·智能汽车全栈环境感知技术专题·

基于改进DETR算法的行人检测研究

张平　岳粹楠　贺艺斌　王龑龙

（长安大学汽车学院，西安 710018）

【摘要】为解决小目标行人检测精度低的问题，基于DETR算法提出面向复杂场景的行人检测模型。在特征提取环节引入SENet注意力机制，强化通道判别能力；结合图像相对位置编码（iRPE）与绝对位置编码，提升模型对全局上下文与局部细节的感知能力。经自建数据集训练，试验结果表明：所提出方案的行人识别准确率达97.2%，误检率和漏检率仅为1.7%和1.1%，在处理多尺度行人、部分遮挡及低可见度等复杂场景时表现优异，检测效果均显著优于传统方法。

主题词：智能驾驶　行人检测　Transformer　DETR　行人安全

中图分类号：U469.79　　**文献标志码：**A　　**DOI:** 10.19620/j.cnki.1000-3703.20250527

Research on Pedestrian Detection Based on DETR

Zhang Ping, Yue Cuinan, He Yibin, Wang Yanlong

（School of Automobile, Chang'an University, Xi'an 710018）

【Abstract】To address the issue of low detection accuracy for small pedestrian targets, a pedestrian detection model tailored for complex scenes is proposed based on the DETR algorithm. In the feature extraction stage, the SENet attention mechanism is introduced to enhance channel discrimination ability. By combining image Relative Position Encoding (iRPE) with absolute position encoding, the model's perception of global context and local details is improved. After training on a self-built dataset, experimental results show that the pedestrian recognition accuracy of the proposed scheme reaches 97.2%, with false detection rate and missed detection rate of only 1.7% and 1.1%, respectively. It performs excellently in handling complex scenes such as multi-scale pedestrians, partial occlusion, and low visibility, and its detection performance is significantly better than traditional methods.

Key words: Intelligent driving, Pedestrian detection, Transformer, DETR, Pedestrian safety

【引用格式】张平, 岳粹楠, 贺艺斌, 等. 基于改进DETR算法的行人检测研究[J]. 汽车技术, 2026(5): 1-6.

ZHANG P, YUE C N, HE Y B, et al. Research on Pedestrian Detection Based on DETR[J]. Automobile Technology, 2026(5): 1-6.

1　前言

根据《中国统计年鉴》数据，行人交通事故的死亡率明显高于其他交通方式，道路行人安全隐患突出[1]。因此，提升车辆高级辅助驾驶系统的行人检测能力具有重要意义。

当前目标检测算法主要分为双阶段算法与单阶段算法。其中，双阶段算法以Faster R-CNN[2]、SPPNet[3]为代表，先生成候选区域，再进行分类与回归；单阶段检测器如SSD[4]、YOLO[5]等摒弃候选框生成步骤，直接在特征图上进行密集采样预测。基于Transformer[6]的DETR[7]网络被提出，该网络无需人工预先设计锚框和非极大值抑制（Non-Maximum Suppression，NMS）等后处理步骤去除冗余检测框，但其训练时间较长，模型收敛速度较慢。

在模型训练方面，Sun等[8]将全卷积单阶段（Fully Convolutional One-Stage，FCOS）目标检测器和Faster R-CNN的优点与DETR相融合，缩短了模型的训练时间；Meng等[9]提出条件交叉注意力机制，从解码器嵌入中学习条件空间查询并定位物体，比原始DETR收敛速度快6.7倍以上；Zhao等[10]通过动态调整注意力范围，结合一对多监督与一对一监督，大幅提高了DETR训练效率；廖峻霜等[11]将多粒度的注意力机制和空间先验监督引入DETR模型，显著提高检测精度的同时，提升了模型的收敛速度。尽管模型训练的收敛速度得到有效缓解，但DETR系列模型在小目标检测任务中难以捕捉局部细节，检测效果欠佳。

