状态，模板集合因而构成对该状态空间的覆盖设计，而非任意枚举。模板之间的语义冗余度可通过各模板提示特征的两两相似度矩阵、或模板所传递语义信息的互信息占比度量。根据语义信息理论，随着同义映射数量增加，集合所覆盖的语义熵单调不减，而新增映射的边际语义增益递减；当模板所覆盖的目标状态维度有限时，语义熵随模板数量增加趋于饱和，因此无需枚举大量模板，采用覆盖主要状态的小规模模板集合即可，为后续候选补充和区域语义校准提供补充语义。

对于类别词表$C$中的类别$c$，本文采用“a photo of {c}”，“a close-up {c}”，“a small {c}”，“a partial {c}”和“{c} object”等$M$个模板，分别提供常规视图、局部放大、小尺度、部分可见和目标物体等描述。经冻结的CLIP文本编码器处理后，得到对应的提示特征$\{e_{c,m}\}_{m=1}^{M}$。对类别$c$在各模板下的提示特征取平均，得到该类别的平均表示$e_c=\frac{1}{M}\sum_{m=1}^{M}e_{c,m}$，以综合不同模板提供的类别信息。随后，将提示特征组成矩阵并执行奇异值分解，取第一右奇异向量$v_c$表示模板集合中的主要语义方向，并按照平均表示对其符号进行对齐。符号对齐的准则为$v_c \leftarrow sgn(\langle v_c,e_c\rangle)\,v_c$，即当第一右奇异向量$v_c$与平均表示$e_c$的内积为负时，将$v_c$整体翻转，以

保证引导表示的主语义方向与平均表示一致。最终，引导类别表示定义为：

$$t_c^g = Norm\left(\alpha e_c + (1 - \alpha) v_c\right) \#(6)$$

其中，$\alpha$ 表示平均类别信息与主要语义方向的融合系数，$Norm$ 表示 $L_2$ 归一化。将平均表示与主要语义方向融合，得到综合不同模板描述的引导类别表示。原始类别表示 $t_c^r$ 采用检测器原有的 CLIP 类别表示，并在 SQE 中保持不变。

所有类别的原始表示与引导表示分别组成 $T^r$ 和 $T^g$。其中，$T^r$ 用于区域类别匹配和最终开放词汇分类，$T^g$ 为候选补充与区域语义校准提供与目标尺度和完整程度相关的补充语义。

### 1.2.2 语义引导的候选补充

设最终解码查询数量为 $K$。SQE 利用 CAFE 增强后的编码特征预测前景得分，选取前景得分最高的 $2K$ 个编码位置构建初始候选池，使部分原本排在前 K 名之外的候选也能参与后续筛选，同时将语义计算限制在这 $2K$ 个候选内。将第 $i$ 个候选的特征和原始前景得分分别记为 $x_i$ 和 $s_i^o$。对于候选池中的每个候选，以其特征作为 Query，以引导类别表示 $T^g$ 作为 Key 和 Value 执行交叉注意力，通过残差更新得到语义增强后的候选特征：

$$x_i^s = x_i + \lambda_s \tanh(\gamma_s)\varphi\left[MHA\left(LN(x_i), LN(T^g), LN(T^g)\right)\right], \#(7)$$

其中，$\varphi(\cdot)$ 表示特征映射，$\lambda_s$ 控制语义残差的作用强度。$\gamma_s$ 为可学习残差系数。

根据增强后的候选特征预测前景得分 $s_i^s$，并将其与原始前景得分 $s_i^o$ 分别转换为对应的前景概率 $p_i^s$ 和 $p_i^o$。将引入引导类别信息前后的前景概率之差定义为语义增益，即 $\Delta_i = p_i^s - p_i^o$。

语义增益越大，表示引入引导类别信息后，该候选的前景概率提升越明显。为使原始前景得分较低、语义增益较大的候选获得更多补充机会，SQE 结合语义增益与原始前景概率计算候选补充权重：

$$w_i^{\text{sem}} = \sigma\left(\frac{\Delta_i - \delta_s}{T_s}\right)\sigma\left(\frac{p_h - p_i^o}{T_o}\right), \#(8)$$

其中，$\delta_s$ 为语义增益阈值，$p_h$ 为前景概率参考值，$T_s$ 和 $T_o$ 为温度系数，分别控制权重随语义增益和原始前景概率变化的平滑程度。在其他条件相同时，语义增益越大、原始前景概率越低，候选补充权重越大。利用该权重调整语义增强前后得分差异对候选排序的影响，得到排序得分：

$$\hat{s}_i = s_i^o + \lambda_r w_i^{\text{sem}}\left(s_i^s - s_i^o\right), \#(9)$$

其中，$\lambda_r$ 控制排序调整的幅度。候选补充权重较小时，$\hat{s}_i$ 接近原始前景得分；权重增大时，语义增强后的前景得分对排序的影响相应增强。

最终，SQE 保留原始前景得分最高的 $\rho K$ 个候选，再从剩余候选中按排序得分 $\hat{s}_i$ 由高到低补充 $(1 - \rho)K$ 个候选。其中，$\rho$ 表示最终 $K$ 个候选中按原始前景得分直接保留的比例。保留部分原始高分候选，可避免语义排序完全替代原始前景排序。最终候选的特征用于构建查询并预测初始参考框。

### 1.2.3 区域语义校准

完成候选补充后，SQE 通过区域类别匹配确定候选对应的类别语义，并根据类别区分度校准语义查询。对于选中的候选框 $b_i$，从冻结的 CLIP 高层视觉特征中采样得到区域表示 $r_i$，以保留 CLIP 原有的开放类别语义。计算区域表示与各类别原始表示的相似度，经温度缩放和 Softmax 得到类别概率，并将概率最高的类别作为候选区域的匹配类别 $c_i$。

