DOI:10.19344/j. cnki. issn1671-5276.2024.05.046

基于改进 YOLOv5 的两阶段抓取检测算法

朱文磊[a,b],董淑宏[a,b],张洪[a,b],于培师[a,b],徐稳[a,b]

(江南大学 a. 机械工程学院; b. 江苏省食品先进制造装备技术重点实验室,江苏 无锡 214122)

摘　要:针对复杂场景中机器人的无序抓取需要,提出一种两阶段的抓取检测算法。改进 YOLOv5 的网络模型,在多尺度特征融合上将浅层位置信息和深层语义信息进行注意力融合,提高多尺度目标的检测能力;将排斥因子引入损失函数中,提高了模型在遮挡环境下的鲁棒性;在目标检测后对抓取目标边界框进行裁切处理,避免了抓取检测过程中其余目标的干扰;改进抓取检测算法,引入 CSP 结构和注意力机制,提高了模型的特征提取能力。在真实环境下针对随意摆放的多目标遮挡物体进行抓取实验,结果表明:机器人抓取成功率为 95%。

关键词:调压阀;目标检测算法;轻量化;重参数化;特征融合

中图分类号:TP391.41　　**文献标志码**:A　　**文章编号**:1671-5276(2024)05-0218-06

A Two-stage Grasp Detection Algorithm Based on Improved YOLOv5

ZHU Wenlei[a,b], DONG Shuhong[a,b], ZHANG Hong[a,b], YU Peishi, XU Wen[a,b]

(a. School of Mechanical Engineering; b. Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment & Technology, Jiangnan University, Wuxi 214122, China)

Abstract: A two-stage grasp detection algorithm is proposed for the disorderly grasping needs of robots in complex scenes. The network model of YOLOv5 is improved by attention fusion of shallow location information and deep semantic information on multi-scale feature fusion to improve the detection of multi-scale targets. The rejection factor is introduced into the loss function to enhance the robustness of the model in occlusion environment. The grasp target bounding box is cropped after the target detection to avoid the interference from the rest of the targets during the grasp detection process. The grasp detection algorithm is improved by introducing the CSP structure and attention mechanism to improve the feature extraction ability of the model. In grasping multi-target obscured objects randomly placed in a real environment, the results show that the robot has a 95% success rate.

Keywords: pressure regulating valve; target detection algorithm; lightweight; re-parameterization; feature fusion

0　引言

伴随着人工智能的快速崛起,智能制造业中机器人的应用深度和广度得到了显著提升。机械手抓取作为智能机器人最重要的技能之一,被广泛应用在工业领域中替代人工进行工件抓取分类、产品包装等工作[1]。目前,在复杂环境下的多目标抓取检测仍具有较大挑战,获取更高精度的抓取姿态成为了抓取控制领域的研究热点[2]。

近年来,基于深度学习的经验法抓取取得了一定的研究成果。LENZ 等[3]首次采用滑动窗口检测框架搭建神经网络,达到了 73.9%的准确率,但是其模型计算量过大,无法进行实时检测;REDMON 等[4]舍弃滑动窗口的检测方法,利用 AlexNet 网络直接回归获得检测结果,达到了 88%的准确率,且可以实时运行;MORRISON 等[5]借鉴语义分割的算法思想,提出了基于像素点检测的轻量化抓取模型 GGCNN,模型运行速度快,但准确率不高;KUMRA 等[6]在其基础上将残差模块添加到特征提取骨干网络,以 RGB-D 融合图像作为输入,提高了模型的准确率;张志康等[7]提出了基于语义分割分阶段特征融合的抓取检测算法,具有较高的检测精度,但网络结构复杂;金欢[8]采用级联式的网络结构,将原始图像先分割后检测,实现了多目标抓取检测。

综上所述,已知的抓取检测算法仅利用特征提取网络中的最后一层输出特征图进行特征预测,对于尺寸多变、形状不同、姿态未知的目标,往往倾向于生成大目标的抓取框,而对小目标的检测性能较差,同时大部分的抓取模型为单目标场景抓取,没有考虑实际工业环境中背景复杂、目标

基金项目:国家自然科学基金项目(11972171)

