2025 年 第 25 卷 第 21 期
2025, 25(21):09018-10
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2406615
引用格式:张森奕, 张雪松, 郭佳琦,等. 基于 Swin Transformer 改进 YOLO 的密集场景行人检测算法[J]. 科学技术与工程, 2025, 25(21): 9018-9027.
Zhang Senyi, Zhang Xuesong, Guo Jiaqi, et al. Improved YOLO based on Swin Transformer for dense scene pedestrian detection algorithm [J]. Science Technology and Engineering, 2025, 25(21): 9018-9027.

基于 Swin Transformer 改进 YOLO 的密集场景行人检测算法

张森奕, 张雪松*, 郭佳琦, 金花, 李光宇
(大连交通大学轨道智能工程学院, 大连 116052)

摘 要 在密集场景中,常常包含众多被遮挡或者小尺度的行人目标。这样的场景对常规的目标检测模型提出了挑战,往往会出现大量的漏检和错检问题。为了解决密集场景中行人检测时出现的高漏报率和误报率问题,提出了一种新的密集场景行人检测框架 ST-YOLO。首先,将 YOLOv5 的骨干网络中的低层小目标检测层融入特征金字塔网络和路径聚合网络结构中,增加了一个检测小目标行人检测层;其次,对 YOLOv5 的颈部网络进行改进,利用基于 Swin Transformer 的多尺度全局信息和卷积神经网络(convolutional neural networks,CNN)所提取的局部信息来构建聚合特征,提高网络的特征提取能力;并且在预测过程中引入了 SIoU(scalable-IoU)损失函数,加快模型的收敛速度和提升检测能力;最后,使用 Soft-NMS(soft non-maximum suppression)代替原非极大值抑制(non-maximum suppression,NMS)算法,减少非最大化抑制阶段误删除检测框问题,降低了检测算法的误报率。在 Wider Person 数据集上的大量实验表明,改进后的 ST-YOLO 算法的精度和 $mAP_{0.5}$ 比目前主流的 YOLOv9 算法分别提升了 5.7% 和 3.6% 。

关键词 行人检测; 密集场景; Swin Transformer; YOLOv5; 特征融合
中图法分类号 TP391.41; **文献标志码** A

Improved YOLO Based on Swin Transformer for Dense Scene Pedestrian Detection Algorithm

ZHANG Sen-yi, ZHANG Xue-song*, GUO Jia-qi, JIN Hua, LI Guang-yu
(School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian 116052, China)

[**Abstract**] In dense scenes, the frequent occurrence of occluded or small-scale pedestrian objects poses significant challenges to traditional object detection models, frequently leading to a high number of missed detections and false positives. In order to solve the problem of high false negative rate and false positive rate in pedestrian detection in such dense scenes, a novel dense scene pedestrian detection framework called ST-YOLO was proposed. Firstly, the low-level small object detection layer in YOLOv5's backbone network was integrated into the feature pyramid network and path aggregation network structure, adding a pedestrian detection layer for detecting small objects. Secondly, the neck network of YOLOv5 was improved by utilizing multi-scale global information based on Swin Transformer and local information extracted by convolutional neural networks (CNN) to construct aggregated features and enhance the network's feature extraction capability. And the SIoU (scalable IoU) loss function was introduced in the prediction process to accelerate the convergence speed of the model and improve detection capability. Finally, Soft NMS (soft non maximum suppression) was used instead of the original non maximum suppression (NMS) algorithm to reduce the problem of mistakenly deleting detection boxes during the non maximum suppression stage and lower the false alarm rate of the detection algorithm. A large number of experiments on the Wide Person dataset have shown that the improved ST-YOLO algorithm has improved accuracy and $mAP_{0.5}$ by 5.7% and 3.6% respectively compared to the current mainstream YOLOv9 algorithm.

[**Keywords**] object detection; dense scenes; Swin Transformer; YOLOv5; feature fusion

收稿日期: 2024-09-03 **修订日期**: 2025-04-14
基金项目: 国家自然科学基金(62276042);辽宁省教育厅科学研究项目(LJKZ0486, LJKMZ20220838)
第一作者: 张森奕(1999—),男,汉族,河南上蔡人,硕士研究生。研究方向:计算机视觉。E-mail:2422272403@qq.com。
* **通信作者**: 张雪松(1980—),男,汉族,湖北襄阳人,博士,副教授。研究方向:计算机视觉,智能优化。E-mail:zhangxuesong@djtu.edu.cn。

在当今社会,随着城市化进程的不断加速,人口密集区域的拥挤场景日益普遍[1]。在这些场景中,行人检测成为了一项至关重要的任务,其在交通管理[2]、安防监控[3]、灾害预警[4]等领域具有重要的应用价值。然而,传统的行人检测算法在面对复杂拥挤场景时往往表现不佳,存在着检测精度低、实时性差等问题。因此,研究密集场景中的行人检测算法,并提高对于这种场景下行人目标的检测精度与效率具有重要意义[5]。

针对密集拥挤场景中的行人检测问题,众多研究者已经对 YOLO(you only look once)系列[6]和快速区域卷积神经网络(faster region convolutional neural networks,Faster R-CNN)[7]等目标检测算法进行了多种优化和改良。高强等[8]基于 YOLOv5 采用加权双向特征金字塔网络改进原始网络中的路径聚合网络,加强多尺度特征的融合能力,提高对行人目标的检测能力。贺宇哲等[9]利用迭代检测(iterative detection,IterDet)对 Faster RCNN 进行改进,有效解决非极大值抑制(non-maximum suppression,NMS)算法及其改进在选择精确度和召回率之间平衡点的难题。文献[10]对交并比(intersection over union,IoU)及其变形的 3 种边界框回归损失函数进行了对比分析,并基于 Alpha-IoU 边界框回归损失函数对模型损失函数进行改进,提出了一种 Focus Multihead Adaptive-YOLO 密集行人检测方法,并通过实验验证了在密集场景中该方法能有效避免被遮挡行人的漏检。王程等[11]将深度可分离卷积代替 YOLOv4 模型中的传统卷积,并在骨干网络的特征融合部分引入通道注意力模块,通过实验验证了该方法能有效提高小目标行人的检测精度。孙杰等[12]在主干网络和颈部网络中加入卷积块注意力模块(convolutional block attention module,CBAM)增强网络对行人重要信息的提取能力,在保证检测精度的同时提升检测速度。文献[13]设计了一种基于全连接的特征尺度均衡模块,通过在特征金字塔的各层级之间构造不同的残差结构来进行特征平衡,辅助模型生成更高质量的特征图,使改进后的模型能够实现对密集场景下的多尺度行人目标进行精准检测。