为衡量区域类别匹配结果的区分程度，将最高类别概率 $p_i^1$ 与次高类别概率 $p_i^2$ 的差值定义为类别区分度，即 $d_i = p_i^1 - p_i^2$。$d_i$ 越小，表示概率最高的两个类别越难区分。SQE 据此设置语义校准权重，调节后续查询校准的强度：

$$u_i = \sigma\left(\frac{\delta_m - d_i}{T_m}\right), \#(10)$$

其中，$\delta_m$ 为类别区分度参考值，$T_m$ 为温度系数，控制校准权重随类别区分度变化的平滑程度。类别区分度越小，校准权重 $u_i$ 越大。根据匹配类别 $c_i$，分别从 $T^r$ 和 $T^g$ 中取出对应的原始类别表示与引导类别表示，并通过文本特征投影 $P_t$ 映射至查询空间，得到原始语义查询 $q_i^r = P_t(t_{c_i}^r)$ 和引导语义查询 $q_i^g = P_t(t_{c_i}^g)$。以原始语义查询为基础，利用两类语义查询的差异进行残差更新，并通过 $u_i$ 调节校准强度：

$$q_i^0 = q_i^r + \lambda_q \tanh(\gamma_q) u_i\left(q_i^g - q_i^r\right), \#(11)$$

其中，$\lambda_q$ 控制语义校准的作用强度，$\gamma_q$ 为可学习的残差系数。在其他条件相同时，类别区分度较低的候选具有更大的校准权重，两类语义查询的差异对更新的影响也更大；类别区分度较高时，校准作用减弱，查询更多保留原始语义查询的信息。

将校准后的语义查询 $q_i^0$ 与选中候选的特征 $x_i^s$ 相加，得到初始查询内容 $z_i^0 = x_i^s + q_i^0$。候选框 $b_i$ 作为

初始参考框，与查询内容共同送入解码器。同时，原始语义查询与引导语义查询组成语义锚点 $A_i = [q_i^r, q_i^g]$，并与语义校准权重 $u_i$ 一同传递至SP。语义锚点在后续解码中保持不变，为SP根据当前查询状态更新类别语义提供参考。训练阶段，与开放世界伪标签匹配的查询继续采用统一的object表示，不参与SP中的具体类别语义更新。

### 1.3　语义保持

SQE在解码前完成候选区域的类别匹配，并构建包含类别语义的初始查询。在多层解码过程中，查询表征随图像特征的融合不断更新，而类别文本语义只存在于初始查询中，查询在每层都与图像特征做交叉注意力，其中的文本成分随层数增加被不断稀释，训练监督又以基类标注为主，使解码前建立的区域—文本联系逐渐减弱，新类语义尤其容易丢失。为此，SP将SQE生成的语义锚点传递至各解码层，根据当前查询状态聚合类别语义，并通过门控残差逐层更新查询，使类别语义持续参与查询更新，保持区域—文本联系。该模块包括语义锚点聚合和逐层语义更新两个部分，细节如图4所示。

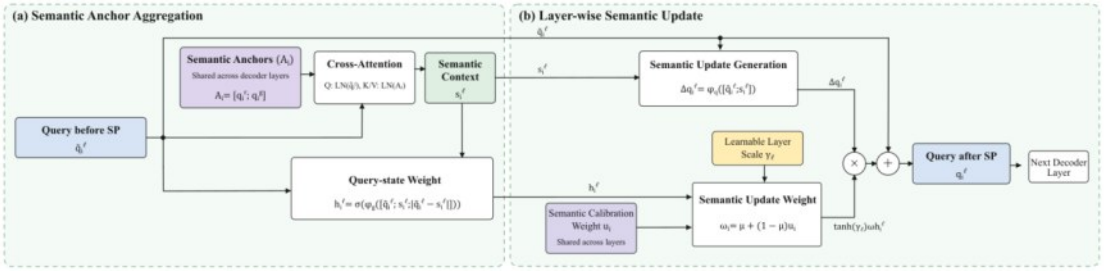



图4　SP模块细节图

#### 1.3.1　语义锚点聚合

对于第 $i$ 个检测查询，SQE生成的语义锚点 $A_i$ 由原始语义查询 $q_i^r$ 和引导语义查询 $q_i^g$ 组成。语义锚点在解码过程中保持不变，为各解码层提供一致的类别语义参照。

设第 $l$ 个解码层完成图像特征交互后、尚未经过SP更新的查询为 $\tilde{q}_i^l$。将当前查询和语义锚点分别归一化为 $\bar{q}_i^l$ 与 $\bar{A}_i$。SP以当前查询作为Query，并以该查询对应的语义锚点作为Key和Value，通过多头交叉注意力得到当前层的语义上下文：

$$s_i^l = MHA\left(\bar{q}_i^l, \bar{A}_i, \bar{A}_i\right), \#(12)$$

各层根据当前查询与两类语义查询的相关性确定注意力权重，从而调整原始类别信息与引导类别信息的聚合比例。为进一步调节语义信息对当前层查询更新的作用，SP将归一化查询、语义上下文及二者的逐维绝对差值进行拼接，计算查询状态权重：

$$h_i^l = \sigma\left(\phi_g\left(\left[\bar{q}_i^l; s_i^l; \left|\bar{q}_i^l - s_i^l\right|\right]\right)\right), \#(13)$$

其中，$\phi_g$ 表示查询状态映射，$[\cdot;\cdot]$ 表示特征拼接。查询状态权重 $h_i^l$ 根据当前查询与语义上下文共同计算，用于调节该层语义更新的强度。