第一作者简介:朱文磊(1998—),男,安徽芜湖人,硕士研究生,研究方向为目标检测和无序抓取,97188654@ qq.com。

间存在相互遮挡等问题。针对以上问题,本文提出了一种两阶段的抓取检测算法。

1　两阶段抓取检测算法

为了满足抓取检测中无序抓取的任务需求,除了生成最优的抓取检测框外,还需要识别出目标种类,本文通过设计并联式的两阶段抓取检测算法,实现在复杂环境中杂乱物体的抓取。整个抓取检测流程如图 1 所示。

2　目标检测网络

YOLOv5 在目标检测领域具有了较好的检测精度和检测效率,其网络框架如图 2 所示。但是在实际应用环境中,机器人检测的目标多样复杂、抓取环境杂乱、物体密集堆叠。针对以上问题,本文在数据预处理、网络结构以及损失函数部分做出改进。

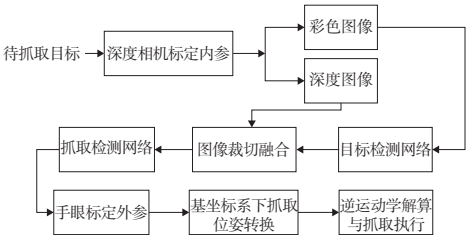


图 1　抓取检测流程框图

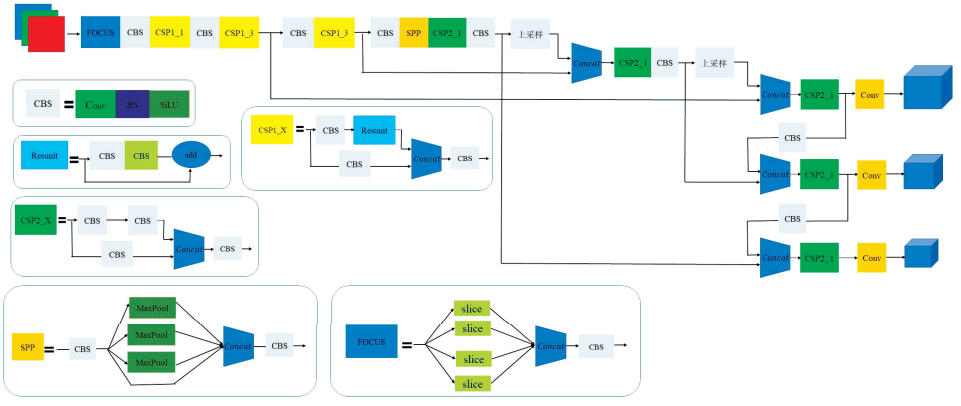
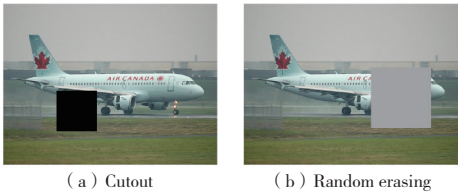


图 2　YOLOv5 网络结构

2.1　数据预处理

在训练的过程中,通过模拟物体遮挡,可以提高模型被遮挡时的抗干扰性,同时对于整体数据集而言是一种正则化处理方式,避免了网络过拟合,对模型的学习能力有所提升。

本文采用多种数据增强方法来模拟物体遮挡的效果,具体效果如图 3 所示。Cutout 和 Random erasing 均通过在图像中随机裁切一个矩形区域,前者直接在此区域内填充 0,后者赋值随机像素值;Hide-and-Seek 为解决弱监督问题中目标定位的精度问题,采用随机裁切若干个区域,从而让模型学习物体的全局信息;GridMask 在 HaS 的基础上采用了等间隔裁切区域的方式,并且对该区域实现一定的旋转。



(a) Cutout　(b) Random erasing

(c) Hide-and-Seek　(d) GridMask

图 3　数据增强

2.2　网络结构优化

在卷积神经网络中,通过上下采样可以获得不同尺寸的特征图。低纬特征图能够包含目标物体的空间特征信息,有利于确定目标的空间位置,而高纬特征图包含更丰富的语义特征信息,具有图像的概括能力,有利于分类任务的完成。

原有的 YOLOv5 采用 SPP(空间金字塔池化)来获取不同感受野的大小,采用统一步长,不同大小卷积核对输入特征图进行卷积操作,没有综合局部信息与全面信息的语义关系。本文结合深度可分离卷积实现 ASPP(空洞空间金字塔池化),降低参数计算量,无需通过减小图片和多个卷积核串联来增加感受野。如图 4 所示,第 1 个分支

