引文格式：孙艳新. 利用Transformer的水体垃圾检测方法研究[J]. 地理空间信息, 2025, 23(6):70-73. doi:10.3969/j.issn.1672-4623.2025.06.015

利用 Transformer 的水体垃圾检测方法研究

孙艳新[1]
（1. 新疆维吾尔自治区第二测绘院，新疆 乌鲁木齐 830001）

摘　要：水体垃圾可能导致传染病的传播，影响公共卫生；同时，堆积的垃圾可能对航行安全构成威胁。基于Transformer的目标检测模型具有更强的全局信息建模和多尺度自适应能力。针对水体垃圾检测问题，分别基于CNN模型PP-YOLOE，Transformer模型DETR、Deformable DETR以及RT-DETR模型对水体垃圾进行检测，对CNN模型和3种Transformer模型在水体垃圾检测方面的能力进行对比分析。通过对比发现，CNN模型PP-YOLOE检测结果优于DETR，略次于Deformable DETR，而RT-DETR取得最优检测结果，在实验数据集上AP_{50}达到0.884，相较于PP-YOLOE，具有7个点的提升。

关键词：水体垃圾；Transformer；全局信息；检测
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2025）06-0070-04

Research on Water Body Garbage Detection Method Based on Transformer

SUN Yanxin[1]
(1. Second Surveying and Mapping Institute of Xinjiang Uygur Autonomous Region, Urumqi 830001, China)

Abstract: The pollution of water body garbage can lead to the spread of infectious diseases and impact public health. Additionally, accumulated garbage can pose a threat to navigation safety. The object detection models based on Transformer possess enhanced capabilities for global information modeling and multi-scale adaptability. Addressing the issue of water body garbage detection, we used the CNN model PP-YOLOE, the Transformer models DETR, Deformable DETR and RT-DETR to detect water garbage, and compared and analyzed the abilities of CNN model and three Transformer models in garbage detection. The comparison reveals that the CNN model PP-YOLOE performs better than DETR, is slightly inferior to Deformable DETR, and RT-DETR achieves the best detection results. On the experimental dataset, RT-DETR achieves an AP50 of 0.884, which is a 7-point improvement over PP-YOLOE.

Key words: water body garbage, Transformer, global information, detection

随着工业发展，水污染已成为一个严峻的问题。水面上漂浮的塑料垃圾不仅影响环境，还对水生生态系统造成了严重的威胁。为有效检测和清理水体垃圾，保护水生生态系统，国内外学者在水体垃圾检测领域开展了深入研究[1-4]。

目前，基于YOLO系列的方法都已获得较好的水体垃圾检测结果，然而相关方法都是基于卷积神经网络CNN的方法[5-6]。随着自注意力机制技术的发展，基于Transformer的目标检测技术已经展现出更优秀的检测精度[7]。基于Transformer模型的水体垃圾检测研究较少，针对这一问题，本文分别基于CNN模型PP-YOLOE[8]，Transformer模型DETR[9]、Deformable DETR以及RT-DETR模型对水体垃圾进行检测[10]，对这4种模型进行对比分析。

1　基于Transformer的水体垃圾检测

图1给出了基于PP-YOLOE的水体垃圾检测流程图，PP-YOLOE采用典型的三段式目标检测网络框架，即骨干网络backbone，颈部网络neck，以及头部网络head。PP-YOLOE采用无锚点方法，减少对预设锚点大小和比例的依赖，简化模型并提高了其适应不同数据集的能力，同时引入CSPRepResNet作为其骨干网，有效降低计算成本的同时保持或提高检测精度。backbone通过堆叠多个卷积层和池化层来逐步缩小图像的尺寸，并提取更高级别的特征。这种逐层的处理方式使得网络能够捕捉到图像中的边缘、纹理、形状等重要特征，然后传递给neck进行特征融合变换。深度网络中，残差连接有助于缓解深层网络训练中的梯度消失问题，而密集连接则聚合了具有多样感受野的中间特征，从而提高了对物体检测任务的性能，为此，PP-YOLOE提出RepResBlock组件，结合了残差连接和密集连接，在不损失精度的情况下降低计算负担，将其用于构建backbone和neck。在采用RepResBlock组件构建backbone时，PP-YOLOE将4个

