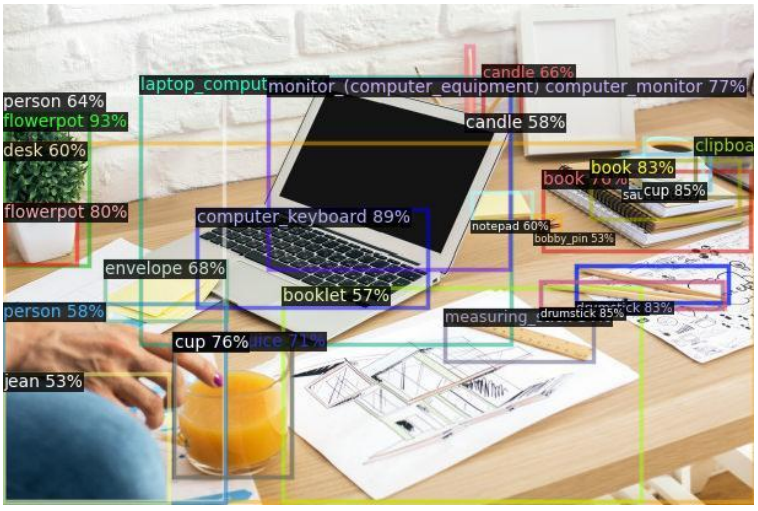
Open Vocabulary Detection 开放世界目标检测竞赛 2023 获胜团队方案分享

王斌
360 人工智能研究院

OVD 技术简介

目标检测是计算机视觉领域中的一项核心任务，其主要目标是让计算机能够自动识别图片中目标的类别，并准确标示每个目标的位置。目前，主流的目标检测方法主要针对闭集目标的开发，即在任务开始之前需要对待检测目标进行类别定义，并进行人工数据标注，通过有监督模型的训练来实现目标检测。这种方法通常适用于待检测目标数量较少的情况，一般限定在几十个类别以内。然而，当待检测目标的类别数量增加到几千甚至万级时，以上述方式进行数据标注已经无法满足需求。同时，已经训练好的模型也无法应对新出现的类别。当新的类别出现时，需要手动进行标注并重新训练模型，整体效率较低。

开放词集目标检测（Open Vocabulary Detection, OVD），亦即开放世界目标检测，提供了解决上述问题的新思路。借助于现有跨模态模型（CLIP[1]、ALIGN[2]、R2D2[3] 等）的泛化能力，OVD 可以实现以下功能：1）对已定义类别的 few shot 检测；2）对未定义类别的 zero-shot 检测。OVD 技术的出现吸引了计算机视觉研究者们的广泛关注，首先，对于已定义类别的 few shot 检测，OVD 的强大泛化能力可以让算法在仅有少量样本的情况下，准确地识别出新的目标类别。其次，对于未定义类别的 zero-shot 检测，OVD 的能力更是令人惊叹。通过学习各种物体的视觉特征和语义信息，OVD 可以在没有见过的类别中进行目标检测，进一步将语言大模型技术引入 OVD，将会进一步提升 OVD 对未知类别的检测能力。OVD 技术有望成为未来目标检测算法开发的新范式。



竞赛介绍

OVD 技术的研究在国内尚处于起步阶段，为了促进国内 OVD 技术的发展，并加强 OVD 技术的生态社区建设，360 人工智能研究院联合中国图象图形学学会于 ICIG2023 大会上开设了 Open Vocabulary Detection Contest - 开放世界目标检测 2023 竞赛。大赛于 4 月 12 日启动报名，报名期间吸引了来自新加坡南洋理工大学、清华大学、北京大学、香港大学、中国科学院自动化研究所紫东太初大模型研究中心、鹏城实验室、华中科技大学、字节跳动、滴滴等知名大学与公司机构共 140 支队伍参加竞赛。此次大赛所使用的赛题数据、竞赛提交平台与赛题设置均由 360 人工智能研究院提供支持。