针对密集场景中物体互相遮挡严重的问题,现提出一种新的密集场景行人检测框架 ST-YOLO,利用 Swin Transformer[14]注意力机制对 YOLOv5s 进行改进优化。改进 YOLOv5 网络模型的颈部网络,添加含有 Swin Transformer 注意力的模块,使其具有更丰富的全局特征信息和目标特征信息,同时抑制背景等无关信息。在 YOLOv5 网络模型的特征融合层增加一个新的预测分支和针对密集场景的小目标检测头。通过增加多尺度感受野,提取图像的全局特征和局部特征,有利于目标推理。利用 SIoU 损失函数替代 YOLOv5s 中的 CIoU(complete intersection over union)损失函数,以加速模型的收敛速度并增强模型的泛化性能力。使用 Soft-NMS 代替 YOLOv5 中原有的 NMS 算法,提高密集场景行人检测的精度。通过在 WiderPerson 数据集上进行消融实验和对比实验,验证所提出方法在密集场景中的有效性和优越性。

1　相关工作

提出的 ST-YOLO 是一种基于 Swin Transformer 注意力机制的密集场景中行人检测方法,通过结合 Swin Transformer 和 YOLO 算法,利用 Swin Transformer 的注意力机制来增强特征表达能力,从而提升在密集场景中的检测能力。

1.1　YOLO 目标检测模型

YOLO 作为单阶段目标检测网络的代表,以其比 Faster RCNN 和 Mask RCNN(mask region-based convolutional neural network)[15]等双阶段检测方法更快的处理速度而闻名。这种方法之所以高效,是因为该方法将图像识别中的分类与定位任务合并处理,通过一步提取特征的方式,同时获得物体的类别和位置信息,显著降低了时间以及计算资源的需求。

YOLOv2[16]采用了改良的锚框机制,在原有基础之上增加了 *K*-means 聚类以生成更贴合特定数据集的锚框;YOLOv3[17]进一步升级,采用 Darknet53 作为骨干网络,增强了网络的特征提取能力;YOLOv4[18]引入了优化版 CSPDarknet53 架构,并结合了特征金字塔网络(feature pyramid network,FPN)的特征融合策略,进一步提升了检测性能。在 YOLO 系列算法的发展进程中,YOLOv5 脱颖而出,引入了一系列革新的改良。该版本根据模型的深度和宽度,提供了一套多样化的复杂度选择,涵盖 YOLOv5n、YOLOv5s、YOLOv5m、YOLOv5l 和 YOLOv5x 共 5 种不同规模的模型以应对不同的应用需求和运算能力。YOLOv7[19]采用了更深的网络结构,并引入了一些新的性能优化技术,如瓶颈注意力模块(bottleneck attention module,BAM)等,从而在精度方面有了提升。YOLOv8 使用了 Anchor-Free 的思想,将跨阶段部分融合模块(cross stage partial network with bottleneck,C3)被替换成卷积融合模块(conv2d with fusion,C2f),实现了模型的轻量化。

YOLOv5 包括 4 个如下主要部分。

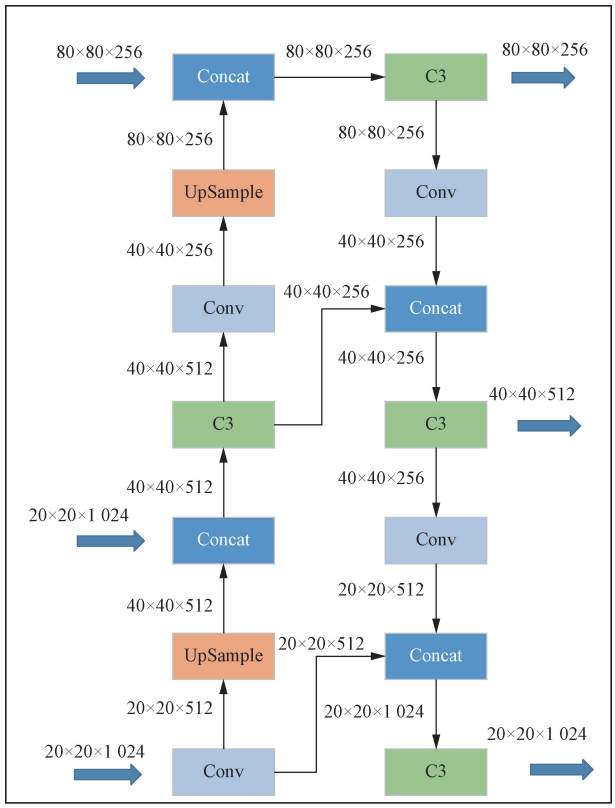
(1)输入层:进行数据预处理。

(2)特征提取骨干网络(Backbone):采用 CSP-Darknet53 加上快速空间金字塔池化(spatial pyramid pooling fast,SPPF)技术以增强特征提取。

(3)特征融合层(Neck):利用路径聚合网络(path aggregation network,PANet)[20]来增强特征之间的联系,YOLOv5 的特征融合层结构。

(4)检测输出层(Head):负责输出识别结果,在 CIoU 边界框损失函数以及 NMS 非极大值抑制的计算约束下输出检测分类结果和置信度。

考虑到在检测实时性和精度之间取得更好的平衡,选择 YOLOv5s 特征融合层作为基础模型,如图 1 所示。



Conv 为卷积块;UpSample 为上采样层;C3 为 3 层卷积层块;Concat 为连接层

图 1 YOLOv5 的特征融合层结构

Fig. 1 The feature fusion layer structure of YOLOv5

1.2 Swin Transformer 注意力机制

近年来,由于 Transformer[21]在自然语言处理领域的强大表现力,在计算机视觉中逐渐崭露头角。Swin Transformer 和 Detection Transformer[22]等基于视觉 Transformer 检测算法陆续出现,其基于全局交互机制的信息提取能力是卷积神经网络所不具备的。Swin Transformer(简称 Swin-T)提出了一项利用滑窗操作的策略,将注意力局限于单个窗口中,使得计算复杂度与输入图片的规模成线性关系,这一举措极大地降低了网络的计算负担。Swin-T 由 4 个基本块组成,这些基本块结构完全相同,如图 2 所示。

