文章编号　1004-924X(2025)18-2929-15

知识嵌入引导的双分支融合增强开放词汇目标检测

金　友[1]，邓　箴[1]，刘立波[1,2*]

(1. 宁夏大学 信息工程学院，宁夏 银川 750021；
2. 宁夏大学 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)

摘要：针对开放场景中检测器新类概念理解弱、标签混淆与新类检测性能不足的问题，提出了一种知识嵌入引导的双分支融合增强开放词汇目标检测(KI-DBFOVD)方法。首先，设计知识嵌入(KI)模块，利用视觉语言模型(VLM)生成的伪标签，嵌入检测器中以促进新类概念的学习；然后，提出标签匹配(LM)模块，通过多级阈值调节和基类-新类独立匹配细化标签匹配过程，缓解检测器训练过程中基类与新类标签混淆现象；最后，将传统视觉分支和视觉语言分支通过几何平均的方式进行融合，构建了一种新颖的双分支融合模块(DBF)，在保持基类检测精度的同时能够更有效地挖掘和定位新类目标，进一步提升了KI-DBFOVD方法整体的检测性能。实验结果表明，本文方法在COCO数据集上对新类的检测精度达到了38.6%，在类别更加繁多且检测难度更高的LVIS数据集上对新类取得了25.4%的检测精度，优于多个主流方法，能够更好地应用在不同的开放场景中。

关　键　词：开放词汇目标检测；知识嵌入；标签匹配；双分支融合

中图分类号：TP391.41　　**文献标识码**：A

doi：10.37188/OPE.20253318.2929　　**CSTR**：32169.14.OPE.20253318.2929

Knowledge integration guided dual-branch fusion enhanced open-vocabulary object detection

JIN You[1], DENG Zhen[1], LIU Libo[1,2*]

(1. *School of Information Engineering, Ningxia University, Yinchuan* 750021, *China*;
2. *Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West, Yinchuan* 750021, *China*)
* *Corresponding author, E-mail: liulib*@163.*com*

Abstract: To address the issues of weak understanding of new class concepts, label confusion, and insufficient detection performance of new classes in open-set scenarios, a Knowledge Integration-guided Dual-branch Fusion Open-Vocabulary Object Detection (KI-DBFOVD) method was proposed in this paper. Firstly, a Knowledge Integration (KI) module was designed, where pseudo-labels generated by a Vision-Language Model were embedded into the detector to learn about new class concepts. Subsequently, a Label Match (LM) module was introduced to refine the label matching process through multi-level threshold

收稿日期：2025-03-24；**修订日期**：2025-06-04.

基金项目：宁夏自然科学基金(No. 2024AAC02010)；国家自然科学基金(No. 62262053)；宁夏科技创新领军人才项目(No. 2022GKLRLX03)；2024年宁夏回族自治区重点研发计划(引才专项)项目(No. 2024BEH04026)

adjustment and independent matching between base and new classes, thereby alleviating the label confusion between base and new classes during detection. Finally, a novel Dual-branch Fusion module (DBF) was constructed by fusing the traditional visual branch and the vision-language branch via geometric averaging. This fusion maintained the detection accuracy of base classes and more effectively detected and localized new class objects, then enhanced the overall detection performance of the KI-DBFOVD method. Experimental results demonstrate that this method achieves a detection accuracy of 38.6% for new classes on the COCO dataset and 25.4% on the more challenging LVIS dataset, which contains a larger number of categories. These results outperform several mainstream methods and indicate that this approach is more suitable for different open-set scenarios. .

Key words: open-vocabulary object detection; knowledge integration; label match; dual-branch fusion

1 引　言

目标检测是计算机视觉的一项基础任务,已广泛应用于环境资源监测[1]、交通安全监控[2]、工业产品检测[3]等领域。传统方法[4-5]在封闭集场景下取得了显著的进展,但其依赖大量成本高昂的人工标注数据,并且对未标注新类的泛化能力受限。为此,Alireza等[6]于2021年首次提出开放词汇目标检测(Open-Vocabulary Object Detection,OVD)范式,通过结合视觉信息和自然语言处理技术,使模型能够检测训练过程中未曾标注的新类别[7]。目前,OVD主要通过基于大规模图像文本数据方法和基于预训练视觉语言VLM[8](Vision-Language Model)模型方法,深入挖掘图像中新类概念信息,提升模型对新类别的泛化能力。

近年来,基于大规模图像与文本数据的研究,逐渐形成了两条主要的研究路线:图像区域-文本对齐和伪标签文本对齐。在图像区域-文本对齐方面[9],Li等[10]提出RcFRCN,Li等[11]提出PDC-VLD以及Chen等[12]提出RWVM,将区域-文本对齐知识注入目标检测器,提升其新类概念理解能力。Wang等[13]提出SIA-OVD,通过形状不变适配器弥合图像-区域差距,提升开放词汇检测性能。然而,由于区域-文本对齐时,图像中常含大量背景和无关信息,且检测器缺乏新类概念的先验知识,导致其对基础类别存在偏好,限制了对新类概念的理解。与图像区域-文本对齐不同,伪标签文本对齐[14]则通过借助VLM,获取新类概念的先验知识注入检测器,以增强检测器的泛化能力。Lu等[15]提出SAN-YOLO模型,融合注意力机制与半监督学习,显著提升目标检测精度。Xu等[16]提出了Dst-det方法,利用VLM零样本分类能力直接发现潜在新类概念。Wang等[17]提出MarvelOVD,Zhang等[18]提出DCS与Xin等[19]提出DART通过挖掘高质量伪标签,提升检测性能。值得注意的是,该类方法输入裁剪图像区域生成伪标签时,丢失了重要的上下文信息,导致对新类概念的理解不深。此外,由于基类和新类之间存在相似性,检测器在训练时对基类和新类产生混淆,产生了标签错误匹配的情况,对总体检测性能产生了影响。

除此之外,基于预训练VLM模型方法也被应用于OVD任务中。Zhong等[20]提出的RegionCLIP,对VLM视觉编码器进行微调,蒸馏知识至检测器以适应不同检测任务。Wu等[21]提出的BARON通过随机掩码网格生成区域包集合,蒸馏视觉特征信息,使得模型学习丰富的新类概念。Gu等[22]提出的ViLD模型,以知识蒸馏的方式将图像和文本知识从VLM中教授到检测器学习。提示学习[23-25]通过提示引导挖掘隐式知识,提升OVD任务的效率。Xu等[26]提出MMC-Det框架,通过多模态上下文知识蒸馏提升OVD性能。Long等[27]提出VTP-OVD通过视觉-文本提示微调实现OVD的特征对齐。然而,VLM模型对新类概念的挖掘能力虽强,但其对目标的定位能力相对较差;而检测器有丰富上下文图像特征信息,且对目标的定位质量高,对基类的检测能力强却对新类概念挖掘能力弱。因此,如何充分结合两者之间的优势,对开放场景中新类检测而言至关重要。

综上所述,在OVD的研究中,现有方法因检

测器对新类概念理解能力不足、训练时基类与新类标签易混淆，以及 VLM 与检测器在视觉检测方面存在差距，导致检测效果欠佳。为此，本文提出知识嵌入引导的双分支融合增强开放词汇目标检测方法（Knowledge Integration guided Dual-branch fusion enhanced Open-Vocabulary Object Detection，KI-DBFOVD）。首先，设计知识嵌入（Knowledge Integration，KI）模块，使检测器学习有效的新类概念知识，增强其上下文信息有效性。其次，提出标签匹配（Label Match，LM）模块，采用多级阈值调节机制与基类-新类独立匹配策略，改善检测流程中标类与新类标签混淆问题。最后，引入一种新颖的双分支融合（Dual-branch Fusion，DBF）模块，其中，传统视觉分支保障基础类别检测性能，视觉语言分支结合 VLM 挖掘新类概念及目标检测器的高质量定位优势，提升新类检测效果，最终通过几何平均融合双分支预测结果，实现整体检测性能提升。

