西北大学学报(自然科学版)
2026 年 10 月,第 56 卷第 5 期,Oct. ,2026,Vol. 56,No. 5
Journal of Northwest University (Natural Science Edition)

·深度学习视觉处理与分析·

面向真实场景的生成数据驱动垃圾分类目标检测优化

杨建锋, 闫　弘, 梅天乐, 王　平, 曹　瑞

西北大学 计算机学院(软件学院),陕西 西安　710127

摘要　针对真实场景中垃圾分类任务背景复杂、目标尺寸小等问题,提出了一种融合生成数据与模型结构优化的目标检测改进方法。构建了包含 12 类的 GTC12 生成数据集,结合大语言模型生成的语义描述和图像生成模型 Stable Diffusion 生成的图像,通过筛选机制提升数据质量,与现有公开数据集相比,生成数据集在图像颜色复杂度和目标框占比 2 个维度上更贴近真实场景。以视觉 Transformer 适配器(ViT-Adapter)为基础,引入动态感受野卷积注意力模块(RFAC)和非线性映射模块,以增强模型的特征提取能力和对复杂背景的适应性。在构建的"真实生成"混合数据集 GTC12-Mix 上,改进模型的平均检测精度 mAP50:95 从 0.77 提升至 0.80;在类别平衡数据集 GTC12-Bal 细粒度二分类实验中,准确率由 0.85 提升至 0.86。结果表明,所构建的数据集具有较高的多样性和真实感,所提模型在复杂背景与小目标识别任务中表现出更强的检测能力。

关键词　垃圾分类;合成数据集生成;目标检测;ViT-Adapter;复杂场景分析

中图分类号:TP391.41　　**DOI**:10.16152/j.cnki.xdxbzr.2026-05-014

Optimized object detection and dataset generation for waste classification in real-world scenarios

Yang Jianfeng, Yan Hong, Mei Tianle, Wang Ping, Cao Rui

College of Computer Science, Northwest University, Xi'an 710127, China

Abstract　To address challenges of complex backgrounds and small targets in real-world waste classification, this study proposes an enhanced object detection framework integrating synthetic data generation with model architecture optimization. The generated trash classification dataset with 12 Categories (GTC12) was constructed by combining semantic descriptions from large language models and images synthesized via Stable Diffusion, employing a filtering mechanism to ensure high quality. Compared with existing public benchmarks, the synthetic dataset exhibits closer alignment with real scenes in color complexity and object-to-image area ratio. For model design, a dynamic receptive-field convolution attention module (RFAC) and a nonlinear mapping module were introduced based on Vision Transformer Adapter (ViT-Adapter) to strengthen feature extraction and robustness against complex backgrounds. On the hybrid "real + synthetic" dataset GTC12-Mix, the enhanced model's mean average precision (mAP50:95) increased from 0.77 to 0.80; accuracy increased from 0.85 to

收稿日期:2026-02-10　　修回日期:2026-03-13

基金项目:国家自然科学基金(62406247)

第一作者:杨建锋,男,高级工程师,从事数据科学及高校应用研究,yjf@nwu.edu.cn。

通信作者:曹瑞,男,博士,讲师,从事多模态表示学习、大型语言模型与扩散模型应用研究,cr@nwu.edu.cn。

0. 86 in fine-grained binary classification experiments on the class-balanced dataset GTC12-Bal. Results demonstrate high diversity and realism of the synthetic dataset, along with superior detection capability of the proposed model for complex backgrounds and small targets.

Keywords　waste classification; synthetic dataset generation; object detection; ViT-Adapter; complex scene analysis

垃圾分类是现代社会面临的重大环境挑战之一，如图 1 所示，我国垃圾总量总体呈持续增长趋势[1]。尽管部分城市已建立初步的垃圾分类体系，但受垃圾种类繁杂、标准不统一以及居民参与意识不足等因素限制，整体推进效果仍不理想。同时，相关法规、配套设施及回收处置体系尚不完善，分类后垃圾的资源化利用效率不高，进一步削弱了公众参与积极性。因此，构建更加精准高效的自动垃圾分类技术，对于减少人为误分、提升资源回收效率、推动城市固体废物治理体系现代化具有重要意义[2]。

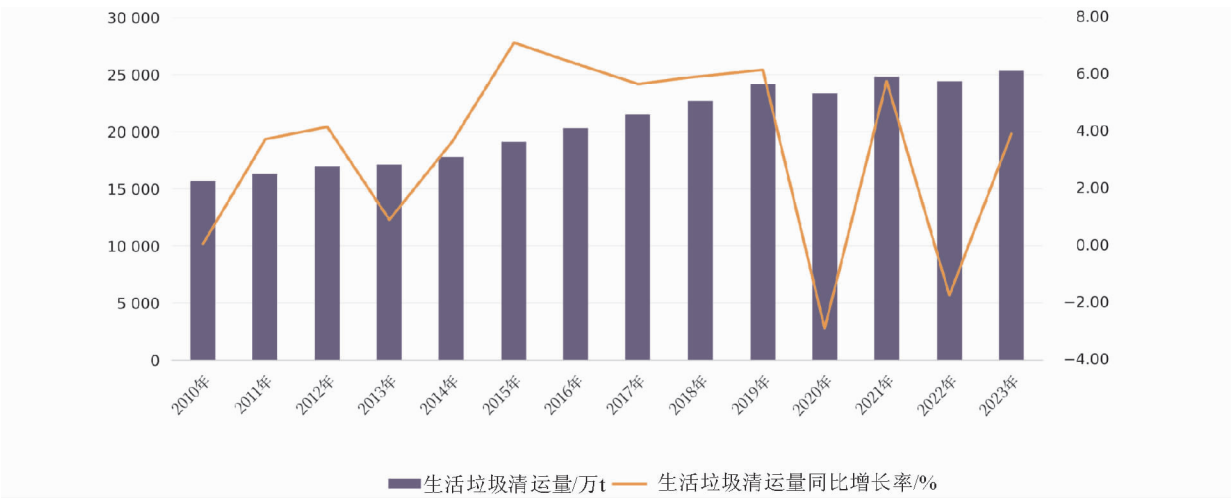

图 1　2010—2023 年我国生活垃圾清运量及同比增长率

Fig. 1　Domestic waste collection and transportation volume and year-on-year growth rate in China from 2010 to 2023

为了解决这些问题，智能化垃圾分类逐渐成为研究热点，其中，目标检测技术作为关键基础，已被广泛应用于图像中的目标识别与定位[3]。目标检测是计算机视觉中的基础任务，其目的是在图像中识别并定位所有感兴趣的对象[4]。传统的目标检测算法主要分为两阶段方法和一阶段方法[5]。前者如 R-CNN[6]、Fast R-CNN[7]、Faster R-CNN[8] 和 Mask R-CNN[9]，先生成候选区域再进行分类，精度高但速度慢；后者如 SSD[10]、RetinaNet[11] 和 YOLO 系列[12-14]，则直接在特征图上回归边界框与类别，能够达到实时检测速度，但在小目标检测或目标密集、遮挡严重的场景中性能仍有提升空间。