采用 1×1 卷积,保留输入特征的感受野;中间 3 个分支分别采用扩张系数为 1、3、5 的空洞卷积,获得不同大小的感受野;第 5 个分支采用全局池化得到全局感受野;最后将各个特征输出 Concat 拼接后经过一个 1×1 卷积,实现多尺度特征提取。经过 ASPP 后的多尺度特征信息包含了大量的冗余信息,可通过添加注意力机制提高其特征提取效率。

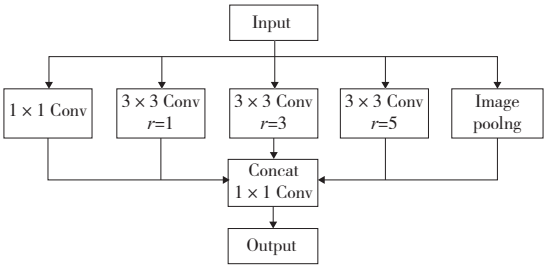


图 4　空洞空间金字塔池化

YOLOv5 原有的 FPN+PAN 在多尺度特征融合上对不同的输入特征图采用了平等的处理方式,而不同尺寸的特征图拥有不同的信息密度,在特征融合过程中所提供的有效特征是不相等的。为了提高多尺度融合中特征复用效率,本文采用 BiFPN[9](加权双向特征金字塔网络),在不同尺度的特征通道上引入了可学习的权重,重复利用自顶向下和自下而上的多尺度特征融合,充分利用不同分辨率中的特征信息。

CBAM[10]作为混合注意力机制,包含两个独立的子模块:通道注意力模块(channel attention module,CAM)和空间注意力模块(spatial attention module,SAM),分别将注意力映射到特征图的通道和空间两个维度,实现自适应特征提取,其网络结构如图 5 所示。

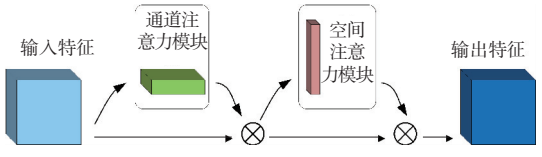


图 5　CBAM 模块

本文将 CBAM 嵌入到 ASPP 和 CSP 模块后,在特征融合之前,对特征图进行加权处理,提升关键特征,并抑制无关特征,使得网络能将重要信息加以融合。这样不仅使融合后的特征图包含更有效的目标信息,提升遮挡目标的定位精度,还达到降低模型的计算量,提升检测速度的目的。

2.3　损失函数改进

为了提高模型的遮挡检测性能,本文在 CIoU 的基础上引入新的损失函数 Repulsion loss[11],通过调整目标预测框与真实框、重叠目标预测框和真实框之间的关系,尽可能让预测框靠近真实框,远离其他目标框,降低 NMS 对阈值的敏感度。Repulsion loss 损失函数如下所示。

$$L=L_{\text{Attr}}+\alpha\times L_{\text{RepGT}}+\beta\times L_{\text{RepBox}} \tag{1}$$

式中:L_{Attr}表示目标预测框与真实框之间的损失,本文采用 CIoU 替换原有的 Smooth_{L1}损失;L_{RepGT}表示目标预测框与周围其他目标真实框之间的损失;L_{RepBox}表示目标预测框与周围其他目标预测框之间的距离;α、β 为权重调节系数。

L_{RepGT}是所有正样本预测框与其最大 CIoU 值的真实框的 IoG 均值,公式如下所示。

$$L_{\text{Rep}}=\frac{\sum_{P\in p}\text{Smooth}_{\text{Ln}}[\text{IoG}(B^P,G^P_{\text{Rep}})]}{|p|} \tag{2}$$

$$G^P_{\text{Rep}}=\underset{G\in g\setminus\{G^P_{\text{Attr}}\}}{\text{argmax}}\,\text{CIoU}(G,P) \tag{3}$$

$$\text{IoG}(B^P,G^P_{\text{Rep}})\triangleq\frac{\text{area}(B^P\cap G^P_{\text{Rep}})}{\text{area}(G^P_{\text{Rep}})} \tag{4}$$

$$\text{Smooth}_{\text{Ln}}=\begin{cases}-\ln(1-x) & x\leqslant\sigma\\ \dfrac{x-\sigma}{1-\sigma}-\ln(1-\sigma) & x>\sigma\end{cases} \tag{5}$$

