基于 YOLO 蒸馏与轻量化的智能垃圾识别算法

本项目旨在介绍如何通过知识蒸馏和模型轻量化，构建一个高精度、可高效部署于边缘设备的智能垃圾识别解决方案。

1. 项目背景与挑战

1

推动绿色社会建设

垃圾分类是城市可持续发展和环保的关键环节。

2

人工分拣效率低下

传统人工分拣劳动强度高、效率低，且易受主观判断影响而出现错误分类。

3

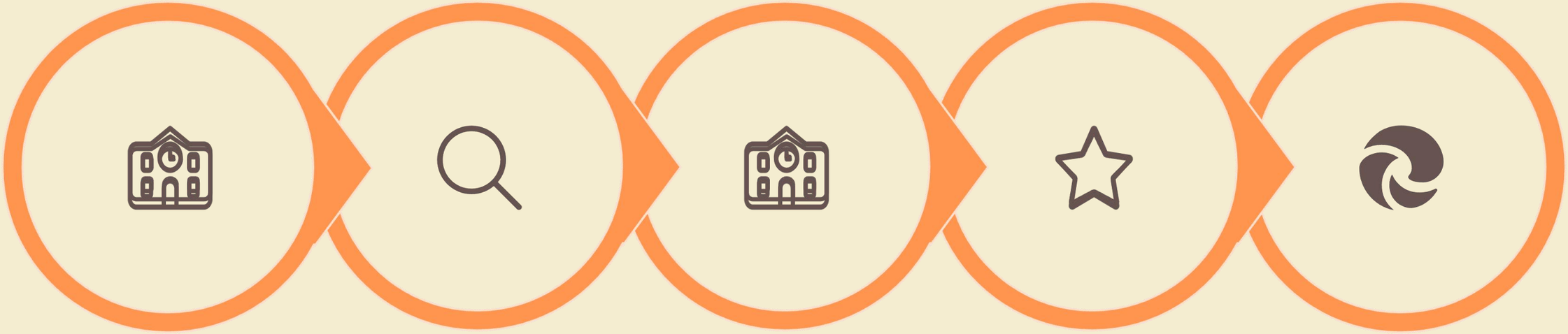
边缘部署的算力约束

深度学习目标检测模型（如 **YOLOv11m**）虽然精度高，但体积庞大、计算资源需求高，难以在树莓派、智能垃圾桶等**边缘设备**上实时运行。

在保持检测精度的前提下实现模型轻量化，可部署在树莓派、智能垃圾桶等边缘设备上。

2. 系统总体流程：Teacher → Student 知识迁移

采用“高精度教师模型”指导“轻量学生模型”的策略，结合剪枝与量化，实现端到端优化。



知识蒸馏

用教师模型的高级特征和伪标签增强学生模型的泛化能力。

模型剪枝

删除模型中对精度贡献较小的冗余通道，减少模型体积。

INT8 量化

将模型精度从浮点数**(FP32)**转换为定点数**(INT8)**，显著加速推理。

3. 算法核心设计：蒸馏与轻量化细节

3.1 知识蒸馏 (Data Distillation)

→ 教师模型生成伪标签

在大量无标注的垃圾图片数据上，使用高精度教师模型 **(YOLOv11m)** 进行推理，生成带有**置信度阈值 (0.35)**的伪标签。

→ 学生模型融合训练

学生模型 **(YOLOv11n)** 在**“真实标签 + 伪标签”**的混合数据集上进行训练，有效学习教师模型的判别边界和泛化能力。

3.2 模型结构轻量化

结构化通道剪枝

基于 `torch-pruning` 库，删除学生模型中 **30%** 的冗余通道，将参数量从 **3.9M** 降至 **2.8M**。

动态量化 (INT8)

将剪枝后的模型导出为 **ONNX** 格式，随后进行动态量化至 **INT8**。这一步使模型体积大幅缩小，并实现约 **1.5** 倍的 **CPU** 推理加速。

4. 数据集与训练增强

数据集来源与重划分

- 基础数据：**Kaggle Garbage Detection 6** 类数据集。

数据均衡：使用 `rebalance_splits.py` 脚本重新划分训练、验证、测试集，确保各类别分布均衡，避免偏斜。

- 无标注数据：收集额外无标注垃圾图片，用于教师模型生成伪标签，构建强大的融合训练集。

训练时的数据增强策略

- 颜色增强：应用颜色抖动 **(Color Jitter)** 提升对光照变化的鲁棒性。
- 几何变换：包含仿射变换 **(Affine Transformations)** 增加样本多样性。

