DOI：10.11830/ISSN.1000-5013.202410016

RGB-D 多模态融合与深度特征增强的固废检测网络

赵雀昊，刘炳辰，杨建红，房怀英

（华侨大学 机电及自动化学院，福建 厦门 361021）

摘要： 针对建筑固废在线识别中因相似特征导致的 RGB 识别准确率不高的问题，搭建双相机采集实验台，同步采集彩色图像和深度图像，提出一种基于彩色图像和深度图像的多模态融合与深度特征增强网络（DFENet）. DFENet 能够有效融合固废的彩色图像特征和深度图像特征。通过设计深度特征加强融合模块 PFPD 平衡并加强深度特征，显著提升了网络的识别精度。实验结果表明：与 RGB＋FPN（特征金字塔网络）方式相比，PFPD 方式在 IoU＝0.50 上的识别精度从 92.4％提高至 94.7％，在 IoU＝0.75 上的识别精度从 90.8％提升至 92.8％；与实例分割网络（Mask R-CNN）相比，DFEnet 识别精度从 86.4％提高至 89.2％；提出的方法有效地提高了固体废弃物识别的目标检测和实例分割模型识别精度。

关键词： 固废分选；深度加强；RGB-D 图像；特征融合；实例分割

中图分类号： TP 183；TP 249　　**文献标志码：** A　　**文章编号：** 1000-5013(2025)02-0133-09

Solid Waste Detection Network With RGB-D Multimodal Fusion and Deep Feature Enhancement

ZHAO Yinhao, LIU Bingchen, YANG Jianhong, FANG Huaiying

(College of Mechanical Engineering and Automation, Huaqiao University, Xiamen 361021, China)

Abstract: Aiming at the problem of low accuracy of RGB recognition due to similar features in online construction identification of solid waste, a dual-camera collection experimental platform is established to collect color images and depth images simultaneously. A multimodal fusion and depth feature enhancement network (DFENet) based on color image and depth image is proposed. DFENet can effectively fuse the color and depth image features of solid waste. By designing a deep feature strengthening fusion module (PFPD), the network balances and enhances depth features, significantly improving recognition accuracy. Experimental results show that compared with RGB＋FPN (feature pyramid network) method, the recognition precision of PFPD method increases from 92.4％ to 94.7％ at IoU＝0.50, and from 90.8％ to 92.8％ at IoU＝0.75. Compared with the instance segmentation network (Mask R-CNN), the recognition precision of DFENet improvs from 86.4％ to 89.2％. The proposed method can effectively improve the recognition precision of object detection and instance segmentation models for solid waste identification.

Keywords: solid waste sorting; depth enhancement; RGB-D image; feature fusion; instance segmentation

智能化分选在固废资源化利用中起到重要作用，非法处理固废会对环境造成破坏[1]，分选的关键技

收稿日期： 2024-10-30

通信作者： 房怀英（1978－），女，教授，博士，主要从事固废分选机器人开发等的研究。E-mail：happen@hqu.edu.cn。

基金项目： 福建省高效产学合作项目（2024H6010）；福建省科技计划项目（2023Y3006）；第 6 批福建省泉州市引进高层次人才团队项目（2023CT003）

术在于固废在线识别,现有的分选系统大多采用破碎、圆盘筛网、磁鼓、人工挑选等多级传统建筑固废分选[2-3],但传统机械结构分选的纯度低,效率无法得到保障,人工捡拾需要投入大量人力,浪费劳动力的同时也难以满足工业自动化的需求。

随着计算机视觉和人工智能技术的快速发展,将相机采集的 RGB 图像输入神经网络,可以对图像中的每个物体进行目标检测[4-5],其中,端到端的单阶段目标检测有 YOLO 系列方法(代表)[6-7]、Segment Anything 方法[8]和 Transformer 方法[9]。通过语义分割[10]划分出轮廓,提出基于颜色特征[11]、MobileNet[12]、pix2pix、残差神经网[13]、YOLOv8[14]的固废分选方法。但真实工况通常比较复杂,如对于破碎后具有相近颜色、纹理、大小的砖块和混凝土,被砂浆包裹的砖块等,RGB 图像无法做到有效地区分[15],Segment Anything 方法及 Transformer 方法识别精度较高但推理速度慢,无法满足在线实时检测需求。

多模态融合的方法得到越来越多的关注,热图像可以补充 RGB 的图像特征,以提高 RGB-T 语义分割性能[16],近红外技术(NIR)解决了复杂工况缺乏纹理信息和照明不足的问题[17],高光谱成像技术可以有效地获得物体的光谱和空间信息的特点[18]。在固废分选领域,利用彩色摄像头和激光轮廓扫描仪采集 RGB 图像和深度图像[19],实例分割网络(Mask R-CNN)采用不同的方式融合 RGB 和深度图像,提高固废检测的性能。利用非对称多尺度特征融合网络(AMFFNet),融合固废 RGB 谱信息[20]、固废检测网络[21]、固废视觉检测方法识别混凝土和灰砖[22],分别对建筑固废的 RGB 图像和深度图像做图像处理,都有效提高建筑固废检测识别精度,但是存在 RGB 特征与深度特征不平衡的问题,双主干网络将 RGB 图像与深度图像进行融合[23],使网络以相同的权重融合两种特征,并在网络中嵌入注意力机制辅助平横特征[24-25]。基于此,本文对 RGB-D 多模态融合与深度特征增强的固废检测网络进行研究。