#### 1.3.2　逐层语义更新

SP复用SQE计算的语义校准权重 $u_i$，将查询的语义更新系数设为 $\omega_i = \mu + (1-\mu)u_i$，其中 $\mu$ 为该系数的下限。类别区分度较低的候选对应较大的 $u_i$ 和 $\omega_i$，二者根据SQE阶段的区域类别匹配结果确定，在解码过程中保持不变。当候选类别区分度较高、初始类别匹配较为明确时，语义校准权重 $u_i$ 较小，对应的 $\omega_i$ 也随之降低；若不加约束，$u_i$ 接近零时 $\omega_i$ 也接近零，使该候选在后续各解码层获得的语义更新过小，语义保持近乎失效。$\mu$ 保证 $\omega_i$ 始终不低于 $\mu$，使每个候选在各解码层都获得最低限度的语义更新。每层实际注入多少语义，再由查询状态权重 $h_i^l$ 和该层残差系数根据当前层状态自适应调节。

对于当前解码层，将归一化查询与语义上下文进行拼接，经特征映射得到语义更新量：

$$\Delta q_i^l = \phi_q\left(\left[\bar{q}_i^l; s_i^l\right]\right), \#(14)$$ 其中，$\phi_q$ 表示语义更新映射。结合查询状态权重 $h_i^l$、语义更新系数 $\omega_i$ 和第 $l$ 个解码层的残差系数，对查询进行更新：

$$q_i^l = \tilde{q}_i^l + \tanh(\gamma_l)\omega_i h_i^l \Delta q_i^l, \#(15)$$

其中，$\omega_i$ 保留候选在SQE阶段的类别区分度信息，$h_i^l$ 根据当前查询状态调整语义更新强度，$\gamma_l$ 为第 $l$ 个解码层独立的可学习残差系数，用于控制该层的残

差强度。更新后的查询用于当前层的类别预测与边界框更新，并作为下一解码层的输入。通过在各层查询更新中持续引入固定锚点所提供的类别语义，SP 将 SQE 建立的区域—文本联系延续至后续解码过程。

### 1.4　损失函数

模型的主要优化目标包括分类损失、边界框定位损失和类别无关前景约束损失。分类与定位损失监督检测查询的预测结果，类别无关前景约束则为 CAFE 中的目标 Token 提供训练指导。分类损失采用 Varifocal Loss，定义为：

$$L_{\text{cls}} = \frac{1}{N_m}\sum_i a_i VFL(p_i, q_i), \#(16)$$

其中，$N_m$ 表示匹配监督框的数量，$p_i$ 表示第 $i$ 个查询的类别预测概率，$q_i$ 为对应的分类监督目标，$a_i$ 为分类权重。对于与监督框匹配的查询，其匹配类别的监督目标取预测框与监督框之间的交并比；对于未匹配查询，$q_i = 0$。

真实标注与开放世界伪标签均参与分类监督，其中开放世界伪标签统一使用类别无关的 object 标签。对于与开放世界伪标签匹配的查询，分类权重设为 $a_i = \sqrt{\xi_i}$，$\xi_i$ 为对应伪标签的置信度；对于与真实标注匹配的查询及未匹配查询，$a_i = 1$。该加权方式减小了较低置信度伪标签对分类损失的贡献。

边界框定位采用 L1 损失和广义交并比 GIoU 损失：

$$L_{\text{reg}} = \lambda_{\text{L1}} L_{\text{L1}} + \lambda_{\text{GIoU}} L_{\text{GIoU}}, \#(17)$$

其中，$L_{\text{L1}}$ 衡量预测框与真实框的坐标差异，$L_{\text{GIoU}}$ 衡量二者的重叠及空间关系，$\lambda_{\text{L1}}$ 和 $\lambda_{\text{GIoU}}$ 分别为两项损失的权重。定位损失仅根据真实标注计算，开放世界伪标签不参与边界框回归，以避免将可能不准确的伪标签位置作为回归目标。

类别无关前景约束损失 $L_{\text{CAFE}}$ 采用第 1.1.3 节的定义，利用真实标注与高置信度开放世界伪标签，引导目标 Token 区分潜在目标与背景，为编码特征增强提供训练指导。上述主要优化目标表示为：

$$L_{\text{main}} = \lambda_{\text{cls}} L_{\text{cls}} + L_{\text{reg}} + \lambda_{\text{CAFE}} L_{\text{CAFE}}, \#(18)$$

其中，$\lambda_{\text{cls}}$ 和 $\lambda_{\text{CAFE}}$ 分别为分类损失与类别无关前景约束损失的权重。SQE 和 SP 不引入额外的语义监督，其可学习参数随检测训练联合优化。

## 2　实验与结果分析

### 2.1　数据集与评估设置

为评估 FESG-OVD 在不同类别规模和数据分布下的开放词汇检测能力，本文在 COCO[24] 和 LVIS[25] 两个主流基准数据集上进行实验。对于 COCO 数据集，本文遵循现有 OV-COCO 标准设置[1]，从 80 个类别中选取 65 类进行实验，并将其划分为 48 个基类和 17 个新类。训练阶段使用包含约 11.8 万张图像的训练集，仅保留基类的人工标注，新类只提供类别名称用于文本编码；测试阶段则使用包含 5000 张图像的验证集，并同时检测全部 65 个类别。

LVIS 数据集包含 1203 个类别，具有更加明显的长尾分布和样本稀疏问题。同时，数据集中存在大量语义相近且外观差异较小的细粒度类别，对模型目标识别与开放词汇泛化能力提出了更高要求。本文遵循现有 OV-LVIS 标准设置，将 337 个稀有类别定义为新类，其余 866 个常见和频繁类别定义为基类。训练阶段使用约 10 万张图像，并且仅保留基类的人工标注；测试阶段在包含约 2 万张图像的验证集上对全部 1203 个类别进行评估。通过在 COCO 和 LVIS 上的实验，分别验证 FESG-OVD 在常规类别划分与大规模长尾场景下的检测性能。