作者简介：孙艳新（1973—），本科，高级工程师，从事工程测量、地图编制工作，E-mail：379356078@qq.com。

RepResBlock堆叠，并采用残差连接组成stage模块，然后将3个卷积层和4个stage堆叠搭建CSPRepResNet，作为backbone。在推理阶段，RepResBlock可以被重参数化为基本的残差块，简化了网络结构，同时保持了训练阶段的性能，使得网络在推理时更加高效。在搭建neck时，PP-YOLOE采用经典的路径聚合网络PAN结构，PAN通过创建自顶向下和自底向上的路径来实现特征层之间的有效信息流。自顶向下的路径帮助将高层的语义信息传递到低层特征图中，而自底向上的路径则将低层的细节信息传递到高层，这种结构可以有效改善不同尺度特征之间的融合，提高模型对小物体的检测能力。对于头部网络head，PP-YOLOE使用ET-head，使用有效挤压和提取（ESE）层替换传统的层注意力机制，达到简化分类和回归分支的对齐方式。

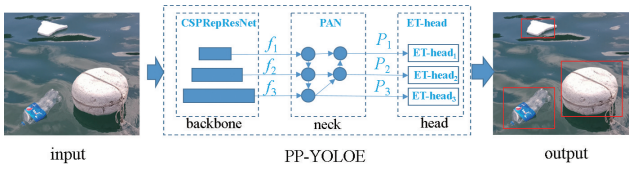


图1　基于PP-YOLOE的水体垃圾检测流程图

对于给定的影像 I，PP-YOLOE将其输入CSPRepResNet得到3个尺度的基础特征向量 $\boldsymbol{f}_1$，$\boldsymbol{f}_2$ 和 $\boldsymbol{f}_3$：

$$\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3=\mathrm{CSPRep\,Res\,Net}(I) \tag{1}$$

在得到 $\boldsymbol{f}_1$，$\boldsymbol{f}_2$ 和 $\boldsymbol{f}_3$ 的基础上，输入路径聚合网络PAN，得到3个多尺度融合后的特征 P_1，P_2 和 P_3：

$$P_1,P_2,P_3=\mathrm{PAN}(\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3) \tag{2}$$

将融合特征分别输入对应尺度的ET-head的3个分支，获得3个尺度上的目标检测结果：

$$\mathrm{cls}_i,\mathrm{bbox}_i=\mathrm{ET-head}_i(P_i) \tag{3}$$

式中，cls_i、bbox_i（$i=1,2,3$）为ET-head第 i 个分支 $\mathrm{ET-head}_i$ 预测的检测框类别和检测框坐标。在得到3个分支的检测框结果之后，利用NMS后处理去掉冗余的检测框。