1 数据与实验方法

1.1 实验台搭建与数据采集

双相机采集系统原理图,如图 1 所示。采集系统包括一个 RGB 成像模块和一个高度成像模块。RGB 成像模块由彩色线阵相机和发光二极管(LED)光源组成,用于采集物体的彩色图像,得到丰富的颜色和纹理信息。高度成像模块为激光轮廓扫描仪,扫描仪包含一个激光发射器和两个单色相机,用于采集物体的深度图像,得到形状信息和深度信息,穹顶光源照明安装在穹顶边缘,指向正上方,使光线从穹顶的曲面反射出去,从而产生均匀反射。抓取模块包括分拣机器人和抓取模组,用于接收检测信息并实时分拣传送带上的物料。

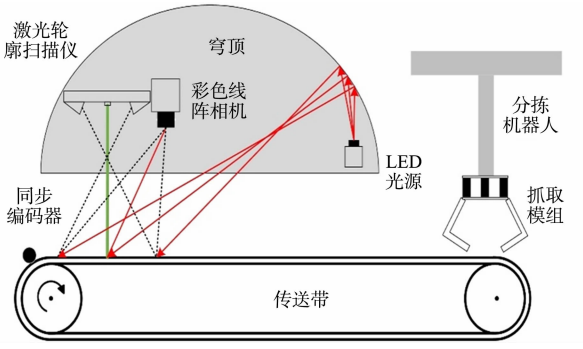
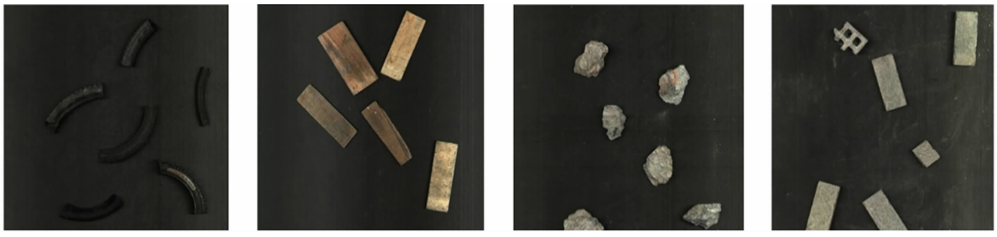


图 1 双相机采集系统原理图

Fig. 1 Schematic diagram of dual camera acquisition system

对黑色橡胶、木头、混凝土和砖块 4 类常见的固废进行实验,其中,1 038 张 RGB 图像和深度图像作为训练集,455 张彩色图像和深度图像作为测试集,深度图像与 RGB 图像均标定并对齐。数据集部分样本,如图 2 所示。



(a) 黑色橡胶 (b) 木头 (c) 混凝土 (d) 砖块

图 2 数据集部分样本

Fig. 2 Partial samples of dataset

1.2　固废检测算法

建筑固废多级破碎后，由皮带传输，建筑固废表面通常被砂浆、粉尘覆盖，颜色特征退化严重，破碎完的建筑固废变得不规则，形状特征无法有效提取。同时，建筑固废在皮带上也会存在堆叠的情况，导致模型会将粘连的同类物体识别为一个物体。

4 类材料中混凝土与砖块同为灰色，在颜色尺度上有相似的特征，会在一定程度上影响分类的准确度；黑色橡胶与传送皮带也同样具有相似的特征，而传统的 RGB 分割算法主要针对颜色和轮廓信息进行提取，因此，难以得到有效的识别结果。针对上述问题，提出 RGB-D 多模态融合与深度特征加强的检测网络 DFENet。

1）特征融合模块。使用双通道卷积神经网络分别提取 RGB 通道的颜色、纹理等特征和深度通道的深度、边缘等特征，通过对应元素叠加的方式融合 RGB 通道和深度通道的特征。

2）注意力机制嵌入模块。使用注意力机制嵌入卷积神经网络，使网络更加关注有用的特征，抑制冗余信息，减小特征信息的损失，得到特征信息含量更高的信息。

3）深度特征加强融合模块 PFPD。通过自上而下的左边特征金字塔网络（L-FPN）网络提取更多的语义信息，再次融合深度特征信息后输入自下而上的右边特征金字塔网络（R-FPN）网络，从而更好地利用不同特征层之间的信息，恢复顶层损失的深度特征信息。