### 2.2　实验参数设置

所有实验均基于 PyTorch[26]、8×NVIDIA GeForce RTX 4090（24 GB）环境。为保证对比公平性，本文沿用 CORA[17] 的基本训练设置。COCO 实验分别采用 CLIP 预训练的 RN50 和 RN50×4[27] 作为视觉骨干，LVIS 实验采用经 CLIPSelf[28] 优化的 ViT-B/16 和 ViT-L/14[29] 作为视觉骨干。模型使用 AdamW[30] 优化器，初始学习率和权重衰减均设为 $1 \times 10^{-4}$，每块 GPU 的批次大小设为 4，总批次大小为 32。COCO 上的训练周期为 30，LVIS 上的训练周期为 35。训练过程采用混合精度计算，并使用衰减系数为 0.99996 的指数滑动平均模型进行评估。除特别说明外，其余参数均与 CORA 保持一致。

### 2.3　评价指标

本文采用平均精度均值（mean Average Precision，mAP）评价模型的检测性能。mAP 表示各类别平均精度的均值，其计算公式为：

$$\mathrm{mAP}=\frac{1}{|\mathcal{C}|}\sum_{c\in C}\mathrm{AP}_c,\#(19)$$

其中，$\mathcal{C}$表示参与评价的类别集合，$\mathrm{AP}_c$表示类别$c$的平均精度。

本文采用开放词汇目标检测的标准评价方式。对于COCO数据集，在IoU阈值为0.5时计算检测精度。根据参与评价的类别范围，分别报告新类精度$\mathrm{AP}_{50}^{\mathrm{Novel}}$、基类精度$\mathrm{AP}_{50}^{\mathrm{Base}}$以及全部类别精度$\mathrm{AP}_{50}^{\mathrm{All}}$。对于LVIS数据集，在IoU阈值从0.5至0.95的范围内计算平均精度。根据类别在训练集中的出现频次，分别报告稀有类、常见类和频繁类的检测精度，依次记为$\mathrm{AP}_r$、$\mathrm{AP}_c$和$\mathrm{AP}_f$，并使用AP表示全部类别的整体检测精度。其中，$\mathrm{AP}_r$主要反映模型对训练时未提供人工标注的新类的检测能力，$\mathrm{AP}_c$和$\mathrm{AP}_f$则用于衡量模型在基类上的检测性能。

### 2.4　消融实验

为验证FESG-OVD中各个模块的有效性及其对整体性能的贡献，本文在COCO数据集的标准开放词汇设置下进行了消融研究。实验以采用RN50×4视觉骨干的CORA为基线，分别引入并组合CAFE、SQE和SP，实验结果如表1所示。

**表1　消融实验结果**

| 方法 | $\mathrm{AP}_{50}^{\mathrm{Novel}}$ | $\mathrm{AP}_{50}^{\mathrm{Base}}$ | $\mathrm{AP}_{50}^{\mathrm{All}}$ |
|---|---|---|---|
| CORA | 41.7 | 44.5 | 43.8 |
| + CAFE | 42.3 | 46.7 | 45.5 |
| + SQE | 42.9 | 45.3 | 44.7 |
| + CAFE+ SQE | 43.1 | 48.9 | 47.4 |
| + SQE+ SP | 44.9 | 47.2 | 46.6 |
| + CAFE+ SQE+ SP | **45.7** | **50.1** | **48.9** |

SP依赖SQE生成的语义锚点，因此不设置单独的SP和CAFE+SP组合。

分别引入CAFE和SQE后，模型的新类检测精度均有所提升。引入CAFE后，$\mathrm{AP}_{50}^{\mathrm{Novel}}$由41.7%提升至42.3%，$\mathrm{AP}_{50}^{\mathrm{Base}}$由44.5%提升至46.7%，表明CAFE及其前景监督能够改善检测性能。单独引入SQE后，$\mathrm{AP}_{50}^{\mathrm{Novel}}$达到42.9%，较基线提高1.2个百分点，$\mathrm{AP}_{50}^{\mathrm{Base}}$达到45.3%。两种设置均取得性能提升，其中CAFE对基类的改善更为明显，SQE则取得了更高的新类检测精度。

同时引入CAFE与SQE时，$\mathrm{AP}_{50}^{\mathrm{Novel}}$、$\mathrm{AP}_{50}^{\mathrm{Base}}$和$\mathrm{AP}_{50}^{\mathrm{All}}$分别达到43.1%、48.9%和47.4%，均高于单独使用任一模块的结果。这说明，在增强编码特征的基础上，进一步利用类别语义进行候选补充与查询构建，能够继续改善检测性能。

由于SP需要SQE生成的语义锚点和语义校准权重，因此在SQE的基础上进一步引入SP后，$\mathrm{AP}_{50}^{\mathrm{Novel}}$达到44.9%，较单独使用SQE提高2.0个百分点，$\mathrm{AP}_{50}^{\mathrm{Base}}$和$\mathrm{AP}_{50}^{\mathrm{All}}$分别达到47.2%和46.6%。该组合在表1所列两模块设置中取得最高的新类检测精度，表明在查询构建之后，于多层解码中持续引入类别语义有助于提高新类检测性能。

最后，集成CAFE、SQE和SP构成完整的FESG-OVD。模型的$\mathrm{AP}_{50}^{\mathrm{Novel}}$、$\mathrm{AP}_{50}^{\mathrm{Base}}$和$\mathrm{AP}_{50}^{\mathrm{All}}$分别达到45.7%、50.1%和48.9%，在表1所列设置中均取得最优结果。相较于CAFE与SQE的组合，进一步加入SP使三项指标分别提高2.6、1.2和1.5个百分点，说明SP在引入前景增强和语义查询增强后仍能带来性能收益。上述结果验证了三个模块组合的有效性。

