中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-12

论文引用格式:Wang Zhuo, Wang Binyi, Xiao Junhao, Huang Kaihong, Li Shiliang, Xie Yunfei, Wang Jia. XXXX. Cross-modal prior driven active learning for object detection. Journal of Image and Graphics, XX(XX):0001-0012(王卓, 王斌翊, 肖军浩, 黄开宏, 李世亮, 谢云飞, 王佳. XXXX. 跨模态先验驱动的目标检测主动学习. 中国图象图形学报, XX(XX):0001-0012)[DOI:10.11834/jig.250504]

跨模态先验驱动的目标检测主动学习

王卓[1,2],王斌翊[1],肖军浩[2],黄开宏[2],李世亮[1,2],谢云飞[1,2],王佳[1,2]

1. 西北机电工程研究所,咸阳　712099; 2. 国防科技大学智能科学学院,长沙　410073

摘　要:**目的**　现有目标检测主动学习的度量方法仅依赖检测器自身输出,难以识别高置信但语义错误的误检结果且易造成采样冗余。为此,本文旨在引入跨模态先验以提升主动采样的可靠性与有效性。**方法**　提出跨模态先验驱动的语义增强主动学习方法(semantic enhanced active learning,SEAL),在不增加额外监督的前提下利用视觉语言模型如对比语言-图像预训练模型(contrastive language-image pre-training,CLIP)的语义对齐能力提升样本选择质量,有效纠正单模态检测器在训练数据匮乏时的判别偏差。在不确定性采样阶段,SEAL使用CLIP对图像候选检测框区域进行特征提取,通过对比检测器与CLIP的类别预测结果,构建融合视觉与语义的一致性指标,实现更鲁棒的实例级不确定性度量。在多样性采样阶段,聚合图像各类别目标的CLIP特征,构建类别级结构特征表示,并据此计算图像间的结构相似性,实现类别对齐的多样性度量,提升采样的类别覆盖和信息表达多样性。**结果**　在MS COCO与Pascal VOC数据集上的实验结果表明,SEAL在多种主动学习基准设置下均优于主流方法,表现出更高的检测精度。在RetinaNet on Pascal VOC基准上(20%标注数据),SEAL方法的mAP@0.5为72.4%,较当前最优方法提升0.8%;在RetinaNet on MS COCO基准上(10%标注数据),AP@[0.5:0.95]为23.9%,提升0.5%。**结论**　本文提出的SEAL方法成功地利用了跨模态先验知识来优化主动学习中的样本选择过程。通过构建更鲁棒的不确定性度量和更具代表性的多样性度量,能够显著减少数据标注成本、提升模型学习效率。

关键词:主动学习;目标检测;跨模态先验;视觉语言模型;对比语言-图像预训练(CLIP)

Cross-modal prior driven active learning for object detection

Wang Zhuo[1,2], Wang Binyi[1], Xiao Junhao[2], Huang Kaihong[2], Li Shiliang[1,2], Xie Yunfei[1,2], Wang Jia[1,2]

1. Northwest Institute of Mechanical and Electrical Engineering, Xianyang 712099, China; 2. School of Intelligent Science, National University of Defense Technology, Changsha 410073, China

Abstract: **Objective** Active learning for object detection commonly estimates sampling uncertainty using detector confidence and localization deviation, which struggles to identify high-confidence yet semantically wrong false positives. Meanwhile, diversity sampling typically relies on global visual descriptors at the image level, making it difficult to capture semantic structures at the category/instance level and often leading to redundant samples. To address these issues, we introduce cross-modal priors so that the selection signals can be informed not only by the detector's visual evidence but also by language-aligned semantics. The goal is to improve the reliability (robustness against semantic mismatch) and effectiveness (utility per labeled image) of active sample selection under constrained annotation budgets. **Method** We propose SEAL (semantic enhanced active learning), a cross-modal prior-driven active learning framework that leverages the CLIP (contrastive language-image pre-training) vision-language model to enhance both uncertainty and diversity criteria without

收稿日期:2025-10-15;**修回日期**:2026-01-26

基金项目:国防科技大学自主科学创新基金项目(No. 23-ZZCX-JDZ-01)

Supported by: Independent Scientific Research Project of NUDT(No. 23-ZZCX-JDZ-01)

any additional supervision. In Uncertainty sampling, For each image, SEAL extracts CLIP features on candidate detection boxes and compares the detector's predicted category with CLIP's category evidence. We design a visual-semantic consistency score that jointly reflects the detector's confidence, its localization quality, and the agreement between visual predictions and cross-modal semantics. Instances with low consistency are prioritized. This yields a more robust instance-level uncertainty measure that explicitly penalizes high-confidence semantic mismatches—precisely the failure mode that frustrates conventional uncertainty heuristics. In Diversity sampling. To reduce redundancy while preserving category coverage, SEAL aggregates CLIP embeddings of all objects within an image and builds a category-level structural representation. We then compute inter-image structural similarity in this semantic space and perform category-aligned diversity selection, encouraging batches that simultaneously cover rare or under-represented classes and span heterogeneous intra-class appearances. This mitigates the tendency of global-feature methods to over-select visually similar images that contribute limited new information. **Results** We conduct comprehensive experiments on MS COCO and Pascal VOC under multiple baselines and annotation budgets. Across settings, SEAL consistently outperforms strong active learning methods, yielding higher detection accuracy per labeled image. Representative results include: RetinaNet on Pascal VOC (20% labeled data): mAP@0.5 = 72.4%, exceeding the strongest prior method by 0.8%. RetinaNet on MS COCO (10% labeled data): AP@[0.50:0.95] = 23.9%, an improvement of 0.5% over the best baseline under the same budget. Faster R-CNN on MS COCO (40% labeled data): AP@[0.50:0.95] = 33.3%, surpassing the best competing approach by 0.22%. These gains are attributable to two complementary effects. First, the consistency-aware uncertainty ranking correctly down-weights detections that are visually confident but semantically unreliable, thereby allocating labels to genuinely informative instances. Second, the category-aligned diversity explicitly controls semantic coverage and inter-image variation, reducing selection redundancy and ensuring that annotation efforts expand both the breadth (more classes covered) and depth (richer intra-class variation) of the training set. In practice, we observe that SEAL improves data efficiency: for a fixed target accuracy, fewer labels are required relative to visual-only strategies, and for a fixed budget, the achieved AP is higher. The approach remains effective across different detectors and budgets, indicating good robustness and transferability. Beyond headline metrics, qualitative analyses show that SEAL preferentially selects images containing semantically ambiguous contexts (e.g., objects with occlusion, atypical viewpoints, or confusing backgrounds). This selection behavior accelerates learning in regions of the data distribution where detectors typically underperform. Because CLIP embodies large-scale language-vision alignment, the semantic signal complements the detector's purely visual evidence and provides a principled handle on semantic correctness, which traditional confidence/localization surrogates cannot fully capture. **Conclusion** — SEAL demonstrates that cross-modal priors can be harnessed to make active learning for object detection both more reliable and more cost-effective. By combining a visual-semantic consistency measure for uncertainty with a category-aligned structural criterion for diversity, SEAL systematically reduces semantic-mismatch errors during selection and curbs redundancy, leading to higher AP under the same labeling budget and, equivalently, lower labeling cost to reach a target accuracy. Overall, our results on COCO and VOC indicate that infusing active selection with language-aligned semantics is a practical and effective path toward scalable, label-efficient detection.