在模型检测方面，王婷婷[12]引入可变形注意力机制并构建能够联系时空上下文关系的行人视频检测模型，提高了模型对小目标行人的检测精度，但不适用于雨雪天气、低光照等场景。吕新尧[13]改进Transformer解码器，通过多数据集渐进式训练增强了模型对细节特征的提取能力，大幅度提高了检测器的性能，但该模型结构复杂且训练周期较长。章东平等[14]融合空间注意力和通道注意力，使用跳跃连接提高了模型对小目标的检测精度，但在层间传递过程中存在丢失小目标信息的问题。李彪等[15]设计了多尺度窗口注意力机制，模型的检测性能与收敛速度均得到提升，但模型参数量显著增加。Ma等[16]将动态可变形卷积和高斯投影通道注意力分别嵌入ResNet-50的低层和高层，但模型结构冗余、检测实时性不足。向南等[17]通过在原DETR加入融合了CSP的DarkNet53模块和特征金字塔网络构建CF-DETR模型，虽然降低了小目标的漏检率和误检率，但因模型的复杂度增加导致检测速度降低。

为解决上述问题，本文通过优化DETR模型，将SENet[18]模块引入特征提取网络，提高网络对特征图中小目标行人的各通道关注程度；引入图像相对位置编码[19]（image Relative Position Encoding，iRPE）使模型关注输入数据的位置信息，提升模型对行人目标的检测精度，并通过自建数据集验证该模型的有效性。

2 DETR算法改进

DETR利用Transformer内部的自注意力机制，整体框架适用于集合预测。该架构主要由CNN骨干网络、Transformer编码器-解码器及预测头构成，模型如图1所示[7]。骨干网络负责特征提取；编码器通过自注意力机制捕获图像中各区域间的长程依赖关系，生成含有全局信息的特征序列；解码器通过目标-特征交互的自注意力，生成预测框并逐步优化，最终输出检测结果；预测头的功能为预测目标类别和边界框。输入图像经过CNN与位置编码处理通过Transformer编码器和解码器架构生成一系列预测框集合。

DETR的主要优势在于其端到端的检测框架和对全局上下文信息的建模能力，在COCO数据集上达到与Faster RCNN相当的性能。但该方法仍存在小目标检测精度不足的问题，因此，本文通过优化DETR框架结构以增强算法对小目标的检测能力，改进后的网络结构如图2所示。

2.1 特征提取网络

特征提取网络主体采用ResNet-50网络，该结构引入残差学习框架（见图3），实现丰富深层特征提取的同时，保证模型训练的稳定性与收敛性，解决深度神经网络训练过程中难以规避的梯度消失与网络退化问题。

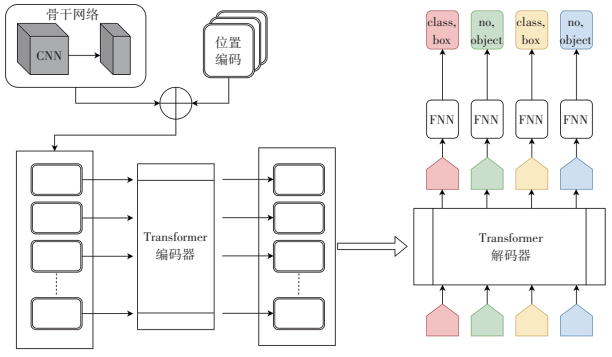


图1 DETR网络结构

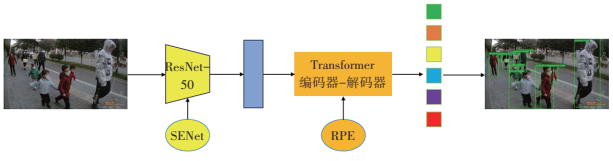


图2 整体框架

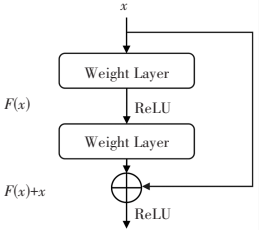


图3 残差结构示意

为了进一步提高特征提取能力，在ResNet-50网络中引入SENet[18]模块。该模块首次通过显式建模通道关系，自适应校准通道权重，显著增强模型的表征能力，强调关键通道抑制无用信息，并揭示各通道与最终特征间的内在关联。