2　KI-DBFOVD 方法设计

2.1　预备知识

本研究遵循 OVD 标准框架设定，利用数据集 $S=\{v_i, a_i\}_{i=1}^{n}$ 以及辅助弱监督数据（例如视觉语义对齐、图像级注释等）训练模型，其中，v_i 表示图像数据，a_i 中包含图像中对象的位置和类别信息。OVD 与传统的目标检测任务不同之处在于，训练时图像的注释只有基类别 C^B 的数据信息，在测试时却需要额外检测新类 C^N。基类和新类二者之间互斥，即 $C^B \cap C^N = \varnothing$，基类的学习遵循检测器的常规回归和分类损失。

为了在开放词汇表标签空间中实现目标检测，目标检测器中的分类头被设计用于比较基于区域之间视觉嵌入和文本嵌入的相似性。区域嵌入 $R=\{r_i\}_{i=1}^{N_r}$ 是通过提取特征和 RoI head（Region of Interest）相应操作获得，其中 N_r 表示图像中区域框的数量。文本嵌入表示为 $T=\{t_i\}_{i=1}^{N_t}$，由类别名称与提示模板组成（例如，“a photo of a/an {category} in the scene.”）输入至 VLM 文本编码器获得，N_t 是类别的数量。基于线性层将区域特征投影的区域嵌入和文本嵌入计算为：

$$p(i,k)=\frac{\exp\langle r_i, t_k\rangle}{\sum_{k'\in N_t}\exp\langle r_i, t_{k'}\rangle}, \tag{1}$$

其中：r_i 为区域图像嵌入，t_k 为类别文本嵌入，$\langle\cdot\rangle$ 表示计算区域和文本嵌入之间的相似度，最终计算得到概率 $p(i,k)$，使检测器能够在开放词汇标签空间中识别对象。

2.2　整体结构

本文提出的 KI-DBFOVD 方法由 CNN 主干网络、知识嵌入模块、标签匹配模块、双分支融合模块构成，具体结构如图 1 所示。

由图 1 可知，首先，将检测图像裁切成区域，传入至 VLM 生成伪标签并构建动态词汇表，将新类通过 KI 模块让检测器学习，以此增强目标检测器对新类概念知识的理解；接着通过主干网络提取多尺度特征图（Feature Map），经过 RPN（Region Proposal Network）获取潜在目标的预测框；然后，由 LM 模块以多级阈值调节来分配预测框和标签，以此改善检测流程中对新类和基类目标之间标签匹配混淆的问题；最后，DBF 模块在保持基类性能的基础上，结合 VLM 新类概念

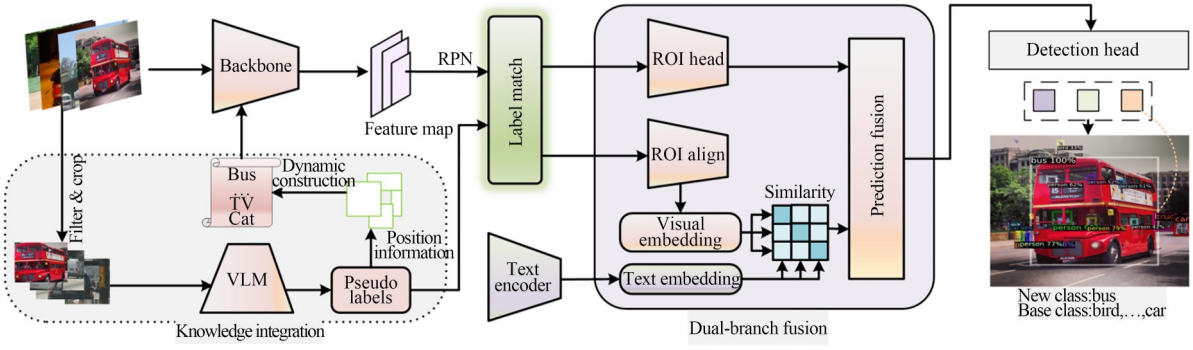


图 1　KI-DBFOVD 方法整体架构图

Fig. 1　Overall Architecture of KI-DBFOVD method

理解能力与检测器上下文特征信息增强对新类的检测效果，将视觉嵌入（Visual Embedding）和文本嵌入（Text Embedding）进行相似度计算，对每一个特征进行预测，通过几何平均给融合双分支预测结果。最终经过检测头（Detection Head）的分类、预测框回归和掩码损失协同优化检测器的分类和定位能力，实现输出高精度目标检测的结果。

2.2.1　知识嵌入模块

增强检测器对新类概念的理解能力是OVD的主要难题。本文提出由伪标签生成与检测器构建两部分构成的知识嵌入KI模块，利用伪标签建立动态词汇表对检测器进行构建，以提升检测器的新类检测能力。

2.2.1.1　伪标签生成

在研究工作VL-PLM中发现，两阶段探测器的RPN对新类别有很强的泛化性。因此，首先使用基类注释训练一个冻结的类别不可知RPN作为预测框生成器（Bounding box Generator，BG），为输入图像生成包含潜在目标的区域预测框。为了减轻计算负担，使用过滤器只选取k个区域预测框。之后，对候选区域进行裁切和缩放操作，以适配VLM视觉编码器（Visual Encoder）的图像输入尺寸要求，使其更聚焦于潜在目标，提升特征提取效果；然后，将处理后的候选区域图像被输入到VLM视觉编码器中生成视觉嵌入。与此同时，将相应包含新类的文本提示模板信息输入到VLM文本编码（Text Encoder）器中计算文本嵌入，计算视觉和文本嵌入之间相似度，计算VLM预测分数，如图2所示。

计算公式如下：

$$\begin{cases} v_i = E_{VE}\left(V_i^{1\times}\right) + E_{VE}\left(V_i^{1.5\times}\right) \in \boldsymbol{R}^{n\times d} \\ t_j = E_{TE}(\boldsymbol{x}) \in \boldsymbol{R}^{1\times d} \end{cases}, \tag{2}$$

其中：v_i是视觉嵌入，$V_i^{1\times}$是由BG生成的区域边界框，$V_i^{1.5\times}$是按$V_i^{1\times}$大小1.5倍的裁剪区域输入至VLM视觉编码器，t_j为文本嵌入VLM文本编码器。之后，使用v_i与t_j进行余弦相似度计算，以计算视觉嵌入与文本嵌入之间的余弦相似度分数$S_{i,j}$。

$$\begin{cases} S_{i,j} = \dfrac{v_i \cdot t_j}{\| v_i \| \| t_i \|} \\ P_{i,j} = softmax(S_{i,j}) \end{cases}, \tag{3}$$

其中：相似度分数$S_{i,j}$计算经过$softmax$操作后，将原始分数S_i转换为概率分布$P_{i,j}$，确保输出是一个有效的新类别上的概率分布。之后，使用多次RoI head对获取到的目标预测框过滤操作，以提升对预测目标的定位精度。通过上述流程计算得到VLM的预测分数，再采用伪标签细化操作。具体来说，把已经得到的候选区域预测分数和VLM预测分数相加，取其平均值与设定的阈值（p_0=0.5）比较，过滤掉小于阈值的冗余预测框。通过这种方法，伪标签能够获得质量更高的边界框b_i定位信息。

$$l = \{(b_i, c_i) | p_i \geqslant p_0\}_{i=1}^{k}, c_i = \arg max_i P_{i,j}, \tag{4}$$
$$p_i = max_i P_{i,j}$$