### 2.5　对比实验

由于不同开放词汇检测方法所采用的骨干网络、监督方式与外部资源并不完全一致。为保证比较的客观性，本文按骨干网络参数量分组，RN50（约24M）、RN50×4（约87M）、ViT-B/16（约86M）和ViT-L/14（约307M）各为一组，仅在相同或相近骨干规模的方法之间直接比较。同时各方法的监督方式（如图像级标签、CLIP蒸馏、图像描述、CLIP类别伪标签、开放世界伪标签及基类标注）及所用外部资源在表2、表3中逐一标注；本文训练设置沿用CORA，对比方法结果均引自其在标准OV设置下的公开报告。

#### 2.5.1　COCO数据集对比实验

为全面评估FESG-OVD在开放词汇目标检测任务中的性能，本文在OV-COCO数据集的标准设置下，将其与不同监督方式和不同骨干规模的代表性方法进行比较。所有方法均采用48个基类和17个新类的类别划分，并报告$\mathrm{AP}_{50}^{\mathrm{Novel}}$、$\mathrm{AP}_{50}^{\mathrm{Base}}$和$\mathrm{AP}_{50}^{\mathrm{All}}$，实验结果如表2所示。

首先，与采用RN50骨干和不同监督方式的代表性方法进行比较。基于知识蒸馏的ViLD[14]和BARON[19]分别取得27.6%和34.0%的$\mathrm{AP}_{50}^{\mathrm{Novel}}$；使用图像描述监督的RegionCLIP[15]取得26.8%；基于CLIP伪标签的VL-PLM[8]和MarvelOVD[10]分别达到32.3%和38.7%。本文前期工作GLCS-Net[9]通过上下文增强和伪标签净化将$\mathrm{AP}_{50}^{\mathrm{Novel}}$提升至40.5%。本

**表2　COCO数据集对比实验结果**

| 方法 | 主干网络(参数) | 监督方式 | $AP_{50}^{Novel}$ | $AP_{50}^{Base}$ | $AP_{50}^{All}$ |
|---|---|---|---|---|---|
| ViLD[22] | RN50(24M) | CLIP蒸馏 | 27.6 | 59.5 | 51.3 |
| BARON[23] | RN50(24M) | CLIP蒸馏 | 34.0 | 60.4 | 53.5 |
| RegionCLIP[22] | RN50(24M) | 图像描述 | 26.8 | 54.8 | 47.5 |
| VL-PLM[22] | RN50(24M) | CLIP伪标签 | 32.3 | 54.0 | 48.3 |
| MarvelOVD[24] | RN50(24M) | CLIP伪标签 | 38.7 | 56.9 | 52.1 |
| GLCS-Net[26] | RN50(24M) | CLIP伪标签 | 40.5 | 57.2 | 53.0 |
| BARON-BFDet[26] | RN50(24M) | 基类标注与场景伪标签 | 37.8 | 58.2 | 52.0 |
| HCC+LoCLIP[26] | RN50(24M) | 基类标注与新类伪标签 | 38.9 | 59.5 | - |
| SSN-OVD[26] | RN50(24M) | 基类标注与开放世界伪标签 | 39.7 | 46.5 | 45.8 |
| **FESG-OVD (Ours)** | RN50(24M) | 基类标注与开放世界伪标签 | **40.0** | **43.2** | **42.4** |
| CORA[23] | RN50x4(87M) | 基类标注与CLIP语义监督 | 41.7 | 44.5 | 43.8 |
| NoOVD[26] | ViT-B/16 (86M) | 基类标注与CLIP语义监督 | 41.9 | 58.3 | 54.0 |
| HCC+LoCLIP[26] | ViT-L/14 (307M) | 基类标注与新类伪标签 | 44.0 | 65.8 | - |
| **FESG-OVD (Ours)** | RN50x4(87M) | 基类标注与开放世界伪标签 | **45.7** | **50.1** | **48.9** |

方法上角标为所提模型发表年份，如BARON[23]为2023年发表;表中监督方式含义为:CLIP蒸馏指以CLIP为教师对区域特征进行蒸馏;图像描述指利用图像描述生成区域级训练数据;CLIP伪标签指由CLIP分类结果生成类别伪标签;场景伪标签、新类伪标签与开放世界伪标签分别表示由场景上下文、类别一致性和开放世界检测器生成的伪标签;本文及SSN-OVD所用开放世界伪标签仅作为类别无关的通用物体监督。

文方法FESG-OVD取得40.0%的$AP_{50}^{Novel}$，与GLCS-Net保持相近的新类检测性能，并分别超过ViLD、BARON、RegionCLIP、VL-PLM和MarvelOVD 12.4、6.0、13.2、7.7和1.3个百分点。该结果表明，FESG-OVD在RN50骨干下仍具有较强的新类泛化能力。

其次，与近期采用RN50骨干的方法进行比较。BARON-BFDet[31]利用场景中的前景与背景信息生成补充监督，其$AP_{50}^{Novel}$为37.8%；HCC+LoCLIP[32]通过类别层级一致性和前景评分筛选新类伪标签，将该指标提升至38.9%。相比之下，FESG-OVD的$AP_{50}^{Novel}$达到40.0%，分别超过上述方法2.2和1.1个百分点。这表明FESG-OVD在相同骨干规模下能够有效提升新类目标的检测性能，并在近期方法中保持竞争力。