SENet的卓越性能主要源于其压缩-激励（Squeeze and Excitation，SE）模块。该模块分为压缩（Squeeze）、激励（Excitation）和缩放（Scale）3个部分，不仅结构简单，还具备高度的可移植性，可集成到所有网络架构中。通过全局平均池化（Global Average Pooling）将各通道全局空间信息压缩为向量z_c：

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{1}$$

式中：u_c为第c个通道的特征图，H、W分别为空间特征的高度和宽度。

使用全连接（Fully Connected，FC）层和激活函数学习通道权重s_c。将输出的一维特征向量输入全连接层FC1，按比例压缩特征通道维度，经修正线性单元（Rectified Linear Unit，ReLU）激活函数完成非线性

变换后，输入全连接层FC2还原通道至初始维度，最后通过Sigmoid函数归一化处理，输出特征图各通道的权重：

$$\boldsymbol{s}_c = f_{\text{Sigmoid}}(\boldsymbol{W}_2 f_{\text{ReLU}}(\boldsymbol{W}_1 \boldsymbol{z}_c)) \tag{2}$$

式中：$\boldsymbol{W}_1$、$\boldsymbol{W}_2$为全连接层权重。

将学习的通道权重$\boldsymbol{s}_c$与原特征图相乘，得到增强后特征：

$$\tilde{\boldsymbol{x}}_c = F_{\text{scale}}(\boldsymbol{u}_c, \boldsymbol{s}_c) = \boldsymbol{s}_c \boldsymbol{u}_c \tag{3}$$

因此，本文在ResNet-50残差部分嵌入SE模块，如图4所示，虽然会略微增加计算工作量，但骨干网络能够根据重要性对卷积后特征图的各通道进行区分和加权，从而提高图像对小目标行人的提取能力。

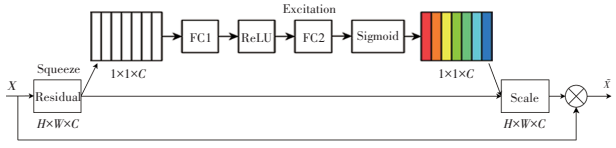


图4　SE-ResNet-50结构

2.2　iRPE位置编码

DETR中Transformer模块的位置编码采用绝对位置编码（Absolute Position Encoding，APE），存在2个局限：一是无法泛化至预设长度之外的序列，二是对平移不变性的感知能力不足。相对位置编码不仅能够适应可变长度的输入序列，且对平移不变模式的归纳偏好更强，更符合实际任务需求。因此，本文结合绝对位置编码与相对位置编码，帮助网络理解图像元素间相对关系，使网络能够精准识别目标位置，从而提升模型性能。

iRPE与自注意力机制的集成可采用偏置和上下文2种模式，选择性能更优的上下文模式接入Transformer自注意力模块的查询向量中，结构如图5所示[19]。该连接方式可保证序列的维度一致性，输入n维序列$x=\{x_1,x_2,\cdots,x_n\}$，输出序列$z=\{z_1,z_2,\cdots,z_n\}$维度不变，引入iRPE的注意力机制：

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j \boldsymbol{V}) \tag{4}$$

$$\boldsymbol{e}_{ij} = \frac{(x_i \boldsymbol{Q})(x_j \boldsymbol{K})^{\mathrm{T}} + \boldsymbol{b}_{ij}}{\sqrt{d_z}} \tag{5}$$

$$\boldsymbol{b}_{ij} = (x_i \boldsymbol{Q}) \boldsymbol{r}_{ij}^{\mathrm{T}} \tag{6}$$