检测头阶段，将特征图中的候选感兴趣区域（ROI）送入 RPN 网络进行过滤，对剩下的 ROI 区域进行 ROIAlign 操作。

1.2.1　特征融合模块　DEFNet 网络结构图，如图 3 所示。图 3 中：$\boldsymbol{C}_i^{\mathrm{Depth}}(i=1、2、3、4)$ 表示为提取到的特征图。RGB 图像和深度图像分别使用对称 ResNet 进行特征提取，RGB 分支图像为三通道输入，图像尺寸为 960 px×1 024 px，提取到的特征图表示为 $\boldsymbol{C}_i^{\mathrm{RGB}}(i=1、2、3、4)$，每一层输出的特征图大小依次为 64、128、256、512。

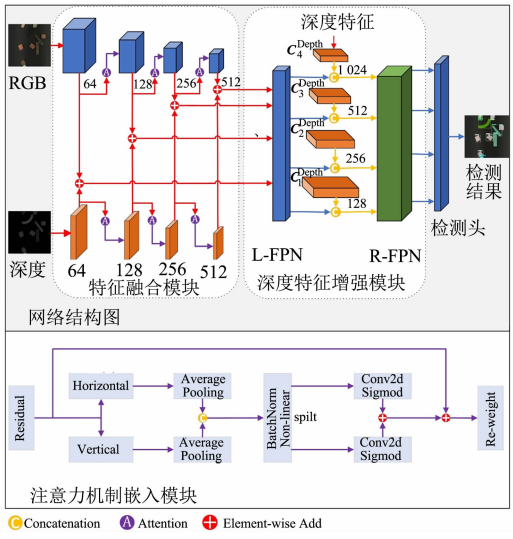


图 3　DEFNet 网络结构图

Fig. 3　DEFNet network architecture diagram

深度分支图像为单通道输入，图像尺寸为 960 px×1 024 px，为了保证提取对称特征并融合，需要将 ResNet 第一层卷积修改为单通道，每一层输出的特征图大小与 RGB 输出的尺寸相同，依次为 64、128、256、512。将最后一个特征层的大小平衡在 7 px×7 px，对输入的 RGB 和深度分支图像进行预处理，归一化再裁剪，尺寸为 224 px×224 px，将其作为第一层卷积的输入。

特征融合部分使用 ReLU 激活函数和最大池化层，ReLu 激活函数可以有效避免梯度爆炸和梯度消失的问题，最大池化层对 RGB 和深度进行下采样，并选择分辨率更高的特征，更好地保留 RGB 纹理特征。

ReLu 激活函数表示为

$$\mathrm{Output}=\max(0,w^{\mathrm{Ther}}x+b)。\tag{1}$$

式(1)中：x 为上一层输入的网络的输入；w^{Ther} 为权重；b 为添加到输入加权总和中的偏置。

通过 Element-wiseAdd 的方式进行一次融合，将特征图对应元素相加，融合后的特征图作为后续 L-FPN 的输入，即

$$\boldsymbol{T}_i=[\boldsymbol{C}_i^{\mathrm{RGB}}\oplus\boldsymbol{C}_i^{\mathrm{Depth}}],\qquad i=1、2、3、4。\tag{2}$$

式(2)中：$\boldsymbol{T}_i$ 表示通过一次 Element-wiseAdd 融合后的特征图。

ResNet[26] 的核心思想是引入了残差连接和残差函数，通过这种方式解决了深层网络训练过程中的梯度消失和梯度爆炸问题。残差连接通过将输入特征与网络的输出进行直接相加，使网络可以更容易地学习残差，从而优化模型的性能。

$$y=F(x,\{W_i\})+W_sx。\tag{3}$$

残差函数的公式为

$$F=W_2\sigma(W_1x)。\tag{4}$$

式(3),(4)中:x,y 分别为输入和输出;$F(x,\{W_i\})$为需要进行残差学习的函数;W_s 为输入 x 的维度。

1.2.2 注意力机制嵌入模块 三通道的RGB固废图像包含的信息更多,每个通道可以独立控制图像中相应颜色的强度,而单通道的深度图像包含的信息少,只记录像素的亮度信息,再将亮度信息转化为实际的深度。RGB图像特征丰富但存在大量冗余,深度图特征单一但存在噪声,因此,需要将注意力机制嵌入RseNet中,使特征提取网络能更好地提取有效特征,排除冗余,加强各层特征之间的联系,帮助模型集中于图像中更重要的部分,从而提高特征提取的效率和准确性。

将RGB图像特征定义为 $\text{input}_i^{\text{RGB}}\in\mathbf{R}^{c\times h\times w}(i=\{1、2、3、4、5\})$,其中,$c$ 表示通道数,h 和 w 分别表示特征图的高度和宽度。将input输入至注意力机制嵌入模块,尺寸为$(h,1)$和$(1,w)$的池化核分别沿水平坐标和垂直坐标两个方向对通道进行编码,从而得到 $X_c^w(w)$和 $X_c^h(h)$,即

$$X_c^w(w)=\frac{1}{h}\sum_{0\leqslant j<h}\text{input}_c^{\text{RGB}}(j,w),\tag{5}$$

$$X_c^h(h)=\frac{1}{w}\sum_{0\leqslant i<w}\text{input}_c^{\text{RGB}}(h,i)。\tag{6}$$

