

DOI:10.19344/j.cnki.issn1671-5276.2025.06.034

YOLOv5s 的改进算法及其在道路车辆行人检测的应用研究

庄曙东^a, 朱文斌^a, 成先明^{a,b}, 田成东^a, 张广杰^a

(河海大学 a. 机电工程学院; b. 材料科学与工程学院, 江苏 常州 213022)

摘要:针对汽车智能化中道路车辆行人检测困难、远景小目标以及目标多尺度等技术难点,基于 YOLOv5s 网络提出一种道路背景下多尺度车辆行人检测识别模型 YOLOv5s-CD,利用 ConvNeXt 网络的先验知识以及 PConv 卷积的性能优势来改进原主干网络,增强特征提取性能;在能够自适应特征变化的可变形卷积中融入通道注意力机制,给予重要空间区域与通道更多的关注,利用浅层特征层构建多尺度级联模块,提升对小目标特征的提取性能。实验表明:在 BDD100K 数据集当中,改进的 YOLOv5s-CD 模型识别查准率达到 96.4%,查全率达到 87%,平均精度达到了 75%,与原网络以及其他主流网络模型相比,识别精度和查全率都有提高。

关键词:车辆行人检测; YOLOv5s; ConvNext; 可变形卷积

中图分类号:TP274 **文献标志码:**B **文章编号:**1671-5276(2025)06-0184-06

Research on Improved Algorithm of YOLOv5s and Its Application in Road Vehicle Pedestrian Detection

ZHUANG Shudong^a, ZHU Wenbin^a, CHENG Xianming^{a,b}, TIAN Chengdong^a, ZHANG Guangjie^a

(a. College of Mechanical and Electrical Engineering;

b. College of Materials Science and Engineering, Hohai University, Changzhou 213022, China)

Abstract: To overcome the technical difficulties such as the difficulty of pedestrian detection, long-term small target and multi-scale target in the intelligent vehicle of automobiles, a multi-scale vehicle pedestrian detection and recognition model YOLOv5s-CD in the road background was proposed based on the YOLOv5s network, which used the prior knowledge of the ConvNeXt network and the performance advantages of PConv convolution to improve the original backbone network and enhance the feature extraction performance. Channel weights are integrated into the deformable convolution that can adapt to feature changes, a multi-scale cascade module is constructed based on more attention to important spatial regions and channels, and the shallow feature layer is used to improve the extraction performance of small target features. Experiments show that in the BDD100K dataset, the recognition accuracy of the improved YOLOv5s-CD model reaches 96.4%, the recall rate reaches 87%, with average accuracy as 75%. Compared with the original network and other mainstream network models, recognition accuracy and recall rate are both improved.

Keywords: vehicle pedestrian detection; YOLOv5s; ConvNext; variability convolution

0 引言

精准高效的车辆行人检测技术是智能驾驶环境感知当中重要的任务之一,伴随着传感器和计算机技术的快速发展,车辆行人检测技术也在飞速进步之中^[1]。在复杂的道路环境当中存在着遮挡物、多变的天气情况以及车辆行人尺度变化等问题,从而加大了道路目标检测难度,容易产生误检和漏检现象,并且也无法适应多变的道路环境^[2]。近年来,随着深度学习在目标检测领域中的不断发展,将神经网络技术运用于道路车辆行人的检测之中,使得道路车辆行人检测技术得到

了重大的突破。通过利用卷积神经网络来自主学习目标特征,相比较于通过手动提取特征和利用滑动窗口遍历的传统检测技术,拥有着更高的鲁棒性、更快的检测速度和更精准的检测精度。基于深度学习的目标检测算法又分为基于候选区域的两阶段算法和基于回归的单阶段算法^[3]。前者继承了传统目标检测算法的滑动窗口思想,选取多个候选框区域来对目标进行检测,检测精度高,定位好,但是检测速度慢;后者通过将整张图像输入到神经网络当中,直接在输出层回归 bounding box 的位置以及所属类别,提高了目标检测的速度,但是精度有所下降。近年来,在深度学习的目