当前主流研究多基于深度卷积神经网络模型开展垃圾图像识别任务，如 MobileNetV3[15]，YOLOv7[16]、Vision Transformer[17]、ResNet-50[18]。然而，这些主流模型在控制环境下表现良好，但在真实场景中，尤其是背景复杂、目标小或形态不规则时，细粒度特征提取能力不足，场景泛化能力较差。例如在区分类似类型的垃圾或在杂乱背景下识别目标仍是难点。本研究选择 ViT-Adapter[19] 作为基线模型，以应对垃圾分类任务中的多样化需求。然而，ViT-Adapter 模型在处理复杂场景下的垃圾分类任务时，仍存在特征提取能力不足的问题。具体而言，由于 Adapter 模块采用随机初始化策略，导致其在复杂背景及小目标检测任务中难以稳定学习到有效特征；同时，模块中传统的卷积操作未能充分考虑到图像中不同空间区域的重要性差异，难以精确捕捉目标物体的显著特征，从而影响整体检测性能。

此外，现有垃圾分类数据集在支持复杂真实场景应用方面仍存在较大局限。以 TrashNet 为例，其包含 2 527 张图像，涵盖 6 类常见垃圾，但均在统一背景下拍摄，目标集中且占比较大，仅提供分类标签，缺乏边界框信息[20]。AquaTrash 虽引入自动化标注以提升准确性，但受限于样本数

量和场景多样性,难以覆盖诸如破碎、潮湿或重叠等复杂形态[21]。GINI 数据集包含 2 561 张图像,其中 956 张为垃圾图像,其余为与垃圾视觉属性相似的非垃圾图像,虽在识别难度上有所增强,但整体规模与类别仍显不足[22]。TACO 聚焦户外环境,包含 1 500 张图像,但其背景复杂度有限,目标分布不均[23]。华为云生活垃圾数据集亦存在场景受限、样本缺乏多样性的情况,如图 2 所示。这些数据集普遍存在背景单一、数据量有限、标注维度不充分等问题,致使训练所得模型在真实复杂环境中泛化能力不足,难以支撑高可靠性的自动化垃圾分类系统部署。

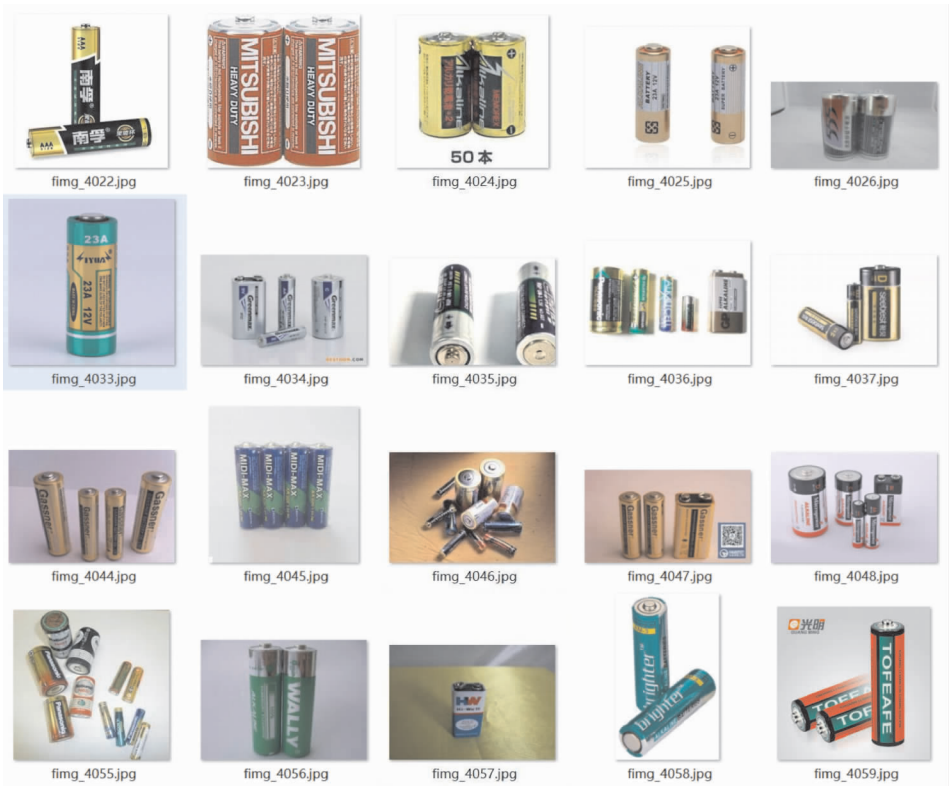

图 2 华为云生活垃圾数据集

Fig. 2 Huawei cloud household waste dataset

针对上述问题,本文从数据集构建和模型优化 2 个角度展开研究。为了构建适用于真实场景的垃圾分类模型,首先,通过 ChatGPT 生成垃圾分类文本提示,并将提示输入 Stable Diffusion 合成多场景、多光照条件下的垃圾图像,构建真实场景下的合成数据集 GTC12(generated trash classification dataset with 12 categories)。针对合成数据与公开数据集在类别和背景上的差异,本文通过模型实验、目标框占比和颜色复杂度进行对比,以评估数据集的多样性和真实性。其次,本文采用预训练 ViT-Adapter 作为基线,并提出了一种改进算法 ViT-Adapter-Det:一是在空间先验模块中用动态感受野卷积注意力模块(RFAC)替代标准卷积,通过感受野级别的动态注意力加权实现对小目标和复杂背景的自适应关注;二是在交叉注意力后插入门控式 MLP 非线性映射,通过并行线性投影与 SiLU 激活的门控机制,增强通道间非线性特征交互。最后,进行 24 类细粒度分类实验,以验证改进带来的性能提升。实验结果显示,所提方法在平均精度和准确率等指标上均优于基线,能够更好地适应复杂场景下的垃圾分类任务。

1 相关工作

1.1 生成式模型研究现状

1.1.1 视觉生成式模型

Stable Diffusion 是一种基于潜在扩散模型(LDM)的图像生成技术,通过将图像压缩到低维隐空间并在其中进行迭代去噪,生成高质量图像[24]。其核心组件包括自动编码器、U-Net 去噪网络[25]、交叉注意力机制和时间步调度,利用文本提示实现条件生成,显著提升生成效率和图像逼真度。在垃圾分类任务中,现有数据集如 TrashNet 和 AquaTrash 因背景单一、目标分布不

均,难以满足真实场景需求。Stable Diffusion 通过生成多样化垃圾图像,有效增强数据集的真实性和多样性。本研究采用 Stable Diffusion Web UI,通过 ChatGPT 生成的文本提示,结合参数优化,构建了 GTC12 数据集,显著提升了数据集的场景复杂度和代表性,为模型训练提供了有力支持。

1.1.2　语言生成式模型

ChatGPT 是 OpenAI 基于生成式预训练变换器架构开发的大型语言模型,以生成流畅、语义准确的文本而闻名[26]。早期模型如 GPT-1[27] 采用单向语言建模,根据上文预测下一个词元,而 GPT-2[28] 和 GPT-3[29] 分别通过多任务学习和 1 750 亿参数增强了泛化能力和文本逼真度。ChatGPT 基于 GPT-3.5[30],通过监督微调和 RLHF[31] 实现更自然、安全的对话生成。GPT-4o 支持多模态输入(文本和图像),上下文窗口达 32 768 词元,推理能力进一步提升。在垃圾分类研究中,ChatGPT 生成详细的垃圾描述,如“传送带上的破碎塑料瓶”,为 Stable Diffusion 提供高质量文本提示,辅助构建包含 12 类垃圾的 GTC12 数据集。这些描述确保了数据集的多样性和真实性,弥补了现有数据集如 TrashNet 和 AquaTrash 的不足。