$\boldsymbol{X}_{i,1}^{\text{RGB}}$对RGB特征宽度方向和高度方向分别进行池化操作并沿着空间方向聚合,对信息在水平方向和垂直方向进行拼接,即

$$\boldsymbol{X}_{i,1}^{\text{RGB}}=(\text{AvePooling}(X_c^w(w)),\text{AvePooling}(X_c^h(h)))。\tag{7}$$

式(7)中:AvePooling为平均池化,表示该窗口的特征;$\boldsymbol{X}_i^{\text{RGB}}$ 为经过平均池化后的特征,使用一个共享的1×1的共享卷积层 F 进行变换,即

$$\boldsymbol{X}_{i,2}^{\text{RGB}}=F(\boldsymbol{X}_{i,1}^{\text{RGB}})。\tag{8}$$

在原始RGB特征图上进行 g^h 和 g^w 的乘法加权计算,输出为

$$\text{Output}_c^{\text{RGB}}(i,j)=X_i^{\text{RGB}}(i,j)\times g_c^h(i)\times g_c^w(j)。\tag{9}$$

首先,通过全局平均池化对每个通道上的特征进行平均池化操作,将特征图的空间维度降为1×1,得到每个通道的全局特征表示。然后,通过全连接层将全局平均池化后的特征输入到一个全连接层中,通过学习每个通道的权重系数确定每个通道的重要性。

1.2.3 深度特征加强融合模块PFPD 深度特征加强的方式采用对浅层卷积与深层卷积一次融合后,再进行深度特征加强,以充分融合位置信息与高度特征,避免一次自上而下的特征金字塔,从而失去整体位置和深度之间的联系。将L-FPN的输出特征图 $\boldsymbol{T}_i$ 与深度图像特征 $\boldsymbol{C}_i^{\text{Depth}}$ 进行聚合,二次融合,即

$$\boldsymbol{D}_i=[\boldsymbol{C}_i^{\text{Depth}},\boldsymbol{T}_i],\qquad i=1、2、3、4。\tag{10}$$

深度特征加强融合模块PFPD,如图4所示。L-FPN从较低分辨率的特征图开始,采用双线性差值算法进行上采样,在原有图像特征图像素的基础上,在像素之间插入新的像素,将主干网络提取的特征图 $\boldsymbol{T}_i(i=1、2、3、4)$尺寸扩大为原来2倍,再依次与前一特征图相加完成,用于整合不同尺度的RGB与深度特征。L-FPN从上而下把包含固废位置和深度等信息的下层特征与包含固废语义信息的上层特征进行融合,不同尺度特征图都包含丰富的信息。R-FPN的输入 $\boldsymbol{D}_i(i=1、2、3、4)$使用尺寸为3×3,步长为2的卷积层进行下采样操作,将特征图缩小为原尺寸的1/2,再依次与前一图相加后完成自下而上R-FPN部分,该部分充分利用深度加强融合后的特征,减少了下层特征信息的传递损失。

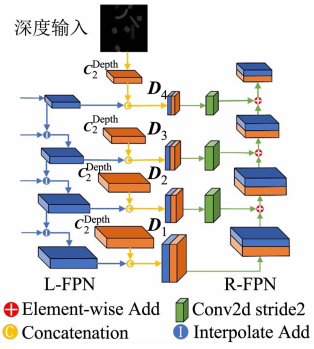


图4 深度特征加强融合模块PFPD

Fig. 4 Deep feature enhancement fusion module PFPD

2 实验结果和分析

2.1 实验参数

训练采用随机梯度下降(SGD),动量参数设置为0.9,偏置 b 设置为0,一共训练100轮,初始学习

率参数设置为 0.004，每迭代 30 次，学习率下降为初始学习率的 1/10，训练损失稳步下降。使用 COCO2014 数据集对提取特征部分的网络进行预训练，得到预训练权重。在经过非极大值抑制（NMS）结构之后，保留 1 000 个 RoI 区域，选择前景分割过程中得分最高的 100 个实例进行分割，实验评价指标选用平均识别精度（P_A）对分割结果进行定量的判定，表示对每一类识别的正确的数量与该类总量之比。P_A、识别精度（P）与召回率（R）之间的关系，即

$$\left.\begin{aligned} P_A &= \int_0^1 P \cdot R \mathrm{d}r, \\ P &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ R &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}。 \end{aligned}\right\} \tag{11}$$

式（11）中：TP 为预测与实际标签相同的正样本数量；FP 为预测与实际标签不同的负样本数量；FN 为以实际为背景但预测为标签的负样本数量。考虑不同的阈值（IoU）对实验结果的影响，选用 0.50、0.75 的 IoU 阈值进行比较。

2.2　基线目标检测模型

为了选择目标检测基线模型，选择双阶段目标检测网络 Faster R-CNN、Mask R-CNN[27]，以及单阶段目标检测网络 YOLOv5、YOLOv8 和 Co-DETR 进行对比。目标检测网络识别精度与推理时间（t）对比，如表 1 所示。

表 1　目标检测网络识别精度与推理时间对比

Tab. 1　Comparison between recognition precision and inference time of object detection model