第一作者简介:庄曙东(1970—),男,江苏常州人,高级实验师,产业教授,博士,研究方向为智能制造与视觉识别, zhsd1970@126.com。

标检测技术当中,已经有众多的学者投入其研究当中,使得该技术在工业检测、农业检测^[4-5]以及道路车辆行人检测得到广泛应用,并取得了众多的研究成果。

目前,基于深度学习的道路车辆行人检测主要研究方向为提升道路车辆行人特征的提取以及提升多尺度目标检测尤其是小目标的检测^[6]。文献^[4]利用 DenseNet 密集连接思想将 YOLOv3 网络当中的深层残差块进行密集连接改进,保留关键特征信息,并引入通道注意力机制对重要通道给予关注,以此来提升网络对道路车辆目标特征的提取。孙备等^[7]通过多尺度融合不同特征层来缓解特征的不变性与等变性,以提升浅层特征层的语义信息;引入空间注意力机制提升了神经网络对小目标特征的提取能力,提高了小目标检测性能。石欣等^[8]在 YOLOv4 基础上引入额外的浅层输出层,并引入引力自适应权重,使拥有小目标信息的特征输出层获得更多的关注,增强更多的小目标细粒度特征,以此来提升小目标行人检测能力。于方程等^[9]通过替换 ResNet-18 作为特征提取主干网络来增强原 CenterNet 网络的特征提取能力,并结合分裂注意力机制来提升对小目标的检测性能;在实际的复杂道路当中,车辆行人受到遮挡物或者同类型物体遮挡以及远景图像中所占像素比例小导致目标检测精度低。此外,在实际道路车辆行人检测当中对于实时性要求较高,网络算法的部署需要借助嵌入式设备为载体,然而车载嵌入式设备算力低,对于高计算量、大模型的算法网络实施部署困难。因此,在提升网络性能的同时需要考虑网络的复杂度,使其能够在算力有限的车载嵌入式设备当中部署。

本文基于 YOLOv5 系列的轻量级网络 YOLOv5s,通过替换原主干网络 CSPDarknet 为 ConvNeXt 网络来提升网络的特征提取能力(包括小目标特征),并且使用 PConv^[10]来替换当中的 DWConv^[11]以减少高维信息的损失;其次,在 PaNet 的由深到浅的上采样路径当中引入注意力可变形卷积来自适应变化卷积核形状加强小目标特征信息,以及通过自适应特征融合将浅层特征层中的小目标特征融合进行预测输出。最后经过实验对比验证了本文算法的合理性。

1 网络模型

在当前流行的轻量级目标检测算法当中, YOLOv5s 检测速度快,检测效率高、易于部署,因

此可以在多个领域当中进行使用。YOLOv5s 由特征提取主干(Backbone)、特征增强部分(Neck)以及检测头(Head)3个部分组成。模型结构如图1所示。其特征提取主干使用的是 CSPDarkNet 网络,其中 N 为模块堆叠层数,其 Neck 部分使用的是 PaNet 结构的网络,将特征主干网络的最后 3 个输出特征层进行多尺度融合,对大中尺寸的目标能够进行有效检测。但对于复杂道路背景下的远景车辆行人小目标,其检测性能不佳。因此,通过对特征提取与特征融合角度进行改进,提升小目标检测性能。

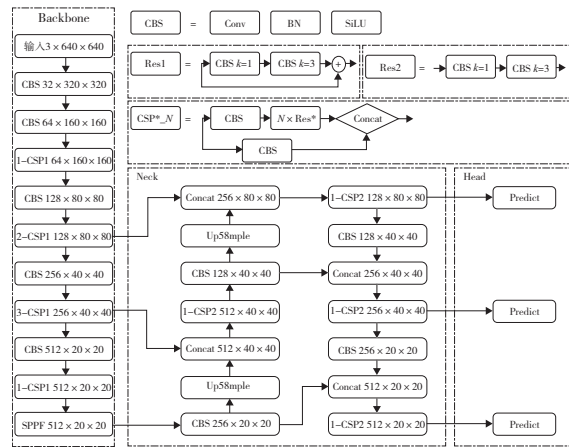


图1 YOLOv5s 网络结构图

2 网络改进改进

2.1 主干改进

YOLOv5s 网络使用 CSPDarkNet 作为主干网络进行特征提取,特征图尺寸以及目标特征信息像素随网络深度加深而减少,小目标信息提取不佳,使用 ConvNeXt 网络替换 CSPDarkNet 主干网络来提升特征提取能力。引入 PConv 弥补 DWConv 空间信息提取能力不足的问题,并命名为 FConvNeXt,其网络结构如图2所示。