最后，在RN50×4骨干下，FESG-OVD取得45.7%的$AP_{50}^{Novel}$，50.1%的$AP_{50}^{Base}$和48.9%的$AP_{50}^{All}$。与直接相关的CORA相比，FESG-OVD的三项指标分别提高4.0、5.6和5.1个百分点。进一步与基于Transformer骨干的方法比较，NoOVD[33]采用ViT-B/16获得41.9%的$AP_{50}^{Novel}$，HCC+LoCLIP采用参数量更大的ViT-L/14获得44.0%，FESG-OVD仍分别领先3.8和1.7个百分点。由此可见，FESG-OVD在未进一步扩大骨干规模的情况下取得了表中最高的新类检测精度，符合开放词汇目标检测重点提升新类泛化能力的任务目标，并在相同骨干设置下保持了良好的基类与整体检测性能。

### 2.5.2　LVIS数据集对比实验

为验证FESG-OVD在大规模类别空间与长尾分布场景下的泛化能力，本文在LVIS数据集上进行了对比实验，并分别采用ViT-B/16和ViT-L/14作为视觉主干。该数据集包含1203个视觉类别。实验将稀有类别设为新类，并在训练时移除其人工标注，以检验模型对缺少人工标注类别的检测能力。实验结果如表3所示。

在ViT-B/16设置下，FESG-OVD在稀有类别指标$AP_r$上取得30.1%的平均精度，达到该组最优结果。相比SSN-OVD[34]和NoOVD，分别提升0.9和1.8个百分点；相较CLIPSelf的25.3%，提升幅度达到4.8个百分点。同时，FESG-OVD在$AP_c$、$AP_f$和整体AP上分别取得27.7%、23.9%和26.6%。上述结果表明，FESG-OVD在较小视觉主干下能够有效提升稀有类别的检测精度，展现出良好的新类泛化能力。

在ViT-L/14设置下，FESG-OVD的$AP_r$进一步达到39.3%，同样取得该组最优结果。相比CLIP-Self、NoOVD和SSN-OVD，分别提升4.4、1.5和1.9个百分点。同时，FESG-OVD在$AP_c$、$AP_f$和整体AP上分别达到35.2%、28.5%和33.3%。在更大规模视觉主干下，FESG-OVD仍然保持了稀有类别

**表3　LVIS数据集对比实验结果**

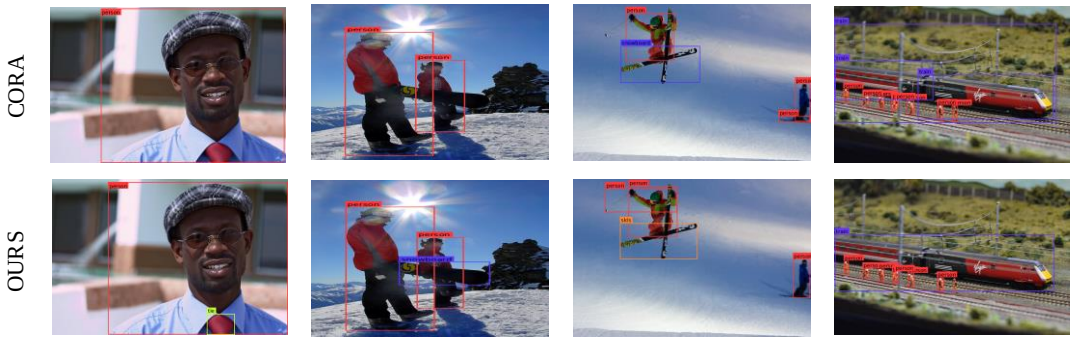
| 方法 | 主干网络(参数) | $AP_r$ | $AP_c$ | $AP_f$ | AP |
|---|---|---|---|---|---|
| RegionCLIP[22] | RN50x4(87M) | 22.0 | 32.1 | 36.9 | 32.3 |
| GLCS-Net[26] | RN50(24M) | 25.4 | 32.9 | 37.7 | 33.5 |
| CORA[23] | RN50x4(87M) | 28.1 | - | - | - |
| CLIPSelf[24] | ViT-B/16 (86M) | 25.3 | 21.8 | 29.1 | 25.2 |
| NoOVD[26] | ViT-B/16 (86M) | 28.3 | 22.6 | 30.7 | 26.7 |
| SSN-OVD[26] | ViT-B/16 (86M) | 29.2 | 27.8 | 25.6 | 27.6 |
| **FESG-OVD (Ours)** | ViT-B/16 (86M) | 30.1 | 27.7 | 23.9 | 26.6 |
| CLIPSelf[24] | ViT-L/14 (307M) | 34.9 | 34.6 | 35.6 | 35.1 |
| NoOVD[26] | ViT-L/14 (307M) | 37.8 | 35.9 | 37.2 | 36.7 |
| SSN-OVD[26] | ViT-L/14 (307M) | 37.4 | 36.0 | 28.5 | 33.7 |
| **FESG-OVD (Ours)** | ViT-L/14 (307M) | **39.3** | **35.2** | **28.5** | **33.3** |

检测优势，说明其能够适应不同规模的视觉特征，并持续提高新类目标的检测性能。

综合来看，FESG-OVD在ViT-B/16和ViT-L/14两种设置下的$AP_r$分别达到30.1%和39.3%，均为表3中对应组的最高结果，表明该方法在两种视觉骨干下均具有稀有类别检测优势。与此同时，模型在频繁类别和整体AP上仍低于部分对比方法。例如，在ViT-L/14设置下，FESG-OVD的频繁类精度与SSN-OVD持平，整体AP则低于SSN-OVD和NoOVD。因此，该方法当前的性能优势主要体现在稀有类别，各类别组之间的性能均衡仍有改进空间。

## 2.6　可视化实验

为直观分析FESG-OVD在开放词汇场景中的目标发现、类别判别和定位能力，本文将其与CORA[17]进行可视化比较。图中第一行为CORA的检测结果，第二行为FESG-OVD的检测结果，边界框上方标注了相应的预测类别，结果如图5所示。