Key words: active learning; object detection; vision-language models; semantic uncertainty; contrastive language-image pre-training(clip)

0 引　言

在大规模目标检测任务中,数据标注始终是限制模型性能扩展的主要瓶颈之一(Deng 等,2009; Roh 等,2019)。主动学习(active learning, AL)作为节省标注成本的策略,旨在从海量未标注样本中选择最具信息价值的子集进行标注,从而在有限预算下实现模型性能的最大化(Gal 等,2017; Settles, 2009)。近年来,随着深度目标检测模型的发展,主动学习方法已被逐步引入目标检测任务,并取得了令人瞩目的初步成效(Brust 等,2018; Kao 等,2019; Wang 等,2022)。

在主动学习领域,不确定性采样(uncertainty sampling)是最基础且应用最广泛的策略,其核心理念是模型对其预测结果最不确定的样本包含最多的

新信息，因而对模型改进的贡献最大。常见的不确定性度量方法包括Lewis（1995）提出的最小置信度方法（least confident，LC），Scheffer等人（2001）提出的最大分类间隔（margin sampling，MS）和Shannon（1948）提出的最大熵方法（entropy sampling，ES）。贝叶斯不一致主动学习方法（Gal等，2017）采用蒙特卡洛Dropout进行多次前向传递来估计不确定性。主动学习的学习损失（Yoo等，2019）使用损失预测作为不确定性的度量。与不确定性采样互补，多样性采样（diversity sampling）旨在选择代表性样本，以便一个小子集能够描述整个数据集。常用的技术包括基于聚类的方法，即从不同的数据簇中选择样本（Nguyen等，2004；Xu等，2003）；核心集方法（Sener等，2017）使用贪心k-中心算法从未标注池中选择一个小的核心集，并使用混合整数规划迭代提高样本多样性。面向主动学习的上下文多样性（contextual diversity for active learning，CDAL）利用预测概率来提高上下文多样性（Agarwal等，2020）。最近，提出了几种结合基于不确定性和基于多样性的主动学习的工作。Prabhu等（2021）和Zhdanov（2019）通过在模型不确定性加权的图像特征上运行k-means算法来平衡不确定性和多样性。在基于多样性梯度嵌入的批量主动学习（Ash等，2019）中，不确定性和多样性通过在模型最后一层的梯度上使用kmeans++算法进行平衡。层次聚类（Citovsky等，2021）首先对未标注样本进行聚类，并以轮询方式查询每个聚类中最不确定的样本。

将主动学习原理应用于目标检测任务（active learning for object detection，ALOD）比应用于图像分类任务更复杂（Budd等，2021）。目标检测不仅需要识别物体类别，还要求通过边界框精确定位物体。此外，单张图像可能包含多个不同类别、不同尺度和不同遮挡程度的目标实例，且背景往往复杂多变（Cao等，2022）。这些特性使得样本不确定性和多样性的定义与量化变得尤为困难。早期尝试将主动学习应用于目标检测时（Agarwal等，2020；Lyu等，2023；Sener等，2017；Yoo等，2019；Yu等，2022），直接采用了图像分类的主动学习算法。然而，这些算法并未考虑到构成检测任务的联合分类与定位，或者图像中可能包含多个不同物体的情况。这促使了专门为检测设计的主动学习算法的出现。混合密度网络（Choi等，2021）修改了一个目标检测器，使其学习用于分类和回归输出的高斯混合模型（gaussian mixture model，GMM），然后从建模的GMM中推导出这两个任务的不确定性。多实例主动学习（multiple instance active learning，MIAL）使用对抗训练来计算模型差异，进而用于计算不确定性（Yuan等，2021）。多样化原型方法（divproto）通过用模型不确定性替换NMS中的检测分数来改进AL算法；它还使用一组多样化的原型来选择最具代表性的样本（Wu等，2022）。即插即用主动学习（plug and play active learning，PPAL）用难度校准的不确定性和类别条件多样性，在不改动模型的前提下，兼顾了信息量和覆盖面（Yang等，2024）。

然而，当前主流方法普遍依赖于检测器自身的预测置信度或定位偏差来衡量样本的不确定性（Aghdam等，2019；Yoo等，2019），这种仅关注检测模型输出信息的策略，虽然直观，但对于目标在复杂场景下的语义合理性、是否存在细粒度类别混淆等深层问题缺乏有效的刻画能力。这种单一模态下的不确定性建模方式，往往导致模型在语义复杂场景中表现不稳，尤其在类别混淆、复杂环境和显著目标干扰的情况下，无法识别“表面自信但语义错误”的预测区域，从而错过了应当优先标注的关键样本。此外，现有的主动学习中的多样性采样策略同样存在明显局限。首先，一些方法采用基于全局视觉特征的聚类或核心集选择策略来确保多样性（Sener等，2017）。这类特征虽然能反映图像整体外观信息，但难以准确刻画目标检测任务中样本间（尤其是实例层面）的相似性。对于目标检测而言，样本间的多样性应当与图像中具体目标实例及其类别结构密切相关，仅依赖全局特征会忽视实例层面的细粒度语义差异。其次，直接采用检测器提取的特征作为多样性依据，容易受到检测器自身训练充分度的影响。在主动学习的早期阶段，检测模型往往尚未获得良好的泛化能力，导致其输出特征难以全面、准确地反映图像的真实语义结构，从而影响多样性采样的可靠性和代表性。因此，如何设计能够体现类别结构层面差异、并对模型初期语义表征不充分具备鲁棒性的多样性度量方法，成为提升主动目标检测采样效率的关键挑战。

近年来，大规模预训练的视觉语言模型通过学习图像与文本描述之间的对应关系，获得了强大的跨模态语义理解和泛化能力（Du等，2022）。其中，

CLIP(contrastive language-image pre-training, CLIP)通过在海量图文对上进行对比学习预训练,获得了强大的零样本图像识别和细粒度语义表征能力(Radford等,2021)。它能够通过自然语言描述与图像区域进行语义对齐,对图像或图像区域进行更接近人类的文本式语义理解。

视觉语言模型的这种能力已被应用于多种下游视觉任务。例如,在零样本目标检测(zero-shot object detection, ZSD)、开放词汇目标检测(open-vocabulary object detection, OVD)和零样本三维模型分类(zero-shot 3D classification)中,研究者利用CLIP的语义对齐能力来检测训练阶段未见过的物体类别(Gu等,2021; Yan等,2025; Zareian等,2021)。这些应用充分证明了视觉语言模型,特别是CLIP,在理解和表征细粒度视觉语义方面的强大潜力,为主动学习中的目标识别与高价值样本判别提供了全新的视角和有力的工具。

然而,将CLIP融合到主动目标检测框架中仍面临挑战,尤其是如何在检测器的结构化输出基础上构建与语义模型一致的评估机制,并将其嵌入不确定性建模与多样性采样流程中,仍是当前研究的空白。

基于上述观察,本文提出了一种融合视觉语言模型的语义增强主动目标检测方法(semantic enhanced active learning, SEAL)。在不引入额外监督的前提下,SEAL联合检测模型与视觉语言模型CLIP,从不确定性和多样性两个维度提升样本选择质量。具体而言,在不确定性评估阶段,SEAL利用CLIP对候选检测框区域进行特征提取,并与检测器输出结果进行对比融合,构建更鲁棒的实例级不确定性指标;在多样性采样阶段,SEAL通过聚合图像中各类别目标的CLIP特征,形成一种类别感知的图像级特征表示,并以此为基础度量图像间的结构化语义相似性,实现对类别结构敏感的多样性采样,优先选择类别覆盖广、语义表达差异显著的代表性样本。