DETR将Transformer架构应用于物体检测任务，与传统的基于区域的方法（如Faster R-CNN系列）或基于锚点的方法（如YOLO系列、SSD）不同，其实现了真正的端到端目标检测，算法逻辑更为简单。它去除了候选区域提议（region proposal）、锚点（anchors）和非最大抑制（NMS）等步骤，通过Transformer的解码器直接输出一组固定数量的预测，每个预测包括目标的类别和边界框。对于其backbone部分，可选用ResNet-50、ResNet-101等任意卷积神经网络，用于提取基础特征。编码-解码模块中的编码器负责处理backbone（如ResNet-50）提取的图像特征。每个编码器层都包含一个自注意力机制，该机制使模型能够关注到输入图像特征的不同部分，并理解这些部分之间的关系。它通过一系列自主注意力和前馈网络层来捕获图像特征中的全局依赖关系。编码-解码模块中解码器负责生成目标检测的最终预测，包括物体的类别和边界框。在解码器中，设置一组数量为固定值 $\boldsymbol{N}$ 的可学习向量为查询对象，每个向量代表一个潜在的检测对象。解码器接收编码器的输出和查询对象，并将这些查询转换成具体的目标检测预测。DETR的前向网络负责最后检测框的坐标回归和类别判定，主要包含3个感知层，一个隐藏层和预测层。前向网络输出预测框的中心点坐标和长宽以及预测框的类别概率分布。对于输入影像 $I\in R^{3\times H_0\times W_0}$（$H_0$，$W_0$ 分别代表影像的行与列），DETR首先将其输入backbone（本文使用ResNet-50），获得基础特征向量 $\boldsymbol{f}\in R^{2\,048\times H\times W}$（$H=\frac{H_0}{32}$，$W=\frac{W_0}{32}$）：

$$\boldsymbol{f}=\mathrm{Res\,Net-50}(I) \tag{4}$$

在将特征 $\boldsymbol{f}$ 输入编码器之前，首先利用一个卷积层对 $\boldsymbol{f}$ 进行变换，将其特征维度变为 d（本文设置 d 为256），得到 $\boldsymbol{f}_c\in R^{256\times H\times W}$，然后将特征 $\boldsymbol{f}_c$ 变形为二维 $\boldsymbol{f}_c\in R^{256\times H\times W}$ 输入编码器，编码器由多个自注意力transformer堆叠构成，对于第一个自注意力层，其计算为：

$$\mathrm{pos_embeding}=\mathrm{pos_encoding}(H,W) \tag{5}$$

$$q=k=\mathrm{pos_embeding}+\boldsymbol{f}_c \tag{6}$$

$$\boldsymbol{f}_{\mathrm{attn}}=\mathrm{Self_Attn}(q,k,\boldsymbol{f}_c) \tag{7}$$

$$\boldsymbol{f}_{\mathrm{an}}=\mathrm{Add_Norm}(\boldsymbol{f}_c,\boldsymbol{f}_{\mathrm{attn}}) \tag{8}$$

$$\boldsymbol{f}_{\mathrm{ffn}}=FFN(\boldsymbol{f}_{\mathrm{an}}) \tag{9}$$

$$\boldsymbol{f}_{\mathrm{out}}=\mathrm{Add_Norm}(\boldsymbol{f}_{\mathrm{an}},\boldsymbol{f}_{\mathrm{ffn}}) \tag{10}$$

式中，$\mathrm{pos_encoding}(H,W)$ 为对特征 $\boldsymbol{f}_c$ 上的每一个位置进行位置编码，以补充特征的空间信息，pos_embeding为位置编码特征；q、k 为 $\boldsymbol{f}_c$ 附加了空间位置编码的特征；Self_Attn(·)为自注意力函数，在得到注意力特征 $\boldsymbol{f}_{\mathrm{attn}}$ 之后，用残差连接Add_Norm(·)进行特征融合得到融合特征 $\boldsymbol{f}_{\mathrm{an}}$，并在随后添加一个前向网络FFN(·)和残差连接Add_Norm(·)进行特征变换。在编码器层，DETR将多个上述自注意力层叠加。经过编码器特征提取，记录得到的特征为 $\boldsymbol{f}_{\mathrm{encoder}}$，DETR将其输入解码层。解码器层主要包含3个主要