赛题数据主要涵盖了服装、数码产品等众多商品类目，对于一件商品，均给出了它的图片以及对应的检测框标注信息作为训练数据。商品数据在互联网搜索、推荐中具有重要价值，是非常贴近业务场景的实用数据。其次商品数据集的难度较大，同类别商品之间普遍存在一些细节差异，而这一点也限制了传统目标检测技术的泛化能力，进而体现出 OVD 技术的优势性。

赛题设置：参赛者运用 OVD 相关的方法，对图像中的商品目标进行检测。对于一件商品，主办方会给出它的图片以及 bbox 作为训练数据。目标类别有两类：base 类和 novel 类。类别均为中文商品词组。base 类的目标提供少量已标注的训练样本，novel 类的目标则没有训练样本。评测分别在 base 类的测试集和 novel 类的测试集上进行，评测指标为 novel 和 base 类的 mAP@50，竞赛按照 novel 和 base 类别的整体 mAP@50 排序。

竞赛共分为初赛与复赛两个阶段，由初赛到复赛，赛题难度逐步提升，考验选手对开放世界目标检测赛题的熟悉程度与灵活应变能力。比赛中，各位选手的方案追逐激烈，最终前三名团队的复赛分数十分接近。经过初赛与复赛的层层选拔，最终有 6 支队伍脱颖而出，由来自南洋理工大学的吴思泽团队摘得桂冠。获得二等奖的是来自华中科技大学的 STAR 团队与来自中国科学院自动化研究所紫东太初大模型研究中心的咱们组有名称吗团队，获得三等奖的是来自北京大学的 OVD 团队、来自哈尔滨工业大学的 wzmwzr 团队与来自武汉邮电科学研究院的蓝色闪团队。Open Vocabulary Detection Contest - 开放世界目标检测竞赛的官网链接：开放世界目标检测竞赛 2023 (360cvgroup.github.io)

在各个竞赛团队的积极参与、中国图象图形学学会与 360 人工智能研究院的大力支持下，Open Vocabulary Detection Contest - 开放世界目标检测竞赛已经正式结束，在征集各个竞赛团队的许可后，我们将部分优胜团队的技术方案汇总并公开分享，详见本文下半部分。

排行	组织	模型	novel	base	all
1	吴思泽	final	56.604	52.12	54.362
2	STAR	OVDEA	52.28	53.176	52.728
3	咱们组有名称吗	再给KFC一次机会	50.075	54.158	52.117
4	OVD	f-huge	47.885	47.146	47.516
5	wzmwzr	Test1	47.635	42.649	45.142
6	蓝色闪	IFEAsT30_100_P1-3++	43.317	45.593	44.455

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, 2021.

[3] Xie C, Cai H, Song J, et al. Zero and R2D2: A Large-scale Chinese Cross-modal Benchmark and A Vision-Language Framework[J]. arXiv preprint arXiv:2205.03860, 2022.

冠军方案讲解

团队介绍

来自南洋理工大学的博士生吴思泽

赛题分析

1 数据集

本次主办方提供的是商品数据集，总共 466 个物体类别，其中训练中可见的有 233 个 base 类别，测试时检测器需要能够同时识别 base 类的物体意见另外 233 个 novel 类别的物体。数据集中图片以网购商品图为主，背景通常较为简单，每张图物体数量不多，存在大量以物体为中心（object-centric）的图片，训练集中平均每张图的物体标注数量<2。