图2中,将 FConvNeXt block 的堆叠层数设置为(1,2,3,1),Down 模块使用了卷积核大小为3和步长都为2和 Padding 为1的常规卷积来进行下采样,并且在此之前使用 BN 层让模型训练更加稳定;FConvNeXt Block 模块采用借鉴 Swin Transformer^[12]中的 block 模块的倒残差结构,能够在减少高维信息损失的同时减少小目标信息的丢失;首先将输入特征图的通道数 split 分割为两条路径,通道比为 1:3,前者使用大卷积核来提升感受野并与后者进行 Concat 操作,之后使用两个 1x1 的卷积进行升维和降维。

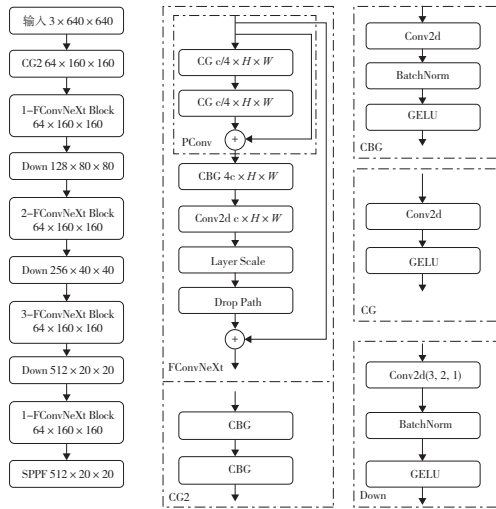


图2 FConvNeXt网络结构图

2.2 注意力可变形卷积

注意力可变形卷积由两个部分组成,如图3所示。图3(a)是调制部分^[13],在常规卷积的基础上使用采样点偏移来应对目标几何变化。通过给采样点施加重重来获取重要输入特征,其表达式如式(1)所示。

$$Y(p) = \sum_{n=1}^N w_n \cdot X(p + p_n + \Delta p_n) \cdot \Delta m_n \quad (1)$$

式中: $Y(p)$ 、 $X(p)$ 分别表示在 p 点处输出与输入的特征; N 为采样点的个数; Δp_n 、 Δm_n 分别表示在第 n 个采样点的偏移和调节权值。

图3(b)为注意力部分,将调制部分输出的特征图通过使用全局最大池化和全局平均池化操作,将两者的输出相加之后进行Sigmoid函数来对每一个通道施加重重,并赋予到原特征通道中使重要通道获取更多关注。

图3(a)部分获取特征图中重要空间信息,图3(b)部分对重要特征通道给予更多关注,注意力可变形卷积从空间和通道两个角度来增强特征提取,并更好地应对小目标特征在上采样时导致的特征形变。