在 Swin-T 网络内部,输入图像经过 4 个阶段的处理后,会逐渐产生不同通道数和分辨率的特征图,随后将其输入若干个堆叠的 Swin-T Blocks 中进行处理。在每个 Swin-T Blocks 内部,根据式(1)~式(4)进行计算。

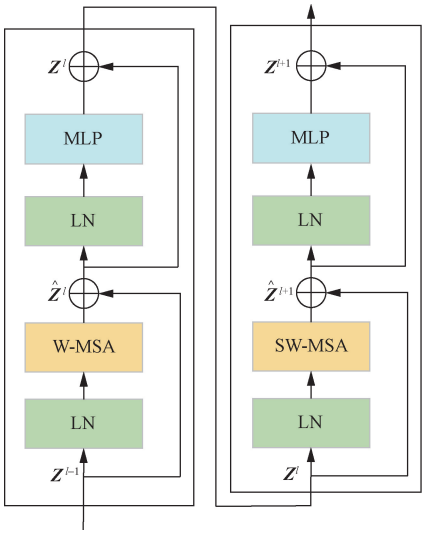
$$\hat{\boldsymbol{Z}}^{l} = \text{W-MSA}[\text{LN}(\boldsymbol{Z}^{l-1})] + \boldsymbol{Z}^{l-1} \tag{1}$$

$$\boldsymbol{Z}^{l} = \text{MLP}[\text{LN}(\hat{\boldsymbol{Z}}^{l})] + \hat{\boldsymbol{Z}}^{l} \tag{2}$$

$$\hat{\boldsymbol{Z}}^{l+1} = \text{SW-MSA}[\text{LN}(\boldsymbol{Z}^{l})] + \boldsymbol{Z}^{l} \tag{3}$$

$$\boldsymbol{Z}^{l+1} = \text{MLP}[\text{LN}(\hat{\boldsymbol{Z}}^{l+1})] + \hat{\boldsymbol{Z}}^{l+1} \tag{4}$$

式中:W-MSA(window multi-head self-attention)为窗口多头自注意力机制;SW-MSA(shifted window multi-head self-attention)为移动窗口多头自注意力机制;在多头自注意力(multi-head self attention,MSA)和多层感知器(multilayer perceptron,MLP)之间使用规范层(layer norm,LN);$\hat{\boldsymbol{Z}}^{l}$ 与 $\boldsymbol{Z}^{l}$ 分别为第 l 个窗口多头自注意力模块和多层感知器模块输出的特征。经过Swin-T对输入图像进行特征提取后,得到了 4 个尺度的特征图,分别表示为 C2、C3、C4、C5,其中输入图像的高和宽分别为 H 和 W。



Z 为特征图;LN 为规范层;W-MSA 为窗口多头自注意力机制;MLP 为多层感知器;SW-MSA 移动窗口多头自注意力机制

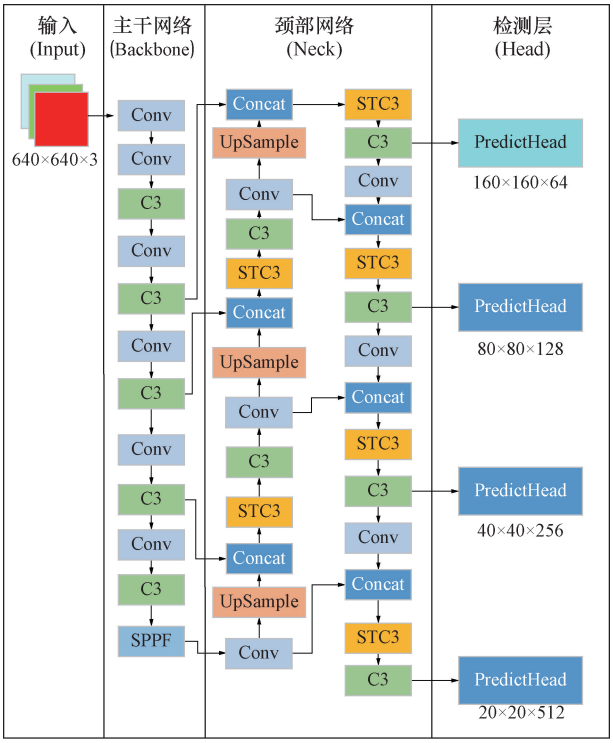
图 2 Swin-T Blocks 结构

Fig. 2 Structure of Swin-T Blocks

2 ST-YOLO 算法

2.1 总体架构

提出的 ST-YOLO 框架如图 3 所示。首先,在 YOLOv5 模型的第 26 层新增大小为 160×160 针对密



Input 为输入;Backbone 为主干网络;Neck 为颈部网络;Head 为头部网络;SPPF 为空间金字塔池化;STC3 为 3 层含有注意力机制的 C3 卷积模块;PredictHead 为检测层

图 3　ST-YOLO 的网络结构

Fig. 3　Network structure of ST-YOLO

集小目标层,新模型分别在第 26 层、第 30 层、第 34 层和第 38 层分别使用大小为 160 × 160、80 × 80、40 × 40 和 20 × 20 的 4 个检测头;其次,对原 YOLOv5 模型的颈部进行修改,通过增加 STC3(swin transformer cross stage partial network with bottleneck)模块,将融合后的特征信息传递给 C3 层,提升网络对特征信息提取的能力;然后,使用 SIoU 提高训练的速度和推理的准确性;最后,引入 Soft-NMS 降低在密集场景下检测行人的漏检率。

2.2　改进的特征融合层

密集拥挤场景中的行人特征在尺度和分辨率都不同,YOLOv5 算法表现不佳。针对这一问题,ST-YOLO 框架中主要对 YOLOv5 的 Neck 进行了改进,如图 4 所示。