式中:G 表示真实框;g 为所有真实框的集合;P 表示预测框;p 为 IoU 大于阈值的正样本预测框的集合;B^P 是根据预测框 P 调整后获得;G^P_{Rep}是目标预测框 p 除与之匹配的最大 CIoU 值的真实框;G^P_{Attr}是与目标预测框 p 相对应具有最大 CIoU 值的真实框。

L_{RepBox}作为相邻但不同目标预测框之间的排斥项,使得预测框和周围的其他预测框尽可能远离,公式如下所示。

$$L_{\text{RepBox}}=\frac{\sum_{i\neq j}\text{Smooth}_{\text{Ln}}[\text{IoU}(B^{p_i},B^{p_j})]}{\sum_{i\neq j}I[\text{IoU}(B^{p_i},B^{p_j})>0]+\grave{o}} \tag{6}$$

3　抓取检测网络

GR-ConvNet 是基于抓取点的抓取位姿检测算法模型。通过 RGB-D 的像素点信息预测出抓取目标的最佳抓取姿态以及每一个抓取点的质量分数,其网络模型如图 6 所示。

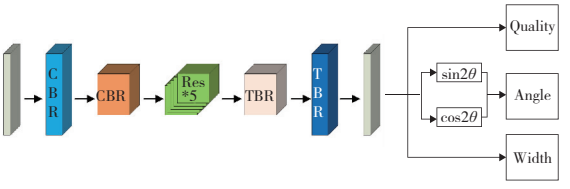


图 6　GR-ConvNet 网络模型

该算法主要应用于单目标的抓取位姿检测,无法对目标对象进行分类处理,抓取受到环境干扰大,同时在多尺度检测上容易忽视小目标的抓取。针对以上问题,本文在数据预处理和网络结构部分做出改进。

3.1　数据预处理

在多目标复杂场景以及目标之间存在堆叠时,输入图片中的背景和其他物体所包含的像素信息对 GR-ConvNet 算法具有一定的干扰性,主要是由于检测过程中只生成一个最佳的抓取框,抓取框选取的是全局图像中抓取置信度最高的点,而部分噪声点会干扰抓取框的选取,导致误检现象的发生。本文采用目标检测算法对抓取检测输入的图像进行预处理,只保留目标物体的边界框,将其余部分填充 0。

3.2　模型结构优化

针对原有模型中的残差模块,本文引进注意力机制 CBAM,如图 7 所示,嵌入在残差模块中的 BN 层后,提高模型的特征提取能力。同时借鉴 CSP 模块对其进行优化,通过采用 CSP 模块将输入特征分为两个分支使得通道数减半,其中一部分通过 5 个改进的残差模块后与另一部分进行通道相加,减少了计算量;在梯度反向传播过程中,同一个梯度在不同的模块中被反复计算,会导致大量的梯度冗余,通过对特征通道的裁剪,使得梯度在不同的分支中独自进行梯度回传,没有重复计算,有效地降低了梯度冗余,提高了模型的运行速度。

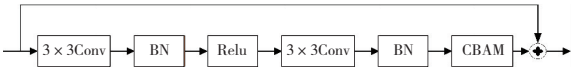


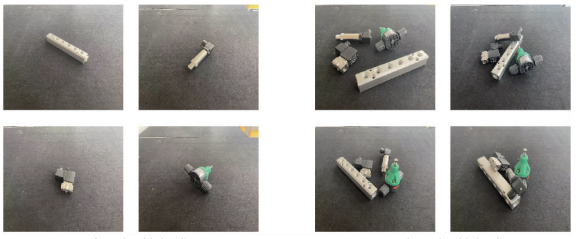
图 7　改进残差结构

4　实验及结果分析

4.1　目标检测

本实验采用自制工件数据集进行模型训练,如图 8 所示,对自动化装备生产中所需的气动工件采用 Kinect V2 深度相机采集。一共选取 4 种工件,采集了 1 200 张图像,以 VOC 格式对其进行标注,按照 8∶1∶1 的比例划分为训练集、验证集和测试集。

为充分验证模型改进的有效性,设置消融实验和对比实验探究不同改进策略对检测算法的性能影响。本文采用 mAP(均值平均精度)和 FPS(帧率)作为评价指标,表示检测算法对目标的平均检测精度和速度。