构件，即自注意力层，交叉注意力层和前向网络层。对于第一层的解码器，DETR将初始化的查询向量输入到自注意力层进行查询向量的自我表达，随后输入至交叉注意力层，计算查询向量和图像特征 $\boldsymbol{f}_{\text{encoder}}$ 的相似度，基于 $\boldsymbol{f}_{\text{encoder}}$ 对查询向量进行加权表达，最后将结果输入前向网络得到新的查询向量：

$$q=k=\boldsymbol{f}_{\text{query}}+\boldsymbol{f}_{\text{query_embeding}} \tag{11}$$

$$\boldsymbol{f}_{\text{d_self}}=\text{Self_Attn}(q,k,\boldsymbol{f}_{\text{query}}) \tag{12}$$

$$\boldsymbol{f}_{\text{d_cross}}=\text{Cross_Attn}(\boldsymbol{f}_{\text{encoder}},\boldsymbol{f}_{\text{d_self}}) \tag{13}$$

$$\boldsymbol{f}_{\text{query}}=\text{FFN}(\boldsymbol{f}_{\text{d_cross}}) \tag{14}$$

式中，$\boldsymbol{f}_{\text{query}}$ 为查询向量，对应第一层编码器；$\boldsymbol{f}_{\text{query}}$ 初始化为0向量，为 $\boldsymbol{f}_{\text{query_embeding}}$ 查询向量对应的可学习位置编码；Cross_Attn(·) 为交叉注意力。为简单起见，本文忽略了 Self_Attn(·)，Cross_Attn(·) 和 FFN(·) 后面跟随的残差连接 Add_Norm(·) 函数。DETR 将多个解码器堆叠，迭代获取新的 $\boldsymbol{f}_{\text{query}}$ 输入至目标检测框位置回归分支和分类分支，获得目标框位置和类别：

$$\text{bbox}=\text{FFN}_{\text{bbox}}(\boldsymbol{f}_{\text{query}}) \tag{15}$$

$$\text{cls}=\text{FFN}_{\text{cls}}(\boldsymbol{f}_{\text{query}}) \tag{16}$$

式中，$\text{FFN}_{\text{bbox}}(\cdot)$ 为检测框预测分支；$\text{FFN}_{\text{cls}}(\cdot)$ 为检测框类别预测分支。

针对DETR注意力机制导致的计算量巨大问题，Deformable DETR引入可变性注意力，对于特征图中的某一点，不再计算与其他所有特征点的相似度，而是基于自身的位置预测若干相关特征点位置及其对应的权重获得加权表达，直接舍弃了相似度计算过程。在提升计算速度的同时，考虑多尺度特征融合，用于提升对小目标的检测精度。Deformable DETR与DETR计算过程相似，其将DETR编码器中的自注意力函数 Self_Attn(·) 替换为可变性自注意力函数，解码器中的交叉注意力函数 Cross_Attn(·) 替换为可变性交叉注意力函数。可变性自注意力函数和可变性交叉注意力函数计算逻辑相似。以可变性自注意力函数为例，对于任意特征点，其对应的特征向量为 $\boldsymbol{f}_q$ 和位置为 (q_x,q_y)，则可变性自注意函数基本计算过程如下：

$$\Delta x,\Delta y=\text{Linear}(\boldsymbol{f}_q) \tag{17}$$

$$w=\text{Linear}(\boldsymbol{f}_q) \tag{18}$$

$$\boldsymbol{f}_{\text{near}}=\text{Linear}(\boldsymbol{f}_{q_x+\Delta x,qx+\Delta x}) \tag{19}$$

$$\boldsymbol{f}_{\text{out}}=w\boldsymbol{f}_{\text{near}} \tag{20}$$

式中，Δx，Δy 为由 $\boldsymbol{f}_q$ 预测的临近点偏移量；w 为临近点对应的权重，根据 Δx，Δy 可以得到临近点位置坐标 $(q_x+\Delta x,q_y+\Delta y)$ 及其对应的特征向量 $\boldsymbol{f}_{q_x+\Delta x,qx+\Delta x}$，对 $\boldsymbol{f}_{q_x+\Delta x,qx+\Delta x}$ 做特征变换得到 $\boldsymbol{f}_{\text{near}}$，最后根据临近点权重 w 和临近点特征 $\boldsymbol{f}_{\text{near}}$ 获得加权表达特征 $\boldsymbol{f}_{\text{out}}$。