在 YOLOv5 算法的特征融合层中增加了 STC3 模块,STC3 模块的结构如图 5 所示。STC3 模块将输入张量“$\boldsymbol{x}$”传递给“self. cv1”进行卷积变换,得到特征图“$\boldsymbol{y}_1$”,同时将同样的输入张量“$\boldsymbol{x}$”传递给“self. cv2”进行另一种通道数不变的卷积变换,得到特征图“$\boldsymbol{y}_2$”,随后将特征图“$\boldsymbol{y}_1$”输入 Swin Transformer 模块,Swin Transformer 模块将输入的特征图“$\boldsymbol{y}_1$”经过 Patch Partition 层进行分割,分割后的数据经过 Linear Embedding 层进行特征映射,随后将特征

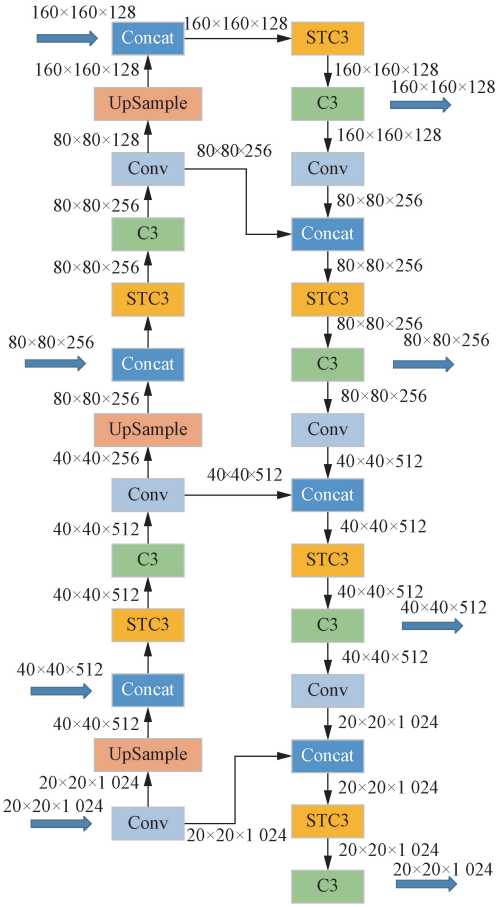
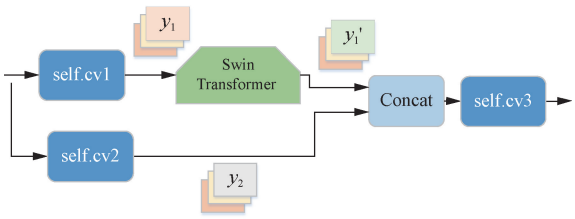


图 4　改进的特征融合层网络结构

Fig. 4　Network structure of the improved feature fusion layer



self. cv 为卷积层;Swin Transformer 为滑动窗口注意力机制模块;Concat 为连接层;y 为特征图

图 5　STC3 模块结构图

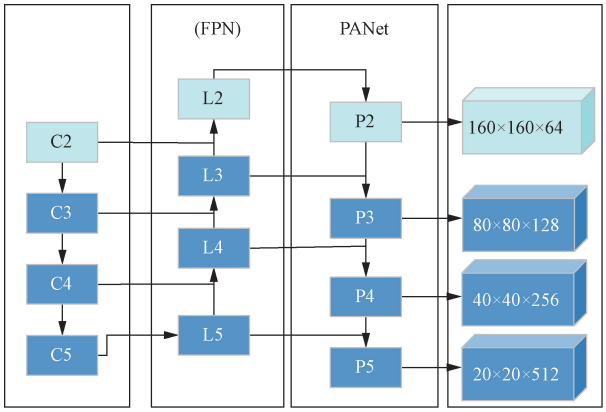
Fig. 5　STC3 module structure diagram

映射后的数据输入 Swin Transformer Block,并与 Linear Embedding 层一起被称为第 1 阶段,与 1 阶段不同,第 2 阶段到第 4 阶段在输入模型前需要进行 Patch Merging 进行下采样,产生分层表示,最终,经过第 4 阶段的数据经过输出模块进行分类。最终得到特征图“$\boldsymbol{y}_1'$”,然后将得到的“$\boldsymbol{y}_1'$”和特征图“$\boldsymbol{y}_2$”在通道维度上进行拼接,拼接后的特征图传递给“self. cv3”进行最终的卷积变换。通过 STC3 模块输出的特征图不仅包含通过卷积操作获取到行人检测目标的局部信息,同时包含通过 Swin Transformer 模块增强后的全局特征,使不同尺度的特征层级之间建

立联系。

2.3 多尺度检测层

YOLOv5 模型构建了 3 个尺度不同的检测层,针对经过 8 倍、16 倍、32 倍降采样的输入进行处理。经过特征金字塔网络和路径聚合网络的处理,这些检测层对应的特征图依次为 P3、P4 和 P5,其尺寸分别为 80×80×128、40×40×256 和 20×20×512。



C 为初级阶段特征图;L 为特征金字塔网络中的特征层;P 为路径聚合网络中的的不同尺度层

图 6 多尺度检测层结构

Fig. 6 Multi-scale detection layer structure

在这些特征图当中,浅层特征图 P3 是从网络的较浅层次提取出来的,其包含了大量的位置信息,因此特别适合于小目标的检测。相比之下,深层特征图 P5 则携带了更多的语义信息,这使其适用于对大目标的检测。为了进一步强化对密集行人检测的特征信息,在颈部网络进行了改进,将骨干网络初级阶段的特征图 C2 与 C3、C4 和 C5 特征进行融合,得到了新的特征图 P2,用于密集行人的探测,如图 6 所示。ST-YOLO 算法添加特征图 C2,这一特征图通过对输入图像进行 4 倍下采样处理来得到,因此它能够保留更多关于小目标的细节信息,通过添加特征图 C2 增强了模型对极小行人目标的检测能力。特征图 C2、C3、C4 和 C5 代表了不同的下采样倍数,具体为输入图像的 4、8、16 以及 32 倍下采样。这些特征图通过特征金字塔网络进行了融合,在融合过程中提供了更加丰富的目标信息,从而有效提高了模型的学习和检测能力。

2.4 优化损失函数

YOLOv5 模型采用了 CIoU 作为边界框回归损失函数,此方法不仅考虑了重叠区域的大小,还包括了中心点距离及宽高比,以此来提高模型预测的准确性。具体公式如下。