目标检测网络	主干网络	$P_{\mathrm{IoU}=0.50}$/%	t/ms
Faster R-CNN	ResNet50＋FPN	83.7	22
Mask R-CNN	ResNet50＋FPN	90.4	27
YOLOv5	CSPDarknet53	89.3	21
YOLOv8	Darknet53	91.7	26
Co-DETR	ResNet50	95.8	315

由表 1 可知：为了能够准确定位建筑固废，确保后续执行机构能够进行抓取和气吹，需要得到目标固废的掩膜和具有较快的检测速度以满足实时性，因此，目标检测网络选择 Mask R-CNN 作为对比。

2.3　注意力机制对比实验

在不同主干网络中分别加入通道注意力机制模块和注意力机制嵌入模块，以强化模型对于特征和位置的关注，将注意力机制嵌入模块主干网络中，对提取的 RGB 图像特征图和深度图像特征图分别编码形成对通道、位置和方向感知敏感的注意力图。不同模式识别精度比较，如表 2 所示。

表 2　不同模式识别精度比较　（单位：%）

Tab. 2　Comparison of different pattern recognition precision　（Unit：%）

模块	主干网络	$P_{\mathrm{IoU}=0.50}$	$P_{\mathrm{IoU}=0.75}$	P
通道注意力机制	ResNet101	93.2	91.4	77.8
	ResNeXt101	93.3	91.8	78.1
注意力机制嵌入	ResNet101	94.7	92.8	77.6
	ResNeXt101	93.8	92.6	78.4

由表 2 可知：注意力机制嵌入模块在不同主干网络上的检测识别精度都高于通道注意力机制模块，通道注意力机制模块只关注通道之间的联系，特征相互分离，位置信息忽略。注意力机制嵌入模块能更好地关注三通道 RGB 图像特征和单通道深度特征之间的联系，沿空间方向捕获特征之间远程依赖关系，并保留精确的位置关系。

不同注意力机制热图，如图 5 所示。由图 5 可知：4 种单类物体工况下，注意力机制嵌入模块能更加聚焦在目标物体区域，对非感兴趣区域抑制能力更强，通道注意力机制模块关注的范围却更加广泛，无法有效的针对目标物体；对于混合类，目标物体种类多，工况更加复杂，通道注意力机制模块仅能重点聚焦一部分感兴趣区域，而注意力机制嵌入模块会对感兴趣区域分区域进行关注，形成多个热点区域。

ResNeXt101[28]的特征提取能力强于ResNet101，因此，网络本身更加关注细节特征，而深度图像仅为单通道灰度图，但加入注意力机制嵌入模块后精准无法聚焦，对于深度图像特征无法起到很好的提取作用，因此使用注意力机制嵌入模块融合ResNet101在特征提取效果上有很好的效果。

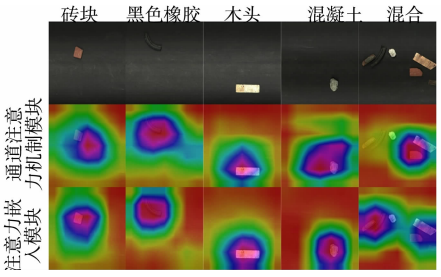


图5 不同注意力机制热图

Fig. 5 Heat maps of different attention mechanisms

2.4 有效性实验

为了验证特征融合与PFPD的有效性，使用MaskR-CNN输入仅为三通道RGB图像(作为基准)，采用ResNet101作为主干网络，分别验证了RGB-D早期融合(RGB-D E)、RGB-D中期融合(RGB-D M)、PFPD的性能。不同融合方式的识别精度比较，如表3所示。

表3 不同融合方式的识别精度比较 (单位：%)

Tab. 3 Recognition precision of different fusion models (Unit: %)

融合方式	$P_{\mathrm{IoU}=0.50}$	$P_{\mathrm{IoU}=0.75}$	P
RGB+FPN	92.4	90.8	77.3
RGB-D E+FPN	93.3	91.2	76.9
RGB-D M+FPN	93.5	91.2	77.0
PFPD	94.7	92.8	77.6

由表3可知：与RGB+FPN(特征金字塔网络)方式相比，PFPD方式在IoU=0.50上的识别精度从92.4%提高至94.7%，在IoU=0.75上的识别精度从90.8%提升至92.8%；相比于仅使用RGB+FPN融合方式，采用PFPD的$P_{\mathrm{IoU}=0.50}$，$P_{\mathrm{IoU}=0.75}$都有提高，这说明深度信息可以作为RGB特征的补充信息，起到有效作用；RGB-D E+FPN融合方式是将RGB图像与深度图像先进行拼接，再输入网络，过早的融合特征信息使特征提取网络不能区分两种信息之间的差别，识别精度低于RGB-D M+FPN融合方式，而PFPD的$P_{\mathrm{IoU}=0.50}$，$P_{\mathrm{IoU}=0.75}$都高于RGB-D E+FPN、RGB-D M+FPN融合方式，这个是因为单通道深度图像特征信息少于三通道RGB图像特征信息，而RGB-D M+FPN融合方式对于RGB特征和深度特征使用相同的权重，因此，只采用一次融合的方式不能有效利用深度特征。PFPD可以更有效地将底层的特征和高层的特征融合起来，在保留高层特征的语义信息的同时，保留低层特征的物体位置信息，有效提升目标检测识别和定位精度。