1 问题描述与算法

1.1 问题描述

根据相关文献(Choi等,2021; Wu等,2022; Yuan等,2021),本文采用批量式主动学习(batch-mode active learning)的设置进行问题建模。与传统的逐样本选择不同,批量式方法在每轮迭代中从未标注数据中选择一批图像进行标注。

设有一个大小为N_{t}的训练集$\boldsymbol{X}_{\mathrm{T}}=\{\boldsymbol{x}_i\}_{i\in[N_{\mathrm{t}}]}$和一个大小为$N_{\mathrm{v}}$的验证集$\boldsymbol{X}_{\mathrm{V}}=\{\boldsymbol{x}_i\}_{i\in[N_{\mathrm{v}}]}$,$\boldsymbol{x}_i$表示图像。在第$r$轮主动学习中,已标注样本集记为$\boldsymbol{X}_{\mathrm{L}}^r$,未标注池记为$\boldsymbol{X}_{\mathrm{U}}^r$,满足$\boldsymbol{X}_{\mathrm{L}}^r\cap\boldsymbol{X}_{\mathrm{U}}^r=\varnothing$且$\boldsymbol{X}_{\mathrm{L}}^r\cup\boldsymbol{X}_{\mathrm{U}}^r=\boldsymbol{X}_{\mathrm{T}}$。

一个参数为$\boldsymbol{\theta}$的目标检测模型$f_{\boldsymbol{\theta}}^r$在$\boldsymbol{X}_{\mathrm{L}}^r$上训练,其在$\boldsymbol{X}_{\mathrm{V}}$上的性能为$Z(\boldsymbol{X}_{\mathrm{V}}|f_{\boldsymbol{\theta}}^r)$,通常通过平均精度均值(mAP)来衡量。给定预算$b$,主动学习算法选择一个大小为$b$的查询集$\boldsymbol{X}_{\mathrm{Q}}^r\subseteq\boldsymbol{X}_{\mathrm{U}}^r$进行标注。

完成标注后,第$r+1$轮的已标注集$\boldsymbol{X}_{\mathrm{L}}^{r+1}=\boldsymbol{X}_{\mathrm{L}}^r+\boldsymbol{X}_{\mathrm{Q}}^r$和未标注集$\boldsymbol{X}_{\mathrm{U}}^{r+1}=\boldsymbol{X}_{\mathrm{U}}^r-\boldsymbol{X}_{\mathrm{Q}}^r$。最后,检测模型在$\boldsymbol{X}_{\mathrm{L}}^{r+1}$上训练以获得新一轮的模型$f_{\boldsymbol{\theta}}^{r+1}$,其性能为$Z(X_V|f_{\boldsymbol{\theta}}^{r+1})$。经过$k$轮主动学习后,可评估主动学习算法的性能提升效果,定义性能增益为

$$\Delta Z^{k,0}=Z(\boldsymbol{X}_{\mathrm{V}}|f_{\boldsymbol{\theta}}^k)-Z(\boldsymbol{X}_{\mathrm{V}}|f_{\boldsymbol{\theta}}^0)\tag{1}$$

式中,$Z(\boldsymbol{X}_{\mathrm{V}}|f_{\boldsymbol{\theta}}^0)$表示初始模型。该指标可以衡量主动学习过程中模型性能的累积改进程度。设计合理的样本查询策略,旨在在相同预算下获得更大的$\Delta Z^{k,0}$,从而实现主动学习算法的最优化。

1.2 整体框架

第r轮主动学习时,在第r轮的未标注池上,检测器进行推理得到候选框的类别分布概率$\boldsymbol{p}^{\mathrm{det}}$并计算基础不确定性。同时,CLIP对候选区域计算语义分布概率$\boldsymbol{p}^{\mathrm{clip}}$,两者融合得到实例级融合不确定性,按得分排序后选取top-k进入候选集。为提升样本多样性,本文聚合图像中各类别的CLIP特征,构建类别级结构矩阵,据此计算图像间的结构相似性,并以最大边际相关算法(maximal marginal relevance, MMR)从候选集中选出代表性子集作为最终查询集。人工标注完成后加入训练集,进行模型训练,进入下一轮主动学习迭代轮次

1.3 语义调制不确定性采样(SMUS)

在主动目标检测任务中,模型在类别边界样本和高置信错误预测样本上的不确定性表达能力,将直接影响样本选择策略的有效性。传统方多基于检测模型的置信度分布信息如最大熵、最小置信度、最小分类间隔等来构建不确定性评分,然而这些方法受限于检测其自身的判别边界,无法准确刻画样本的不确定性,从而错过应采样的重要目标。

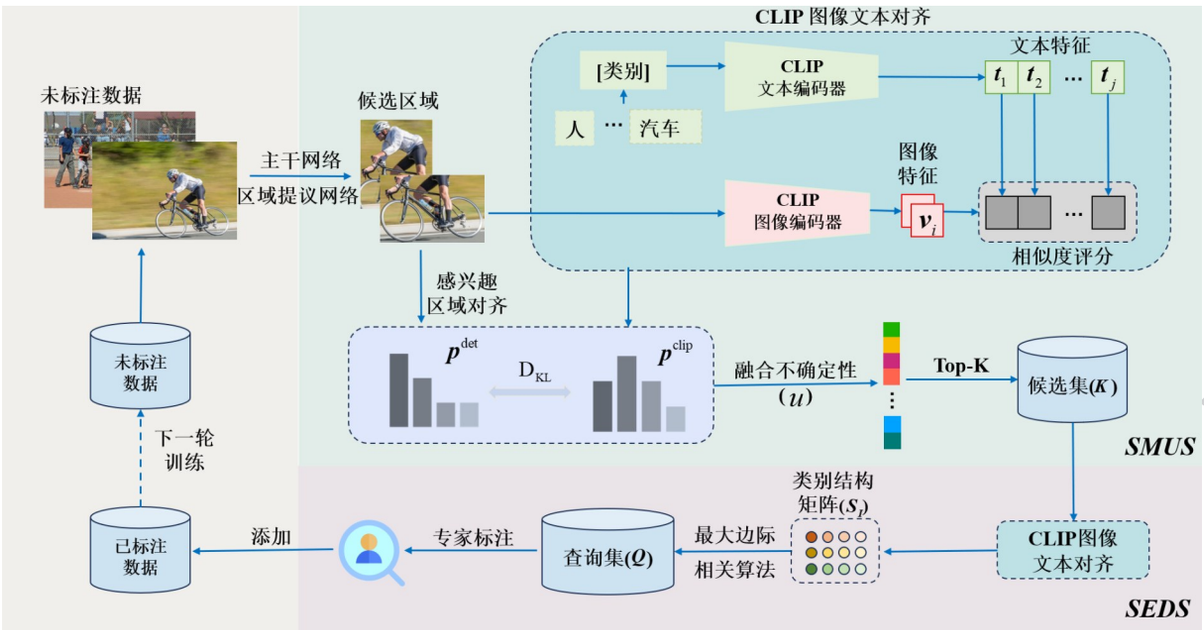


图1　整体框架

Fig. 1　Overall framework

为克服上述不足，本文引入预训练视觉语言模型CLIP的语义感知能力，创新性地构建一种融合检测模型与CLIP语义一致性的跨模态不确定性评估机制，名为语义调制不确定性采样（semantic-modulated uncertainty sampling，SMUS）。该机制通过衡量检测模型与语言模型在类别分布预测上的语义分歧，校正模型在极端情况下的不确定性评估，显著提升主动采样阶段对语义难例的判别能力，从而增强主动学习的样本选择精度。