其中：p_i, c_i分别表示预测分数和类别标签，得到最终的伪标签集合l。

2.2.1.2　检测器构建

高质量的区域预测框生成对有效检测新类

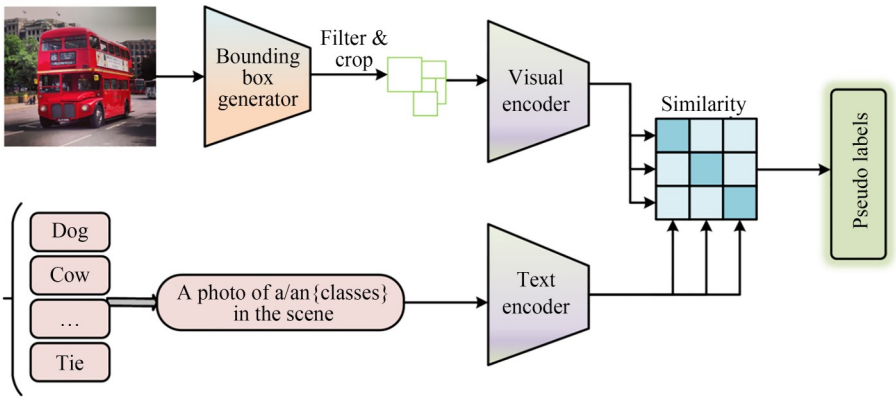


图2　伪标签生成

Fig. 2　Pseudo Label Generations

目标至关重要，会直接影响到后续检测器对预测框的定位质量和目标分类的准确性。然而，由于检测器缺乏对新类概念的学习，难以处理检测图像中新类的任务。为此，本文在继承SAN-YOLO方法阈值过滤思想的基础上，进一步构建可持续更新动态词汇表，在训练迭代中不断优化新类概念上下文信息，提升区域预测框质量与新类检测精度。

动态词汇表用于训练模型的检测器网络，增强其理解新类概念的能力，以帮助检测器在后续迭代中快速聚焦新目标精准定位的特征信息。其中，设输入图像集合 $X=\{x_i\}_{i=1}^{N}$，伪标签集合为 $L=\{l_i\}_{i=1}^{N}$，每个 l_i 包含边界框坐标、类别概率和相应的置信度得分。将伪标签中的区域位置信息输入至检测器，遵循公式(1)为每张图像 x_i 中的区域生成的提案预测得分为 x_i^{sc}，再与伪标签 l_i 中置信度得分 l_i^{sc} 进行计算：

$$\begin{cases} \alpha=\alpha_0\cdot\left(1-\dfrac{e}{E}\right) \\ G=\alpha\cdot l_i^{sc}+(1-\alpha)\cdot x_i^{sc} \end{cases}, \tag{5}$$

其中：α 为基于训练迭代数的自适应平衡参数，用于调节VLM生成的伪标签置信度分数 l_i^{sc} 与检测器预测分数 x_i^{sc} 的融合权重，其值随训练迭代数 e 线性衰减，初始值 α_0 设定为0.5，e 为当前训练迭代数，E 表示总的训练迭代数，G 表示动态词汇表。上述公式使检测器早期依赖VLM提供的新类概念的先验知识，后期充分学习到新类概念后，随着 e 逐渐增大同时 α 减小，使检测器逐步增加对自身预测的权重。通过这种渐进式权重变换，检测器在后期训练中更加注重利用有效的上下文信息，平衡了新类概念的先验知识与局部特征学习之间的关系，确保检测器从语义引导逐步到自身的优化。通过公式(5)对目标预测框进行加权融合处理后，可初步得到动态词汇表 $G=\{g_i\}_{i=1}^{N}$。之后再进一步执行基于置信度候选框的筛选：

$$f(g_i)=\begin{cases} 1, & \text{if } s(g_i)\geqslant\tau \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

其中：$s(g_i)$ 表示第 i 提案 g_i 的预测得分，$\tau=0.8$ 为经验阈值，过滤得到高置信度候选框 $f(g_i)$。通过逐元素筛选得到过滤后动态词汇表 G_f：

$$G_f=\{g_i|f(g_i)=1,\forall g_i\in G. \tag{7}$$

为提升检测器对新类概念的学习能力，之后对每个训练批次中动态词汇表样本 G_f 进行如下处理：

$$G_f'=G_f\oplus G_f\in G^{2n}. \tag{8}$$

获得最终的动态词汇表 G_f'，其中 G^{2n} 表示复制扩展后的高置信度伪标签空间，使得在保持原始数据分布的前提下实现训练样本量翻倍。其核心目标是通过增加高质量实例的学习频次，缓解因新类样本稀缺导致的检测器对基类的偏差问题。之后使用RoI Heads对每个高质量提案进行进一步的分类和边界框回归，用于优化整个检测器网络，如图3所示。

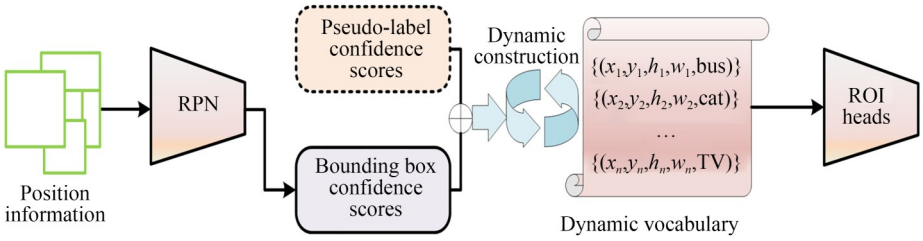


图3　检测器构建

Fig. 3　Detector Construction

由图3可知，这个动态过程随着检测器不断训练迭代，学习到更丰富、更准确的新类信息后，动态词汇表内的信息也会持续更新数据。检测器从多样化、高质量的伪标签中学习，增强了检测器在更广泛检测中上下文信息的有效性，进而输出高效的区域预测框。

2.2.2　标签匹配模块

检测器经过KI强化后虽已经具备理解新类概念的能力，但对具有相似特征的新类和基类目标仍会出现标签分配混淆的问题，导致新类检测

性能受限。SIA-OVD方法通过将不同形状区域特征进行优化分类，改善了新类的特征表达。受此启发，本文提出LM模块，通过动态多级阈值与基类-新类独立匹配的方式，进一步增强检测器新类与基类特征区分能力，提升对新类检测效果。

在匹配阶段，该模块引入一个匹配质量矩阵 $Q\in\boldsymbol{R}^{M\times N}$，其中，$M$表示真实框的数量，$N$表示预测框的数量。矩阵元素 $\boldsymbol{Q}_{i,j}$ 代表第 i 个真实框（gt_i）和第 j 个预测（$pred_j$）之间的匹配质量，其值通过 IoU（gt_i, $pred_j$）计算，用于衡量它们的空间重合程度。对于每个预测框 j，进行如下处理：

$$\begin{cases}\boldsymbol{Q}_{i,j}=\dfrac{\left|gt_i\cap pred_j\right|}{\left|gt_i\cup pred_j\right|}\\ \boldsymbol{q}_j=\max\limits_i\boldsymbol{Q}_{i,j}\\ \boldsymbol{m}_j=\arg\max\limits_i\boldsymbol{Q}_{i,j}\end{cases},\tag{9}$$

其中：$|\cdot|$表示边界框的面积，由上式计算得到最大质量值 q_j 及真实标注索引 m_j 后，对应标签列表 $L=\{l_1,\ldots l_i,l_n\}$，为预测框 j 进行如下匹配：

$$L(n)=\begin{cases}1, & \text{if } q_j\geqslant 0.5\\ -1, & \text{if } 0.3<q_j<0.5.\\ 0, & \text{otherwise}\end{cases}\tag{10}$$

该过程实现预测的初步分类。其中，对于阈值大于0.5高质量预测作为正样本，参与后续的检测器训练；阈值大于0.3且小于0.5的预测作为忽略样本，不参与训练，避免干扰检测器正常的学习；其他情况预测作为负样本，提升检测器的稳定性。之后，在传统IoU匹配基础上引入真实标注的置信度信息 $C\in[0,1]^M$，构造复合的质量评价指标。定义修正质量矩阵 $\tilde{Q}\in\boldsymbol{R}^{M\times N}$，其元素计算为：

$$\tilde{\boldsymbol{Q}}_{i,j}=\begin{cases}\lfloor 10\boldsymbol{C}_i\rfloor+\boldsymbol{Q}_{i,j}, & \text{if } \boldsymbol{Q}_{i,j}\geqslant 0.5\\ \boldsymbol{Q}_{i,j}, & \text{otherwise}\end{cases}.\tag{11}$$