2.5 消融实验

将DFENet嵌入通用网络Mask R-CNN中，并使用不同深度的主干网络进行目标检测和实例分割，以评价其通用性和有效性。分别使用ResNet50、ResNet101、ResNeXt50和ResNeXt101作为主干网络。主干网络识别精度比较，结果如表4所示。

表4 主干网络识别精度比较 (单位：%)

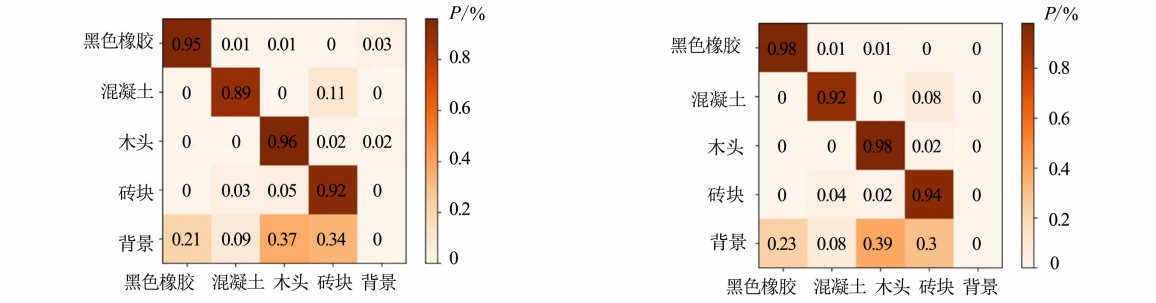
Tab. 4 Comparison of recognition precision of backbone networks (Unit: %)

主干网络	Mask R-CNN			DFENet		
	$P_{\mathrm{IoU}=0.50}$	$P_{\mathrm{IoU}=0.75}$	P_{A}	$P_{\mathrm{IoU}=0.50}$	$P_{\mathrm{IoU}=0.75}$	P_{A}
ResNet50	90.4	88.5	76.2	93.1	89.8	77.9
ResNet101	92.4	90.8	77.3	94.7	92.8	77.6
ResNeXt50	93.2	91.4	77.8	92.5	91.0	77.4
ResNeXt101	93.0	90.8	76.7	93.8	92.6	78.4

由表4可知：对于ResNet50，DFENe的P_{A}比Mask R-CNN提高1.7%，$P_{\mathrm{IoU}=0.50}$比Mask R-CNN提高2.7%，$P_{\mathrm{IoU}=0.75}$比Mask R-CNN提高1.3%；随着网络层数的加深，提取特征能力加强，对于ResNet101，DFENet比Mask R-CNN的P_{A}提高0.3%，$P_{\mathrm{IoU}=0.50}$，$P_{\mathrm{IoU}=0.75}$比Mask R-CNN分别提高2.3%和2.0%。

4类固废检测结果热力图，如图6所示。图6中：列表示真实类别标签；行表示预测类别标签。由图6(a)可知：混凝土(0.89)和砖块(0.92)的检测识别精度较低，主要原因是将混凝土误识别为砖块，黑

色橡胶类的识别结果中，将黑色橡胶误识别为背景，说明只使用 RGB 作为输入，在特征提取阶段，无法区分有相似纹理的信息，从而在后续识别阶段出现误识别。由图 6(b)可知：DFENet 融合了深度信息并对深度特征进行加强，可以有效避免与背景误识别的情况，在混凝土类中，识别精度有所提高，砖块的误识别率也有所下降。



(a) Mask R-CNN 检测结果　　(b) DFENet 检测结果

图 6　4 类固废检测结果热力图

Fig. 6　Thermal diagrams of detection results for 4 types solid waste

实例分割平均识别精度可以有效地显示模型对每一类物体的分割情况，实例分割平均识别精度比较，如表 5 所示。由表 5 可知：与 Mask R-CNN 相比，DFEnet 识别精度从 86.4%提高至 89.2%；相比于 MaskR-CNN，DFENet 在与黑色传送带有相同颜色特征的黑色橡胶平均识别精度提高 1.3%，在有相似颜色形状特征的混凝土和砖块平均识别精度都提高 2.8%，木头类平均识别精度提高 1.9%；相比 YOLOv8，DFENet 在黑色橡胶的平均识别精度有减少 0.8%，但其他三类固废提高 0.2%～4.0%。这证明了 DFENet 能够很好地融合 RGB 特征和深度特征的优点，对于轮廓的划分更加的精细和准确，更有利于固废识别检测。