1.2　目标检测模型

1.2.1　Vision Transformer

ViT(Vision Transformer)是由 Google Brain 团队于 2020 年提出的一种基于 Transformer 架构[32] 的神经网络模型,成功将自然语言处理中的自注意力机制应用于图像分类任务。ViT 通过将图像分割为固定大小的块,并将其转换为线性嵌入序列,结合可学习的位置嵌入以保留空间信息,随后输入标准的 Transformer 编码器进行特征提取[33]。编码器由多头自注意力机制[34]、归一化模块和多层感知机模块组成。在分类阶段,ViT 引入一个分类标记(class token),通过自注意力机制与其他块特征交互,最终经全连接层预测类别。ViT 能有效捕捉图像的局部和全局特征,预训练于大规模数据集后,在图像识别任务中展现出优于传统卷积神经网络的性能。其减少了对图像特有归纳偏见的依赖,适合复杂场景,但对预训练数据需求较高。ViT 不仅限于分类,还广泛应用于目标检测、图像分割等任务,为计算机视觉研究提供了新方向。

1.2.2　ViT-Adapter

ViT-Adapter 是一种为密集预测任务(如目标检测和语义分割)设计的框架,旨在提升视觉变换器(ViT)的性能。ViT 在图像分类中表现出色,但因缺乏局部归纳偏置,在密集预测任务中效果有限。ViT-Adapter 通过无需预训练的适配器模块引入图像相关偏置,保持 ViT 架构纯粹性。该模块包括空间先验模块、空间特征注入器和多尺度特征提取器。空间先验模块利用卷积生成多尺度特征金字塔,空间特征注入器通过交叉注意力融合特征,多尺度特征提取器生成适合任务的输出。相较于 Swin Transformer 等模型[35],其保持 ViT 的通用性,适合跨模态学习,广泛应用于 3D 占用预测和垃圾分类检测等任务,为计算机视觉提供高效解决方案。

2　方法

2.1　数据集生成方法

本研究构建了一个多模态协同生成框架,以解决现有公开垃圾分类数据集中背景单一、目标占比过高、类别分布不均等问题。整个生成流程如图 3 所示,由“ChatGPT 文本描述→Stable Diffusion 文生图与图生图→自动化与人工筛选”3 个阶段组成。

2.1.1　文本描述生成

文本描述生成是构建合成垃圾分类数据集的关键环节。首先,根据预设的 12 类垃圾,利用 ChatGPT 生成每类垃圾在真实场景中可能呈现的多种形态描述,具体描述方式如图 3 所示。随后,对每种形态补充对应的背景与光照条件、位置信息、质量控制等,其生成方式与形态描述类似。最后,将所有输出整理为正向与负向提示词。

以“glass bowls”为例,正向描述用于定义垃圾在图像中理想的表现形式,描述内容不仅包括垃圾的形态、环境背景、目标在图片的位置等信息,还有一些如质量控制和风格控制的信息,具体包括以下:

```
垃圾形态特征(括号表示增强提示词的权重):
(((glass bowls))),
背景描述:glass bowls sit on a neat table,
环境参数:sunlight streams through the window,
          real scene,
质量控制:high quality.
```

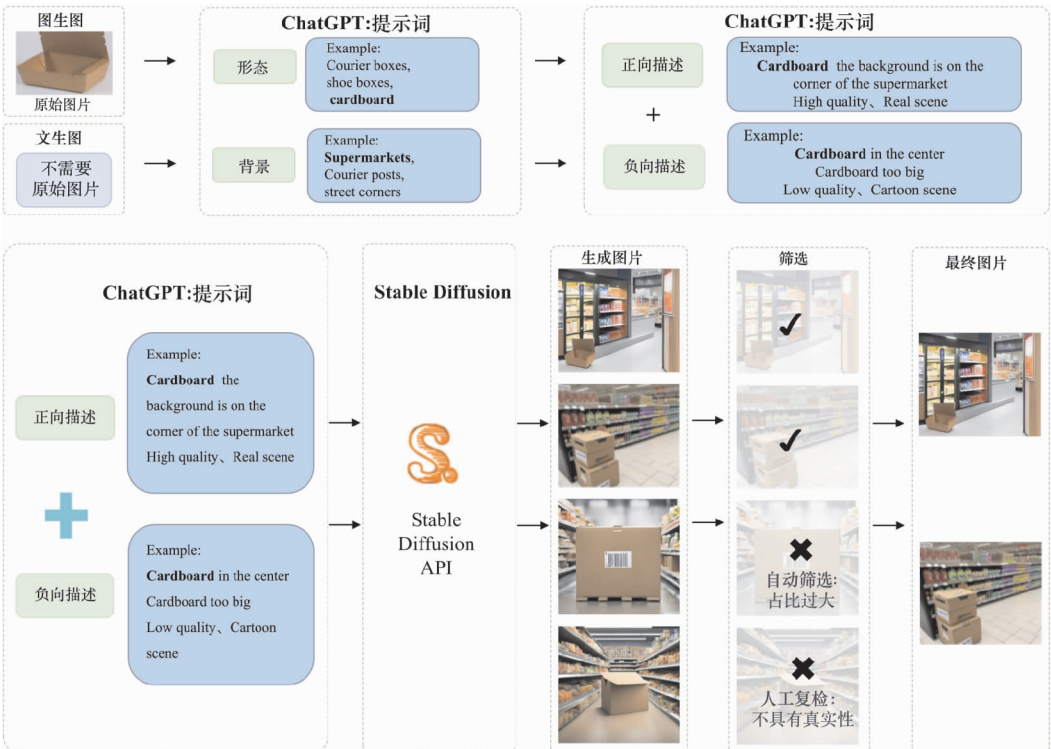

图 3 数据集生成流程图

Fig. 3 Dataset generation flowchart

负向描述用于指定在生成图像时需要避免的场景或特征,以确保图像质量和真实性,其主要通过人为设置描述语言。例如,负向描述可以限定物体的大小、位置、光照条件或风格等。负向描述的引入有效提升了生成图像与实际场景需求的匹配度,同时避免了偏离研究目标的无效内容。负向描述的提示词具体包括以下:

物理属性限制:glass bowls too big,
风格排除:cartoon scene,
质量控制:low quality.