使用当前目标检测模型f_θ^r在未标注图像集$\boldsymbol{X}_{\mathrm{U}}^r$上进行前向推理，对图像$\boldsymbol{x}$生成预测框集合$\{b_i\}_{i=1}^{M_x}$，$M_x$表示预测框个数，每个目标框的类别概率分布为$\boldsymbol{p}^{\mathrm{det}} \in \mathbf{R}^C$，其中$C$为类别数。则这个目标框的基础不确定性由香农熵表示为

$$u^{\mathrm{det}} = -\sum_{c=1}^{C} \boldsymbol{p}_c^{\mathrm{det}} log \boldsymbol{p}_c^{\mathrm{det}} \tag{2}$$

式中，$\boldsymbol{p}_c^{\mathrm{det}}$表示目标实例为类别$c$的概率，该指标越大表示模型越不确定，通常在预测置信度分布接近均匀时达到最大。

同时，利用CLIP模型获取预测框的语义概率分布。具体地，将预测框区域图像块输入CLIP的视觉编码器获得视觉特征$\boldsymbol{v}_i$。同时，将所有类别标签$\boldsymbol{y}_j$转换为提示文本如（如"a photo of [CLASS]"），经文本编码器得到语义向量$\boldsymbol{t}_j$。计算图像块视觉特征与所有类别文本之间的余弦相似度，并经softmax归一化，得到语义概率分布$\boldsymbol{p}^{\mathrm{clip}} \in \mathbf{R}^C$。使用检测器输出分布$\boldsymbol{p}^{\mathrm{det}}$和CLIP的语义概率分布$\boldsymbol{p}^{\mathrm{clip}}$构建语义分歧评分$u^{\mathrm{clip}}$

$$u^{\mathrm{clip}} = \mathrm{D}_{\mathrm{KL}}(\boldsymbol{p}^{\mathrm{det}} \| \boldsymbol{p}^{\mathrm{clip}}) \tag{3}$$

式中，D_{KL}表示KL散度，这个评分可以反映两种模型在类别判别上的语义偏离程度，敏感地捕捉模型在语义模糊、类别混淆或结构误判等场景下的不一致性，不一致性越高，得分越高。

为增强模型对语义错误预测的判别能力，本文采用语义一致性调制机制，将检测模型自身的不确定性与语言模型的语义支持一致性进行耦合。最终融合不确定性表示为

$$u = u^{\mathrm{det}} \cdot (1 + \gamma \cdot u^{\mathrm{clip}}) \tag{4}$$

式中，$\gamma > 0$为响应强度因子，控制语言模型对最终评分的放大效应。这个机制可以在检测器和CLIP不一致时显著提升不确定性评分，从而增强对于潜在难例的敏感性。

在目标检测任务中，由于类别分布严重不均或类别学习难度存在较大差异，直接使用统一度量可能会导致采样过程偏向高频或易分类别。为此，本文参考PPAL（Yang等，2024）方法引入类别难度动态加权机制。基本思想是对每类样本的训练难度d_c进行动态建模，据此调整该类样本的不确定性

权重w_c。

$$w_c = 1 + \alpha\beta \cdot \log(1 + (e^{1/\alpha} - 1) \cdot d_c) \tag{5}$$

式中，α控制不确定性权重的变化速度，β控制不确定性权重的上限。

最终，图像x的不确定性得分由所有预测框的不确定性加权求和而得，表示为

$$\mathrm{Unc}(\boldsymbol{x}) = \sum_{i=1}^{M_x} w_{c(i)} \cdot u_i \tag{6}$$

式中，$w_{c(i)}$为预测框b_i所属类别c的加权系数。

随后，依据图像级不确定性得分$\mathrm{Unc}(\boldsymbol{x})$对所有未标注图像进行排序，选取前$\delta \cdot b$个图像构建候选集$\boldsymbol{K}$，其中$\delta > 1$为候选集扩展系数，$b$为每轮标注预算。本文提出的方法在提升信息密度的同时为多样性采样阶段提供语义分布更均衡的采样空间。

1.4 语义增强多样性采样(SEDS)

在主动学习任务中，单纯依赖不确定性指标进行样本选择可能导致候选样本之间的高度冗余，降低模型的学习效率。为此本文在不确定性采样基础上引入多样性采样机制，以确保选取的样本在特征空间中具有良好的覆盖性与代表性。

在目标检测场景中，传统多样性采样通常直接使用检测器全局特征衡量样本间相似度。然而，检测器在早期迭代中的特征表达能力有限，难以准确反映样本的真实语义多样性。同时，常见的图像整体特征平均化或最大池化方法只能捕捉全局趋势，难以刻画图像中目标实例间的组合结构差异。为此，本文提出语义增强多样性采样方法(semantic-enhanced diversity sampling, SEDS)，这个方法使用CLIP将每张图像表示为一个结构化的类别-特征矩阵$\boldsymbol{S}_I \in \mathbf{R}^{C \times d}$，其中$C$是类别数，$d$是特征维度，矩阵第$c$行对应图像中类别$c$的平均特征向量中心。具体过程如下

使用当前目标检测模型f_θ^r在不确定性候选集上进行前向推理，对图像$\boldsymbol{x}$生成预测框集合$\{b_i\}_{i=1}^{M_x}$，M_x表示预测框个数。随后，本研究将每个预测框区域图像块输入CLIP视觉编码器获得其视觉特征$\boldsymbol{v}_i$，并进一步与预定义的类别模板文本作余弦相似度匹配，得分最高的类别文本作为类别标签$\boldsymbol{c}_i$。

本研究对图像中所有目标实例的特征，按类别进行分组，并对每类求均值，从而构建图像级结构表示

$$\boldsymbol{S}_I[c] = \frac{1}{|\boldsymbol{V}_c|} \sum_{\boldsymbol{v}_i \in \boldsymbol{V}_c} \boldsymbol{v}_i \tag{7}$$

式中，$\boldsymbol{V}_c = \{\boldsymbol{v}_i | \boldsymbol{c}_i = c\}$，表示图像中类别$c$的所有实例，若图像中未出现类别$c$，则对应向量为零向量，最终得到统一结构的矩阵$\boldsymbol{S}_I \in \mathbf{R}^{C \times d}$。

为实现图像间结构感知的匹配，本研究基于类别对齐策略定义图像相似度。在实际任务中，由于大多数图像仅包含部分类别，若在所有类别上平均匹配，易导致缺失类别引起的相似度稀释。因此，本研究采用仅在图像间共同出现类别上计算平均余弦相似度的策略

$$\mathrm{Sim}(\boldsymbol{I}_a, \boldsymbol{I}_b) = \frac{\sum_{c \in \boldsymbol{C}_a \cap \boldsymbol{C}_b} \cos(\boldsymbol{S}_{I_a}[c], \boldsymbol{S}_{I_b}[c])}{|\boldsymbol{C}_a \cup \boldsymbol{C}_b| + \varepsilon} \tag{8}$$