式中：α_{ij}为权重系数，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别为查询矩阵、键矩阵、值矩阵，d_z为矩阵维度，$\boldsymbol{b}_{ij}$为额外权重，$\boldsymbol{r}_{ij}$为二维向量，$\boldsymbol{e}_{ij}$为缩放点积注意力。

为实现不同相对位置间的编码共享，iRPE引入一个多对一分段函数$g(x)$，将相对距离转换到一个有限的整数集合中，模型根据此整数索引对应的$\boldsymbol{r}_{ij}$：

$$g(x) = \begin{cases} [x] & , \ [x] \leqslant \alpha \\ \operatorname{sgn}(x) \times \min(\beta, \chi) & , \ [x] > \alpha \end{cases} \tag{7}$$

$$\chi = \alpha + (\beta - \alpha) \frac{\ln\left(\frac{[x]}{\alpha}\right)}{\ln\left(\frac{\gamma}{\alpha}\right)} \tag{8}$$

$$r_{ij} = P_{I^x(i,j) I^y(i,j)} \tag{9}$$

式中：$[x]$为对x取整，sgn()为符号函数，α、β、γ为分段函数的核心参数。

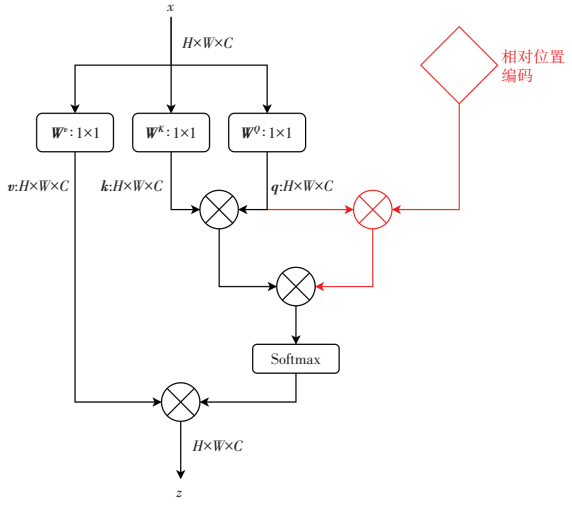


图5　iRPE与自注意力机制结合方式

使用考虑方向性的Product方式计算iRPE中相对距离，确定水平方向距离分量$I^{\tilde{x}}(i,j)$和垂直方向距离分量$I^{\tilde{y}}(i,j)$，通过分段函数将其映射至指定整数：

$$I^{\tilde{x}}(i,j) = g(\tilde{x}_i - \tilde{x}_j) \tag{10}$$

$$I^{\tilde{y}}(i,j) = g(\tilde{y}_i - \tilde{y}_j) \tag{11}$$

通过结果可视化可知（见图6），iRPE对邻近图块赋予更高权重。与传统Transformer的全局注意力机制不同，图像相对位置编码iRPE的引入使DETR在局部细节捕捉方面获得显著增益，从而更好地分析总体和局部特征的相对关系。

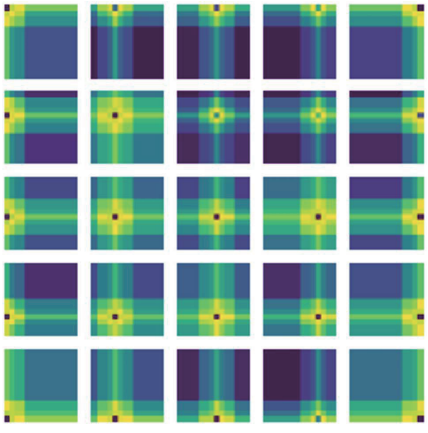


图6　Product方式计算iRPE可视化

3 试验环境及评价

3.1 数据集

本文自建数据集A由2个部分构成：Caltech[20]、INRIA、COCO2017[21]等公开行人数据集的道路行人子集，共计5 000张）；行车记录仪实际采集图像，共计3 214张，补充了儿童类型图像，弥补了当前数据集中儿童数据量不足的缺陷。数据集A共计8 214张图像，按照8∶1∶1划分训练集、验证集和测试集。小型自建数据集B包含600张涵盖多种场景与姿态的行人图像并附加50张背景图像。