在初步生成后,对所有文本进行严格筛选和优化,剔除逻辑矛盾或场景不符的描述(例如"黑暗环境下的强光照"),删除过于奇幻或无关的内容,并通过同义词替换和表达调整(如将"bright"改为"vivid")增强描述准确性。为模拟真实环境中的动态变化,还在提示中加入"多个塑料瓶散落在地面"或"纸张被水浸湿"等动态场景,从而进一步丰富生成图像的复杂性。通过这一系列步骤,最终获得既具多样化特征,又能涵盖多种场景和细节的高质量文本描述,为 Stable Diffusion 提供了可靠、精细的输入支持。

2.1.2 图像生成阶段

在图像生成阶段,采用 Stable Diffusion 的文生图与图生图 2 种模式①,对上一阶段生成的同一提示词集进行并行采样。文生图模式主要负责生成全新场景图像,而图生图模式则在给定模板图像的基础上,通过控制去噪强度与高分辨率修复等参数,实现场景细节与构图的一致性。

2.1.3 生成数据集后处理

生成后的原始图像会进入后处理环节,如图 3 所示,其主要流程如下。

1) 自动化质量筛查。利用 Python 脚本实现目标区域占比分析,对生成图像进行初步筛选。针对部分图像存在目标区域过大的问题(例如占据整个画面),采用了 OpenCV 和 Pillow 库②进行自动化检测。

step 1 标注数据解析:读取标注文件,提取目标框坐标信息。

step 2 面积占比计算:基于图像尺寸和目标框坐标计算面积占比。

① 软件来源:https://github.com/Stability-AI/stablediffusion

② 软件来源:https://github.com/python-pillow/Pillow

step 3　阈值筛选:面积占比阈值设置为 0.5,超过此阈值的样本将会被标记为异常。

step 4　异常样本管理:生成异常样本清单,记录异常类型和对应的生成参数,此清单便于后续参数调整和问题追溯。

2)人工质量校验。在自动化筛选之后,通过人工进一步复核,剔除可能存在噪声、模糊、颜色异常或背景失真的样本。同时建立人工复核机制,对自动化筛选的图像进行多维度评估:视觉质量方面检查图像清晰度、色彩还原度、背景真实性;场景合理性方面评估光照条件和物体摆放位置是否符合现实场景。

3)数据增强策略。为了提升模型泛化能力,对部分生成样本较少的垃圾进行了数据增强操作。具体而言,选择了以下几种增强方法。

① 几何变换:通过对图像进行旋转(例如随机 90°翻转)和水平翻转,模拟垃圾在不同位置和角度的分布,从而增强数据的空间多样性。

② 颜色调整:对图像的亮度和对比度进行随机调节,生成多种光照条件下的样本,以模拟真实场景中可能存在的环境变化。

③ 随机裁剪:对目标区域和背景进行裁剪,并重新组合生成不同目标框大小和背景比例的样本,从而增强数据的复杂性。

4)标注数据集。对筛选和增强后的数据集,使用 LabelImg 工具[①]进行标注。标注结果保存为 XML 格式,标注过程确保每个目标区域的坐标和类别信息均被准确记录。

通过自动化筛选、人工复核和数据增强的协同作用,生成数据的后处理流程在数量和质量上均得到了全面提升。这一体系不仅高效剔除了不符合要求的样本,同时,显著增强了数据集的多样性和复杂性。生成的训练数据不仅适用于垃圾分类任务,还能为其他依赖复杂场景的计算机视觉研究提供参考价值,流程的模块化设计也便于在不同任务中复用和调整。

2.1.4　生成数据集的整体规模与混合策略

本文基于 Stable Diffusion 与 ChatGPT 生成了一套高质量的垃圾分类数据集 GTC12,并对其总体规模和具体组成情况进行了统计。数据集规模统计显示,生成数据集 GTC12 共包含 7 920 张图像,覆盖 12 类垃圾及其 3 ~ 5 种细分形态,且在生成时严格控制了 8 种典型场景和 4 种光照条件,以保证样本的多样性和真实性。样本数量如表 1 所示。

表 1　各数据集中不同类别的图像数量

Tab. 1　Per-class image counts across datasets

类别	数量		
	GTC12	GTC12-Mix	GTC12-Bal
生物垃圾	930	934	200
纸板	1 018	1 018	200
烟蒂	257	257	200
衣物	449	449	200
砧板	717	717	200
花盆	506	506	200
玻璃	1 018	1 051	200
口罩	72	72	200
纸张	731	731	200
塑料	928	1 095	200
电线插头	240	240	200
铁罐	1 054	1 132	200
总数	7 920	8 202	2 400

为全面评估生成数据集的性能,本研究提出了生成数据集与公开数据集的混合使用方案。将生成数据集与经过筛选的 TACO 数据集融合,其中从 TACO 数据集中筛选了 282 张真实场景图像,融合后形成了总计 8 202 张图像的实验数据集 GTC12-Mix。这种混合方式在保持样本规模的基础上平衡了合成数据与真实数据的特征分布。

为验证生成数据的有效性,本节设计了平衡对比数据集 GTC12-Bal,从生成数据集中按类别均匀抽取 1 200 张样本,每类选取 100 张,同时,从华为云生活垃圾分类数据集中随机抽取对应类别的 1 200 张真实样本,最终构建包含 2 400 张图像的对比数据集。等比例抽样确保了 2 组数据在类别分布和数量上的一致性,为客观评估生成数据的应用价值提供了标准化的实验基准。

2.2　改进 ViT-Adapter 的垃圾图像检测网络

在传统模型优化思路的基础上,基于 2.1 节提出的混合数据集 GTC12-Mix 处理生成数据集与公开数据集,分析其对模型性能的影响,在此基

① 软件来源:https://github.com/tzutalin/labelImg

础上优化模型结构,提升其在复杂背景下的垃圾分类性能。合成数据生成与模型优化的垃圾分类研究技术框架图如图 4 所示,上半部分是 2.1 节的数据集生成流程,阐明了数据合成的技术路径,下半部分详细展示了本节的改进方法,包括对基线模型的改进以及细粒度二分类的内容,这些方法提高了模型对复杂背景和小目标的捕捉能力。

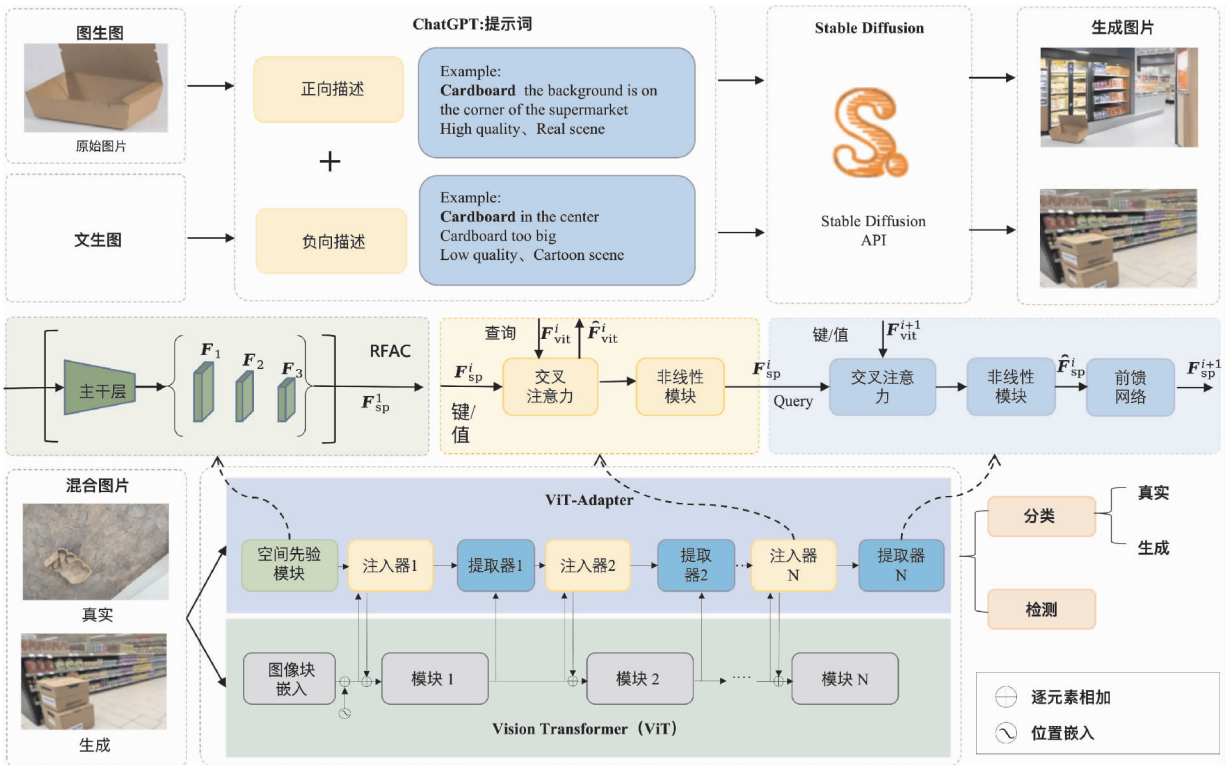

图 4 基于合成数据生成与模型优化的垃圾分类研究技术框架图

Fig. 4 Technical framework for waste classification research based on synthetic data generation and model optimization