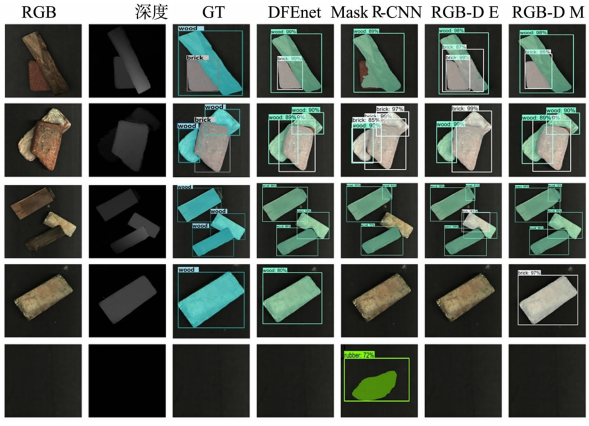
表 5　实例分割平均识别精度比较　　(单位：%)

Tab. 5　Comparison of average recognition precision in instance segmentation　　(Unit: %)

网络	P_A(黑色橡胶)	P_A(混凝土)	P_A(木头)	P_A(砖块)
Mask R-CNN	95.1	86.4	94.6	92.2
YOLOv5	91.6	89.6	96.0	93.1
YOLOv8	97.2	85.2	95.3	94.8
DFENet	96.4	89.2	96.3	95.0

2.6　可视化结果

网络检测结果，如图 7 所示。图 7(a)中：第 1～3 列分别为 RGB 图像、深度图像、真实标签(GT)，第 4～7 列分别为 DFENet 的检测结果、Mask R-CNN 的检测结果、RGB-D E 和 RGB-D M 的检测结果。



(a) 不同融合方式检测结果

(b) 真实工况下 DFENet 检测结果

图 7　网络检测结果

Fig. 7　Detection results of networks

由图7(a)可知:第1～3行均存在木头与砖块堆叠的情况,Mask R-CNN会只识别为一个物体或误识别成多个物体,无法有效区分被遮挡部分;第4行RGB-D M将木头误识别为砖块,说明RGB-D中期融合的方式对于深度特征提取能力不够;第5行Mask R-CNN将空皮带误识别成黑色橡胶,而采用RGB-D融合的方法均能避免该类情况发生;相比于RGB输入的Mask R-CNN,融合深度信息能有效避免漏检、误检的问题,黑色橡胶与深色传送带之间的区分,对于堆叠的情况,DFENet也能有效区分不同的物体。由图7(b)可知:对于堆叠、粘连情况,DFENet可以有效识别并分割。

综上所述,DFENet可以在实验测试集中更加准确地进行目标检测和实例分割,其检测结果优于基准Mask R-CNN网络。

3 结论

1) 通过提出DFENet,在特征提取网络中加入了注意力机制嵌入模块以增加特征提取能力,PFPD先用自下而上的结构,融合RGB图像和深度图像的特征,深度特征加强融合后自上而下进行多尺度特征融合,DFENet融合方式显著提升了固废目标检测的性能,相较于传统的单模态方法,DFENet能使目标检测识别精度提高0.3%,$P_{\mathrm{IoU}=0.50}$,$P_{\mathrm{IoU}=0.75}$分别提高2.3%和2.0%。这证明了融合RGB和深度信息对于改善目标检测的效果具有显著的积极影响。

2) 在实例分割任务上也取得了显著的改进,通过融合RGB图像和深度图像信息,能够更好地捕捉目标的边界和细节信息,提高了实例分割的准确性和鲁棒性,并且相较于单模态方法,在实例分割任务中表现出更好的性能,DFENet在单类实例分割识别精度上最高提高2.8%。

模型不足之处在于实验室工况存在少量污染、大量堆叠等情况,在运用于真实工况任务中会表现不稳定,有较高的误识别率。下一步将针对不同工况进行研究,提高模型的泛化能力。

参考文献:

[1] FRATERNALI P,MORANDINI L,GONZÁLEZ S L H. Solid waste detection, monitoring and mapping in remote sensing images: A survey[J]. Waste Management,2024,189:88-102. DOI:10.1016/j.wasman.2024.08.003.

[2] BONIFAZI G,SERRANTI S. Recycling technologies[C]// Encyclopedia of Sustainability Science and Technology. New York: Springer,2019:1-57. DOI:10.1007/978-1-4939-2493-6_116-4.

[3] JANK A,MÜLLER W,SCHNEIDER I,*et al*. Waste separation press: A mechanical pretreatment option for organic waste from source separation[J]. Waste Management,2015,39:71-77. DOI:10.1016/j.wasman.2015.02.024.

[4] ROSS T Y,DOLLÁR G. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press,2017:2980-2988. DOI:10.1109/ICCV.2017.324.

[5] LIN T Y,DOLLÁR P,GIRSHICK R,*et al*. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2017:2117-2125. DOI:10.1109/CVPR.2017.106.

[6] WANG C Y,YEH I H,LIAO H Y M. Yolov9: Learning what you want to learn using programmable gradient information[C]// European Conference on Computer Vision. Cham:Springer,2025:1-21. DOI:10.1007/978-3-031-72751-1_1.

[7] WANG C Y,BOCHKOVSKIY A,LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2023:7464-7475.