式中，$\boldsymbol{C}_a, \boldsymbol{C}_b$表示图像$\boldsymbol{I}_a, \boldsymbol{I}_b$中出现的类别集合。$\varepsilon = 10^{-6}$为平滑项。该策略能够准确地反映图像在共享类别结构上地匹配程度，避免冗余类别位对结构对齐结果的干扰。

在样本选择阶段，传统多样性采样多基于当前轮候选集内部评价，忽视了跨轮次的全局分布约束，易导致样本在多轮主动学习过程中分布聚集，降低整体数据覆盖率。因此，本研究基于上述结构相似度引入最大边际相关性方法，兼顾模型不确定性与样本之间的多样性信息。定义候选图像$\boldsymbol{x}_i$的边际相关性为

$$MR_i = \lambda \cdot \overline{\mathrm{Unc}(\boldsymbol{x}_i)} - (1 - \lambda) \max_{\boldsymbol{x}_j \in \boldsymbol{Q}} \mathrm{Sim}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{9}$$

式中，$\overline{\mathrm{Unc}(\boldsymbol{x}_i)}$由$\mathrm{Unc}(\boldsymbol{x})$经Min-Max归一化处理后得到，$\boldsymbol{Q}$为已选集合，包含之前轮次已标注和当前轮次已选择，$\lambda \in (0, 1)$为融合权重参数。

每轮采样过程中，每次从候选池中贪婪地选取MR_i得分最高的样本加入采样集并同步更新$\boldsymbol{Q}$，直至选满b张，该策略实现了跨轮全局多样性约束，同时兼顾样本的信息增益。

2 实验设置

2.1 数据集设置

本实验使用两个主流目标检测数据集对提出的SEAL方法进行系统性实验验证，分别为MS COCO和Pascal VOC。在与之前的工作进行比较时，本实验遵循他们的数据集划分设置(Wu等，2022；Yang等，2024；Yuan等，2021)，以确保公平比较。

对于COCO数据集，本实验使用train2017集进行训练，并在mini-val集上评估模型。训练初始阶段随机选取2.5%的图像作为初始标注集，主动学习共进行4轮，每轮查询2%的样本。评估指标采用IoU阈值为0.5~0.95的平均精度(AP)。

对于Pascal VOC数据集，本实验使用train2007+train2012进行训练，使用test2007进行测试。所有的消融实验在Pascal VOC上进行，采用统一的设置：训练初始阶段随机选取5%的图像作为初始标注集，主动学习共进行6轮，每轮查询2.5%的样本。评估指标采用交并比(IoU)阈值为0.5的全类平均精度(mAP)。

为了减弱初始标注集随机性的影响，本实验对COCO以及Pascal VOC都是使用三个不同的初始训练集运行所有实验，并记录平均性能。

2.2 模型设置

本文默认采用RetinaNet(Lin等，2017)为基础目标检测模型，骨干网络为ResNet-50(He等，2016)。同时还使用以ResNet-50为骨干网络的Faster R-CNN(Ren等，2015)作为检测模型。语义评估模块引入CLIP模型，采用ViT-B/32作为视觉编码器，并基于开源目标检测框架MMDetection(Chen等，2019)实现全部功能模块。

在训练过程中，对COCO和Pascal VOC两个数据集，检测模型均训练26个epoch，初始学习率为0.001，第20个epoch后将学习率衰减为原始的0.1。式(4)中的γ设置为2，式(5)中的α,β参考原论文设置为0.3和0.2，候选集扩展系数δ设置为4，式(9)中的λ设置为0.7。所有实验均在单张NVIDIA RTX 4090D GPU上完成。

3 实验及结果分析

3.1 对比实验

为验证本文所提出的SEAL方法在不同基线和数据集下的性能，本研究在COCO和Pascal VOC两个主流目标检测数据集上，分别采用Faster R-CNN和RetinaNet两种检测器进行系统性实验。具体而言，有以下三个设置：Faster R-CNN on COCO(Wu等，2022)、RetinaNet on COCO(Yuan等，2021)、RetinaNet on Pascal VOC(Yuan等，2021)。对于第一个基准，本研究参考(Wu等，2022)中的实验设置，对于后两个基准，参考(Yuan等，2021)中的实验设置。

在每组设置下，本研究与多种现有主动学习方法进行了对比，包括随机采样(Random)、熵方法(Entropy)(Shannon，1948)、CoreSet(Sener等，2017)、CDAL(Yu等，2022)、MIAL(Yuan等，2021)、DivProto(Wu等，2022)、PPAL(Yang等，2024)方法。各方法均设置相同的采样轮数与训练参数，确保对比的公平性。其中对于熵方法，本实验将所有检测到的对象的分类熵求均值作为图像不确定性。

3.1.1 Pascal VOC数据集的实验结果

图2中记录了本文所提出的方法在RetinaNet on Pascal VOC基准上，不同标注比例下各方法的mAP@0.5，并与其他主流的主动学习算法进行了比较。可以看到，SEAL的检测性能优于其他方法。具体的，当使用7.5%、10%、和12.5%的样本时，它分别比最先进的方法高出0.82%、0.94%和0.91%。在使用20%的样本下时，SEAL实现了72.4%的检测mAP，优于最先进的方法0.8%。曲线整体位于所有其他方法之上，这些结果验证了SEAL方法的有效性。

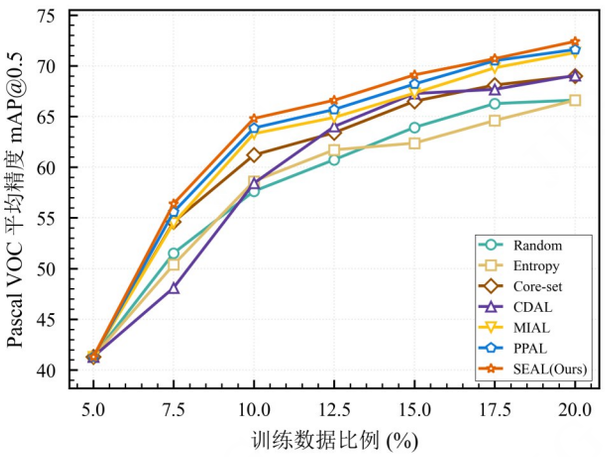


图2 Pascal VOC上基于RetinaNet的实验

Fig. 2 Experiments based on RetinaNet on Pascal VOC

3.1.2 MS COCO数据集的实验结果

MS COCO是一个具有更多类别、更密集对象和更大尺度变化的数据集，图3中记录了本文所提出的方法在RetinaNet on COCO基准上AP@[0.5:0.95]随标注比例的变化，并与其他主流的主动学习算法进行了比较。可以看到，SEAL的检测性能优于其他方法。具体的，当使用4%、6%和8%的样本时，分别比最先进的方法高出0.84%、1%和0.79%。在