（a）对比1　（b）对比2　（c）对比3　（d）对比4

图5　可视化对比结果图

图5（a）和图5（b）展示了部分可见及遮挡目标的检测结果。在图5（a）中，CORA检测出了人物，但遗漏了位于图像边缘且仅部分可见的领带，FESG-OVD则同时检测出了人物和领带。在图5（b）中，单板滑雪板（snowboard）与人物重叠，部分区域被人体遮挡。CORA仅检测出人物，FESG-OVD进一步检测出了单板滑雪板。这两组结果展示了FESG-OVD对容易被忽略目标的检测能力。

图5（c）展示了外观相近类别的判别结果。CORA将双板滑雪板（skis）错误识别为单板滑雪板，FESG-OVD则给出了正确的类别预测，减少了该样例中的类别混淆。

图5（d）展示了列车的检测框预测结果。CORA对同一列车产生了多个重叠且覆盖范围不准确的检测框，FESG-OVD则以单一检测框较完整地覆盖列车，减少了重复检测，并改善了检测框的覆盖范围。

综合来看，前两组结果体现了FESG-OVD对小尺度、遮挡及不完整目标的发现能力，第三组结果展示了其对相似类别的判别能力，第四组结果则反映了定位稳定性的改善。上述结果表明，FESG-OVD能够从候选发现、语义判别和边界框定位三个方面提升开放词汇目标检测性能。

### 2.7　模型局限性分析

为客观评估FESG-OVD的适用范围，本文结合实验结果，从视觉语言模型知识迁移和类别性能均衡两个方面分析模型的局限性，并讨论后续研究方向。

本文的开放语义知识主要来源于CLIP，其语义向区域级检测的迁移主要通过两条路径实现：一是文本路径，即CLIP文本编码器输出的类别文本表示直接用于区域类别匹配；二是视觉路径，即通过区域特征采样与蒸馏，将CLIP整图级对比语义迁移到区域级。然而，这种迁移存在粒度失配问题：CLIP在整图级图像—文本对比任务上训练，而区域级检测以局部候选区域为推理对象，当目标尺度较小或仅部分可见时，区域特征与CLIP整图级语义的对齐效果下降，容易造成漏检或类别混淆。此外，训练中CLIP保持冻结，在细粒度类别、长尾类别等特定场景下可能存在域偏移。尽管CAFE的类别无关监督、SQE的多模板引导表示和SP的逐层语义注入在一定程度上缓解了上述问题，但整图级语义向区域级迁移的粒度失配和冻结表示的域偏移无法完全消除。

在类别性能均衡方面，本文与NoOVD代表了对预训练CLIP的两种不同处理方式。NoOVD倾向于保持CLIP的原始能力，它冻结CLIP，由不含可学习参数的K-FPN直接构建检测特征金字塔，并在R-RPN中明确保证基类在K-FPN与RPN两个分支都维持高置信度，因此其能力分布基本继承自CLIP；而频繁类在CLIP海量图文预训练中样本最充分、语义最稳定，这一方法自然在频繁类和整体精度上更占优势。本文则采取了更主动的策略：针对CLIP检测框架中新类前景被当作背景、稀有类召回不足的薄弱环节，通过类别无关的前景监督与开放世界伪标签主动重构特征空间，聚焦被抑制的稀有类，这也是本文在稀有类上达到39.3%、反超NoOVD的37.8%的原因；但这种主动重构会改变CLIP原本均衡的分布，相应地扰动了频繁类，使其精度为28.5%。

从任务本质看，这一取舍是有必要的。频繁类在训练中样本最充足，而且已被CLIP充分学习，属于能力冗余区，但是稀有类在训练中缺乏框标注，是CLIP检测的盲区，也正是开放词汇检测需要解决的核心问题，本文将有限的模型容量优先投向这一目标。需要说明的是，该权衡并非普遍现象：在COCO设置下，本文的基类与整体精度相对基线全面提升，说明它主要出现在大规模长尾、类别无关监督占比较高的场景。与此同时，本文在LVIS上的整体AP仍低于NoOVD约3.4个百分点，说明当前取舍尚偏激进，如何在保持稀有类优势的同时恢复整体均衡，是后续研究的重点。

综上，FESG-OVD在常规开放词汇场景下能够有效检测新类目标，但在区域级语义对齐和类别组性能均衡方面仍有改进空间。未来工作可从三个方面展开：一是研究区域感知的提示学习与轻量微调，增强视觉语言模型向区域级检测的语义适配，缓解粒度失配与域偏移；二是研究类别平衡的伪标签筛选与损失加权，在提升稀有类性能的同时保持频繁类性能；三是结合多尺度特征增强与细粒度建模，提高模型对小目标、部分可见目标及外观相近类别的检测能力。

## 3　结论

本文针对开放词汇目标检测中新类目标前景监督不足、通用类别表示难以充分体现候选区域中目标尺度和完整程度差异，以及多层解码中区域—文本联系可能减弱的问题，提出前景增强与语义引导的开放词汇目标检测方法FESG-OVD。CAFE利用开放世界伪标签提供的通用物体信息补充前景监督，减少具体类别噪声的直接干扰，并通过目标Token聚合多尺度目标结构、引导编码特征增强，以减少潜在目标在候选生成阶段的遗漏。SQE在保留原始CLIP类别表示的基础上，利用多模板引导类别表示补充候选并校准语义查询，增强文本语义在候选筛选和查询构建中的作用。SP利用SQE生成的语义锚点，根据各层查询状态更新查询，将这种语义引导延续至多层解码过程。实验结果表明，FESG-OVD在COCO数据集上的新类检测指标$AP_{50}^{Novel}$达到45.7%，在LVIS数据集上的稀有类检测指标$AP_r$达到39.3%，验证了该方法在新类目标检

测中的有效性。该方法为缓解基类标注覆盖有限对物体特征学习的限制、促进CLIP语义知识由整图级预训练向区域级检测的迁移提供了思路。

## 参考文献：


[1] ZAREIAN A, ROSA K D, HU D H, et al. Open-vocabulary object detection using captions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14393-14402.