组合增强：采用 **Mosaic (0.2)** 和 **MixUp (0.1)** 等复杂数据增强技术，进一步提升模型的泛化能力和鲁棒性。

5. 训练结果：性能逼近与超越

通过知识蒸馏，轻量级的学生模型在关键指标上逼近甚至超越了其教师模型。

模型	预训练权重	mAP@0.5	mAP@0.5:0.95
教师模型 (YOLOv11m)	**yolo11m.pt**	**0.64**	**0.46**
学生模型 (YOLOv11n) - 蒸馏	无	**0.66**	**0.48**

> 学生模型在融合数据集上训练了 **140** 个 **epoch**，其 **mAP@0.5** 达到了 **0.66**，比教师模型提升了 **2** 个百分点，充分证明了伪标签蒸馏的有效性。

学生模型检测结果展示

经过知识蒸馏与轻量化优化的 **YOLOv11n** 学生模型在垃圾检测任务中的实际效果。

垃圾目标检测示例图

精准识别与鲁棒性

模型能够准确识别出**纸板、玻璃、塑料、金属**等多种垃圾目标，并在复杂背景与光照变化下保持稳定检测性能。

轻量高效，性能卓越

通过蒸馏学习，学生模型充分继承了教师模型的识别能力。剪枝与量化后，模型体积缩减约 **45%**，但检测精度几乎无损，依旧实现 **mAP@0.5 ≈ 0.66**。

6. 模型压缩与性能评估

在保持蒸馏所得高精度的基础上，进一步实施了剪枝与量化，实现了极致的轻量化。

3.9M	2.8M	3.0MB	100+
原始参数量	剪枝后参数量	INT8 最终体积	CPU 推理速度
YOLOv11n (FP32) 原始参数量。	剪除 **30%** 冗余通道后的参数量，体积已大幅减小。	动态量化至 **INT8** 后，模型文件大小的最终结果。	**INT8** 量化模型在标准 **CPU** 上的推理帧率 **(FPS)**。

剪枝+微调结果： 剪枝后微调 **50** 个 **epoch**，模型精度为 **mAP@0.658**，几乎无损失。

7. 性能对比与消融实验总结

下表清晰展示了从原始教师模型到最终 **INT8** 量化模型在性能、体积和速度上的迭代优势：

模型阶段	mAP@0.5	模型大小	FPS (CPU)	核心特点
YOLOv11m 教师	**0.64**	**39MB**	**45**	高精度但庞大
YOLOv11n 学生 (蒸馏)	**0.66**	**5.3MB**	**85**	蒸馏提升泛化能力
剪枝 + 微调	**0.658**	**5.3MB**	**100**	精度几乎无损
INT8 量化模型	**0.658**	**3.0MB**	**100+**	最轻量高效的边缘部署方案