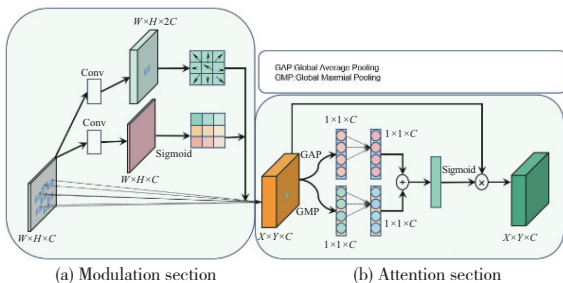


图3 注意力可变形卷积结构图

2.3 级联融合

YOLOv5s的Neck结构使用了PaNet网络结构,结合了特征金字塔网络(feature pyramid network, FPN)^[14]和像素聚合网络(pixel aggregation network, PAN)来进行多尺度融合。将主干网络中FConvNeXt Block所对应的4个特征层分别记作 P_1 、 P_2 、 P_3 和 P_4 ,原Neck将 P_2 、 P_3 和 P_4 3个包含高级特征信息特征层作为输入。通过将不同层级的特征层信息进行融合,以此来缓解特征的不变性与其等变性矛盾,使得输出特征图拥有浅层特征层的位置信息与深层特征层的语义信息。特征融合模块额外引入 P_1 层,并对其进行Focus操作,在尽可能保存多的特征位置信息的前提下进行下采样,并且提升维度,对 P_3 特征层当中先进行注意力可变形卷积后进行上采样,最后将 P_1 、 P_2 、 P_3 进行加权融合,当CSP模块中 N 为3时简称 C_3 ,特征融合模块如图4所示。

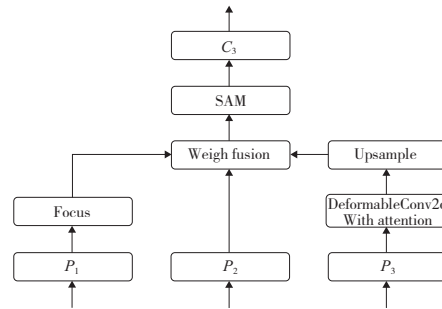


图4 特征融合模块结构图

加权融合按每层特征层所提供的贡献来进行权值相加,权值计算规则如式(2)所示。

$$O = \sum_i \frac{W_i \times I_i}{e + \sum_j W_j} \quad (2)$$

式中: I_i 表示输入特征; W_i 、 W_j 表示学习到的权值参数; e 表示约束参数; O 为输出。

Weight Fusion输出计算公式如式(3)所示。

$$O = \text{Conv} \left(\frac{W_1 F(P_1) + W_2 P_2 + W_3 R_E(P_3)}{W_1 + W_2 + W_3 + e} \right) \quad (3)$$

式中: $F()$ 表示Focus操作; R_E 表示注意力可变形卷积加上采样操作。通过加权融合能够高效提取更多道路车辆行人特征,并且学习不同输入特征的权重,获取不同输入对于输出特征的贡献情况,有效解决车辆行人尺度变化问题,并且获取 P_1 特征层中的小目标样本信息,提升输出层的小目标信息的贡献。

3 实验

3.1 数据集选取

本实验所用数据集通过在 BDD100K 当中随机选取 10 000 张图像作为训练集,并从中选取 1 000 张图片作为测试集。将目标检测对象分为 person、car、bus、truck 4 种并且包含它们当中的小目标信息。

3.2 实验环境

本文网络模型在笔记本上完成训练,其配置如下:12th Gen Intel(R) Core(TM) i7-12700H CPU,

NVIDIA GeForce RTX 3070 Ti Laptop GPU。使用 PyTorch 深度学习框架,语言使用 python3.7。

3.3 实验分析

1) 识别效果展示

将本文改进的网络模型 YOLOv5s-CD 对 BDD100k 当中的道路图像进行检测和识别,其检测结果如图 5 所示。本文对 person、car、bus 和 truck 4 种目标进行训练和检测,从检测结果图 5 当中可以看出,对于白天和黑夜场景当中的道路目标能够起到很好的检测识别效果,并且对远景小目标和重叠目标检测也同样拥有着不错的效果。