虽然Deformable DETR等基于Transformer的目标检测模型速度已经得到极大提升，然而与YOLO系列的目标检测器相比，仍较为逊色。针对这一问题，RT-DETR对DTER中的编码器进行了改造，提出了一种混合编码器，分离内部尺度交互和跨尺度融合以便更有效地处理多尺度特征。其中跨尺度融合模块将多尺度融合，内部尺度交互模块仅在单一尺度上的特征进行自注意力特征变换，忽略其他尺度特征，这样可以在保证检测精度的情况下降低计算量。

2 实验结果与分析

本文训练集样本包含1 302张影像，图像大小为2 048×2 048，测试集包含230张影像，共包含12个类别，分别为球形塑料浮标（pla_ba）、塑料瓶（pet_b）、塑料浮标（pla_b）、泡沫板（sty_p）、金属瓶（metal）、玻璃瓶（glass）、其他塑料瓶（o_pla）、绳子（rope）、网（net）、泡沫盒（sty_b）、泡沫浮标（sty_bu），图2给出了每类目标样例。在本文中，对于每一张影像降采样为800×600，训练学习率为0.000 3，训练遍历次数Epoch设置为70，采用配置单张NVIDIA 4090显卡的服务器训练模型。

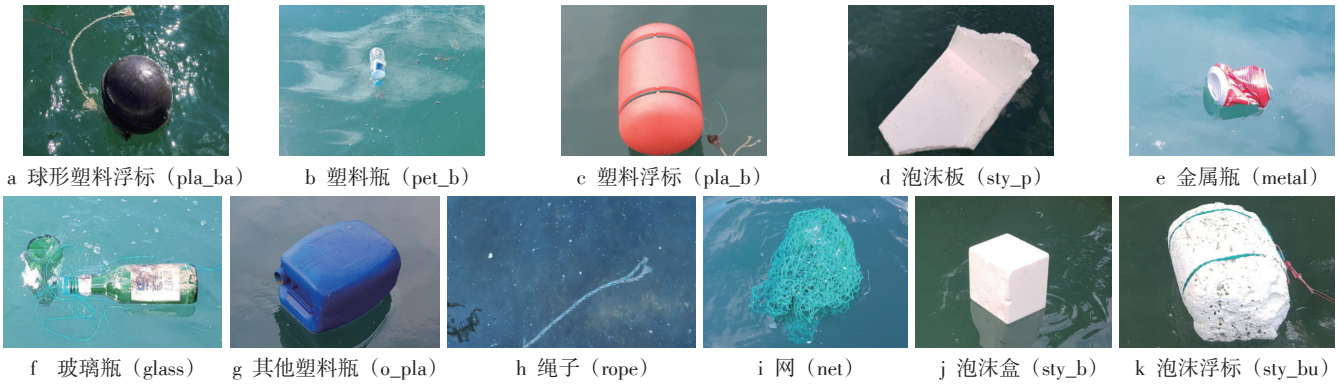
表1给出了PP-YOLOE，DETR、Deformable DETR和RT-DETR在测试数据上的检测精度。

表1 PP-YOLOE，DETR、Deformable DETR和RT-DETR在测试数据集上的检测精度

方法	AP_{50}	glass	pet_b	pla_e	sty_bu	metal
PP-YOLOE	81.3	63.2	79.2	71.7	91.5	78.8
DETR	66.7	35.0	58.3	44.1	67.9	49.6
Deformable DETR	81.8	58.2	73.3	71.8	93.1	69.5
RT-DETR	88.4	71.1	82.9	83.2	93.2	79.8

表1显示，在4种方法中，融合多尺度的RT-DETR具有最优的检测精度，在所有的类别中都取得最好的检测结果。

图3给出了PP-YOLOE，DETR、Deformable DETR和RT-DETR的可视化检测结果，其中红色框为类别识别错误的检测框，其真实类别为其他塑料瓶，从图中可以看出，PP-YOLOE，DETR、Deformable DE-