精度保持，体积减小 45%，速度提升 2 倍。

8. 实际应用与创新成果

部署与演示场景

边缘设备集成

模型可部署于 **Raspberry Pi**等嵌入式系统，实现实时识别。

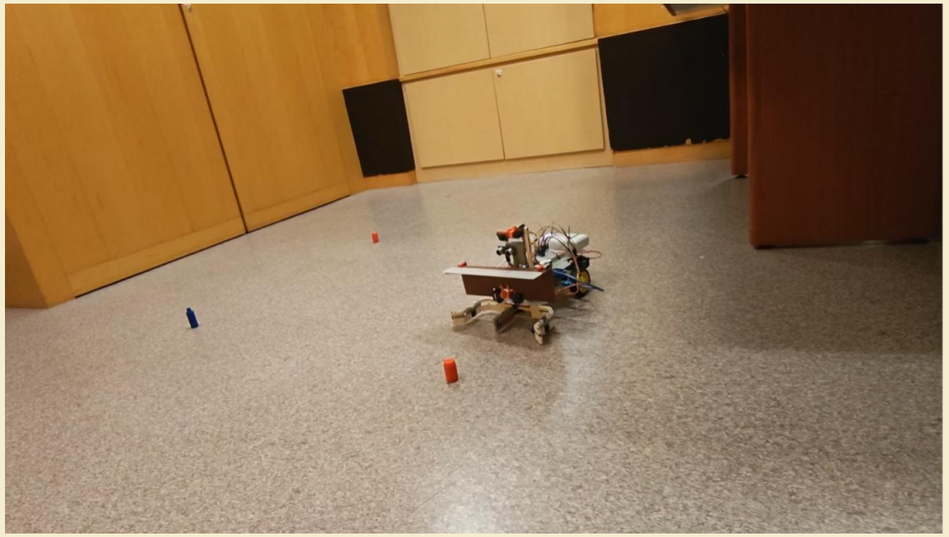
远程与网页演示

已实现支持网页实时视频检测与远程控制，方便用户交互和系统监控。

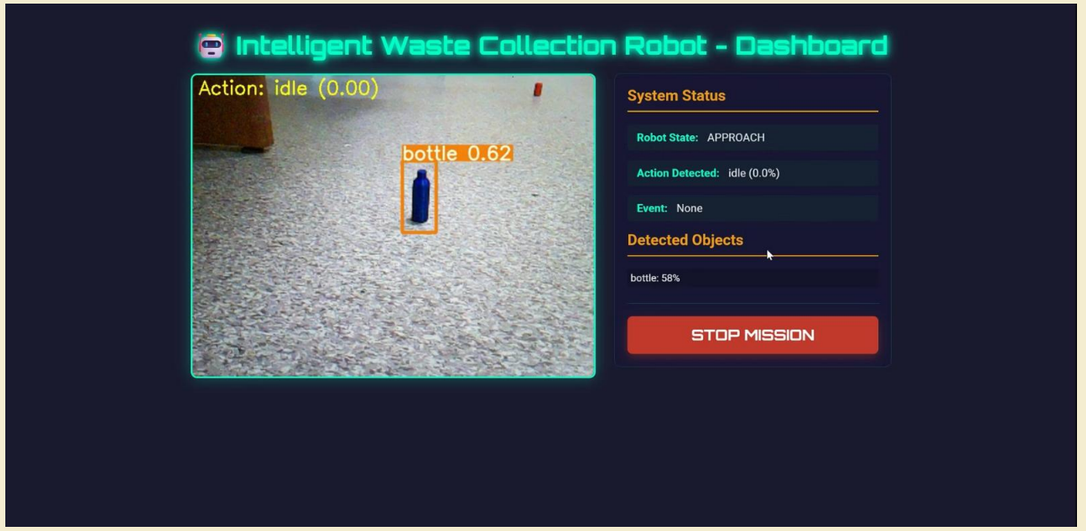
核心创新点

- **半监督伪标签蒸馏：** 显著提升了轻量级学生模型的识别精度和泛化能力。
- **联合优化策略：** 结合结构化通道剪枝和动态量化，实现极致的体积压缩和推理加速。
- **全流程可复现部署：** 提供了基于 **Ultralytics YOLO**、**Torch-Pruning** 和 **ONNX Runtime** 的完整、可复现的轻量化工作流程。

边缘部署展示



树莓派小车演示视频截图



网页实时检测截图

模拟实现了 **智能垃圾分类小车**，并简单实现了 **网页端实时预览 + 小车控制**。小车能够识别并检测 **bottle、can、paper** 三类目标（为模拟演示，采用 **3D** 打印瓶罐）。

在保持 **YOLO + 轻量化优化框架** 不变的前提下，我们将原**6**类数据集精简为**3**类并部署至树莓派，成功实现了低功耗条件下的实时检测与可视化展示。

未来展望

持续优化方向

深度蒸馏： 引入特征层（**Feature Map**）与温度蒸馏 **(Temperature Distillation)**，以更精细地迁移教师知识。

量化感知训练 (QAT)： 采用 **QAT** 替代动态量化，进一步降低端侧部署时的精度损失。

应用扩展： 将轻量化方案推广至 **IoT** 回收网络和自动分拣机械臂等更广阔的工业场景。

最终目标是打造一个**“可落地的轻量AI垃圾识别方案”**，为智慧环保提供强劲的边缘计算支撑。

谢谢观看

项目仓库与复现说明已提交至比赛平台，欢迎查阅完整技术细节。