2.2.1 动态感受野卷积注意力模块(RFAC)

在真实场景下,垃圾分类任务中目标形态多变且背景复杂,传统卷积在固定感受野和参数共享的机制下,难以突出小目标或差异化区域的特征,影响检测精度。ViT-Adapter 的空间先验模块虽通过多层卷积和多尺度特征金字塔提供了局部信息,但其标准卷积对所有位置一视同仁,无法根据区域重要性分配权重,也不具备动态调整能力。为克服这一局限,本文在空间先验模块中引入了动态感受野卷积注意力模块(RFAC)。设输入特征图为 $\boldsymbol{X} \in \mathbf{R}^{B\times C\times H\times W}$,其中:$B$ 表示批量大小;C 表示通道数;H、W 为空间尺寸。RFAC 的具体计算流程如下。

首先,将 $\boldsymbol{X}$ 通过分组卷积(group convolution)映射到感受野空间特征 $\tilde{\boldsymbol{X}}$。一共划分 G 个组,对应 G 个不重叠的滑动窗口,每个滑动窗口视作一个“感受野”。映射过程可写为

$$\tilde{\boldsymbol{X}} = \mathrm{GroupConv}(\boldsymbol{X};\{W_g\}_{g=1}^{G}),\quad \tilde{\boldsymbol{X}} \in \mathbf{R}^{B\times G\times C'\times H'\times W'} \tag{1}$$

式中:C' 是感受野空间特征的通道数;(H',W') 是经分组卷积后对应每个感受野的空间大小。

RFAC 模块通过分组卷积提取各感受野的空间特征后,对每个感受野执行全局平均池化,得到 $\boldsymbol{z} \in \mathbf{R}^{B\times G\times C'}$:

$$\boldsymbol{z}_{b,g,c'} = \frac{1}{H'W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\hat{\boldsymbol{X}}_{b,g,c',i,j} \tag{2}$$

接着,通过一个 1×1 分组卷积得到未归一化的注意力得分 $\boldsymbol{s} \in \mathbf{R}^{B\times G\times 1}$:

$$\boldsymbol{s}_{b,g} = \mathrm{Conv}_{1\times 1}(\boldsymbol{z}_{b,g,:}) \tag{3}$$

之后,对 $\boldsymbol{s}$ 应用 Softmax,得到归一化后的感受野注意力权重 $\propto \in \mathbf{R}^{B\times G\times 1}$:

$$\boldsymbol{\alpha}_{b,g} = \frac{\exp(\boldsymbol{s}_{b,g})}{\sum_{g'=1}^{G}\exp(\boldsymbol{s}_{b,g'})} \tag{4}$$

将注意力权重 $\propto$ 与 $\tilde{\boldsymbol{X}}$ 相乘,得到加权后的感受野特征 $\tilde{\boldsymbol{X}} \in \mathbf{R}^{B\times G\times C'\times H'\times W'}$:

$$\tilde{\boldsymbol{X}}_{b,g,c',i,j} = \boldsymbol{\alpha}_{b,g}\hat{\boldsymbol{X}}_{b,g,c',i,j} \tag{5}$$

最后,将 $\tilde{\boldsymbol{X}}$ 按照感受野进行汇聚,并通过常规模块卷积(Conv)映射回原始通道数 C:

$$\boldsymbol{X}_{\mathrm{RFAC}} = \mathrm{Conv}(\mathrm{Concat}_{g=1}^{G}(\widetilde{\boldsymbol{X}}_{b,g,:,:,:})) \in \mathbf{R}^{B\times C\times H''\times W''} \tag{6}$$

通过此种方式,卷积核不再是对所有位置共享同一组权重,而是根据区域重要性动态分配关注度,显著增强对小目标和复杂背景中关键特征的捕捉能力。

在实现中,本文将原 SPM 模块中用于将通道数从$\{C,2C,4C,4C\}$投影到 embed_dim 的 4 个 1 × 1 卷积层,均替换为上述 RFAC 模块操作。记原始的 1 × 1 卷积为

$$\mathrm{Conv}_{1\times 1_t}(\boldsymbol{X}_t),t \in \{1,2,3,4\} \tag{7}$$

改进后替换为

$$\mathrm{RFAConv}_t(\boldsymbol{X}_t),t \in \{1,2,3,4\} \tag{8}$$

这样将显著增强对小目标和复杂背景中关键特征的捕捉能力,使多尺度特征融合更加精准、高效,并在后续消融实验中验证了其对分类性能的显著提升。

2.2.2　门控特征映射模块

在处理多样化且背景复杂的垃圾图像时,ViT-Adapter 的 Adapter 模块因线性变换方式难以捕捉细微差异,表现出特征表达和学习稳定性的不足。为此,本文在交叉注意力之后嵌入了门控特征映射模块(gated feature mapping module)。

该模块由 2 个并行的全连接层和 SiLU 激活函数构成,通过元素级乘法完成“门控” 式特征交互,并在最后将结果映射回原始维度。因此,假设第l层交叉注意力的输出为$\widetilde{\boldsymbol{F}}_{\mathrm{att}}^{(1)} \in \mathbf{R}^{N_l\times D}$,其中:$N_l$表示该层展开后的词元数量;$D$ 表示特征维度;则该 MLP 模块可以形式化为

$$\boldsymbol{A}_1^{(l)} = \widetilde{\boldsymbol{F}}_{\mathrm{att}}^{(l)}\boldsymbol{W}_1^{(l)},\boldsymbol{A}_2^{(l)} = \widetilde{\boldsymbol{F}}_{\mathrm{att}}^{(l)}\boldsymbol{W}_2^{(l)} \tag{9}$$

式中:$\boldsymbol{W}_1^{(l)}$、$\boldsymbol{W}_2^{(l)} \in \mathbf{R}^{D\times D'}$,通常取$D' = \frac{D}{2}$,这样并行的 2 组线性映射可以分别提取不同通道的特征。接着对$\boldsymbol{A}_2^{(l)}$施加SiLU激活,并与$\boldsymbol{A}_1^{(l)}$做元素级乘积,以“门控” 方式融合信息:

$$\boldsymbol{G}^{(l)} = \boldsymbol{A}_1^{(l)} \odot \sigma(\boldsymbol{A}_2^{(l)}),\boldsymbol{G}^{(l)} \in \mathbf{R}^{N_l\times D'} \tag{10}$$