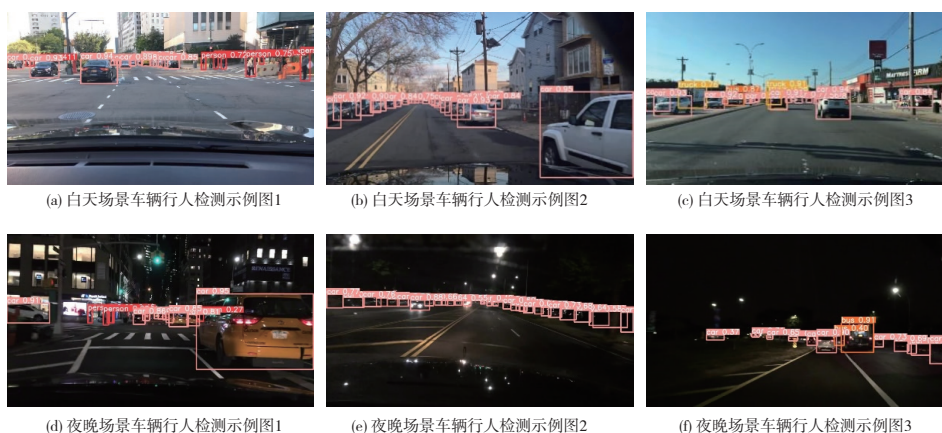


图 5 道路车辆行人图像识别结果

为了进一步验证本文改进网络的性能,将当前网络和主流轻量化网络 YOLOv5s、YOLOv6s、YOLOv7-tiny^[15-16] 和 YOLOv8s 模型以及具有代表性的二阶段网络 Faster R-CNN 模型进行对比实验,对比结果如图 6 所示。所选图 6 中存在 person、car 和 truck 3 种目标,且存在远景小目标和密集重叠目标。YOLOv5s-CD 能够完美识别检

测出图 6 中的目标并且定位无误;其他网络存在了漏检误检以及对重叠目标存在重复检测的问题。根据检测效果对比结果展示,改进之后的 YOLOv5s-CD 能够有效提取道路目标特征,对于远景目标和重叠目标都能起到很好的检测效果,提升了道路目标的检测精度和识别准确率。



图 6 不同网络识别结果

2) 网络模型性能对比

为了更好地对上述网络模型进行性能评估,分别使用查准率(precision)、查全率(recall)和平均精度(mAP)作为评价指标与其他轻量化网络进行比较。其比较结果如表1所示。

表1 网络模型评价指标 单位:%

网络模型	评价指标		
	$P_{\text{precision}}$	R_{recall}	A_{mAP}
Faster R-CNN	62.3	61	55.3
YOLOv5s	96.0	85	71.8
YOLOv6s	91.5	83	68.5
YOLOv7-tiny	92.7	88	62.8
YOLOv8s	95.3	86	72.9
YOLOv5s-CD	96.4	87	75.0

由表1可以看出,改进之后的网络在平均精度和召回率方面均优于经典的二阶段网络Faster R-CNN;并且超越了原网络,召回率与最新的YOLOv7-tiny和YOLOv8s不相上下,但在检测精度方面均有提升^[17]。为了更好地验证算法对道路各个种类目标的检测效果,将各类道路目标的查准率和查全率训练结果制成单目标精度表,如表2所示。

表2 不同目标平均精度 单位:%

网络模型	平均精度			
	person	car	bus	truck
Faster R-CNN	44.9	68.4	55.8	56.1
YOLOv5s	57.5	77.2	75.1	77.1
YOLOv6s	58.3	77.6	68.9	69.3
YOLOv7-tiny	51.1	76.3	62.5	61.2
YOLOv8s	61.2	79.7	74.4	76.5
YOLOv5s-CD	61.1	79.1	78.6	81.2

由表2可知,改进之后的网络在bus和truck两种类别使所有网络当中检测精度最高的, person和car两种类别略低于YOLOv8s,其他均高于其他网络类型。

在实际道路中,不仅仅需要考虑道路目标的检测精度,还需要考虑网络模型的大小所产生的推理时间。通过选用100张道路图像,将训练好的上述6种模型进行推理时间对比,将100张图像分成5份,每份中含有20张道路图像,通过对

每一份当中的图像进行检测,并求得平均时间,记为 T_1 — T_5 ,结果如表3所示。

表3 推理耗时对比 单位:ms

网络模型	推理耗时				
	T_1	T_2	T_3	T_4	T_5
Faster R-CNN	48.7	48.8	48.1	47.4	47.6
YOLOv5s	9.5	9.6	9.2	8.6	8.7
YOLOv6s	13.8	13.6	13.7	13.5	13.8
YOLOv7-tiny	6.8	7.1	7.0	7.0	6.6
YOLOv8s	16.6	16.4	16.5	16.3	16.6
YOLOv5s-CD	11.0	11.3	10.7	10.5	10.8