3.2 试验环境

模型训练时，使用AMD Ryzen 9 8945H处理器，主频为4.00 GHz，图像处理器为NVIDIA GeForce RTX 4060，显存为8 GB。模型训练框架使用PyTorch深度学习框架，CUDA版本为12.7，Python版本为3.8.10。试验采用迁移学习策略，模型权重继承COCO数据集上预训练的DETR模型。

关键超参数设置：学习率为0.000 1，随机失活率（Dropout rate）为0.1，使用随机梯度下降（Stochastic Gradient Descent，SGD）优化器，Transformer解码器位置编码数为100，多头注意力头数为8，训练共100个迭代轮（Epoch）。

3.3 评价指标

使用准确率（Accuracy）A、误检率（False Positive Rate）A_{FPR}、漏检率（False Negative Rate）A_{FNR}、平均精度（Average Precision，AP）和帧率作为模型检测性能的评估指标。其中，准确率为所有识别正确的样本占总样本的比例，误检率为将负样本识别为正样本的数量占负样本数量的比例，漏检率为将正样本识别为负样本的数量占正样本数量的比例：

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$A_{FPR} = \frac{FP}{TN + FP} \tag{13}$$

$$A_{FNR} = \frac{FN}{TP + FN} \tag{14}$$

式中：TP、TN分别为正确识别的正、负样本数量，FP、FN分别为错误识别的正、负样本数量。

平均检测精度根据精确率-召回率（PR）曲线下面积的计算数值，反映不同置信度下的模型表现，取值范围为(0,1)，其值越大，模型性能越佳。PR曲线遍历不同置信度阈值，获取多组对应的精确率与召回率并拟合生成。其中，AP50和AP75分别表示置信度为0.5和0.75时的平均精度，APS、APM、APL分别为小目标、中目标、大目标的平均精度。

帧率指每秒播放或采集的画面帧数，其数值高低决定画面流畅程度，同时也是视频采集、监控摄像及计算机视觉任务中影响时序信息完整性与算法效果的重要基础参数。

4 试验结果分析

本文模型训练的损失曲线如图7所示：在前60个迭代轮内，模型损失快速下降；当迭代轮超过60个时，损失曲线趋于平缓，仅在小范围内轻微波动，未见明显下行或上行趋势，表明模型已进入收敛状态，训练过程平稳且性能达到最优。

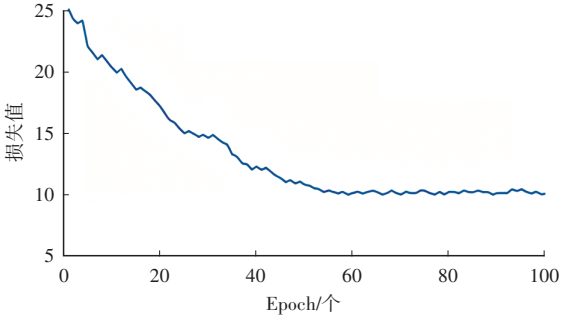


图7 Product方式计算iRPE可视化

以原始DETR为基线模型，通过消融试验进一步分析SE和iRPE模块对于检测结果的影响。同时，将本文模型与现有典型检测算法进行对比分析，以融入特征金字塔网络（Feature Pyramid Network，FPN）的Faster R-CNN与YOLOv8作为两阶段和端到端算法代表。在小型自建数据集B上的检测结果如表2所示。