这一操作等价于让每个 token 的第 i 个通道特征在被$\sigma(\boldsymbol{A}_2^{(l)})$加权时,自适应地强调或抑制不同位置的响应。最后,将$\boldsymbol{G}^{(l)}$通过一个线性映射恢复到原始维度:

$$\boldsymbol{F}_{\mathrm{mlp}}^{(l)} = \boldsymbol{G}^{(l)}\boldsymbol{W}_3^{(l)} + b_3^{(l)},\quad \boldsymbol{W}_3^{(l)} \in \mathbf{R}^{D'\times D},\boldsymbol{b}_3^{(l)} \in \mathbf{R}^{D} \tag{11}$$

得到与$\widetilde{\boldsymbol{F}}_{\mathrm{att}}^{(l)}$维度一致的非线性特征。最终,将该输出与原始交叉注意力残差相加并进行 LayerNorm 归一化:

$$\hat{\boldsymbol{F}}^{(l)} = \mathrm{LayerNorm}(\widetilde{\boldsymbol{F}}_{\mathrm{att}}^{(l)} + \boldsymbol{F}_{\mathrm{mlp}}^{(l)}) \tag{12}$$

该设计在不显著增加计算开销的前提下,通过 SiLU 激活引入非线性映射,使得跨通道特征能够以“门控”方式完成更复杂的交互和重构,因而显著提升了模型对细微纹理、与背景相近物体以及少量像素分布的辨别能力,并在训练初期有效缓解了梯度不稳定的问题。

2.2.3　损失函数

为了深度验证改进版 ViT-Adapter 模型在垃圾分类任务中的实际应用效果,本研究精心规划了细粒度二分类实验。该实验聚焦于测试改进版模型区分生成数据与真实数据的能力,同时借助细粒度分类任务,对生成数据集的真实性和多样性展开评估。

实验选取了 2.1.4 节中生成的平衡对比数据集 GTC12-Bal。在这个数据集中,每个样本均明确标注了其数据来源,区分为生成数据或真实数据的具体类别,从而成功将任务转化为细粒度二分类任务。即依据既定规则把 12 类原始标签转换为 24 类标签。将真实数据标签映射为偶数,新生成的数据标签映射为奇数,以此确保 2 类样本在训练过程中的均衡性。

数据加载器在预处理阶段自动完成标签映射,训练时仅冻结主干网络的部分参数,选用交叉熵损失(cross entropy loss)作为损失函数。利用 Adam 优化器更新所有可训练参数,同时冻结主干网络(ViT 部分)的参数,仅对新增的 Adapter 模块和分类头进行训练。模型初始化时,仅加载主干网络部分的预训练权重,由于分类头维度扩展,对其进行随机初始化。训练过程中,通过打印模型结构,确认分类头输出维度为 24,保证改进流程的准确性。

通过上述方法,改进后的 ViT-Adapter 实现了对复杂背景中小目标和细粒度特征的更强适应。

3　实验及结果分析

3.1　实验设置

本研究中不同数据集上的模型性能对比实验以及数据集特性对比实验均在配备 24 GiB 显存的 NVIDIA RTX 3090 GPU 和 24 核 CPU 的平台上完成,以量化评估 Stable Diffusion 所生成数据集与公开真实数据集在目标检测任务中的性能差

异。其余所有实验则统一在以下硬件与软件环境中开展:操作系统为 Ubuntu 20.04;CPU 为 Intel(R) Xeon(R) Gold 6133;GPU 为 NVIDIA V100;开发语言 Python 3.8;深度学习框架 PyTorch 1.9.0;CUDA 版本 11.1。值得说明的是,模型改进中的 2 项实验采用了 4 张 NVIDIA V100 显卡,而细粒度二分类的实验因资源限制仅使用了 2 张 V100。

所有针对模型改进与消融的实验均采用相同的超参数设置:批量大小为 16;优化器选用 AdamW;初始学习率设为 1×10^{-4},权重衰减系数为 0.05。模型训练策略包括 1×(12 epoch)和 3×(36 epoch)2 种方案,涵盖 Faster R-CNN 与 ATSS 2 种目标检测框架,以及 tiny 与 base 2 种模型规模,以全面考察改进模块在不同配置下的适用性和性能表现。

本研究构建了 2 种数据集:一是平衡对比数据集 GTC12-Bal,用于模型性能对比、数据特性分析及细粒度二分类实验;二是混合数据集 GTC12-Mix,用于 ViT-Adapter 改进模块的消融实验以及在多框架、多尺度设置下的性能与参数对比。

3.2 评价指标

为全面衡量方法有效性,本文构建了涵盖目标检测与细粒度分类的统一评价体系,主要包括目标框占比、颜色复杂度、平均精度(mAP)及其衍生指标、模型参数量,以及分类任务中的精确率、召回率、F1 score 和准确率。目标框占比通过将所有标注框面积之和与图像总面积比较,反映了目标大小对检测难度的影响;颜色复杂度则以各像素颜色直方图熵值的均值衡量图像色彩分布的丰富性,从而评估背景复杂程度。mAP 作为检测任务的综合得分,分别在 IoU = 0.5 和 0.75 阈值下得到 mAP50 和 mAP75,而 mAP50:95 则对 0.5 ~ 0.95 范围内多个阈值的 mAP 求平均,以更全面地反映模型在不同重叠要求下的检测能力。模型参数量衡量可训练参数总数,用于评估模型的复杂度与推理效率。分类任务中,精确率衡量预测正样本的准确性,召回率衡量对所有正样本的覆盖率,F1 score 则为二者的调和平均,用以平衡精度与召回的偏差,准确率代表所有样本预测的总体正确性。这些指标既刻画了数据集本身的难度分布,也为不同模型或不同数据集间的性能对比提供了可比且一致的量化标准。

1)目标框占比。在垃圾分类数据集中,目标区域面积(S_t)与图像总面积(S)的比值反映了目标框占比,该占比的高低直接反映了垃圾目标的大小,比例越低,说明垃圾目标越小,比例高则提示目标偏大。

其计算方法如下:根据标注信息确定每张图像中目标区域的边界框,累加所有目标区域面积,并将此总和除以图像总面积及数据集图片数量,结果即为总目标框占比。数据集目标框占比(P)计算公式为

$$P = \frac{\sum \frac{S_t}{S}}{N} \tag{13}$$

式中,N 为数据集中图片的总数量。

2)颜色复杂度。其用于量化图像中色彩分布的丰富性和多样性。在垃圾分类任务中,颜色复杂度越高说明垃圾目标图像中色彩变化大,背景可能越复杂,也越可能接近于真实场景,反之则表明背景较简单。

其计算方法如下:通过图像的颜色直方图统计图像中各颜色通道的分布情况,利用熵值(Entropy)对颜色直方图进行计算,得到一个反映图像色彩复杂程度的数值。数据集颜色复杂度(O)计算公式为

$$O = -\frac{\sum_{i=1}^{N}\left(\sum_{j} p_j(i)\log p_j(i)\right)}{N} \tag{14}$$

式中:$p_j(i)$ 表示在第 i 张图片中颜色 j 的概率分布;$\sum_{i=1}^{N}$ 表示对数据集中所有图片的颜色复杂度求和;$\sum_{j}$ 表示对第 i 张图片中所有颜色的概率分布求和。