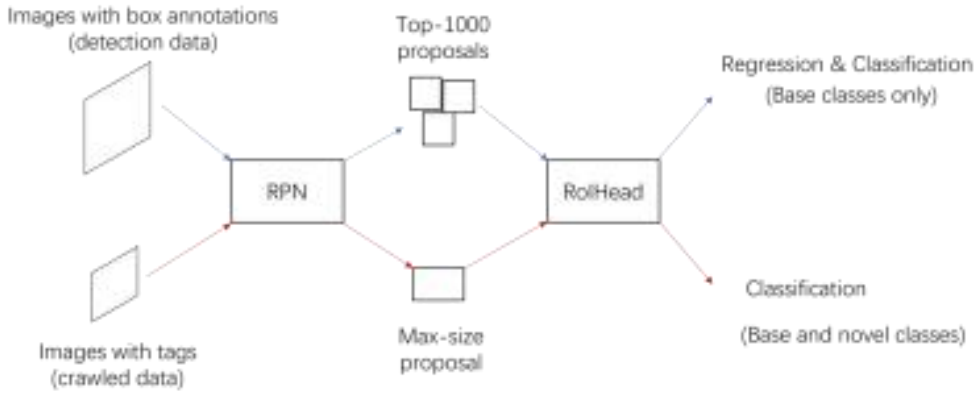
2 解决思路

根据数据集属性，可知互联网中存在大量包含新类别的商品图片，由于图片场景简单，物体单一，在图像层级（image-level）上学习新类别的表征，可很好

泛化到检测上。因此选择基础方案为 Detic，使用爬虫获取带有新类别 tag 的图片，用于 image-level 的训练。

方案总览

采用 Detic[1]的训练策略，同时使用目标检测数据（base 类）和图像分类数据（base 类+novel 类）训练检测器。



方案流程介绍

1 数据处理

选择百度图片为爬取对象，索引关键词为”[中文名称] 商品图片”，为保证类别平衡，novel 和 base 类别均爬取 40 页(大约 1000 张)。每个类别爬取到的图片存到一个路径下，这些图片只有类别 Tag，没有物体框标注。

2 类别名称翻译

为方便使用现有的开源模型（CLIP），需要将 466 个中文名称均翻译成英文，我们使用 google translator 翻译每个名称并人工校对

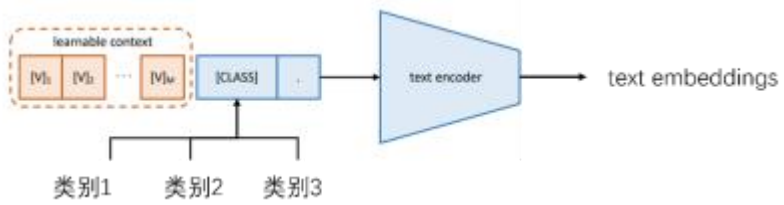
```
> 000 = {dict: 5} {'id': 1, 'supercategory': '', 'name': 'Antiques', 'split': 'seen', 'ch_name': '古董文玩'}
> 001 = {dict: 5} {'id': 2, 'supercategory': '', 'name': 'Berets', 'split': 'seen', 'ch_name': '贝雷帽'}
> 002 = {dict: 5} {'id': 3, 'supercategory': '', 'name': 'chicken', 'split': 'seen', 'ch_name': '鸡肉'}
> 003 = {dict: 5} {'id': 4, 'supercategory': '', 'name': 'CPU', 'split': 'seen', 'ch_name': 'CPU'}
> 004 = {dict: 5} {'id': 5, 'supercategory': '', 'name': 'cup', 'split': 'seen', 'ch_name': '茶杯'}
> 005 = {dict: 5} {'id': 6, 'supercategory': '', 'name': 'eyebrow pencil', 'split': 'seen', 'ch_name': '眉笔'}
> 006 = {dict: 5} {'id': 7, 'supercategory': '', 'name': 'fishing supplies', 'split': 'seen', 'ch_name': '垂钓用品'}
```

3 模型介绍

选择 ResNet50 和 SwinB 作为检测器 backbone，检测器结构为 CenterNet2，使用 Detic 公开的在公开数据集 LVIS 和 ImageNet 上预训练的模型权重作为初始化。CLIP 模型选择 ViT-L-14（只用 text encoder）来得到类别名称的 embeddings。分类的损失函数为 BCE Loss。

4 Learnable Prompt

为了获取类别名称的 text embeddings，在训练过程中学习一组长度为 4 的 learnable prompt 以获得更好的 text 表征。具体方案参考了 coop[2]。