上述设计置信度项 $\lfloor 10\boldsymbol{C}_i\rfloor$ 将质量值提升至整数区间 $[0,10]$，使得低质量匹配维持原始值，高质量匹配预测获得置信度加权，通过计算修正质量矩阵 $\tilde{\boldsymbol{Q}}_{i,j}$ 最大值，作为更新后的标注索引 m_j'，保证检测器在更加可靠的标注中学习。

在开放集场景识别任务中，数据集中包含基类 C^B 和新类 C^N，且 $C^B\cap C^N=\varnothing$，将基类掩码 $B\subseteq\{1,\cdots,M\}$ 与新类掩码 $N=B^c$（基类的补集），分别生成掩码矩阵：

$$\begin{cases}\boldsymbol{Q}_i^B=\boldsymbol{Q}_i\cdot\boldsymbol{I}(i\in B)\\ \boldsymbol{Q}_j^N=\boldsymbol{Q}_j\cdot\boldsymbol{I}(i\in N)\end{cases},\tag{12}$$

其中：$\boldsymbol{Q}_i^B$ 表示第 i 个基类掩码矩阵，$\boldsymbol{Q}_j^N$ 表示第 j 个新类掩码矩阵，之后，并行计算两类匹配结果：

$$\begin{cases}\boldsymbol{m}_i^B=\arg\max\limits_i\boldsymbol{Q}_{i,j}^B\\ \boldsymbol{m}_j^N=\arg\max\limits_i\boldsymbol{Q}_{i,j}^N\end{cases},\tag{13}$$

其中：$\boldsymbol{m}_i^B$ 表示第 i 个基类的匹配结果，$\boldsymbol{m}_j^N$ 表示第 j 个新类的匹配结果。式（13）的处理在得到两类匹配结果的同时，也抑制了背景等无关因素对检测器的影响，之后对基类与新类的标签处理方法为：

$$\boldsymbol{m}_j=\begin{cases}\boldsymbol{m}_j^B, & \text{if } q_j^B\geqslant 0.5\\ \boldsymbol{m}_j^N, & \text{otherwise}\end{cases}.\tag{14}$$

当基类最大质量值 $q_j^B<0.5$ 时，被匹配为基类目标，其他情况以匹配新类目标处理。该方法不仅确保了检测器在基础类与新类学习中的独立匹配，缓解了类别间的竞争，同时避免了新类与基类预测混淆。最后，为确保每个真实标注至少匹配一个预测，针对低质量预测集合 $S_i=\left\{j|\boldsymbol{Q}_{i,j}=\max_{j'}\boldsymbol{Q}_{i,j'}\right\}$，对于 $\forall j\cup_i S_i$，强制 $l_i=1$。该操作表示为：

$$l_j\leftarrow\max\left(l_j,\boldsymbol{I}\left(j\in\bigcup_i S_i\right)\right).\tag{15}$$

通过这种方式能够确保任意一个真实框都至少被一个预测框匹配，避免真实标注遗漏。动态词汇表经LM优化后迭代更新，以便融入更准确的新类信息，确保词汇表中的高质量伪标签能被高效利用，增强了检测器对新类概念的学习能力。

2.2.3 双分支融合模块

经KI与LM对检测流程改进后，检测器已经具备了一定的新类检测能力。然而，由于检测时VLM和检测器之间结合不充分，造成了对新类目标定位质量不高，并出现了预测框重叠的现象。针对上述问题本文提出了一个DBF模块，在沿用RegionCLIP利用视觉与文本编码器协同特征提取机制思想的同时，通过耦合VLM语义理解与检测器精确定位，借双通道特征交互和融合提高新类别辨识力，如图4所示。

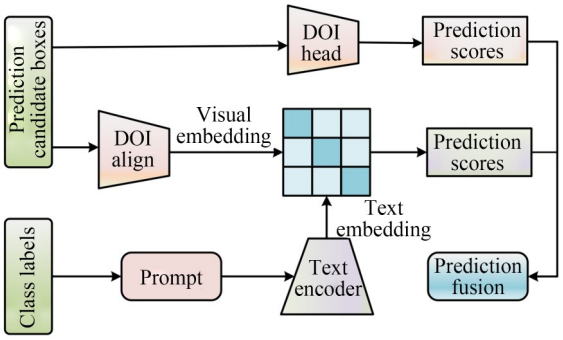


图 4　双分支融合

Fig. 4　Dual-branch Fusion

由图 4 可知，DBF 由传统视觉（Traditional Branch，TD）和视觉语言分支（Vision-Text Branch，VT）构成，前者遵循传统目标检测中对检测网络进行的常规学习，将预测候选框（Prediction Candidate Box）输入至 ROI Head 计算预测分数。而 VT 旨在结合目标检测器与 VLM 之间的优势，更好地挖掘和定位新类目标，将类别标签（Class Labels）以提示词（Prompt）的方式输入至 VLM 的文本编码器中，计算文本嵌入与视觉嵌入预测分数，最后融合得到最终的预测值（Prediction Fusion）。在知识嵌入和标签匹配的增强下，检测器捕获到了更加丰富且有效的全局上下文信息。然而，VLM 依赖基于直接裁剪图像区域的方法，导致其在局部区域推断时缺乏对上下文信息的有效利用。尽管 VLM 在裁切区域的定位推断能力相对有限，但其在真实框的预测上表现出较高的准确性。因此，通过检测器给出精确定位的预测框，结合 VLM 的语义推断信息得到 VT 的预测结果，同时联合 TD 预测结果，最终实现双分支预测的有效融合，弥补了单一分支检测的局限性。

该模块输入 $\boldsymbol{I}\in\boldsymbol{R}^{H\times W\times 3}$ 的图像至 ResNet50-FPN 主干网络提取多尺度特征图 $Features=\{p_2,p_3,p_4,p_5,p_6\}$，然后通过全局最大池化操作和在不同 FPN 层的所有张量特征的拼接，将这些特征图送入 RPN 中，生成预测候选框。经过标签匹配后，将基类和新类特征传输至 RoI head 中获得全局图像特征。全局图像特征随后在两个分支中使用：一个输入至 TD 分支，将提取的全局图像特征传入 RoI head 中，进行传统的训练和推理；一个传入 VT 分支，结合目标检测器和 VLM 模型之间各自的优势，增强对新类概念的检测。

首先，将给定的所有类别名称 $C=\{C_1,C_1,\cdots,C_n\}$，按照预设计好的模板输入到 VLM 文本编码器中。通过与相同类别的嵌入进行平均值计算后，生成的不同类别的文本嵌入（$T\in\boldsymbol{R}^{d\times c}$），$d$ 是嵌入的维度，c 是类别数。

$$\begin{cases} T=E_{TE}(C), & T\in\boldsymbol{R}^{d\times c} \\ \boldsymbol{R}=Linear(V), & \boldsymbol{R}\in\boldsymbol{R}^{d\times hw} \end{cases}, \tag{16}$$

其中：T 为 VLM 文本编码器处理的文本嵌入，$\boldsymbol{R}$ 为使用一个线性层投影经过 RoI Align 机制处理后的视觉嵌入 $\boldsymbol{V}$，该线性层将区域特征 $\boldsymbol{V}$ 映射到同一纬度下的文本嵌入空间中，为每个预测候选框中区域的对象提供丰富的语义信息。得到预测候选框嵌入和文本嵌入后，根据得到的信息计算每个候选框的新颖性：

$$P_i^{VT}=\frac{\sum_{k}^{C^N}\exp(r_i,t_k)}{\sum_{j}^{C^B}\exp\left(r_i,t_j\right)+\sum_{k}^{C^N}\exp(r_i,t_k)}, \tag{17}$$