(a) 单目标数据集　　(b) 多目标数据集

图 8　自制工件数据集

消融实验如表 1 所示,A 表示替换 ASPP 模块后模型在不增加计算量的前提下,扩大了感受野,增强了模型识别不同尺寸目标的能力,检测速度和检测精度均有所提升;B 表示增加注意力机制 CBAM,加强了目标对象的关注度,有效降低了背景的干扰,增强了模型的鲁棒性,提高了模型的检测精度;C 表示在特征融合阶段采用了 BiFPN,在不同深度的特征图中采用不同的权重进行特征裁切,检测速度有小幅度降低但精度有所提升;D 表示在损失函数中引入排斥因子,使得模型在复杂环境中对遮挡物体的检测能力和精度得到了提升。相比原有模型,改进后的 YOLOV5 检测算法 mAP 提高了 4.3 个百分点,而检测速度基本没有受到影响。

表 1　YOLOv5 及其改进模型性能对比

YOLOv5	A	B	C	D	mAP/%	帧率/(帧/s)
√					88.8	58.8
√	√				89.4	59.4
√	√	√			90.6	59.3
√	√	√	√		91.0	59.1
√	√	√	√	√	93.1	59.1

为探究改进算法在各类别检测上的影响,本文将原算法 A 与改进算法 B 进行类别性能测试实验,如表 2 所示。

表 2　不同类别检测性能对比　　单位:%

物体	R		P		AP	
	算法 A	算法 B	算法 A	算法 B	算法 A	算法 B
压力传感器	89.0	94.8	90.2	93.5	88.7	92.8
背压阀	84.6	95.5	88.0	90.4	87.5	91.3
流量控制器	82.9	94.4	87.5	91.5	86.1	93.6
分气排	90.7	96.3	86.2	94.2	90.8	94.7

由表 2 可知,改进模型在准确率、召回率和平均精度上均有所提升。在多目标遮挡环境下,流量控制器的模型较小,且表面特征不明显,部分特征与压力传感器相似,当遮挡情况严重时便会导致误检或漏检,而改进后的模型显著提高了对流量控制器的特征提取能力以及遮挡情况下的召回率。

将改进后的 YOLOv5 算法与目前主流的目标检测算法进行性能对比(表 3),检测速度和检测精度均有了提升。在背景环境复杂、检测目标存在遮挡的情况下依然可以识别出目标并精准定位,减少了漏检、误检的概率。

表 3　不同目标检测算法性能对比

算法	Input	mAP/%	帧率/(帧/s)
Faster-R-CNN	1 000×600	73.1	16.1
SSD	512×512	68.4	38.2
YOLOv4	640×640	83.2	52.6
YOLOv5	640×640	88.8	58.8
改进 YOLOv5	640×640	93.1	59.1

4.2　抓取位姿检测

由于 cornell 数据集全部为单目标场景,缺少多目标堆叠场景下的数据集,本文选用 cornell 数据集和自制单目标工件数据集作为训练集和验证集,自制多目标工件数据集作为测试集,验证不同遮挡程度下抓取检测算法的性能。

测试数据集首先通过目标检测算法进行图像预处理,裁切出抓取目标区域并将其余背景部分进行填 0 处理,再输入到抓取位姿检测模型中。不同的检测算法实验结果如表 4 所示。

表 4　不同抓取检测算法性能对比

算法	输入图像	准确率/%	每张检测时间/ms
GGCNN	Depth	73.0	4
GGCNN	RGB-D	77.8	5
GR-ConvNet	RGB-D	85.6	13
改进 GR-ConvNet	RGB-D	94.7	15

由表 4 可知,在抓取位姿检测模型中,采用轻量化设计的 GGCNN 在检测速度上优势较大,但是检测准确率较低,通过将彩色图像和深度图像融合进行多模态输入,模型的检测精度有所提升;本文改进的 GR-ConvNet 采用 RGB-D 图像作为输入,引入 CSP 模块和 CBAM 注意力机制对残差结构进行优化,减少了梯度冗余,提高了模型的特征提取能力,解决了模型推理速度和检测精度不平衡的问题,在增加少量推理时间前提下获取了较高的准确率,相比原有模型提高了 9.1 个百分点。