使用10%的样本时，SEAL方法实现了23.9%的检测AP，优于最先进的方法0.5%。达到了100%样本监督训练结果(36.5%)的65.48%。

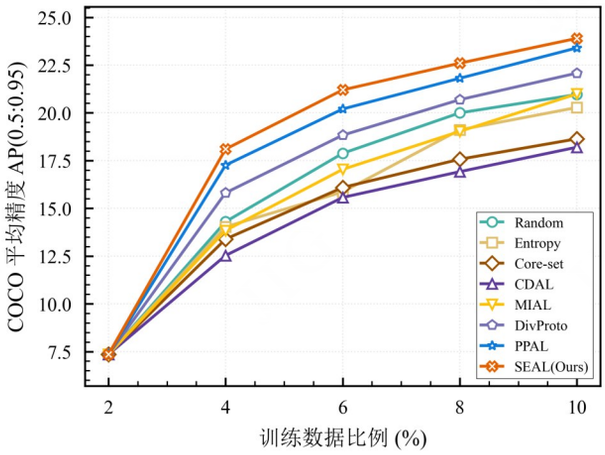


图3　MS COCO上基于RetinaNet的实验

Fig. 3　Experiments based on RetinaNet on MS COCO

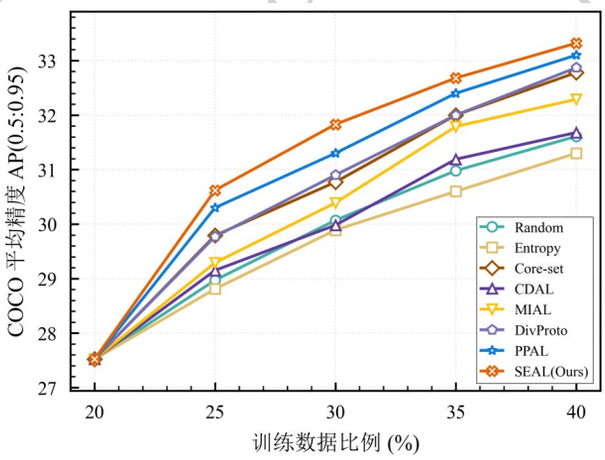


图4　MS COCO上基于Faster R-CNN的实验

Fig. 4　Experiments based on Faster R-CNN on MS COCO

图4记录了本文所提出的方法在Faster R-CNN on COCO基准上AP@[0.5:0.95]随标注比例的变化，并与其他主流的主动学习算法进行了比较。可以看到，SEAL的检测性能优于其他方法。具体的，当使用25%、30%、和35%的样本时，它分别比最先进的方法高出0.32%、0.53%和0.28%。在使用40%的样本下时，SEAL实现了33.32%的检测AP，优于最先进的方法0.22%。达到了100%样本监督训练结果(37.4%)的89.1%。

可以观察到，在三种典型设置下，所提出的SEAL方法在各个标注比例下均显著优于现有方法，始终保持最高的检测性能。上述实验结果充分验证了所提方法在不同基线和数据集上的普适性和优势。

3.2　消融实验

所有的消融实验均在Pascal VOC数据集上开展，采用RetinaNet为目标检测模型，骨干网络为ResNet-50。

3.2.1　SEAL方法的有效性

在本小节，将对语义调制不确定性采样(SMUS)和语义增强多样性采样(SEDS)进行消融研究。实验结果如表1所示，其中第一轮为主动学习的启动轮次，mAP均为41.3%。

本实验首先将SMUS方法和随机采样(Random)以及基于熵的采样方法(Entropy)进行了对比。这里Entropy方法使用图像中实例熵的均值作为图像级不确定性。可以看到，在全程随机采样、Entropy+None和SMUS+None三种组合中，SMUS方法采样的样本训练出的模型性能最佳。然后，本实验分别固定多样性方法为SEDS，固定不确定性采样方法为SMUS。

表1　Pascal VOC上的模块消融实验

Table 1　Module ablation experiments on Pascal VOC.

阶段1	阶段2	在不同比例标注样本上的mAP(%)					
		7.5%	10%	12.5%	15%	17.5%	20%
Random	None	51.3	57.7	60.4	63.8	66.1	66.4
Entropy	None	50.4	58.1	61.7	62.4	64.6	66.4
SMUS	None	53.9	62.8	66.4	68.6	69.6	70.3
Random	SEDS	57.6	62.8	66.5	68.5	69.9	70.8
Entropy	SEDS	52.9	61.4	64.3	65.1	68.1	69.5
SMUS	Random	**57.7**	62.9	66.3	67.8	69.7	70.4
SMUS	Global-det	56.1	63.2	65.6	67.5	68.5	70.2
SMUS	Global-clip	56.3	62.5	65.4	68.0	69.0	70.6
SMUS	SEDS	56.4	**64.8**	**66.6**	**69.1**	**70.7**	**72.4**

注：粗体为最优结果

当多样性采样方法固定为SEDS时，本实验对Random、Entropy、SMUS三种方法进行了对比，结果表示SMUS方法是最优的。

当不确定性采样方法固定为SMUS时，本实验将SEDS方法和随机采样(Random)、检测器全局相

似性(Global-det)、CLIP全局相似性(Global-clip)

进行了比较，这里Global-det方法的相似性度量方法为骨干网络最后一层特征图平均池化后的余弦相似度；Global-clip使用CLIP的视觉编码器提取的图片视觉特征的余弦相似度进行图像间的相似性度量。实验结果表明，本文所提出的多样性采样方法SEDS优于其他多样性采样方法。

消融实验进一步证明，所提出的语义调制不确定性采样方法SMUS与语义增强多样性采样方法SEDS均对整体性能提升起到关键作用，为实际应用的低成本标注方案提供了有力支持。

3.2.2 超参数

本小节讨论式(4)中的响应强度因子γ，候选集扩展系数δ，式(9)中的融合权重参数λ对整体性能的影响。需要说明，下面的实验中第一轮为主动学习的启动轮次，mAP均为41.3%。

表2展示了式(4)中的响应强度因子γ设置为不同值时对整体性能的影响，本实验将γ分别设置为0.5、1、2、3、5，实验结果表明，当$\gamma=2$时，主动学习性能达到最优。过小难以突出语义分歧，过大则容易放大噪声。

表2 响应强度因子γ对结果的影响

Table 2 Effect of γ on the results

γ	在不同比例标注样本上的mAP(%)					
	7.5%	10%	12.5%	15%	17.5%	20%
0.5	58.4	64.2	67.7	69.5	71.1	71.1
1	59.3	64.2	67.9	68.7	70.5	71.6
2	56.4	64.8	66.6	69.1	70.7	72.4
3	57.9	64.1	67.1	69.0	70.5	71.7
5	58.3	64.1	66.9	68.9	70.6	71.8

表3展示了候选集扩展系数δ设置为不同值时对整体性能的影响，本实验将δ分别设置为2、3、4、5、6，实验结果表明，当$\delta=4$时，主动学习性能达到最优。合理的候选集扩展有助于提升样本选择的多样性和代表性。δ过小导致样本空间受限，δ过大则可能引入噪声样本，影响性能。

表4展示了融合权重参数λ对整体性能的影响，本实验将λ分别设置为0.1、0.3、0.5、0.7、0.9，当$\lambda=3$时，主动学习算法达到最优，此时两种信息源的融合最为协调。若λ过大或过小，模型对某一信息源的依赖性增强，可能导致整体性能下降。

表3 候选集扩展系数δ对结果的影响

Table 3 Effect of δ on results

δ	在不同比例标注样本上的mAP(%)					
	7.5%	10%	12.5%	15%	17.5%	20%
1	53.9	62.8	66.4	68.6	69.6	71.3
2	56.5	63.2	66.1	69.0	70.6	71.3
3	55.5	63.1	66.9	68.6	70.1	71.6
4	56.4	64.8	66.6	69.1	70.7	72.4
5	56.4	64.3	67.5	69.4	71.1	71.5
6	57.5	64.2	68.2	69.4	71.3	71.1