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{p^2(b,b^{gt})}{c^2} - av \tag{5}$$

$$L_{\mathrm{CIoU}} = 1 - \mathrm{CIoU} \tag{6}$$

式中:IoU 为预测框与真实框的交并比;b^{gt} 为真实框中心点坐标;b 为预测框中心点坐标;c 为预测框与真实框最小区域的对角线长度;v 为两框长宽比的一致性;$p(\cdot,\cdot)$ 为欧氏距离度量函数;a 为平衡参数。由于 CIoU Loss 未涉及边界框回归方向的调整,会导致模型在训练过程中收敛的速度较慢。

针对 CIoU 存在的不足,ST-YOLO 模型在确定物体位置的损失计算中采用了 SIoU 损失函数,以取代原有的损失函数,SIoU 损失函数引入了一个新的考虑因素,即预测框与真实框之间向量的夹角,对损失函数进行了重新定义。SIoU 损失函数的数学表达式为

$$L_{\mathrm{SIoU}} = 1 - \mathrm{IoU} + \frac{\Delta + \Omega}{2} \tag{7}$$

式(7)中:Ω 为形状函数;Δ 为距离函数。其中距离函数公式为

$$\Delta = \sum_{t=x,y}(1 - \mathrm{e}^{-\gamma p_t}) = 2 - \mathrm{e}^{-\gamma p_x} - \mathrm{e}^{-\gamma p_y} \tag{8}$$

式(8)中:p_x 为预测框和真实框中心点宽度差与最小外接框宽度比值的平方;p_y 为预测框和真实框中心点高度差与最小外接框高度比值的平方;$\gamma = 2 - \Lambda$,SIoU 损失函数中的角度函数公式为

$$\Lambda = 1 - 2\sin^2\left[\arctan(x) - \frac{\pi}{4}\right] \tag{9}$$

式(9)中:x 为预测框和真实框中心点的高度差与其欧氏距离的比值。SIoU 损失函数中的形状函数公式为

$$\Omega = \sum_{t=w,h}(1 - \mathrm{e}^{w_t})^{\theta} = (1 - \mathrm{e}^{w_w})^{\theta} + (1 - \mathrm{e}^{w_h})^{\theta} \tag{10}$$

式(10)中:w_w 为预测框与真实框的宽度差绝对值与两者最大宽度比值;w_h 为预测框与真实框的高度差绝对值与两者最大高度比值;θ 为常量,取值范围为区间[2, 6]。

2.5 改进 NMS

在目标检测算法中,在对物体进行预测时,常常会生成多个重叠的预测框。为了解决这一问题,保留最优的预测框并删除其余的冗余框,常采用的技术是非极大值抑制,表达式为

$$S_i = \begin{cases} S_i, & \mathrm{IoU}(A,B_i) < N_{\mathrm{t}} \\ 0, & \mathrm{IoU}(A,B_i) \geqslant N_{\mathrm{t}} \end{cases} \tag{11}$$

式(11)中:S_i 为检测框置信度;N_{t} 为设定阈值;A 为真实框;B_i 为预测框。NMS 算法的关键在于两个主要测量指标:首先,边框的预测得分决定了它们的优先级,得分高的边框将被首次选取;其次,边框之间的重叠度,通常用 IoU 来度量,是确定是否排除某些边框的依据。IoU 表达式为

$$\mathrm{IOU} = \frac{A \cap B}{A \cup B} \tag{12}$$

由式(12)可知,在物体检测算法中,如果物体的边界框与得分排名最高的边界框存在相交,且它们之间的交并比不小于预定的叠加阈值,那么这个物体的边界框将会被移除。这一过程可能导致即使物体存在,检测算法也无法正确识别。在拥挤场景下的行人检测中,行人较为密集,检测目标之间相互遮挡严重,相近的其他物体的预测框当作自身的冗余检测框,将其他预测框的 score 强制置 0,从而导致对相近的其他物体检测框误删。因此 ST-YOLO 引入 Soft-NMS 算法针对该情况进行改进。Soft-NMS 算法的表达式为

$$S_i = \begin{cases} S_i, & \mathrm{IoU}(A,B_i) < N_t \\ S_i[1 - \mathrm{IoU}(A,B_i)], & \mathrm{IoU}(A,B_i) \geqslant N_t \end{cases} \tag{13}$$

由式(13)可知,当两检测框重合度越高时 S_i 的取值会越小。当 $\mathrm{IoU}(A,B_i) \geqslant N_t$ 时,原 NMS 算法会直接删除该检测框,与原算法不同,Soft-NMS 算法对同一目标识别出的多个检测框进行处理,该算法通过降低重叠检测框的置信度而不是直接消除这些框,有效规避了原方法中删除重叠框可能导致的目标漏检问题,进而增强了目标检测的准确性。

3　实验与实验结果分析

3.1　数据集概述与评估标准

3.1.1　数据集概述

选用 WiderPerson 数据集[23],该数据集是一个密集场景的行人检测 Benchmark 数据集,共包含 5 类标注数据,分别为普通行人、骑自行车的人、身体有遮挡的人、人群和假人。由于实验设定不需要对假人类别进行识别,因此将该类别从数据集中剔除。同时,将正常行人、部分遮挡的人和人群数据整合成一个类别,骑行自行车的人作为单独的一类。WiderPerson 数据集共有 133 820 张图像,每张图像平均有 29.87 个标签,数据集中被标注的行人无重复,保证了每个人物的唯一性。同时,整个数据集在场景和人物方面展现出了丰富的多样性和巨大的变化。一共选取 10 000 张图像,其中训练集 8 000 张,测试集 1 000 张,验证集 1 000 张。

3.1.2　评估标准

选用准确率(precision,P)、召回率(recall,R)和平均精度(average precision,AP)作为评估标准来对模型的性能进行评估[24]。P 表示识别正确的样本数占所有被识别为正样本的比例,R 表示识别正确的样本数占所有正样本的比例,其计算分别如式(14)和式(15)所示。

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{14}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{15}$$