5 重要参数

（1）模型初始化：使用 LVIS 和 ImageNet 上预训练的模型作为初始化

（2）总迭代次数：18000

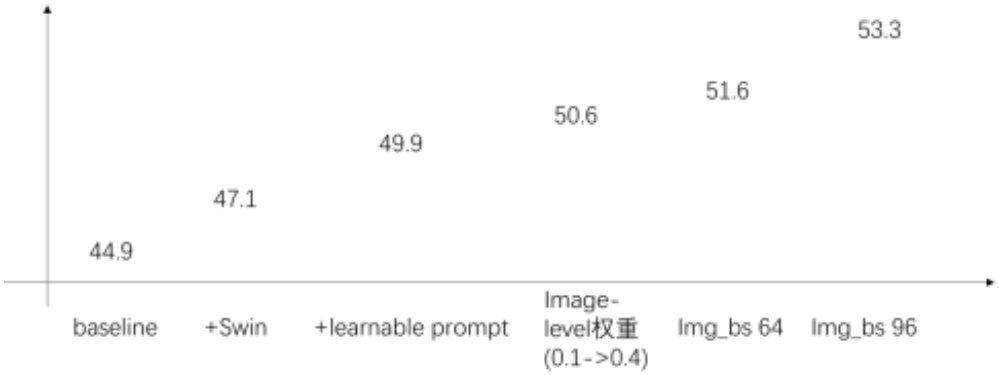
（3）image-level 分支的 batch size: 8x96，检测分支 batch size: 8x4

（4）image-level 的权重：1.2，det 分支权重：1.0

（5）图像分辨率：image-level 分支 448， 检测分支 896

6 测试结果

这里介绍的测试结果是随着我们模块和参数改变的变化，我们初始使用 R50 backbone 作为 baseline, image-level 分支的 batch size 为 32，训练资源 8xV100，增加到 64 之后需要 8xA100 （或者整体 batch size 缩小，迭代数增加）。以下结果均来自初赛。



[1] ADetecting Twenty-thousand Classes using Image-level Supervision, Zhou et.al ECCV 2022.

[2] Prompt Learning for Vision-Language Models, Zhou et.al IJCV 2022.