3.3 实验结果对比分析

实验采用 5 种主流目标检测模型(YOLOv8、YOLOv10、FasterNet[36]、DETR[37]和 RT-DETR[38])进行训练和测试,训练轮数统一设定为 200 轮。不同数据集上的模型性能对比实验结果如表 2 所示,其中,生成数据集的检测性能显著低于公开数据集。以 YOLOv8 为例,生成数据集的 mAP50:95 为 65.0%,较公开数据集的 72.7% 下降 7.7 百分点;RT-DETR 的性能差异尤为显著,mAP50:95 从 34.9% 降至 18.3%。所有模型在生成数据集上的各类指标均呈现类似趋势,这一现象表明,生成数据集的复杂场景显著增加了模型检测难度,验证了其贴近真实应用环境的特性。

表 2　生成与真实数据集上的目标检测模型性能对比

Tab. 2　Performance comparison of object detection models on synthetic and real-world datasets

模型	数据集类别	mAP50:95	mAP50	mAP75	Precision	Recall	FPS	Params/10^6
YOLOv8	生成	0.650	0.778	0.522	0.725	0.729	233	3.1
	真实	0.727	0.797	0.657	0.769	0.739	233	3.1
YOLOv10	生成	0.566	0.683	0.448	0.651	0.609	277	2.3
	真实	0.680	0.742	0.618	0.768	0.654	277	2.3
FasterNet[36]	生成	0.607	0.736	0.479	0.639	0.639	226	3.9
	真实	0.695	0.765	0.626	0.746	0.698	226	3.9
DETR[37]	生成	0.644	0.760	0.528	0.693	0.669	166	69.0
	真实	0.722	0.786	0.658	0.712	0.736	166	69.0
RT-DETR[38]	生成	0.183	0.223	0.143	0.241	0.381	251	20.0
	真实	0.349	0.383	0.315	0.420	0.392	251	20.0

表 3 展示了目标框占比和颜色复杂度的对比数据,生成数据集的目标框占比均值为 0.32,显著低于公开数据集的 0.59。这一差距表明,生成数据集中目标物体尺寸更小且分布更分散,这种分布特性增加了模型定位与识别的难度。颜色复杂度指标显示,生成数据集的复杂度达到 7.55,较公开数据集的 6.26 提升了 20.6%。这说明了生成数据集的背景和目标颜色更多样,特征提取的挑战也更大。

表 3　数据集特性对比

Tab. 3　Comparison of data set characteristics

数据集	目标框占比	颜色复杂度
真实	0.59	6.26
生成	0.32	7.55

综上所述,实验充分验证了生成数据集在多样性和真实性上的突出优势。在相同模型的条件下,不同数据集性能的差异反映了数据集复杂度的高水平,目标框占比和颜色复杂度的量化分析进一步证实了生成数据集在模拟真实场景的有效性。这些特性为模型训练创造了更具挑战性的条件,也为后续进一步完善算法提供了坚实的数据基础。实验结果不仅体现了生成数据集的价值,还指明了优化方向。

接下来,从公开数据集和生成数据集中各选取典型样本进行对比(见表 4)。公开数据集样本背景多为纯色或固定场景,垃圾目标尺寸较大且居中,而生成数据集样本背景更多样(超市、水边、厨房等)、光照条件更复杂,目标在画面中随机分布且相对更小。这样的定性差异直观展现了生成方法在场景多样性和目标随机性上的优势,补充了前文的量化分析。

表 4　公开与生成数据集样本对比

Tab. 4　Comparison of samples of public and generated datasets

	华为云生活垃圾数据集	TACO	GTC12-Mix
训练集数量	8 869	1 200	6 561
测试集数量	2 956	300	1 641
类别数量	40	60	12
数据是否平衡	否	否	否
数据样例			

表 5 为 ViT-Adapter 改进模块消融实验的结果,本实验基础设置为:Faster R-CNN 为目标检测框架,tiny 版本以及 3 × 训练策略。实验表明,RFAC 和 MLP 模块共同作用,能够全方位提升模型性能。其中,RFAC 更侧重于增强空间信息,MLP 则着重于特征的非线性组合。二者相结合,赋予了 ViT-Adapter 更强的表达能力,使其在不同的交并比阈值下都能有更优的表现。

表 5 ViT-Adapter 改进模块消融实验

Tab. 5 ViT-Adapter modified module ablation experiments

算法	mAP50:95	mAP50	mAP75
ViT-Adapter	0.77	0.90	0.64
ViT-Adapter-RFAC	0.78	0.90	0.66
ViT-Adapter-Det	0.80	0.92	0.68

为了评估改进版 ViT-Adapter 模型的通用性,表 6 的实验研究了多框架多尺度下性能及参数对比,通过该实验,全面检验改进后的模型在不同场景和模型配置中的有效性。实验结果显示,改进版模型在绝大多数配置下均优于原始模型,尤其是在 tiny 规模与 3 × 训练策略下,mAP50:95 提升最为显著,表明 RFAC 和 MLP 模块通过加强空间特征提取与非线性表达,能有效弥补小模型在特征学习上的不足,并在更长训练时间下实现更好的收敛和性能提升;此外,不同框架中也存在差异:在资源充足的 base 配置中,Faster R-CNN 表现最优,而在资源受限的 tiny 配置中,ATSS 的检测精度更高;虽然改进后参数量有所增加,带来更高的计算成本,但性能增益显著,需根据实际场景在高性能与轻量化之间进行权衡。

表 6 多框架多尺度下性能及参数对比

Tab. 6 Comparison of performance and parameters in multiple frameworks and multi-scale

检测框架	模型规模	训练策略	模型版本	mAP50:95	mAP50	mAP75	Params/10^6
ATSS	tiny	1 ×	原始	0.83	0.91	0.76	15.8
			改进	0.84	0.91	0.76	18.0
		3 ×	原始	0.81	0.90	0.72	15.8
			改进	0.84	0.90	0.77	18.0
	base	1 ×	原始	0.76	0.86	0.66	107.5
			改进	0.78	0.86	0.70	178.2
		3 ×	原始	0.78	0.84	0.71	107.5
			改进	0.78	0.83	0.72	178.2
Faster R-CNN	tiny	1 ×	原始	0.76	0.89	0.64	25.1
			改进	0.78	0.91	0.65	27.3
		3 ×	原始	0.77	0.90	0.64	25.1
			改进	0.80	0.92	0.68	27.3
	base	1 ×	原始	0.85	0.94	0.76	117.2
			改进	0.86	0.94	0.78	188.0
		3 ×	原始	0.84	0.91	0.76	117.2
			改进	0.85	0.92	0.77	188.0

实验结果见表 7,改进版 ViT-Adapter 模型在 Precision、Recall 和 F1 score 等指标上均优于原始模型。Precision 的提升表明改进版模型在预测正样本时更加精准;Recall 的提高体现其在覆盖所有真实正样本方面表现更佳。这 2 个指标的优化,带动 F1 score 和 Accuracy 也有所上升,有力地

验证了改进模块在细粒度分类任务中的优势。

表 7　细粒度二分类实验对比

Tab. 7　Comparison of fine-grained dichotomous classification experiments

算法	Precision	Recall	F1 score	Accuracy
ViT-Adapter	0. 86	0. 85	0. 85	0. 85
ViT-Adapter-Det	0. 87	0. 86	0. 86	0. 86