式中:TP 为识别正确的正样本数量;FP 为错误识别为正样本的数量;FN 为识别为负样本但实际上为正样本的数量。以准确率 P 和召回率 R 为横纵坐标,其围成的面积即为该类别的 AP,表达式为

$$I_{\mathrm{AP}} = \int P \mathrm{d}R \tag{16}$$

mAP 是衡量模型性能的指标,代表了多个类别平均精度的均值,表达式为

$$I_{\mathrm{mAP}} = \frac{1}{n}\sum_{i=1}^{n} I_{\mathrm{AP}_i} \tag{17}$$

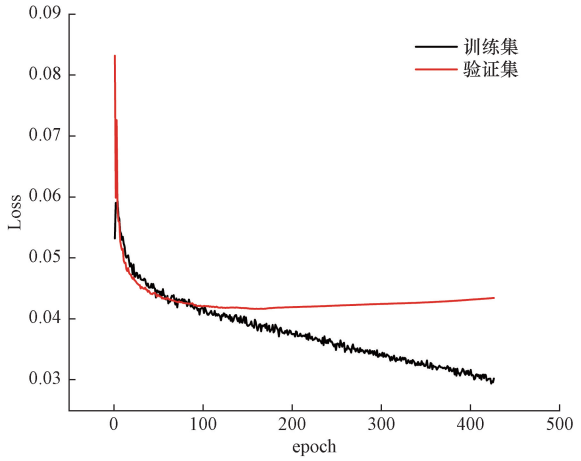
式(17)中:n 为模型需要识别的总类别数。

采用交并比(IoU)标准,阈值定为 0.5,以此来评估模型的平均精度(mAP)及其平均值。普遍而言,较高的 IoU 阈值意味着模型预测的目标与实际目标间有更高程度的重叠,这反映了模型在识别和定位目标方面具备较强的性能。在人群密集场景下,每个行人在图像中所占的像素比例较小,如果采用更高的 IoU 阈值来评估,可能会导致较低的检测精确度,这不利于准确评价行人检测算法在密集场景下的性能。因此,选取 IoU 为 0.5 作为衡量标准。

3.2　实验环境与参数设置

实验所用的硬件包括:具有 80 个核心的 Intel Xeon(R) Gold 5218R CPU,主频为 2.10 GHz,内存容量为 503.4 GB,以及 NVIDIA GeForce RTX3090 GPU。在软件方面,选择 Ubuntu 18.04 操作系统,使用 Python 3.7 作为编程语言,采用了 PyTorch 1.8 作为深度学习框架,并搭配 CUDA 10.2 以利用 GPU 进行加速。

图 7 为训练集和验证集回归损失函数值的变化曲线,在进行训练时,BatchSize 设置为 16,为了有效优化模型的参数,选择了 Adam 优化器。初始学习率被设定为 0.01,动量参数设定为 0.9。在学习率的调整方面,采取了余弦退火策略,该策略会随着训练的进行周期性地调整学习率,有助于平衡全局和局部优化,从而提高模型的性能并促进收敛。实验中所有模型的超参数都设置为默认[25](不一定为最佳参数),并在此设置下进行训练、验证和测试。从图 7 可以看出,训练集损失函数在 425 个 epoch 内呈平滑下降趋势,验证集损失函数曲线在经过约 200 个 epoch 后有一定程度的上升趋势,说明模型有一定程度的过拟合,所以选取 epoch 为 200。



epoch(周期)为模型训练过程中数据集的完整遍历次数;Loss 为损失函数值

图 7 训练集和测试集损失曲线

Fig. 7 Loss curves for training and testing sets

3.3 消融实验及分析

为了验证改进方法对 YOLOv5 模型的影响,对提出的方法进行了消融实验,包括增加检测层、引入 Swin Transformer、SIoU 损失函数以及 Soft-NMS。实验结果如表 1 所示,基线算法为 YOLOv5s,具体内容如下。

(1)增加多尺度小目标检测层的影响。表 1 中,模型 a 为基线模型,未经过任何改动。模型 b 在基线模型 a 的基础上新增了一层针对密集场景中小目标的的检测层。根据表 1 中实验结果表明,相较于 a 模型,b 模型的 $mAP_{0.5}$ 增加了 3.7%,说明增加小目标检测层对于该场景下的行人检测是有效可行的。

(2)改进 Neck 网络的影响。c 模型是在 b 模型的基础上通过 STC3 模块为模型引入了 Swin Transformer 注意力机制,该机制可以捕获全局信息的特征,对于拥挤密集的人群在提取局部特征的同时,也提取到全局特征。由表 4 可知,c 模型相对于 b 模型的 mAP0.5 增加了 2.1%,说明引入 Swin Transformer 主干网络有利于在拥挤的场景下对行人进行检测。

(3)更换 SIoU 损失函数的影响。d 模型是在 c 模型的基础上,将基准模型的 CIoU 损失函数更换为了 SIoU 损失函数,由表 1 可知,d 模型相较于 c 模型的 mAP0.5 提升了 1.8%,说明更换 SIoU 损失函数对模型的检测性能的提升是有利的。

(4)更换 Soft-NMS 的影响。e 模型是在 d 模型的基础上,增加了 Soft-NMS 算法,相较于 d 模型中的原 NMS 算法,Soft-NMS 算法避免了因密集场景中重合问题删除检测框造成的漏检问题。由表 1 可知,e 模型的 mAP0.5 相较于 d 模型提升了 2.1%,说明了 Soft-NMS 增强了模型对行人检测的能力。

综上所述,通过实验评估了每项改进方法对整体模型性能的影响。提出的 ST-YOLO 模型相较于基线模型 YOLOv5 的 mAP0.5 提升了 9.7%。结果表明,每个单独的改进方法都对模型性能有正向贡献,同时改进方法的组合应用进一步提高了算法的性能。

表 1 ST-YOLO 模型消融实验

Table 1 ST-YOLO model ablation experimental results

序号	基线	新增检测层	Swin Transformer	SIoU	Soft-NMS	准确率/%	召回率/%	$mAP_{0.5}$/%
a	√	—	—	—	—	73.5	61.2	65.4
b	√	√	—	—	—	74.5	64.0	69.1
c	√	√	√	—	—	79.4	63.2	70.5
d	√	√	√	√	—	80.2	65.4	71.8
e	√	√	√	√	√	80.4	65.6	75.1