表4 融合权重参数λ对结果的影响

Table 4 Effect of λ on results

λ	在不同比例标注样本上的mAP(%)					
	7.5%	10%	12.5%	15%	17.5%	20%
0.1	56.6	64.2	67.1	69.1	70.7	71.4
0.3	56.4	64.8	66.6	69.1	70.7	72.4
0.5	54.9	63.8	66.9	68.6	70.3	70.8
0.7	55.1	61.3	66.0	68.9	70.0	71.2
0.9	54.7	62.9	65.2	67.7	69.3	70.9

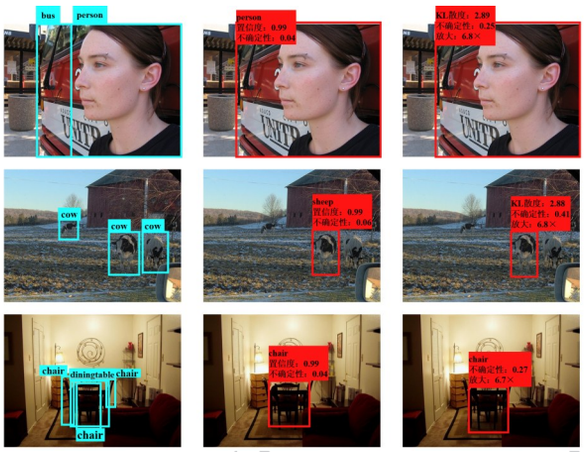
整体实验结果说明，不同超参数的调整对最终性能有一定影响，适当选择可有效提升主动学习的样本选择效率和检测精度。但总体上，本文提出的方法的性能是相对稳定的。

3.3 可视化

图5展示了三个典型场景下的检测结果，分别对应(1)前景干扰与背景遮挡、(2)类别混淆、(3)室内多目标复杂场景。左为人工标注(仅供参考)，中为检测器的错误预测，右侧为经CLIP调制的不确定性结果。这里仅展示了检测器产生错误预测的目标框，以便突出分析模型的不确定性表达能力。

第一行中，前景的人物目标非常突出，导致检测器忽略了背景中相对难以识别的“bus”，最终只对“person”做出了高置信度低不确定性的预测。这一现象常见于目标明显度悬殊、背景信息复杂的场景，检测器往往对显著目标过于自信，忽略了难以分辨的背景目标。

第二行中，在户外多目标环境下，检测器对本应为“cow”的目标错误地识别为“sheep”。虽然检测器



(a)真实标注 (b)检测器错误预测 (c)调制后结果

图5 三类典型场景的目标检测与不确定性分析结果

Fig. 5 Detection and Uncertainty Analysis Results in Three Typical Scenarios((a)ground truth;(b)detector erroneous predictions;(c)CLIP modulated results)

输出的置信度依然很高,不确定性很低,但实际上该场景存在较强的类别混淆,说明检测器难以区分细粒度类别,尤其在外观相似的情况下。

第三行,展示了室内多目标场景,其中检测器将“diningtable”误报为“chair”。模型对小目标、遮挡目标或布局复杂场景可能产生错误,并表现出低不确定性分数,反映出模型对困难目标缺乏足够警觉。

在上述场景中,传统检测器面对类别混淆、复杂环境或显著目标干扰时,往往会产生高置信度但低不确定性的“过度自信”现象。引入CLIP调制后,不确定性分数显著提升,能够更加敏感地捕捉模型在类别判断上的犹豫与风险。该方法可以有效暴露主动学习任务中的难例,有助于指导采样算法更有针对性地选择需重点标注的高价值样本,从而提升主动学习的整体效率和模型泛化能力。

4 结 论

针对主动学习目标检测的两类瓶颈,高置信度误检难以被不确定性度量感知、基于全局视觉特征的多样性采样并不能很好的衡量目标检测背景下图片之间的相似度,本文提出跨模态先验驱动的两阶段主动学习框架SEAL。其核心思路是引入CLIP的视觉—语言对齐能力,第一阶段以检测器与CLIP的一致性构造实例级不确定性,从语义维度抑制“高置信但错误类别”的预测;第二阶段在CLIP表征上形成类别—样本结构,按结构化语义差异优先选择类别覆盖广的样本,从而提高标注数据的信息量与代表性。多个基准上的实验结果表明,SEAL在不依赖模型改造的条件下,SEAL能够以较低的标注成本显著提升模型在检测精度与鲁棒性。SEAL方法在处理语义易混淆目标、初期表征不充分以及目标密集分布的场景时具有显著的适用优势。然而,本研究仍存在一定的局限性。首先,跨模态先验的效能受限于预训练分布与目标任务领域的对齐程度,在处理具有显著分布偏移的专业垂直领域(如遥感或医疗影像)时,其引导作用可能减弱。其次,候选框级别的特征提取与结构相似性度量在处理大规模未标注池时会引入额外的计算开销。此外,文本提示的构造方式对最终的对齐结果具有一定的敏感性。后续工作将围绕以下方面展开,其一,与增量学习机制进行联合设计,面向类别演化场景建立更稳健的覆盖与记忆;其二,引入近似检索与特征缓存以及自动化提示词工程等工程优化进一步降低系统的计算成本并提升模型在特定领域的自适应泛化性能。

参考文献(Reference)

Agarwal S, Arora H, Anand S and Arora C. 2020. Contextual diversity for active learning // Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part XVI 16. Berlin: Springer: 137-153[DOI: 10.1007/978-3-030-58517-4_9]

Aghdam H H, Gonzalez-Garcia A, Weijer J v d and López A M. 2019. Active learning for deep detection neural networks // IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, USA: IEEE: 3672-3680[DOI: 10.1109/iccv.2019.00377]

Ash J T, Zhang C, Krishnamurthy A, Langford J and Agarwal A. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds // International Conference on Learning Representations (ICLR). OpenReview.net

Brust C-A, Käding C and Denzler J. 2018. Active learning for deep object detection[EB/OL]. [2018-09-26]. https://arxiv.org/pdf/1809.09875

Budd S, Robinson E C and Kainz B. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical image analysis, 71: 102062[DOI: 10.1016/j.media.2021.102062]

Cao J, Li Y, Sun H, Xie J, Huang K and Pang Y. 2022. A survey on deep learning based visual object detection. Journal of image and graphics, 27(6): 1697-1722 (曹家乐, 李亚利, 孙汉卿, 谢今,

黄凯奇，庞彦伟. 2022. 基于深度学习的视觉目标检测技术综述. 中国图象图形学报，27(6)：1697-1722[DOI：10.11834/jig.220069]

Chen K，Wang J，Pang J，Cao Y，Xiong Y，Li X，Sun S，Feng W，Liu Z and Xu J. 2019. MMDetection：Open mmlab detection toolbox and benchmark[EB/OL]. [2019-06-17]. https://arxiv.org/pdf/1906.07155

Choi J，Elezi I，Lee H-J，Farabet C and Alvarez J M. 2021. Active learning for deep object detection via probabilistic modeling // IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，USA：IEEE：10264-10273[DOI：10.1109/iccv48922.2021.01010]