深入探究实验结果背后的原因,生成数据集因其多样化的垃圾形状和背景信息,为模型提供了高质量的训练资源,使改进后的模型在区分生成数据与真实数据的细微差异时,展现出更强的特征提取能力和分类鲁棒性。其中,RFAC 模块在空间特征提取方面发挥了关键作用,使得改进版模型能够更精确地区分生成数据与真实数据之间的细微差异。

表 7 的实验结果不仅证明了改进后的模型在区分生成数据与真实数据的细微差异上表现更佳,显著提升了特征提取和分类鲁棒性,还间接证实了生成数据集在垃圾分类任务中的重要应用价值,从而验证了改进策略在复杂场景下的有效性。该结论为后续进一步优化生成数据集和改进分类模型提供了重要参考依据。

4　结语

本研究首先通过多模态协同框架(借助 ChatGPT 生成的语义提示与 Stable Diffusion 的文生图/图生图技术)构建了涵盖 12 类、3 ~ 5 种细分形态、8 种场景及 4 种光照条件的高质量合成数据集 GTC12,并结合双阶段自动化与人工筛选以及旋转、裁剪、亮度调整等增强策略,大幅提升了样本的多样性与真实性,为模型泛化奠定了基础。随后针对 ViT-Adapter 在复杂背景和小目标场景下的特征提取不足,引入了动态感受野卷积注意力模块(RFAC)与交叉注意力后的 MLP 非线性映射模块,并在混合数据集 GTC12-Mix 上通过多框架、多尺度、不同训练策略的消融实验,以及在 GTC12-Bal 上的细粒度二分类测试,验证了两者协同可将模型的 mAP50:95 由 0. 77 提高至 0. 80,并将细粒度二分类准确率由 0. 85 提高至 0. 86,有效实现了对小目标和复杂背景场景的高效识别。

本文从数据构建与模型优化两方面,为复杂场景下的垃圾识别提供了改进思路。后续将结合生成模型与检测框架的更新,进一步验证所提方法在真实场景中的适用性,并探索其在视频检测与轻量化部署中的应用。

参考文献

[1] United Nations Environment Programme. Global waste management outlook 2024: Beyond an age of waste: Turning rubbish into a resource[R/OL]. Nairobi: United Nations Environment Programme, 2024. [2026-09-15]. https: // doi. org/10. 59117/20. 500. 11822/44939.

[2] Pongrácz E, Pohjola V J. Re-defining waste, the concept of ownership and the role of waste management [J]. Resources, Conservation and Recycling, 2004, 40(2): 141-153.

[3] 黄健,张钢. 深度卷积神经网络的目标检测算法综述[J]. 计算机工程与应用,2020,56(17):12-23.
Huang J, Zhang G. Survey of object detection algorithms for deep convolutional neural networks [J]. Computer Engineering and Applications, 2020, 56(17): 12-23.

[4] Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: A review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[5] Zaidi S S A, Ansari M S, Aslam A, et al. A survey of modern deep learning based object detection models [J]. Digital Signal Processing, 2022, 126: 103514.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580-587.

[7] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440-1448.

[8] 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.

[9] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017:2980-2988.

[10] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot MultiBox detector [M] // Computer Vision-ECCV 2016. Cham: Springer, 2016: 21-37.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for

dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999-3007.

[12] Wang A, Chen H, Liu L, et al. YOLOv10: Real-time end-to-end object detection[C]//Advances in Neural Information Processing Systems 37. Vancouver: NeurIPS Foundation, 2024: 107984-108011.

[13] Wang C Y, Yeh I H, Mark Liao H Y. YOLOv9: Learning what you want to learn using programmable gradient information[C]//Computer Vision-ECCV 2024. Cham: Springer, 2024: 1-21.

[14] Wang C Y, Bochkovskiy A, Liao H M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[EB/OL]. (2022-07-06)[2025-12-15]. https://arxiv.org/abs/2207.02696.

[15] Chen Z C, Yang J, Chen L F, et al. Garbage classification system based on improved ShuffleNet v2[J]. Resources, Conservation and Recycling, 2022, 178: 106090.

[16] 朱道恒, 朱洽灵, 李志强, 等. 基于改进 YOLOv7 的海滩垃圾检测方法[J]. 海洋环境科学, 2024, 43(5): 684-695.
Zhu D H, Zhu H L, Li Z Q, et al. A beach litter detection method based on improved YOLOv7[J]. Marine Environmental Science, 2024, 43(5): 684-695.

[17] Prova N N I. Garbage intelligence: Utilizing vision transformer for smart waste sorting[C]//2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). Coimbatore: IEEE, 2024: 1213-1219.

[18] Zhou Y, Wang Z Q, Zheng S R, et al. Optimization of automated garbage recognition model based on ResNet-50 and weakly supervised CNN for sustainable urban development[J]. Alexandria Engineering Journal, 2024, 108: 415-427.

[19] Chen Z, Duan Y C, Wang W H, et al. Vision transformer adapter for dense predictions[EB/OL]. (2022-05-17)[2025-12-15]. https://arxiv.org/abs/2205.08534.

[20] Tong Z, Gao J, Han Z Q, et al. Recognition of asphalt pavement crack length using deep convolutional neural networks[J]. Road Materials and Pavement Design, 2018, 19(6): 1334-1349.

[21] Panwar H, Gupta P K, Siddiqui M K, et al. AquaVision: Automating the detection of waste in water bodies using deep transfer learning[J]. Case Studies in Chemical and Environmental Engineering, 2020, 2: 100026.

[22] Mittal G, Yagnik K B, Garg M, et al. SpotGarbage: Smartphone app to detect garbage using deep learning[C]//Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg: ACM, 2016: 940-945.

[23] Proença P F, Simões P. TACO: Trash annotations in context for litter detection[EB/OL]. (2020-03-16)[2025-12-15]. https://arxiv.org/abs/2003.06975.

[24] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10674-10685.

[25] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[26] Liu P F, Yuan W Z, Fu J L, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9):1-35.

[27] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[R/OL]. San Francisco: OpenAI, 2018.[2025-12-15]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[28] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[R/OL]. San Francisco: OpenAI, 2019.[2025-12-15]. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[29] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. ACM, 2020: 1877-1901.

[30] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. New Orleans: NeurIPS, 2022, 35: 27730-27744.

[31] Knox W B, Stone P. Augmenting reinforcement learning with human feedback[C]//Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. Toronto: IFAAMAS, 2010: 439-446.

[32] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2025-12-15]. https://arxiv.org/abs/1706.03762.

[33] Hu L C, Zhang W S. Orthogonal neighborhood preserving discriminant analysis with patch embedding for face recognition[J]. Pattern Recognition, 2020, 106: 107450.

[34] Zhou Y, Wang X Y, Yao L, et al. LDAformer: Predicting lncRNA-disease associations based on topological feature extraction and Transformer encoder[J]. Briefings in Bioinformatics, 2022, 23(6): bbac370.

[35] Liu Z, Lin Y T, Cao Y T, et al. Swin transformer:Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021:9992-10002.

[36] Chen J R, Kao S H, He H, et al. Run, don't walk: Chasing higher FLOPS for faster neural networks[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 12021-12031.

[37] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Computer Vision. Cham: Springer, 2020: 213-229.

[38] Zhao Y A, Lyu W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 16965-16974.

(编　辑　李　静)