注:a 为原 YOLOv5 模型;b 为在原模型加入了新的检测层;c 为在原模型的基础上加入检测层和 Swin Transformer 注意力机制;d 为在 c 模型的基础上加入了 SIoU;e 为在 d 模型基础上加入 Soft-NMS;Swin Transformer 为添加滑窗注意力机制;SIoU 为添加 SIoU 损失函数;Soft-NMS 为添加非极大值抑制;$mAP_{0.5}$ 为 IoU 阈值为 0.5 的平均精度均值;“√”表示在模型中使用了此方法,“—”表示在模型中未使用该方法。

3.4 对比实验与分析

为了对比模型改进后的效果,选用经过训练的模型在 WiderPerson 数据集上进行测试,改进前后的模型性能评估结果如表 2 所示。因为选取数据集中“riders”类别的图像数量少于“pedestrians”类别的图像,所以算法对“pedestrians”类别的精度整体高于“riders”类别。由表 2 可以看出,改进后的算法在检测两种类别时,检测的精度都高于原算法的检测精度,相较于原始 YOLOv5,改进后的 ST-YOLO 模型算法的精度、召回率、mAP0.5 平均值分别提升了 6.9%、4.4%、9.7%。为进一步验证改进后算法的性能,将其与 Faster-RCNN、YOLOv3、YOLOv5、YOLOv7、YOLOv8、YOLOv9 和 YOLOv10 进行了基于同一数据集的比较实验。

所有模型的输入图像尺寸均为 640×640,并保持超参数及训练参数设置的一致性。WiderPerson 数据集上的实验结果如表 3 所示,改进后的算法相较于上述算法的 $mAP_{0.5}$ 分别提升了 13.65%、4.6%、

表 2 改进前后模型性能评估结果

Table 2 Model performance evaluation results before and after improvement

类别	准确率/%		召回率/%		$mAP_{0.5}$/%	
	YOLOv5	ST-YOLO	YOLOv5	ST-YOLO	YOLOv5	ST-YOLO
pedestrians	80.1	87.3	81.9	83.6	80.5	88.6
riders	67.0	73.4	40.5	47.6	50.3	61.7
平均值	73.5	80.4	61.2	65.6	65.4	75.1

注:pedestrians 为行人类别;riders 为骑自行车的人类别。

表 3　多算法性能对比结果

Table 3　Multi-algorithm performance comparison results

算法	准确率/%	召回率/%	$mAP_{0.5}$/%
Faster RCNN	—	—	61.5
YOLOv3	74.7	63.6	70.5
YOLOv5	73.5	61.2	65.4
YOLOv7	77.3	62.6	70.3
YOLOv8	74.4	60.9	66.3
YOLOv9	74.7	65.5	71.5
ST-YOLO	80.4	65.4	75.1

9.7%、4.8%和8.8%,相较于最新的 YOLO 系列算法 YOLOv9 和 YOLOv10 分别高出 3.6%和 10.4%,从实验结果可以看出改进后算法对于密集行人目标的检测具有显著效果。

3.5　实验效果对比

为了更加直观地看出改进后的模型和基准模型之间的检测差距,使用 WiderPerson 数据集中具有代表性并且检测较为困难的行人图片进行测试,部分检测结果如图 8 所示。图 8(a)、图 8(d)和图 8(g)中行人密集程度依次增加,可以有效观察出原基线算法和改进后的算法在不同密集程度场景中的性能表现对比。

由图 8(b)可以看出 YOLOv5 原算法将两行人后的行人目标忽视,从而产生漏检现象,由图 8(c)可知改进后的模型检测出原模型漏检的行人目标;图 8(d)相较于图 8(a)背景更为丰富,图 8(e)所示 YOLOv5 漏检被广告牌遮挡的白色上衣行人目标,由图 8(f)可以观察到,ST-YOLO 成功检测到了该行人目标,降低了漏检现象的发生;在检测图像 3 这种行人更加密集、背景更加复杂时,由图 8(h)可知,原模型将左侧的黑色广告牌错检成了行人,同时将左边白色上衣行人和白色行人中的阴影部分错检为行人,图 8(i)所示改进后的 ST-YOLO 算法展现出较好的抗干扰能力,整体降低了错检率。由上述多种拥挤密集场景下,改进后的 ST-YOLO算法在密集场景下对行人检测效果更好。

(a) 图像1　(b) YOLOv5检测结果1　(c) ST-YOLO检测结果1

(d) 图像2　(e) YOLOv5检测结果2　(f) ST-YOLO检测结果2

(g) 图像3　(h) YOLOv5检测结果3　(i) ST-YOLO检测结果3

图 8　算法改进前后的检测效果图

Fig. 8　Detection effect before and after algorithm improvement

4 结论

针对密集拥挤场景中难以准确进行行人检测的问题,提出了一种基于改进 YOLOv5 的密集场景行人检测模型。通过对该模型进行一系列实验可以得出以下结论。

(1)在模型中增加了针对拥挤场景的目标检测层,让改进后的模型在特定的高密度区域对细粒度信息更加敏感,从而提升检测性能。

(2)针对密集场景行人目标特征信息复杂这一问题对 Neck 网络进行改进,通过 STC3 模块将获取到的特征信息进行融合,有效提升了网络提取特征信息的能力。

(3)该模型采用 SIoU 损失函数来替代 CIoU 损失,这样的改进让模型在训练过程中更加重视预测框与真实框之间的角度对齐,有助于提升在密集场景中对行人目标的定位精度。

(4)向模型引入 Soft-NMS,减少模型在拥挤场景行人密集造成检测框误删而产生的漏检误检现象。

改进后的模型的准确度、召回率和 $mAP_{0.5}$ 相较于原始模型分别提高了 6.9%、4.4% 和 9.7%,$mAP_{0.5}$ 高于最新的 YOLO 系列算法 YOLOv9 算法。但本文模型仍有改进空间,由于复杂拥挤场景中的有些行人过于密集且检测目标过小,对于过小的行人目标,该模型仍然存在一些漏检问题。未来的研究工作计划专注于进一步改进这一方面,比如优化模型结构,在确保检测精度的前提下,可能会运用模型剪枝等技术来减少模型的参数量,使得网络结构更加简洁、高效,达到轻量化的目的,以适应更多的应用场景并提高实用性。

参 考 文 献

[1] 王宏, 韩晨, 袁伯阳, 等. 基于改进 YOLOv5 的拥挤行人检测算法[J]. 科学技术与工程, 2023, 23(27): 11730-11738.
Wang Hong, Han Chen, Yuan Boyang, et al. Crowded pedestrian detection algorithm based on improved YOLOv5[J]. Science Technology and Engineering, 2023, 23(27): 11730-11738.