[2] 聂秀山, 赵润虎, 宁阳, 等. 开放词汇目标检测方法综述[J]. 山东大学学报(工学版), 2025, 55(1): 1-14.
NIE X S, ZHAO R H, NING Y, et al. Survey of open vocabulary object detection methods[J]. Journal of Shandong University(Engineering Science), 2025, 55(1): 1-14.

[3] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

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

[6] 单飞龙, 吕鹏远, 李梦晨. 基于CNN-Transformer半监督交叉学习的遥感图像场景分类方法[J]. 宁夏大学学报(自然科学版), 2024, 45(3): 325-332.
SHAN F L, LV P Y, LI M C. Remote sensing image scene classification based on semi-supervised cross learning of CNN and transformer[J]. Journal of Ningxia University (Natural Science Edition), 2024, 45(3): 325-332.

[7] ZHOU X, GIRDHAR R, JOULIN A, et al. Detecting twenty-thousand classes using image-level supervision[C]//European Conference on Computer Vision. Cham: Springer, 2022: 350-368.

[8] ZHAO S, ZHANG Z, SCHULTER S, et al. Exploiting unlabeled data with vision and language models for object detection[C]//European Conference on Computer Vision. Cham: Springer, 2022: 159-175.

[9] 赵镛清，张若楠，等。全局 - 局部上下文协同的开放词汇目标检测 [J]. 物联网学报，录用待刊.
ZHAO Y Q, ZHANG R N, et al. Open vocabulary object detection via global-local context synergy[J]. Chin. J. Internet Things, in press.

[10] WANG K, CHENG L C, CHEN W K, et al. MarvelOVD: marrying object recognition and vision-language models for robust open-vocabulary object detection[C]//European Conference on Computer Vision. Cham: Springer, 2024: 106-122.

[11] ZHANG H L, GUAN D Y, KE X R, et al. Open-vocabulary object detection via debiased curriculum self-training[J]. Expert Systems with Applications, 2024, 255: 124762.

[12] XU S L, LI X T, WU S Z, et al. DST-Det: open-vocabulary object detection via dynamic self-training[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(5): 5037-5050.

[13] 刘立波, 郗思宇, 邓箴. 结合改进ConvNeXt网络与知识蒸馏的天气识别[J]. 光学精密工程, 2023, 31(14): 2123-2134.
LIU L B, XI S Y, DENG Z. Weather recognition combining improved ConvNeXt models with knowledge distillation[J]. Opt. Precision Eng., 2023, 31(14): 2123-2134.

[14] GU X, LIN T Y, KUO W, et al. Open-vocabulary object detection via vision and language knowledge distillation[C]//The Tenth International Conference on Learning Representations. 2022: 1-21.

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

[16] FU S H, YAN J K, YANG Q Z, et al. A hierarchical semantic distillation framework for open-vocabulary object detection[J]. IEEE Transactions on Multimedia, 2025, 27: 8835-8846.

[17] WU X S, ZHU F, ZHAO R, et al. CORA: adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7031-7040.

[18] WANG H, REN P Z, JIE Z Q, et al. OV-DINO: unified open-vocabulary detection with language-aware selective fusion[J]. arXiv preprint arXiv:2407.07844, 2024.

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

[20] CHEN Y, WANG C, LI Z H, et al. Enhancing open-vocabulary object detection through region-word and region-vision matching[J]. Multimedia Systems, 2025, 31(3): 232.

[21] 金友, 邓箴, 刘立波. 知识嵌入引导的双分支融合增强开放词汇目标检测[J]. 光学精密工程, 2025, 33(18): 2929-2943.
JIN Y, DENG Z, LIU L B. Open-vocabulary object detection enhanced by knowledge embedding-guided dual-branch fusion[J]. Opt. Precision Eng., 2025, 33(18): 2929-2943.

[22] 牛凯，张平。语义通信的数学理论 [J]. 通信学报，2024, 45 (6): 7-59.
NIU K, ZHANG P. A mathematical theory of semantic communication [J]. Journal on Communications, 2024, 45 (6): 7-59.

[23] NIU K, ZHANG P. The mathematical theory of semantic communication[M]. Singapore: Springer Nature, 2025.

[24] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[25] GUPTA A, DOLLÁR P, GIRSHICK R. LVIS: a dataset for large vocabulary instance segmentation[C]//Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition. 2019: 5351-5359.

[26] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems. 2019, 32: 8024-8035.

[27] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[28] WU S, ZHANG W W, XU L M, et al. CLIPSelf: vision transformer distills itself for open-vocabulary dense prediction[C]//The Twelfth International Conference on Learning Representations. 2024.

[29] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [C]//The Ninth International Conference on Learning Representations. 2021.

[30] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]//The Seventh International Conference on Learning Representations. 2019.

[31] LI Y Q, NIU J W, GU N B, et al. From scene to object: enhancing open-vocabulary object detection via foreground-background context reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2026, 40(8): 6601-6609.

[32] LEE S, LEE G, PARK H, et al. Exploring hierarchical consistency and unbiased objectness for open-vocabulary object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings. 2026: 6819-6828.

[33] ZHANG Y P, HAN R Z, CHEN Z W, et al. NoOVD: novel category discovery and embedding for open-vocabulary object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026: 6304-6313.

[34] 金友, 张若楠, 邓箴, 等. 慧眼识“新”开放词汇目标检测[J]. 光学精密工程, 2026, 34(5): 830-846.
JIN Y, ZHANG R N, DENG Z, et al. Sharp eyes spot the "novel" in open-vocabulary object detection[J]. Opt. Precision Eng., 2026, 34(5): 830-846.