由表3可知,YOLOv7-tiny是所有网络中参数数量和计算量最少的,因此也是所有网络当中推理速度最快的网络,但其检测精度较低,并不适合运用于真实道路;YOLOv5s-CD在YOLOv5s基础网络上进行改进,增加了网络参数,提高了检测精度的同时,虽然增加了网络推理耗时,但依然快于其他网络。因此改进之后的YOLOv5s-CD平衡了检测精度和检测速度,能够更好地满足车载嵌入式部署。

3) 消融实验对比

为了进一步验证改进之后网络模型的有效性,对其进行消融实验并对比分析,实验结果如表4所示。模型1是没有经过改进的YOLOv5s原模型;模型2是仅仅改进了主干网络的YOLOv5s;模型3是仅仅改进了Neck网络模型;模型4是同时改进了主干网络和Neck网络。

表4 消融实验模型对比 单位:%

网络模型	评价指标		
	$P_{\text{precision}}$	R_{recall}	A_{mAP}
模型1	96.0	85.0	71.8
模型2	95.0	86.0	74.7
模型3	96.1	85.2	72.7
模型4	96.4	87.0	75.0

通过表4可知,在对主干网络以及Neck网络单独进行改进时,在召回率与平均精度方面均有提升型,说明改进之后的主干网络能够更好地提取目标特征信息,并且改进之后的Neck在原程度上更好地应对了小目标特征变化。再将两种改进融合之后,提升明显。说明改进之后的网络模型

能够更好地提取特征,识别多目标以及小目标信息,提高道路目标检测的准确性。

4 结语

道路行人车辆识别技术有着实际应用与研究价值,其可以运用于各式车载系统之中。本文算法以YOLOv5s为基础网络模型,设计了一种检测识别性能更好的检测网络。一是基于ConvNeXt网络模块结构对原主干网络进行改进,提升空间信息提取能力;二是在原Neck的基础上引入注意力可变形卷积加权融合机制,提升了对于小目标检测能力。通过对比其他主流网络,验证了本文算法的先进性。

参考文献:

- [1] WANG Z G, ZHAN J, DUAN C G, et al. A review of vehicle detection techniques for intelligent vehicles[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8): 3811-3831.
- [2] 金旺, 易国洪, 洪汉玉, 等. 基于卷积神经网络的实时车辆检测[J]. 计算机工程与应用, 2021, 57(5): 222-228.
- [3] 杨伟, 杜学峰, 张勇, 等. 基于深度学习的车辆目标检测算法综述[J]. 汽车实用技术, 2022, 47(2): 24-26.
- [4] 梁继然, 陈壮, 董国军, 等. 结合注意力机制和密集连接网络的车辆检测方法[J]. 电子测量与仪器学报, 2022, 36(3): 210-216.
- [5] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2261-2269.
- [6] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 6517-6525.
- [7] 孙备, 左震, 吴鹏, 等. 面向无人艇环境感知的改进型SSD目标检测方法[J]. 仪器仪表学报, 2021, 42(9): 52-61.
- [8] 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.
- [9] 于方程, 张小俊, 张明路, 等. 基于改进CenterNet的自动驾驶小目标检测[J]. 电子测量技术, 2022, 45(15): 115-122.
- [10] 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 (CVPR). Vancouver, BC, Canada: IEEE, 2023: 12021-12031.
- [11] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 11966-11976.
- [12] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002.
- [13] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 764-773.
- [14] KUANG H W, WANG B, AN J P, et al. Voxel-FPN: multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds[J]. Sensors, 2020, 20(3): 704.
- [15] WANG C, BOCHKOVSKIY A, LIAO H. YOLOv7: trainable bag of freebies sets new state of the art for real-time object detections[EB/OL]. (2020-07-06) [2022-05-10] arXiv: <https://arxiv.org/abs/2207.02696>.
- [16] WANG C, LI B, LIU Y, et al. YOLOv6: a single-stage object detection framework for industrial applications[C]//2023 IEEE International Conference on Multimedia and Expo(ICME). Brisbane, Australia: IEEE, 2023: 1-6.
- [17] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 7464-7475.

收稿日期: 2023-12-05