其中：r_i 表示第 i 视觉嵌入 R，t_k 表示第 k 文本嵌入 T。C^N 代表新类的集合，C^B 代表基类的集合。通过上述计算，模型得到了对预测候选框的新颖性得分 P_i^{VT}，即对新类的预测分数，基类候选框的得分为 P_i^{TD}。本文采用了与 BARON 类似的几何平均融合策略，在推理阶段将预测结果经几何平均值融合，计算公式如下：

$$P'_{i,k}=\begin{cases} (P_{i,k}^{TD})^{(1-\alpha)}\cdot\left(P_{i,k}^{VT}\right)^{\alpha}, & \text{if } k\in C^B \\ (P_{i,k}^{TD})^{\alpha}\cdot\left(P_{i,k}^{VT}\right)^{(1-\alpha)}, & \text{if } k\in C^N \end{cases}, \tag{18}$$

其中：$\alpha\in[0,1]$ 平衡以两个分支的权重，P_i^{TD}，P_i^{VT} 分别是传统视觉分支和视觉-文本分支的预测分数，i,k 索引和公式（1）中相同，计算最终预测分数 $P'_{i,k}$。

3　实验与结果分析

3.1　数据集与实验细节

3.1.1　数据集介绍

本文选用公开的 OVD 基准数据集 COCO[28] 和 LVIS[29]，进行对比实验以评估模型具体表现。COCO 数据集类别数量相对较少（共 65 类），遵循公开基准设置将类别划分为 48 个基类和 17 个新类，该数据集中的目标对象主要从复杂的日常

场景中截取，具有相对清晰的视觉特征的特性。COCO训练集中有118 287张图片，评估集5 000张图片中包含28 538个基本类别实例和4 614个新类别实例。LVIS作为一个大规模、类别丰富的细粒度数据集（共1 203类），依据标准设置将337个罕见类（rare）设置为新类，将常见类（custom）和频繁类（frequent）保留为基类，其细粒度类别数量远超COCO。与COCO相比，LVIS因严重的类别不平衡与高度语义重叠的特性，使其能更真实地反映现实世界复杂场景的检测挑战。其训练集有100 000张图像，验证集包含20 000张图像。

3.1.2　实验细节

实验使用的硬件及深度学习环境配置信息，如表1所示。

表1　实验细节配置表

Tab. 1　Experiment details configuration table

配置项	参数
操作系统	Ubuntu18.04
GPU	4×NVIDIA GeForce RTX 3090（24 GB）
CUDA/Python	11.3/3.8
深度学习框架	PyTorch 1.10、Detectron2 v0.6[30]
优化器	AdamW
Batch size	8
训练迭代步数	90 000
VLM模型	CLIP ViT-B/32

本文实验针对COCO和LVIS数据集的不同特性，选用ResNet50-FPN结合Mask R-CNN处理COCO数据集，利用RoI Align和FPN提升检测精度；LVIS数据集则使用CenterNet2[31]为检测器，其解耦分类与定位、Probabilistic Two-stage框架及注意力机制，避免高频类主导特征学习，有效提升新类召回率。

3.1.3　评价指标

COCO数据集采用AP_{base}，AP_{novel}，$AP50$作为评价指标，分别代表基类、新类和所有类别在IoU为0.5时的均值平均检测精度（mean Average Precision，mAP）。LVIS数据集进行评估时，将罕见、常见、频繁和所有类别的IoU从0.5到0.95均值平均检测精度（mAP），记为AP_r，AP_c，AP_f和AP。mAP计算公式如下：

$$mAP=\frac{\sum_{i=1}^{n}AP_i}{n}, \tag{19}$$

其中：AP_i为检测时每个类别的平均精度，n代表总类别数，i表示检测的类别号。$AP50$即$mAP_{0.5}$表示当IoU阈值为0.5时，所有目标类别的平均检测精度；$mAP_{0.5\sim0.95}$时当IoU阈值为0.5～0.95时，步长为0.05，在0.5～0.95的10个阈值下，计算阈值的检测精度的平均值。

3.2　消融实验

为验证本文提出的KI，LM和DBF模块在OVD中的有效性及协同效果，本节在COCO评估集上设计了递进式消融实验。首先，依次添加各模块至检测器来评估其对检测性能的独立贡献；然后，通过各模块间的交互作用综合验证KI-DBFOVD方法的整体性能。

3.2.1　知识嵌入模块消融实验

本节对KI模块在检测器训练时对性能的影响进行评估，实验结果如表2所示。

表2　知识嵌入模块实验结果

Tab. 2　Experimental results on the KI module（%）

Steps	AP_{novel}	AP_{base}	$AP50$
10 000	10.9	23.2	20.3
30 000	24.8	36.1	30.2
60 000	35.5	47.7	42.6
90 000	35.7	53.1	49.4

由表2可知，训练初期，检测器对新类的识别能力较弱。然而，随着动态词汇表不断引入高质量伪标签，检测器逐步掌握新类的语义及空间位置信息，从而实现更精准的检测。最终，在训练完成时，KI模块将新类检测精度提升至35.7%，表明其在增强检测器对新类目标理解能力方面的有效性。

3.2.2　标签匹配模块消融实验

本节对加入LM模块的效果进行验证，实验结果由表3所示。分析可知，训练阶段仅使用基类标签监督学习时，检测器对基类的检测精度为56.5%。在此基础上引入KI模块后，新类检测效果显著提升，但基类检测精度下降了3.4%。原因在于检测器学习新类概念时，容易将基类和

新类具有相似特征的目标混淆，导致标签分配冲突，影响检测性能。进一步加入LM模块后，对新类与基类预测框采用独立匹配避免二者竞争，优化了标签匹配流程。实验结果表明，经LM改进的检测器不仅改善了基类检测精度下降的问题，还提升了对新类的检测精度。

表3　标签匹配实验结果

Tab. 3　Experimental results on the LM module (%)

训练配置	AP_{novel}	AP_{base}	$AP50$
仅基类标签监督	—	56.5	—
＋KI模块	35.7	53.1	49.4
＋KI＋LM模块	36.5	56.2	50.6

3.2.3　双分支融合模块消融实验

为验证DBF模块的有效性，在引入KI和LM模块后，本节进行了不同分支配置下的消融实验，结果如表4所示。

表4　双分支融合实验结果

Tab. 4　Experimental results on the DBF module(%)

方法	TD	VT	AP_{novel}	AP_{base}	$AP50$
A	√	×	36.7	56.5	50.7
B	×	√	37.5	55.9	50.2
C	√	√	38.6	56.8	52.0

由表4可知，DBF采用不同分支时的性能差异显著。方案A仅启用TD时，新类检测精度为36.7%，已达性能上限。方案B仅启用VT分支时，新类精度提升了0.8%，体现了VT分支在挖掘新类语义概念上的优势。然而，其基类精度较方案A下降了0.6%，反映出该分支在利用基类上下文信息方面存在不足。而方案C融合TD和VT后，新类检测精度达到38.6%，较方案A和B分别提升了1.9%，1.1%，基类精度提升至56.8%，较方案A和B分别提升了0.3%，0.9%，整体精度显著提升至52.0%。这凸显了双分支融合(DBF)策略的优势，TD保持基类的判别能力与定位质量，VT增强对新类概念的语义理解，二者互补融合，提升了检测器的泛化能力。