Citovsky G，DeSalvo G，Gentile C，Karydas L，Rajagopalan A，Rostamizadeh A and Kumar S. 2021. Batch active learning at scale // Advances in Neural Information Processing Systems.（NeurIPS）Cambridge，USA：MIT Press：11933-11944

Deng J，Dong W，Socher R，Li L-J，Li K and Fei-Fei L. 2009. Imagenet：A large-scale hierarchical image database // IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，USA：IEEE：248-255[DOI：10.1109/cvpr. 2009.5206848]

Du Y，Liu Z，Li J and Zhao W X. 2022. A survey of vision-language pre-trained models // Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence，IJCAI-22. Vienna，Austria：International Joint Conferences on Artificial Intelligence Organization：5436－5443[DOI：10.24963/ijcai.2022/762]

Gal Y，Islam R and Ghahramani Z. 2017. Deep bayesian active learning with image data // International Conference on Machine Learning（ICML）. New York，USA：ACM：1183-1192[DOI：10.48550/arXiv.1703.02910]

Gu X，Lin T-Y，Kuo W and Cui Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation // International Conference on Learning Representations（ICLR）. OpenReview.net

He K，Zhang X，Ren S and Sun J. 2016. Deep residual learning for image recognition // IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，USA：IEEE：770-778[DOI：10.3390/app12188972]

Kao C-C，Lee T-Y，Sen P and Liu M-Y. 2019. Localization-aware active learning for object detection // Asian Conference on Computer Vision（ACCV）. Berlin，German：Springer：506-522[DOI：10.1007/978-3-030-20876-9_32]

Lewis D D. 1995. A sequential algorithm for training text classifiers：Corrigendum and additional data // Acm Sigir Forum. New York，USA：ACM：13-19[DOI：10.1145/219587.219592]

Lin T-Y，Goyal P，Girshick R，He K and Dollár P. 2017. Focal loss for dense object detection // IEEE/CVF International Conference on Computer Vision（ICCV）. Piscataway，USA：IEEE：2980-2988[DOI：10.1109/iccv.2017.324]

Lyu M，Zhou J，Chen H，Huang Y，Yu D，Li Y，Guo Y，Guo Y，Xiang L and Ding G. 2023. Box-level active detection // IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，USA：IEEE：23766-23775[DOI：10.1109/cvpr52729.2023.02276]

Nguyen H T and Smeulders A. 2004. Active learning using pre-clustering // International Conference on Machine Learning（ICML）. New York，USA：ACM：79[DOI：10.1145/1015330.1015349]

Prabhu V，Chandrasekaran A，Saenko K and Hoffman J. 2021. Active domain adaptation via clustering uncertainty-weighted embeddings // IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，USA：IEEE：8505-8514[DOI：10.1109/iccv48922.2021.00839]

Radford A，Kim J W，Hallacy C，Ramesh A，Goh G，Agarwal S，Sastry G，Askell A，Mishkin P and Clark J. 2021. Learning transferable visual models from natural language supervision // International Conference on Machine Learning（ICML）. New York，USA：ACM：8748-8763

Ren S，He K，Girshick R and Sun J. 2015. Faster r-cnn：Towards real-time object detection with region proposal networks // Advances in Neural Information Processing Systems（NeurIPS）. Cambridge，USA：MIT Press[DOI：10.1109/tpami.2016.2577031]

Roh Y，Heo G and Whang S E. 2019. A survey on data collection for machine learning：a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering，33(4)：1328-1347[DOI：10.1109/tkde.2019.2946162]

Scheffer T，Decomain C and Wrobel S. 2001. Active hidden markov models for information extraction // International Symposium on Intelligent Data Analysis（IDA）. Berlin，German：Springer：309-318[DOI：10.1007/3-540-44816-0_31]

Sener O and Savarese S. 2017. Active learning for convolutional neural networks：A core-set approach // International Conference on Learning Representations（ICLR）. OpenReview.net

Settles B. 2009. Active learning literature survey.

Shannon C E. 1948. A Mathematical Theory of Communication. Bell System Technical Journal，27(3)：379-423[DOI：10.1002/j.1538-7305.1948.tb01338.x]

Wang P，Yan Z，Rong X，Li J，Lu X，Hu H，Yan Q and Sun X. 2022. Review of multimodal data processing techniques with limited data. Journal of image and graphics，27(10)：2803-2834（王佩瑾，闫志远，容雪娥，李俊希，路晓男，胡会扬，严启炜，孙显. 2022. 数据受限条件下的多模态处理技术综述. 中国图象图形学报，27(10)：2803-2834)[DOI：10.11834/jig.220049]

Wu J，Chen J and Huang D. 2022. Entropy-based active learning for object detection with progressive diversity constraint // IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，USA：IEEE：9397-9406[DOI：10.1109/cvpr52688.2022.00918]

Xu Z，Yu K，Tresp V，Xu X and Wang J. 2003. Representative sam-

pling for text classification using support vector machines // European Conference on IR Research (ECIR). Berlin, German: Springer: 393-407[DOI: 10.1007/3-540-36618-0_28]

Yan H, Bai J and Zheng H. 2025. Consistency constraint guided network for zero-shot 3D classification. Journal of image and graphics, 30(5): 1450-1465 (晏浩, 白静, 郑虎. 2025. 一致性约束引导的零样本三维模型分类网络. 中国图象图形学报, 30(5):1450-1465)[DOI: 10.11834/jig.240397]

Yang C, Huang L and Crowley E J. 2024. Plug and play active learning for object detection // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, USA: IEEE: 17784-17793[DOI: 10.1109/cvpr52733.2024.01684]

Yoo D and Kweon I S. 2019. Learning loss for active learning // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, USA: IEEE: 93-102[DOI: 10.1109/cvpr.2019.00018]

Yu W, Zhu S, Yang T and Chen C. 2022. Consistency-based active learning for object detection // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, USA: IEEE: 3951-3960[DOI: 10.1109/cvprw56347.2022.00440]

Yuan T, Wan F, Fu M, Liu J, Xu S, Ji X and Ye Q. 2021. Multiple instance active learning for object detection // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, USA: IEEE: 5330-5339[DOI: 10.1109/cvpr46437.2021.00529]

Zareian A, Rosa K D, Hu D H and Chang S-F. 2021. Open-vocabulary object detection using captions // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, USA: IEEE: 14393-14402[DOI: 10.1109/cvpr46437.2021.01416]

Zhdanov F. 2019. Diverse mini-batch active learning[EB/OL]. [2019-01-17]. https://doi.org/10.48550/arXiv.1901.05954

作者简介

王卓,男,硕士研究生,主要研究方向为主动学习和增量学习。E-mail: wangzhuo@nudt.edu.cn

黄开宏,通信作者,男,讲师,主要研究方向为特种机器人的智能传感与控制。E-mail: kaihong.huang@nudt.edu.cn

王斌翊,男,研究员,主要研究方向为计算机控制技术。E-mail: wangbyxx@Outlook.com

肖军浩,男,教授,主要研究方向为具身智能和特种机器人技术。E-mail: junhao.xiao@nudt.edu.cn

李世亮,男,硕士研究生,主要研究方向为模式识别与智能系统。E-mail: 2606236279@qq.com

谢云飞,男,硕士研究生,主要研究方向为4D毫米波雷达目标追踪。E-mail: 1539698951@163.com

王佳,男,硕士研究生,主要研究方向为偏振视觉图像处理与定位。E-mail: wjoffice_anton@163.com