亚军方案讲解（第二名）

团队介绍

来自华中科技大学的团队，成员有冷福星，易成龙。

赛题分析

1 数据集

训练数据：233 类已知类别的目标检测框

初赛：7401 张图像

复赛：14802 张图像

数据特点：

1 全部是电商类的商品图像

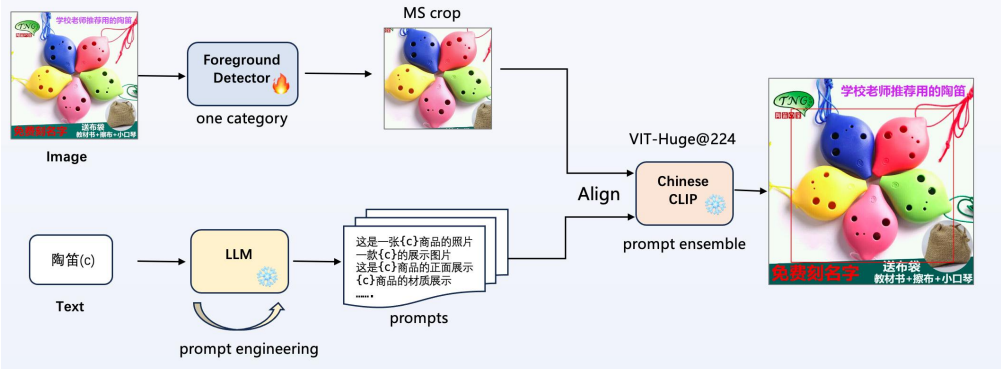
2 单张图像中的目标类别相同

3 存在部分有效的 OCR 信息

2 解决思路

利用前景检测器对图片进行目标定位，利用 LLM 来扩充文本信息，最后结合 ChineseCLIP 进行多模态图文对齐生成类别信息。

方案总览



如图所示，是本次比赛中提出的算法 pipeline，不需要使用提供的类别信息，不引入额外的数据，即可进行任意商品类别的目标检测：

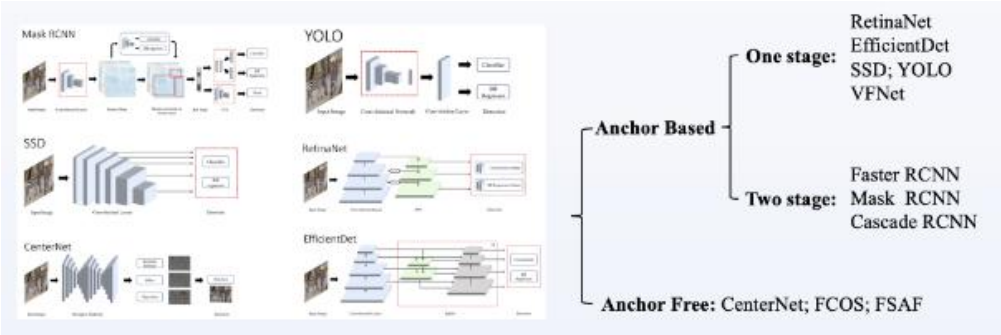
● 前景检测器（Foreground Detector）：不需要使用提供的 233 类类别信息，只使用位置坐标训练一个前景检测器，整个 pipeline 中只有这里进行梯度更新；

● 提示词工程（prompt engineering）：使用大语言模型（LLM）进行半自动化的提示词工程，输入类别 c，给定模板规范，生成更多风格多样的提示词；

● 多模态图文对齐：使用 Chinese CLIP 进行图文特征对齐，进行类别分类，使用提示词集成（prompt ensemble）提高性能；

方案流程介绍

1 前景检测器



当前主流的检测器如图所示，主要包括 Anchor Based 和 Anchor Free 两类检测器，前者精度高但速度慢，后者精度略差但速度快；

● 前景 proposal 使用 WBF（Weighted Boxes Fusio）集成了 CBNetV2_Swin，CascadeRCNN_Convnext，CascadeRCNN_Hornet，CascadeRCNN_resnext101，DetecotoRS_r101，VFNet_resnext101；实际使用 CBNetV2_Swin 单个检测器分数不低，集成在分数提升大概 1 个点；

● 使用训练好的前景检测器检测目标，进行多尺度裁剪（外扩+0，+30 像素），并加入全图（利用有效的 OCR 信息，如图 2 中右上角的陶笛文本）一起进行图文对齐，将 3 个尺度的输出 logits 进行平均；

2 提示词工程

CLIP 模型是双塔结构，直接使用类别信息进行文本对齐的效果不是最佳的，为了充分挖掘文本 encode 的潜力，需要进行一定的提示词工程；在实验中，使用 “c” 和 一张“c”的图片，验证集上后者分数高 5 个点；

可以使用 ChahtGPT/LLMA 2 进行交互，逐步引导 LLM 生成想要的提示词模板；最后得到多条 prompts，可以进行 prompt ensemble，ensemble 的方法有以下三种，实际只使用了最简单的 Uniform averaging；