3.2.4　综合消融实验

为充分验证KI，LM与DBF模块的交互作用与协同提升效应，本节通过逐步引入或移除各个模块至检测器的方式，评估模块间的相互作用及对检测性能的影响，如表5所示。分析结果可知，仅添加KI模块使新类检测精度提升2.3%，表明KI显著增强了检测器对新类概念的学习，但因新类学习引发的特征混淆导致基类精度降低1.8%；联合LM模块后，新类的检测精度进一步提升3.1%，同时基类精度显著回升2.8%，验证了LM通过独立匹配策略有效消除基类-新类标签冲突；而移除KI仅保留LM＋DBF时，新类精度较KI＋LM下降1.8%，证实KI提供的新类先验知识是该方法的核心基础；而KI＋DBF组合虽提升新类精度4.4%，基类精度较LM＋DBF组合显著下降2.2%，反映出LM在平衡新类和基类性能方面的重要作用；最后，引入KI、LM和DBF三个模块后，新类、基类和所有类别的检测精度分别提升至38.6%，56.8%和52.0%，较基线分别提升了5.2%，3.4%，3.0%。综上分析可知，实验结果充分验证了KI-DBFOVD方法的有效性，以及不同模块之间的协同优势，表明其能够更好地适应于开放场景中新类的检测任务。

表5　COCO数据集上实验结果

Tab. 5　Experimental Results on the COCO Dataset(%)

方法	KI	LM	DBF	AP_{novel}	AP_{base}	$AP50$
KI-DBFOVD	×	×	×	33.4	53.4	49.0
	√	×	×	35.7	53.1	49.4
	√	√	×	36.5	56.2	50.6
	×	√	√	34.7	56.4	50.5
	√	×	√	37.8	54.2	50.3
	√	√	√	38.6	56.8	52.0

3.3　对比实验

3.3.1　数据集对比试验

为深入评估本文提出的KI-DBFOVD方法在OVD任务中的性能，本小节选用COCO数据集与其他先进的OVD方法进行了比较。其中，包括利用伪标签增强检测方法VL-PLM，与五种当前最先进的方法RWVM，SIA-OVD，Marvel-OVD，LBP，FFPLearning，以及四种经典基线方法 OVR-CNN，RegionCLIP，BARON，ViLD 在内的共十种OVD方法。为了实验对比的公平性，每种方法除了特殊的参数保持了与原文设置一致以外，其他设置如迭代次数、学习率与批次

大小等超参数值设置，均与3.1.3节中相同。此外，还包含了不同方法对使用检测器的对比，实验结果如表6所示。

表6　对比实验结果

Tab. 6　Experimental performance of comparison(%)

方法	Detector	AP_{novel}	AP_{base}	$AP50$
OVR-CNN	Faster RCNN	22.8	46.0	39.9
RWVM	Faster RCNN	36.4	54.6	49.8
SIA-OVD	DAB-DETR[31]	35.5	40.3	39.3
VL-PLM	Mask RCNN	33.4	53.4	49.0
MarvelOVD	Mask RCNN	37.8	56.3	51.6
RegionCLIP	Faster RCNN	31.4	57.1	50.4
BARON	Faster RCNN	35.8	54.2	50.2
ViLD	Faster RCNN	27.6	59.5	51.2
LBP	Faster RCNN	35.9	60.8	54.3
FFPLearning	Faster RCNN	32.6	57.6	52.4
KI-DBFOVD	Mask RCNN	**38.6**	56.8	52.0

由表6可知，本文所提KI-DBFOVD方法表现卓越，其中在关键评价指标新类检测精度(AP_{novel})上达到最优水平38.6%。与同样采用了伪标签技术的VL-PLM方法相比，新类的检测精度大幅提升了5.2%，这表明本文方法在增强检测器对新类概念的理解方面具有明显优势，能够更有效地利用伪标签进行训练，从而提高新类目标的检测性能。与目前最先进的五种OVD方法RWVM，SIA-OVD，MarvelOVD，LBP以及FFPLearning相比，本文方法在新类检测精度上均展现出显著优势，提升幅度分别达到了2.2%，3.1%，0.8%，2.7%与6.0%。值得注意的是，与SIA-OVD相比取得的3.1%提升，特别验证了本文方法在处理不同形状区域特征分类问题上的有效性。同时，相较于经典方法OVR-CNN，RegionCLIP，BARON与ViLD，本文方法在新类检测精度方面同样表现更为出色。其中，对RegionCLIP高达7.2%的显著提升，有力证明了检测器与VLM协同融合策略在充分利用图像上下文信息、增强新类目标检测能力方面的优越性。另一方面，本文方法在保持基类检测高竞争力(AP_{base}=56.8%)的同时，实现了整体检测精度($AP50$=52.0%)的优化，达成了新类检测能力提升与基类性能稳固之间的平衡。

此外，为进一步验证本文方法在长尾分布的复杂开放场景下，对新类目标的泛化能力，在更具有挑战性的LVIS数据集上进行实验，如表7所示。由表7可知，与当前最先进的RWVM，LBP与FFPLearning方法对比显示，本文方法对关键指标新类检测精度(AP_r)分别提升了1.6%，1.3%与5.0%，达到了最优的25.4%。这是因为通过动态词汇表机制，使模型在新类样本缺失情况下仍能学习精准概念表征。此外，采用多级阈值解耦策略，缓解了基类与新类间语义混淆，提升了对新类的检测能力。在常见类(AP_c=30.2%)和频繁类(AP_f=37.2%)检测精度上，本文方法同样保持竞争力，最终整体精度(AP=32.6%)表现优异。表7结果表明，在具有严重长尾分布和高语义重叠特性的LVIS数据集上，本文方法展现出对新类目标卓越的泛化能力。其核心优势在于能够更有效地学习区分新类与基类的特征，利用几何融合实现定位精度与语义理解，从而显著提升了对新类目标的检测性能。

表7　LVIS数据集实验结果

Tab. 7　Experimental results on the LVIS dataset(%)

方法	AP_r	AP_c	AP_f	AP
RWVM	23.8	30.0	34.6	30.7
VL-PLM	17.2	23.7	35.1	27.0
RegionCLIP	22.0	32.1	36.9	32.3
BARON	23.2	27.8	32.4	28.4
ViLD	19.8	27.1	34.5	28.7
LBP	24.1	29.5	32.8	29.9
FFPLearning	20.4	29.6	33.2	29.7
KI-DBFOVD	**25.4**	30.2	37.2	32.6

3.3.2　可视化对比试验

为直观展示本文方法在OVD任务中对新类实际的检测效果，对COCO评估数据集中图像进行了可视化展示，如图5所示。

由图5可知，KI-DBFOVD方法增强后的检测器在OVD任务中展现出卓越的性能。在使用COCO评估集对模型检测性能分析时，新类(novel classes)集合为novel classes={cat, airplane, cow, bus, cup, cake, skateboard, dog, couch, tie, sink, umbrella, scissors, eleph-ant, keyboard,