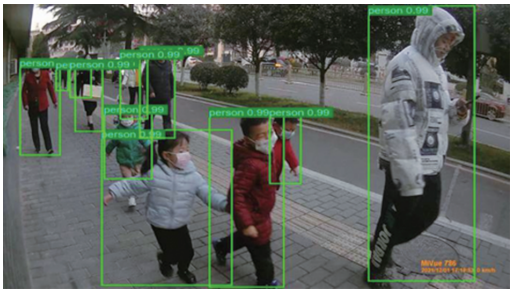
表2 自建数据集B上各模型性能对比 %

模型	准确率	误检率	漏检率
Faster R-CNN	92.0	4.3	3.7
YOLOv8	96.5	2.2	1.4
DETR	94.3	2.5	3.2
DETR+SE	96.5	2.5	1.0
DETR+iRPE	96.8	2.1	1.1
DETR+SE+iRPE	97.2	1.7	1.1

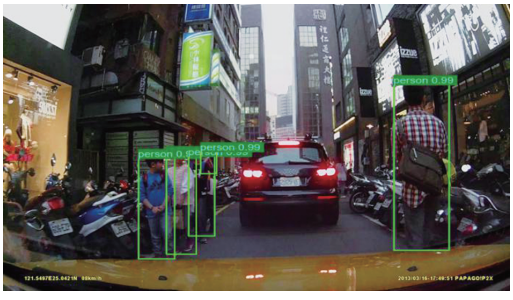
由表2可知，相较于基线模型，引入SE模块后，本文模型的准确率提高2.2百分点，漏检率降低2.2百分点。鉴于数据集中存在较多儿童目标，说明SE模块成功提升了模型对小目标行人的关注度。引入iRPE后，模型准确率提高2.5百分点，误检率、漏检率分别降低0.4百分点和2.1百分点，表明iRPE有助于提升模型对行人目标识别的精度。因此，本文模型的准确率高达

97.2%，同时误检率和漏检率均有降低，具有良好的检测性能。

对本文模型检测结果进行可视化分析，如图8所示，该方法对多尺度行人、遮挡行人均可实现较好的检测效果。



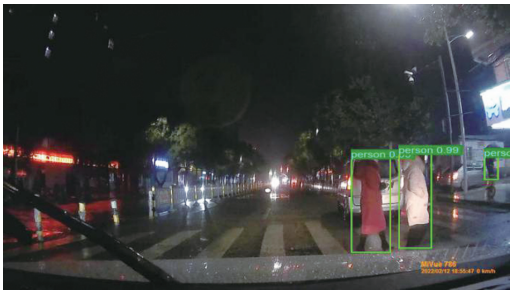
（a）轻微遮挡场景



（b）遮挡行人场景



（c）多尺度行人场景



（d）雨天行人场景

图8　本文方案行人检测结果可视化

为全面评估模型性能，本文在Caltech与自建数据集A上横向对比YOLOv8、Faster R-CNN模型，评估指标包括AP系列及帧率，结果如表3所示。DETR的整体精度与小目标（APS）检测性能提升显著，但检测效率仅与Faster R-CNN相近，主要源于在ResNet-50的残差部分嵌入SE模块，产生了额外的运算开销，故本文模型的实时检测性能略低于YOLOv8。

表3　各模型检测效果和速度对比

模型	AP /%	AP50 /%	AP75 /%	APS /%	APM /%	APL /%	帧率 /帧·s^{-1}
Faster R-CNN	89.22	92.35	85.34	72.31	85.56	95.33	24
YOLOv8	92.41	94.24	87.92	76.06	89.01	97.25	44
本文	93.44	95.21	88.21	77.59	89.24	97.39	28

5　结束语

本文面向复杂场景的行人检测任务，提出基于改进DETR算法。在特征提取网络中，利用SENet的通道注意力机制动态校准特征响应，通过强化网络对特征图通道的注意力，有效提升了模型对小尺寸目标的检测能力。融合相对位置编码与绝对位置编码，使模型能够同时捕捉全局上下文信息与局部细节间的相对关系，从而增强模型对全局信息与局部关系的协同建模能力，在保持全局上下文感知优势的同时，提升模型局部感知能力。未来，考虑将测试场景扩展至雪、雾、沙尘等恶劣天气，以提升模型的鲁棒性与泛化能力，同时，对模型进行轻量化设计，以降低计算开销。

参　考　文　献

[1] 国家统计局. 中国统计年鉴2024[M]. 北京: 中国统计出版社, 2024.
National Bureau of Statistics of China. China Statistical Yearbook 2024[M]. Beijing: China Statistics Press, 2024.