[2] 高昕, 甄国涌, 储成群, 等. 基于改进 YOLOv5 的自动驾驶目标检测方法[J]. 科学技术与工程, 2024, 24(16): 6757-6765.
Gao Xin, Zhen Guoyong, Chu Chengqun, et al. Autonomous driving target detection method based on improved YOLOv5[J]. Science Technology and Engineering, 2024, 24(16): 6757-6765.

[3] 刘春雷, 李志华, 王超, 等. 一种融合 RepVGG 和 YOLOv5 的行人检测方法[J]. 科学技术与工程, 2023, 23(7): 2945-2951.
Liu Chunlei, Li Zhihua, Wang Chao, et al. A pedestrian detection method of integrated RepVGG and YOLOv5[J]. Science Technology and Engineering, 2023, 23(7): 2945-2951.

[4] 何明杰, 刘德方, 张猛, 等. 融合 YOLOX 和 ASFF 的高原山地灾害检测模型[J]. 防灾减灾工程学报, 2023, 43(6): 1215-1223.
He Mingjie, Liu Defang, Zhang Meng, et al. A plateau mountain disaster detection model by Integrating YOLOX and ASFF[J]. Journal of Disaster Prevention and Mitigation Engineering, 2023, 43(6): 1215-1223.

[5] 刘城道, 何涛, 景嘉宝. 基于双目视觉的部分遮挡行人检测算法[J]. 科学技术与工程, 2024, 24(13): 5465-5472.
Liu Chengxiao, He Tao, Jing Jiabao. Pedestrian detection method in front of vehicle based on binocular vision[J]. Science Technology and Engineering, 2024, 24(13): 5465-5472.

[6] Redmon J, Divvala K S, Girshick B R, et al. You only look once: unified, real-time object detection[C]//Computer Vision & Pattern Recognition. Las Vegas: IEEE, 2016: 779-788.

[7] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[8] 高强, 唐福兴, 李栋, 等. 基于改进 YOLOv5 的密集场景行人检测方法研究[J]. 国外电子测量技术, 2023, 42(4): 125-130.
Gao Qiang, Tang Fuxing, Li Dong, et al. Research on pedestrian detection method in dense scenes based on improved YOLOv5[J]. Foreign Electronic Measurement Technology, 2023, 42(4): 125-130.

[9] 贺宇哲, 徐光美, 何宁, 等. 迭代 Faster R-CNN 的密集行人检测[J]. 计算机工程与应用, 2023, 59(21): 214-221.
He Yuzhe, Xu Guangmei, He Ning, et al. Iterative Faster R-CNN for dense pedestrian detection[J]. Computer Engineering and Applications, 2023, 59(21): 214-221.

[10] 蒋博文. 基于改进 YOLOv5 的密集行人检测方法[D]. 合肥: 安徽理工大学, 2022.
Jiang Bowen. Dense pedestrian detection method based on improved YOLOv5[D]. Hefei: Anhui University of Science and Technology, 2022.

[11] 王程, 刘元盛, 刘圣杰. 基于改进 YOLOv4 的小目标行人检测算法[J]. 计算机工程, 2023, 49(2): 296-302, 313.
Wang Cheng, Liu Yuansheng, Liu Shengjie. Small target pedestrian detection algorithm based on improved YOLOv4[J]. Engineering with Computers, 2023, 49(2): 296-302, 313.

[12] 孙杰, 吴绍鑫, 王学军, 等. 基于 Sophon SC5 + 芯片构架的行人搜索算法与优化[J]. 计算机应用, 2023, 43(3): 744-751.
Sun Jie, Wu Shaoxin, Wang Xuejun, et al. Pedestrian search algorithm and optimization based on Sophon SC5 + chip architecture[J]. Journal of Computer Applications, 2023, 43(3): 744-751.

[13] 于范, 张菁. 滑窗注意力多尺度均衡的密集行人检测算法[J]. 计算机科学与探索, 2024, 18(5): 1286-1300.
Yu Fan, Zhang Jing. Dense pedestrian detection algorithm based on sliding window attention multi-scale equalization[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(5): 1286-1300.

[14] 石欣, 卢灏, 秦鹏杰, 等. 一种远距离行人小目标检测方法[J]. 仪器仪表学报, 2022, 43(5): 136-146.
Shi Xin, Lu Hao, Qin Pengjie, et al. A method for detecting small pedestrian targets at long distances[J]. Instrumentation,

2022, 43(5): 136-146.

[15] Kaiming H E, Gkioxari G, Dollár P. Mask R-CNN[C]// Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2980-2988.

[16] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7263-7271.

[17] Redmon J, Farhadi A. YOLOv3: an incremental improvement [J]. ArXiv Preprint ArXiv, 2018: 1804.02767.

[18] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. ArXiv Preprint ArXiv, 2020: 2004.10934.

[19] 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. Vancouver: IEEE, 2023: 7464-7475.

[20] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8759-8768.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. ArXiv, 2017: 1706.03762.

[22] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 10012-10022.

[23] Zhang S, Xie Y, Wan J, et al. Widerperson: a diverse dataset for dense pedestrian detection in the wild[J]. IEEE Transactions on Multimedia, 2019, 22(2): 380-393.

[24] 李佳东, 张丹普, 范亚琼, 等. 基于改进 YOLOv5 的轻量级船舶目标检测算法[J]. 计算机应用, 2023, 43(3): 923-929.
Li Jiadong, Zhang Danpu, Fan Yaqiong, et al. Lightweight ship object detection algorithm based on improved YOLOv5[J]. Journal of Computer Applications, 2023, 43(3): 923-929.

[25] 秦强强, 廖俊国, 周弋荀. 基于多分支混合注意力的小目标检测算法[J]. 计算机应用, 2023, 43(11): 3579-3586.
Qin Qiangqiang, Liao Junguo, Zhou Yigou. Small object detection algorithm based on multi branch mixed attention[J]. Journal of Computer Applications, 2023, 43(11): 3579-3586.