(a) 样本1
(a) Sample 1

(b) 样本2
(b) Sample 2

(c) 样本3
(c) Sample 3

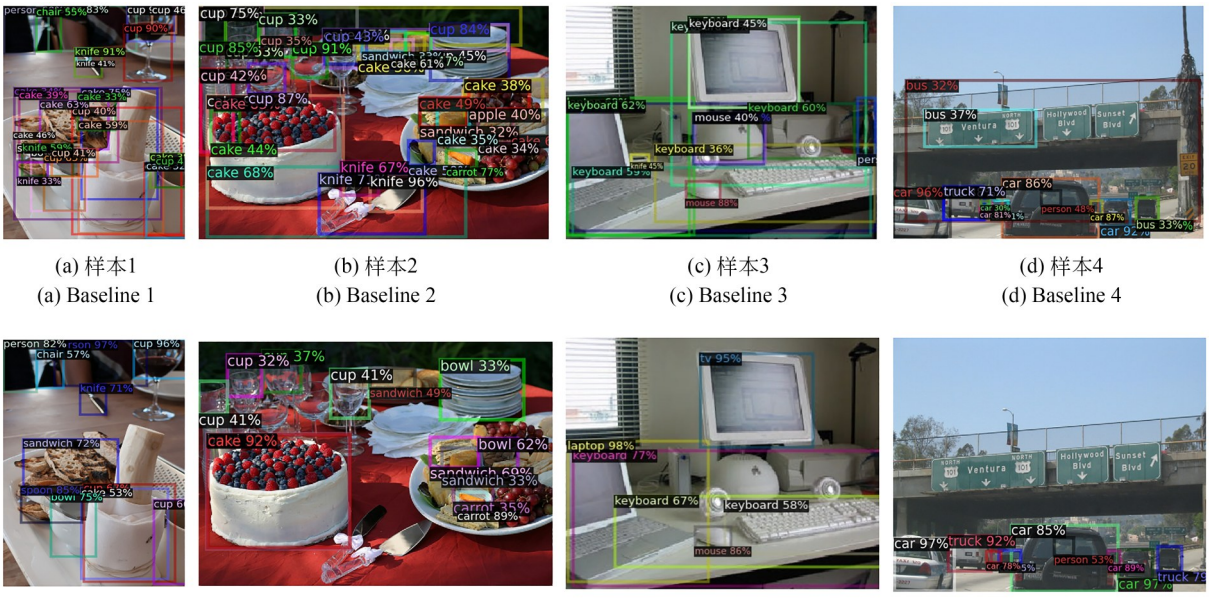
(d) 样本4
(d) Sample 4

图 5　可视化结果

Fig. 5　Results of visualization

knife, snowboard}。可视化结果表明，该方法能够精确识别并准确定位多种复杂场景中不同类别的目标对象。对于新类目标，如狗(dog)、大象(elephant)、公交车(bus)等，检测器均实现了高精度的检测。值得注意的是，如图 5(a)所示，即使在部分遮挡的挑战性条件下，该方法依然能够有效检测目标。这一鲁棒性的提升可归因于所提出的双分支协作机制，传统视觉分支(TD)有效保留了目标的空间定位能力，而视觉语言分支(VT)则通过语义信息补全，显著增强了模型对局部遮挡目标的识别能力。

为更加深入评估本文方法的性能，并全面展示其优势，下面选取了 COCO 数据集中四组不同场景下的图像，对比未使用本方法与应用本方法时的检测效果，如图 6 所示。图 6 可视化对比表明：未采用本文方法时(第一行)，检测结果存在显著的类别混淆与定位缺陷。具体表现为：6(a)图中基类“三明治”(sandwich)被误判为新类“蛋糕”(cake)，且新类间出现“蛋糕”(cake)与“杯子”(cup)误检；6(c)图中基类“笔记本电脑”(laptop)和“电视显示器”(tv)被错误归类为新类“键盘”。此类混淆验证了基类与新类标签分配冲突是 OVD 中不可忽视的挑战。而采用 KI-DBFOVD 方法后(第二行)：6(e)图正确识别所有基类与新类目标；6(g)图消除误检情况，证实 LM 模块通过独立分配策略有效解决类别混淆问题；同时，



(a) 样本1
(a) Baseline 1

(b) 样本2
(b) Baseline 2

(c) 样本3
(c) Baseline 3

(d) 样本4
(d) Baseline 4

(e) 对比1
(e) Comparison 1

(f) 对比2
(f) Comparison 2

(g) 对比3
(g) Comparison 3

(h) 对比4
(h) Comparison 4

图 6　可视化对比结果图

Fig. 6　Visualization of Comparative Results

如6(e),(f),(g)与(h)图所示,所有检测框重叠现象显著减少,并且被遮挡目标(如“蛋糕”、“杯子”、“笔记本电脑”等)也得到了准确地检测,凸显DBF机制通过整合检测器定位优势与VLM语义理解,提升了目标定位精度。

为验证本文方法对每个新类的检测性能,分别统计了在COCO评估集中使用与不使用该方法对各新类的具体检测效果,如表8所示。分析结果可知,对比中共选取了10类在各个场景中具有代表性的新类目标。结果显示,KI-DBFOVD方法总体上提升了新类目标检测的精度,尤其是对狗、雨伞、沙发及键盘这种特征容易区分的4种新类目标,其检测精度分别提升了12.3%,13.9%,6%和19.4%。这是因为通过基类-新类独立匹配策略,显著提升了检测器的特征区分能力。另外,在对特征不容易区分的新类目标如水槽、大象,检测器经过DBF融合机制协同定位与语义特征,强化纹理明确目标的判别性,也分别提升了2.3%,1.7%。其中,对蛋糕的检测性能下降了2.8%,揭示了目标特征(颜色/纹理)与背景相似性混淆仍是OVD任务的固有挑战[18]。综上,通过对各类新目标检测精度的综合评估结果,充分证明了本文方法的有效性和优秀的检测性能。

表8　新类检测实验结果

Tab. 8　Experimental results on the novel class detection　(%)

目标类别	蛋糕 Cake	公交车 Bus	狗 Dog	雨伞 Umbrella	单板滑雪板 Snowboard	沙发 Couch	水槽 Sink	飞机 Airplane	大象 Elephant	键盘 Keyboard
未使用	35.0	72.0	57.8	11.8	7.8	33.4	7.1	61.8	75.8	23.4
KI-DBFOVD	32.2	75.2	70.1	25.7	8.7	39.4	9.4	66.8	77.5	42.8

3.3.3　超参数对比试验

为确定所选择的超参数能够使检测器达到最优性能,本小节对预测框过滤阈值p_0与公式(5)中自适应平衡参数α在不同取值下的变化进行展示,如图7所示。

由图7可知,图7(a)与图7(b)分别展示了在α为自适应平衡和固定值0.5时,p_0取[0,1]区间不同数值下新类检测精度变化情况。当伪标签过滤阈值$p_0\in[0.3,0.7]$时,新类检测精度波动小于±0.5%(α自适应与固定下均成立),证实该区间可平衡有效预测框保留和噪声过滤。当$p_0=0.5$时,新类检测精度达到了最优,这是因为在此阈值下最大程度保留高置信度正样本,检测器能够充分学习新类概念。而当$p_0=0$时,由于低阈值过滤使得噪声预测框倍增,检测器学习了大量的无效信息,从而导致精度下降。当$p_0=1$时,则因阈值太高导致潜在正样本被过滤,检测器无法充分学习新类特征信息致使检测精度降

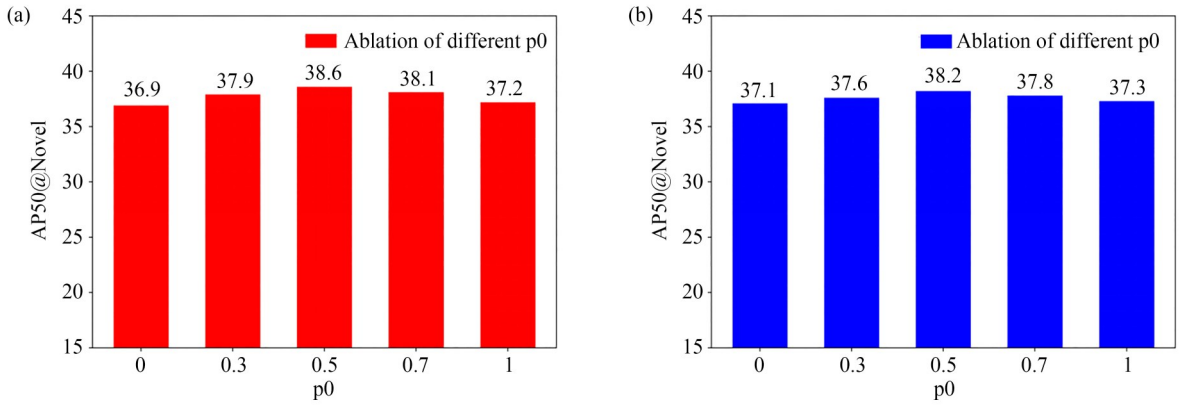


图7　超参数可视化对比结果图

Fig. 7　Hyperparameter Visualization and Comparison Results

低。此外，动态权重 α（式 5）较固定值 $\alpha=0.5$ 提升 0.4%，归因于其衰减机制协调 VLM 先验置信度与检测器自学习权重，强化了检测器对新类概念表征能力，使其能更精准地识别和定位新类目标。