[2] 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.

[3] HE K M, ZHANG X Y, REN S Q, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 21-37.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 779-788.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach California, USA: NIPS, 2017: 6000-6010.

[7] CARION N, MASSA F, SYNNAEVE G, et al. End-to-End Object Detection with Transformers[C]// European Conference on Computer Vision. Glasgow, UK: IEEE, 2020: 213-229.

[8] SUN Z Q, CAO S C, YANG Y M, et al. Rethinking Transformer-Based Set Prediction for Object Detection[C]// International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 3611-3620.

[9] MENG D P, CHEN X K, FAN Z J, et al. Conditional DETR for Fast Training Convergence[C]// International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 3651-3660.

[10] ZHAO C Y, SUN Y F, WANG W H, et al. MS-DETR: Efficient DETR Training with Mixed Supervision[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 17027-17036.

[11] 廖峻霜, 谭钦红. 多粒度空间注意力与空间先验监督的DETR[J]. 计算机科学, 2024, 51(6): 239-246.
LIAO J S, TAN Q H. DETR with Multi-Granularity Spatial Attention and Spatial Prior Supervision[J]. Computer Science, 2024, 51(6): 239-246.

[12] 王婷婷. 面向道路交通场景的行人智能检测方法研究[D]. 贵阳: 贵州大学, 2022.
WANG T T. Research on Intelligent Pedestrian Detection Methods for Road Traffic Scenarios[D]. Guiyang: Guizhou University, 2022.

[13] 吕新尧. 基于视觉Transformer的行人检测算法研究[D]. 济南: 山东师范大学, 2023.
LÜ X Y. Research on Pedestrian Detection Algorithm Based on Visual Transformer[D]. Jinan: Shandong Normal University, 2023.

[14] 章东平, 魏杨悦, 何数技, 等. 特征融合与层间传递: 一种基于Anchor DETR改进的目标检测方法[J]. 图学学报, 2024, 45(5): 968-978.
ZHANG D P, WEI Y Y, HE S J, et al. Feature Fusion and Inter-Layer Transmission: An Improved Object Detection Method Based on Anchor DETR[J]. Journal of Graphics, 2024, 45(5): 968-978.

[15] 李彪, 陈春雨. 基于窗口注意力机制的端到端目标检测算法[J]. 计算, 2025, 52(1): 181-188.
LI B, CHEN C Y. End-to-End Object Detection Algorithm Based on Window Attention Mechanism[J]. Computing Magazine of the CCF, 2025, 52(1): 181-188.

[16] MA C J, ZHUO L, LI J F, et al. Cascade Transformer Decoder Based Occluded Pedestrian Detection with Dynamic Deformable Convolution and Gaussian Projection Channel Attention Mechanism[J]. IEEE Transactions on Multimedia, 2023, 25: 1529-1537.

[17] 向南, 潘传忠, 虞高翔. 融合优化特征提取结构的目标检测算法[J]. 计算机应用, 2022, 42(11): 3558-3563.
XIANG N, PAN C Z, YU G X. Target Detection Algorithm Fused with Optimized Feature Extraction Structure[J]. Journal of Computer Applications, 2022, 42(11): 3558-3563.

[18] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

[19] WU K, PENG H W, CHEN M H, et al. Rethinking and Improving Relative Position Encoding for Vision Transformer[C]// International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 10033-10041.

[20] GRIFFIN G, HOLUB A, PERONA P. Caltech-256 Object Category Dataset[EB/OL]. (2007-03-10) [2025-10-29]. https://authors.library.caltech.edu/records/5sv1j-ytw97.

[21] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[C]// European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 740-755.

（责任编辑　瑞　秋）

修改稿收到日期为2025年10月29日。