[8] KIRILLOV A,MINTUN E,RAVI N,*et al*. Segment anything[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE Press,2023:4015-4026. DOI:10.48550/arXiv.2304.02643.

[9] ZONG Zhuofan,SONG Guanglu,LIN Yu. Detrs with collaborative hybrid assignments training[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press,2023:6748-6758. DOI:10.48550/arXiv.2211.12860.

[10] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2015:3431-3440. DOI:10.1109/TPAMI.2016.2572683.

[11] 郑龙海，袁祖强，殷晨波，等. 基于机器视觉的建筑垃圾自动分类系统研究[J]. 机械工程与自动化，2019(6)：16-18. DOI：10. 3969/j. issn. 1672-6413. 2019. 06. 006.

[12] XU Xiong，ZHAO Beibei，TONG Xiaohua，*et al*. A data augmentation strategy combining a modified pix2pix model and the copy-paste operator for solid waste detection with remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2022，15：8484-8491. DOI：10. 1109/JSTARS. 2022. 3209967.

[13] DAVIS P，AZIZ F，NEWAZ M T，*et al*. The classification of construction waste material using a deep convolutional neural network[J]. Automation in Construction，2021，122：103481. DOI：10. 1016/j. autcon. 2020. 103481.

[14] LI Pan，XU Jiayin，LIU Shenbo. Solid waste detection using enhanced YOLOv8 lightweight convolutional neural networks[J]. Mathematics，2024，12(14)：2185. DOI：10. 3390/math12142185.

[15] LU Weisheng，CHEN Junjie，XUE Fan. Using computer vision to recognize composition of construction waste mixtures：A semantic segmentation approach[J]. Resources，Conservation and Recycling，2022，178：106022. DOI：10. 1016/j. resconrec. 2021. 106022.

[16] DENG Fuqin，FENG Hua，LIANG Mingjian，*et al*. FEANet：Feature-enhanced attention network for RGB-thermal real-time semantic segmentation[C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway：IEEE Press，2021：4467-4473. DOI：10. 1109/IROS51168. 2021. 9636084.

[17] XIAO Wen，YANG Jianhong，FANG Huaiying，*et al*. A robust classification algorithm for separation of construction waste using NIR hyperspectral system[J]. Waste Management，2019，90：1-9. DOI：10. 1016/j. wasman. 2019. 04. 036.

[18] LU Bing，DAO P D，LIU Jianggui，*et al*. Recent advances of hyperspectral imaging technology and applications in agriculture[J]. Remote Sensing，2020，12(16)：2659. DOI：10. 3390/rs12162659.

[19] LI Jiantao，FANG Huaiying，FAN Lulu，*et al*. RGB-D fusion models for construction and demolition waste detection [J]. Waste Management，2022，139：96-104. DOI：10. 1016/j. wasman. 2021. 12. 021.

[20] CAI Zhenxing，FANG Huaiying，JIANG Fengfeng，*et al*. AMFFNet：Asymmetric multi-scale feature fusion network of RGB-NIR for solid waste detection[J]. IEEE Transactions on Instrumentation and Measurement，2023，72：1-10. DOI：10. 1109/TIM. 2023. 3300445.

[21] LI Yangke，ZHANG Xinman. Multi-scale context fusion network for urban solid waste detection in remote sensing images[J]. Remote Sensing，2024，16(19)：3595. DOI：10. 3390/rs16193595.

[22] ZHUANG Jiangteng，FANG Huaiying，XIAO Wen，*et al*. Recognition of concrete and gray brick based on color and texture features[J]. Journal of Testing and Evaluation，2019，47(4)：3224-3237. DOI：10. 1520/JTE20180523.

[23] HU Xinxin，YANG Kailun，FEI Lei，*et al*. Acnet：Attention based network to exploit complementary features for rgbd semantic segmentation[C]// IEEE International Conference on Image Processing. Piscataway：IEEE Press，2019：1440-1444. DOI：10. 1109/ICIP. 2019. 8803025.

[24] HE Kaiming，GKIOXARI G，DOLLÁR P，*et al*. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Xision. Piscataway：IEEE Press，2017：2961-2969. DOI：10. 1109/ICCV. 2017. 322.

[25] HU Jie，SHEN Li，SUN Gang. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE Press，2018：7132-7141. DOI：10. 1109/CVPR. 2018. 00745.

[26] HOU Qibin，ZHOU Daquan，FENG Jiashi. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE Press，2021：13713-13722. DOI：10. 1109/CVPR46437. 2021. 01350.

[27] MA Wanqi，CHEN Hong，ZHANG Wenkang，*et al*. DSYOLO-trash：An attention mechanism-integrated and object tracking algorithm for solid waste detection[J]. Waste Management，2024，178：46-56. DOI：10. 1016/j. wasman. 2024. 02. 014.

[28] XIE S，GIRSHICK R，DOLLÁR P，*et al*. Aggregated residual transformations for deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE Press，2017：1492-1500. DOI：10. 1109/CVPR. 2017. 634.

（**责任编辑**：陈志贤　　**英文审校**：陈婧）