4　结　论

针对检测器对新类概念理解弱、训练时基类与新类标签之间分配混淆，以及对新类的检测性能不足的问题。本文提出了 KI-DBFOVD 方法。通过设计了知识嵌入模块构建动态词汇表，嵌入高质量伪标签以增强检测器的新类概念理解能力；标签匹配模块利用基类-新类独立匹配策略，有效缓解标签混淆问题；最后提出的双分支融合模块更进一步检测器提升了对新类的检测能力。在 COCO 数据集上，新类检测精度达 38.6%，较主流方法表现更优；在更具挑战性的 LVIS 数据集上，新类检测精度达 25.4%。实验证明了 KI-DBFOVD 方法在 OVD 任务中的有效性，展现出了优秀的检测性能。

在研究的过程中发现，本文方法在背景区分方面仍存在不足，干扰了新类概念的学习。未来工作将聚焦于更高效的检测器与 VLM 融合方式，以解决背景干扰问题，提升开放词汇目标检测在实际应用中的性能。

作者贡献声明：

金友：论文方法的提出，论文构思和撰写，实验的设计与验证、数据整理和分析、可视化呈现；

邓箴：论文写作指导与理论分析，论文审核与编辑写作；

刘立波：提供实验所需资源；整体项目管理；论文审核与编辑写作。

参考文献：

[1] 王勰，殷高方，赵南京，等. 基于可变荧光统计法的藻类活体细胞密度宽范围检测实验[J]. 光学学报，2024，44(12)：172-181.
WANG X，YIN G F，ZHAO N J，*et al*. Experimental investigation of wide-range detection of live algal cell density based on variable fluorescence statistical analysis[J]. *Acta Optica Sinica*，2024，44(12)：172-181.（in Chinese）

[2] 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.（in Chinese）

[3] 陈俊英，李朝阳，黄汉涛，等. 并行特征提取和渐进特征融合的计算机主板装配缺陷检测[J]. 光学精密工程，2024，32(10)：1622-1637.
CHEN J Y，LI Z Y，HUANG H T，*et al*. Computer motherboard assembly defect detection using parallel feature extraction and progressive feature fusion[J]. *Opt. Precision Eng.*，2024，32(10)：1622-1637.（in Chinese）

[4] HE K M，GKIOXARI G，DOLLÁR P，*et al*. Mask R-CNN[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). *October* 22-29，2017，*Venice*，*Italy*. IEEE，2017：2980-2988.

[5] 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.

[6] ZAREIAN A，DELA ROSA K，HU D H，*et al*. Open-vocabulary object detection using captions[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25，2021. *Nashville*，*TN*，*USA*. IEEE，2021：14393-14-402.

[7] 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.（in Chinese）

[8] RADFORD A，KIM J W，HALLACY C，*et al*. Learning transferable visual models from natural langu-age supervision[C]. *International conference on mac-hine learning*. PMLR，2021：8748-8763.

[9] CHEN K Y，JIANG X L，WANG H C，*et al*. OVDAR：open-vocabulary object detection and attributes recognition[J]. *International Journal of Com-*

puter Vision, 2024, 132(11): 5387-5409.

[10] YANG L X, CHEN D P, CHEN Y F, *et al*. A neuroinspired contrast mechanism enables few-shot object detection[J]. *Pattern Recognition*, 2024, 156: 110766.

[11] LI J Y, ZHAO F T, ZHAO H M, *et al*. A multimodal open object detection model for tomato leaf diseases with strong generalization performance using PDC-VLD [J]. *Plant Phenomics*, 2024, 6: 220.

[12] 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.

[13] WANG Z S, ZHOU W H, XU J L, *et al*. SIA-OVD: shape-invariant adapter for bridging the image-region gap in open-vocabulary detection[C]. *Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne VIC Australia*. ACM, 2024: 4986-4994.

[14] ZHAO S Y, ZHANG Z X, SCHULTER S, *et al*. Exploiting unlabeled data withvision and language models forobject detection[C]. *Computer Vision-ECCV* 2022. *Cham: Springer Nature Switzerland*, 2022: 159-175.

[15] 卢汉, 崔博伦, 万华洋, 等. 半监督式野生动物夜间目标端到端检测[J]. 光学 精密工程, 2025, 33(5): 789-801.
LU H, CUI B L, WAN H Y, *et al*. End-to-end recognition of nighttime wildlife based on semi-supervised learning[J]. *Opt. Precision Eng.*, 2025, 33(5): 789-801. (in Chinese)

[16] 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.

[17] WANG K, CHENG L C, CHEN W K, *et al*. *MarvelOVD: Marrying Object Recognition and Vision-Language Models for Robust Open-Vocabulary Object Detection*[M]. Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 106-122.

[18] 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.

[19] XIN C, HARTEL A, KASNECI E. DART: an automated end-to-end object detection pipeline with data Diversification, open-vocabulary bounding box Annotation, pseudo-label Review, and model Training[J]. *Expert Systems with Applications*, 2024, 258: 125124.

[20] ZHONG Y W, YANG J W, ZHANG P C, *et al*. RegionCLIP: region-based language-image pretraining[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 16772-16782.

[21] WU S Z, ZHANG W W, JIN S, *et al*. Aligning bag of regions for open-vocabulary object detection [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 15254-15264.

[22] GU X, LIN T Y, KUO W, *et al*. Open-vocabular-y object detection via vision and language knowl-edge distillation[C]. *The Tenth International Conf-erence on Learning Representations* (*ICLR*). 2022:1-21.

[23] LI J M, ZHANG J C, LI J C, *et al*. Learning background prompts to discover implicit knowledge for open vocabulary object detection[C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). June 16-22, 2024, Seattle, WA, USA. IEEE, 2024: 16678-16687.

[24] SONG H, BANG J. Prompt-guided DETR with RoI-pruned masked attention for open-vocabulary object detection[J]. *Pattern Recognition*, 2024, 155: 110648.

[25] LI Y. Federated fine-grained prompts for vision-language models based on open-vocabulary object detection [J]. *Applied Intelligence*, 2025, 55(7): 626.

[26] XU Y F, ZHANG M D, YANG X S, *et al*. Exploring multi-modal contextual knowledge for open-vocabulary object detection[J]. *IEEE Transactions on Image Processing*, 2024, 33: 6253-6267.

[27] LONG Y X, HAN J H, HUANG R H, *et al*. Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(11): 16277-16287.

[28] LIN T Y, MAIRE M, BELONGIE S, *et al*. *Mi-*

crosoft COCO: *Common Objects in Context*[M]. Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[29] GUPTA A, DOLLÁR P, GIRSHICK R. LVIS: a dataset for large vocabulary instance segmentation [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2019: 5351-5359.

[30] WU Y X, KIRILLOV A, MASSA F, *et al*. Detectron2 [EB/OL]. https://github. com/facebookresear-ch/detectron2, 2019.

[31] ZHOU X, KOLTUN V, KRÄHENBUHL P. Proba-bilistic two-stage detection [J]. *arXiv preprint* arXiv: 2103. 07461, 2021.

[32] ZHU X, SU W, LU L, *et al*. Deformable DETR: deformable transformers for end-to-end object detection[C]. *Proceedings of the* 9*th International Conference on Learning Representations* (ICLR 2-021), 2021: 1-16.

作者简介：

金　友(1999－)，男，宁夏吴忠人，硕士研究生，目前就读于宁夏大学，主要研究基于深度学习的目标检测。E-mail：1210791546@qq. com

通讯作者：

刘立波(1974－)，女，宁夏平罗人，博士，教授，博士生导师，2012年于北京理工大学博士后出站，主要从事智能信息处理、计算机视觉方面的研究。E-mail：liulib@163. com