● Uniform averaging

● Weighted averaging

● Majority Voting


3 消融实验与实验结果

验证集：初赛训练集，（训练中没有使用类别信息，用来评测 CLIP 模型分类能力）

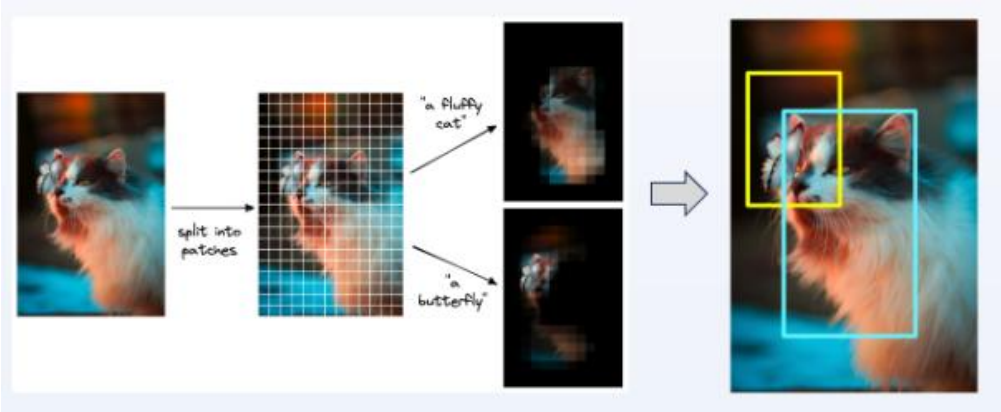
● PE：prompt engineering

● CME：CLIP model ensemble（0.7*VIT-H@224+0.3*VIT-L@336）

Method	No PE	PE	LLM PE	CME	score
VIT-H@224	✓				0.522
VIT-H@224		✓			0.597
VIT-H@224			✓		0.640
VIT-H@224			✓	✓	0.645

4 拓展思路

上述提出的 pipeline 使用了位置信息进行训练，使用 CLIP 也可不进行训练进行任意目标检测：



将图像分成小 patch，滑动窗口 crop 图像送入 CLIP 模型提取图文相似性。每个窗口根据阈值判断目标类别，也可以将当前窗口图像置 0，看整图类别相似性哪个下降最多。但该方法，滑动窗口的方式替代 proposal 的检出比较耗时，实测精度也没有上述方法高；

亚军方案讲解（第三名）

团队介绍

“我们组有名称吗”团队来自中国科学院自动化研究所紫东太初大模型研究中心，紫东太初大模型研究中心致力于构建低功耗万亿突触多模态认知大模型，建立面向开放复杂环境的可解释、可信、可演化的多模态人工智能基础平台，建成新一代人工智能重大基础设施，形成创新应用生态。比赛团队由两名博士生（詹宇飞、杨帆）、一名硕士生（赵弘胤）和一名本科生（王天琦）组成，在朱优松老师指导下共同完成本次比赛，目前团队主要研究方向为视觉大模型、目标检测、开放词汇目标检测及长尾目标检测等。

赛题分析

1 数据集

在开放词汇目标检测的研究中，端到端训练方法由于其在训练速度的优势和公平对比的要求获得了更为广泛的使用。在本次商品场景下的开放世界目标检测竞赛中，主要存在以下四个问题：

（1）噪声大---数据标注噪声大，各类别均存在误标、漏标等情况，标注方式不统一；

（2）定位难---少样本训练设定下，端到端微调精准定位和分类效果差；

（3）易混淆---类别细粒度程度高，且多为商品数据，类内方差大，通用中文图文模型无法有效区分；

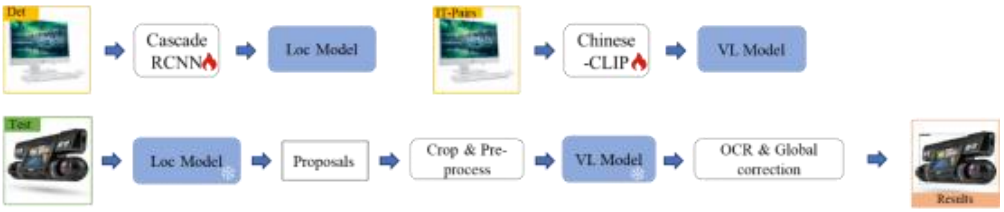
（4）主体判断难---该场景设定下，每张图中只可识别出主要商品，共同出现的其他商品需被抑制。

2 解决思路

为解决上述问题，通过对数据的类别分布和实例位置分布的分析，我们发现图片的实例以单类别形式出现，且居中分布，具备任务解耦的基础。因此，我们选择双阶段的方法，将框回归和商品物体分类进行剥离，分别实现类别无关的框回归以解决定位难和主题判断难得问题，和基于 CLIP 特征的零样本和少样本分类以解决噪声大和易混淆得问题。且将任务拆分为两个子任务，分别迭代，有效提高了优化速度。

方案总览

方案整体框架下图所示，按照子任务拆分，我们将训练分为检测器训练和图文模型优化两部分，将最终优化好的模型在推理阶段进行组合，在推理规则的辅助下完成对场景中的少样本和零样本类别的检测。