4.3　真实机械臂抓取实验

本文采用 UR5 机械臂、Robotiq 机械夹爪和 Kinect V2 深度相机搭建抓取实验平台,采用眼在手外的方式固定相机,如图 9 所示。

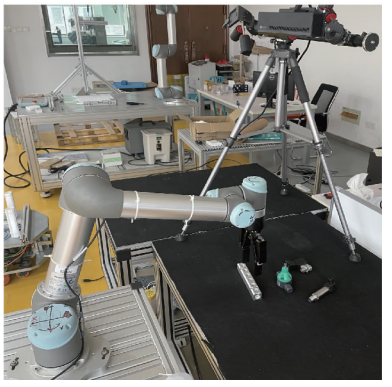


图 9　抓取实验平台

实验采用多目标场景,在平台上随机摆放工件,部分工件之间存在遮挡现象,重复实验 50 次,以实际抓取的成功率作为评价指标。抓取效果如表 5 所示。

表 5　实际抓取检测结果

物体	目标检测次数	检测准确率/%	抓取次数	抓取成功率/%
流量控制器	48/50	96	46/48	96
背压阀	50/50	100	46/50	92
压力传感器	48/50	96	44/48	92
分气排	50/50	100	50/50	100

由表 5 可知,本文提出的双阶段抓取检测算法在遮挡条件下具有较高的抓取成功率,通过目标检测算法获取遮挡目标的局部信息,提高了模型的抗干扰性,但依然存在检测失败和抓取失败的案例。高度遮挡环境下,目标间重叠面积过高,导致检测对象的特征不明显,对后续的抓取检测也有着较大的挑战。分气排模型较大且形状简单,在抓取过程中具有最佳的抓取表现;压力传感器由于其表面存在金属光泽,在图像中有效像素较少,抓取成功率相比其他种类较低;背压阀的最佳抓取位姿较少且表面光滑,使用二指夹爪在抓

取过程中容易脱落导致抓取失败。

5　结语

为实现工业环境中工件无序分拣,针对环境中背景复杂和目标间存在堆叠,难以实现高效分类抓取的问题,提出了一种两阶段抓取检测算法对工件进行抓取位姿估计。在第一阶段采用了目标检测算法,通过修改网络结构和损失函数,增强了对遮挡目标的检测能力,在多目标密集遮挡环境下获得了良好的性能,降低了模型漏检、误检的概率。第二阶段中,算法利用第一阶段生成的目标检测结果,对最佳抓取范围进行裁切,抑制甚至消除了环境背景对检测的干扰,再对目标物体进行细粒度的姿态估计和抓取框生成,实现了最佳的抓取效果。该算法在实际的机器人抓取场景中得到了广泛应用和验证,具有较强的通用性和鲁棒性,能够适应各种不同形态和大小的物体,并实现高效、精确的抓取操作。今后将进一步研究多个目标之间的顺序抓取问题以及被遮挡物体的信息补全,进一步提升抓取成功率。

参考文献:

[1] 陈苗苗,叶文华,马庭田,等. 不规则金属物料的抓取位姿实时检测方法研究[J]. 机械制造与自动化,2022,51(1):177-180,191.

[2] DU GG, WANG K, LIAN S G, et al. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review[J]. Artificial Intelligence Review, 2021, 54(3): 1677-1734.

[3] LENZ I, LEE H, SAXENA A. Deep learning for detecting robotic grasps[J]. The International Journal of Robotics Research, 2015, 34(4/5): 705-724.

[4] REDMON J, ANGELOVA A. Real-time grasp detection using convolutional neural networks[C]//2015 IEEE International Conference on Robotics and Automation. Seattle, WA, USA: IEEE, 2015: 1316-1322.

[5] MORRISON D, CORKE P, LEITNER J. Learning robust, real-time, reactive robotic grasping[J]. The International Journal of Robotics Research, 2020, 39(2/3): 183-201.

[6] KUMRA S, JOSHI S, SAHIN F. Antipodal robotic grasping using generative residual convolutional neural network[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, NV, USA: IEEE, 2021: 9626-9633.

[7] 张志康,魏赟. 基于语义分割的两阶段抓取检测算法[J/OL]. 计算机集成制造系统. (2022-05-11)[2022-12-11]. http://lkns.cnki.net/kcms/detail/11.5946.TP.20220517.1009.008.html.

[8] 金欢. 基于卷积神经网络的机器人抓取检测研究[D]. 哈尔滨:哈尔滨工业大学,2019.

[9] TAN MX, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 10778-10787.

[10] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[11] WANG X L, XIAO T T, JIANG Y N, et al. Repulsion loss: detecting pedestrians in a crowd[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7774-7783.

收稿日期:2023-03-14