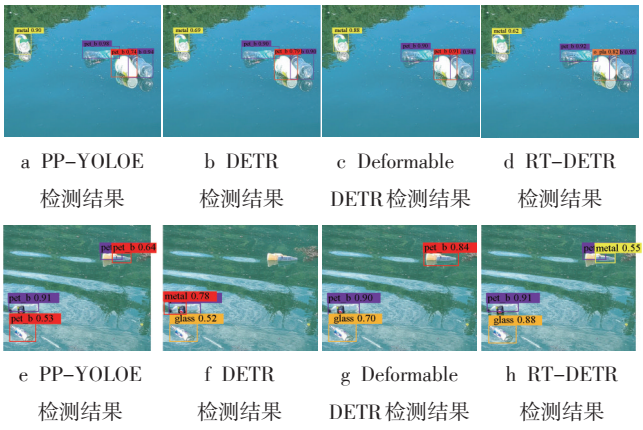


a 球形塑料浮标（pla_ba）　b 塑料瓶（pet_b）　c 塑料浮标（pla_b）　d 泡沫板（sty_p）　e 金属瓶（metal）

f 玻璃瓶（glass）　g 其他塑料瓶（o_pla）　h 绳子（rope）　i 网（net）　j 泡沫盒（sty_b）　k 泡沫浮标（sty_bu）

图2　实验检测影像

TR均将其识别错误，而RT-DETR因为采用了更高效的注意力机制和跨尺度特征融合策略，具有更强的识别能力，能正确判断其类别。



a PP-YOLOE检测结果　b DETR检测结果　c Deformable DETR检测结果　d RT-DETR检测结果

e PP-YOLOE检测结果　f DETR检测结果　g Deformable DETR检测结果　h RT-DETR检测结果

图3　PP-YOLOE、DETR、Deformable DETR和RT-DETR的可视化检测结果

3　结　语

本文探讨了基于Transformer的水体垃圾检测方法，并与传统的基于CNN的方法进行了比较分析。实验结果表明，CNN模型PP-YOLOE在某些情况下显示出良好的性能，较Transformer模型DETR展示出更好的检测结果，但基于Transformer的RT-DETR模型在多尺度特征融合和全局信息处理方面展现出更为显著的优势，取得最好的检测结果。RT-DETR在实验数据集上的AP_{50}达到了0.884，相较于PP-YOLOE有7个百分点的提升。

参考文献

[1] Girshick R. Fast R-CNN[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1 440-1 448

[2] Ren S,He K,Girshick R,et al. Faster R-CNN:Towards Ae Al Time Object Detection with Region Proposal Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015:91-99

[3] He K,Gkioxari G,Dollár P,et al. Mask R-CNN[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017:2 961-2 969

[4] Cai Z,Vasconcelos N. Cascade R-CNN:Delving into High Quality Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6 154-6 162

[5] Chen K,Wang J,Pang J,et al. Hybrid Task Cascade for Instance Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2019:4 974-4 983

[6] Carion N,Massa F,Synnaeve G,et al. End-to-end Object Detection with Transformers[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10 781-10 790

[7] 石露露,廖光忠.改进YOLOv5s的明渠漂浮垃圾实时检测方法[J].计算机技术与发展,2023,33(9):83-90

[8] Yang W,Che J,Zhang L,et al. Research of Garbage Salvagesystem Based on Deep Learning[C]//International Conference on Computer Application and Information Security (ICCAIS 2021),SPIE,2022,12 260:292-298

[9] Shan H Z,Li P,Ning S F. Real-time Detection Method of Surface Floating Objects Based on Deep Learning[C]// 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES),IEEE,2020:174-177

[10] Li X,Tian M,Kong S,et al. A Modified YOLOv3 Detection Method for Vision-based Water Surface Garbage CaptureRobot[J]. International Journal of Advanced Robotic Systems,2020,17(3):17-29