方案流程介绍

1 用于目标定位的数据补充

为抑制模型产生大框的倾向和纠正在部分情况下错误产生部件框造成得定位难问题，我们额外爬取 659 张 Base 类别商品图片，利用训练好的模型打伪标签的形式构建，选取置信度大于 0.8 的预测框并采用人工校验的方式进行清洗过滤，去掉其中的局部框等，构建了包含 659 张图片的纠正数据子集，用于模型的微调。

2 目标定位模块

在商品目标定位部分，考虑到在开放词汇目标检测任务下，检测器首先应当定位出所有可能的物体，其中包括不具备检测标注的 novel 类别。因此，我们选择将检测器训练为二分类商品检测器，用于提取图片中可能存在的商品。我们

选择 Cascade-RCNN 训练二分类的商品检测模型，利用多个级联的回归分支提升模型对于物体的识别与定位能力。为提高模型的特征提取能力，我们选择以 Swin-Transformer Small 为骨干网络，Neck 默认使用了 FPN 融合高层语义特征与低层的细节特征，最后输出物体得分大于 0.1 的候选框中选择排名前 100 个检测框。

3 用于目标分类的数据补充

为解决低数据量下的噪声和混淆问题，在开放词汇任务设定的启发下，我们分别采用关键字“类别名称 商品图片”搜索和相似图片搜索的方式，从百度、谷歌、电商平台等网络数据中收集了 70w 的数据用于模型的微调，并利用 ChatGLM 对类别和图片生成描述，提高图文对的语义丰富度，进而增强模型的判别能力，如图 2 所示。通过对微调方式的对比，我们对比了目前较优的三种微调方式 Finetune、Lora 及 LiT，如表 1 所示，发现 Lora 进行微调时能够更准确的识别 novel 类别，当采用全量微调时能够，能够获得更好的 base 类别识别效果，因此在最终的模型中我们将这两者进行融合。

微调方式	图像编码器	文本编码器	AP50-novel	AP50-base	AP50-all
LORA	√		45.81	52.18	48.99
LORA	√	√	46.18	52.22	49.20
LiT		√	44.80	52.60	48.70
Finetune	√	√	44.50	52.96	48.73

4 目标分类模块

在商品目标分类部分，通过对当前开源的中文图文模型的调研，我们选择目前性能最优的中文图文模型 Chinese-CLIP，该模型继承于 OpenCLIP，视觉分支采用 ViT 结构，文本分支采用 RoBERTa 结构，我们选择 ViT-H-224 的模型进行微调。

5 推理优化

在推理阶段，我们将数据先验（单一类别、图文并茂）以规则的形式加入其中，设计了全局概率融合、OCR 辅助推理和类别一致性校正三条规则，进一步解决数据的易混淆和主体判断难问题。我们将规则和模型整理为如下的推理流程：

（1） RPN 候选框推理：使用训练好的定位模型，对测试集中的每张图进行推理，得到每张图的候选框；

（2） 图文特征提取：对每一张图，根据（1）中产生的候选框结果，截取对应的感兴趣区域，与全图共同送入训练好的 CLIP 模型中提取区域特征和类别文本特征；

（3） 全局特征融合：对每一个候选框产生的区域特征，按照 8：2 的比例与全局特征相加，校正得到最终的区域特征，与文本特征计算余弦相似度；

（4） OCR 辅助校正：对于每一个候选框的分类概率，结合全图的 OCR 结果，根据所设计的 OCR 规则进行类别概率校正；

（5）类别一致性校正：对所有的候选框的分类结果和全图的分类结果进行对比，若候选框中存在与全图类别一致的候选框，则输出一致候选框，若无则输出所有框中分数最高的候选框作为该图片的最终结果。

推理流程

6 测试结果

通过模型优化和规则设计，我们的方案在零样本类别上实现了 50.08%的 AP50，在少样本类别上实现了 54.16%的 AP50，最终识别效果如下：

开源链接

冠军方案：https://github.com/wusize/OVD_Contest

亚军方案(第二名)：https://github.com/FX-STAR/OVD2023

季军方案(第四名)：https://github.com/xuliu-cyber/OVD2023

季军方案(第六名)：https://github.com/thunderstudying/OVD_Contest