控制与决策
Control and Decision

基于RGB-D融合的密集遮挡抓取检测

李明, 鹿朋, 朱龙, 朱美强, 邹亮

引用本文:
李明,鹿朋,朱龙,朱美强,邹亮. 基于RGB-D融合的密集遮挡抓取检测[J]. 控制与决策, 2023, 38(10): 2867-2874.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0259

基于RGB-D融合的密集遮挡抓取检测

李 明[1,2], 鹿 朋[1,2], 朱 龙[1,2], 朱美强[1,2†], 邹 亮[1,2]

(1. 地下空间智能控制教育部工程中心，江苏 徐州 221116；
2. 中国矿业大学 信息与控制工程学院，江苏 徐州 221116)

摘 要: 针对当前抓取检测模型对密集遮挡物体的检测效果差以及人工数据标注工作量大的问题,提出基于RGB-D图像融合的目标检测与抓取检测分步骤进行的改进方案. 新方案支持将单物体图像训练的抓取检测模型直接应用于密集遮挡的多物体图像场景中. 首先,考虑到密集遮挡场景下抓取物具有多尺度的特点,提出子阶段路径聚合(SPA)的多尺度特征融合模块,用于丰富RGB-D特征级别融合的目标检测模型SPA-YOLO-Fusion的高维语义特征信息,以便于检测模型定位所有的抓取物;其次,使用基于RGB-D像素级别融合的GR-ConvNet抓取检测模型估计每个物体的抓取点,并提出背景填充的图像预处理算法来降低密集遮挡物体的相互影响;最后,使用机械臂对目标点进行抓取. 在LineMOD数据集上对目标检测模型进行测试,实验结果表明SPA-YOLO-Fusion的mAP比YOLOv3-tiny与YOLOv4-tiny分别提高了10%与7%. 从实际场景中采集图像制作YODO_Grasp抓取检测数据集并进行测试,结果表明增加背景填充预处理算法的GR-ConvNet的抓取检测精度比原模型提高了23%.

关键词: RGB-D融合；密集遮挡检测；多尺度检测；抓取检测；机械臂；深度学习

中图分类号: TP911.73；TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0259

引用格式: 李明,鹿朋,朱龙,等. 基于RGB-D融合的密集遮挡抓取检测[J]. 控制与决策, 2023, 38(10): 2867-2874.

Densely occluded grasping objects detection based on RGB-D fusion

LI Ming[1,2], *LU Peng*[1,2], *ZHU Long*[1,2], *ZHU Mei-qiang*[1,2†], *ZOU Liang*[1,2]

(1. Engineering Research Center of Intelligent Control for Underground Space Ministry of Education，Xuzhou 221116，China；2. School of Information and Control Engineering，China University of Mining and Technology，Xuzhou 221116，China)

Abstract: Current grasp detection algorithms suffer from the poor accuracy and time-consuming or expensive data annotation in densely occluded scenes. To address this concern, a step-by-step improved solution for object detection and grasp detection based on RGB-D fusion is proposed. The grasp detection model trained on single-object can be directly applied to densely occluded multi-object scenes. Firstly, considering the multi-scale characteristics of objects in densely occluded scenes, the sub-stage and path aggregation (SPA) multi-scale feature fusion module is proposed to enrich the high dimensional feature characterization of middle fusion detector SPA-YOLO-Fusion, so as to locate all objects. Then the GR-ConvNet equipped with RGB-D pixel-level fusion outputs the optimal grasp points of all detected objects. At the same time, the background padding preprocessing algorithm is proposed to reduce the interference of other objects in the GR-ConvNet. The mAP of SPA-YOLO-Fusion is 10% and 7% higher than that of YOLOv3-tiny and YOLOv4-tiny on the LineMOD dataset, respectively. The grasp detection accuracy of the GR-ConvNet equipped with the padding algorithm is improved by 23% compared with the original model on the YODO_Grasp dataset, which is collected from the actual scene.

Keywords: RGB-D fusion；densely occluded detection；multi-scale detection；grasp detection；manipulator；deep learning

0 引 言

近年来,随着人工智能的不断发展,在工厂环境以及家居场所中,机器人发挥着越来越重要的作用. 机械手臂的视觉伺服抓取是智能机器人的一项重要技能,已广泛地应用于固定场所或结构化环境下的物品分拣、产品包装等工作中[1-2]. 然而,在服务机

收稿日期: 2022-02-17；**录用日期:** 2022-05-17.
基金项目: 国家自然科学基金项目(51904297, 61901003).
责任编委: 方勇纯.
†通讯作者. E-mail: zhumeiqiang@cumt.edu.cn.
*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

器人工作的非结构化环境中,实现对物品的准确识别与精准抓取依然是一项具有挑战性和复杂性的工作,尤其是对于密集遮挡物品的识别与抓取. 目前,基于深度学习的视觉抓取工作可分为两类: 从视觉感知到动作执行的端到端方法,包含物品识别、抓取点检测及机械臂规划抓取的分阶段方法[3-4]. 显然,端到端方法存在需要大数据样本、可解释性较差且泛化性有待提高等不足[5],因此,在实际应用中人们更青睐使用分阶段的方法.

抓取物密集遮挡摆放是生活中常见的情形. 该场景下,由于不同形状、尺寸的物体密集摆放,致使物体之间存在相互堆叠与遮挡的情况,使得服务机器人仅使用RGB相机进行目标检测、抓取检测变得异常困难. 近年来,随着RGB-D深度相机的普及,其基本上成为机械臂视觉抓取任务中的标配,而且在深度学习中利用两种图像的特点进行融合,用于物体检测与抓取检测的方法也逐渐成为研究热点[6]. 在RGB-D相机中,RGB图像的成像范围广泛,可以获取物品的纹理信息和颜色细节信息,而Depth图像则记录着物体与相机之间的距离,因此,该类图像中各个物体都具有非常显著的边界轮廓信息,并且能够捕捉到其可测范围内的几乎所有物体. 基于RGB-D的融合方式主要分为像素级别和特征级别的融合,两种融合方式的区别主要在于融合操作与特征提取操作的顺序上: 前者是将Depth图像作为RGB图像的第4维数据进行图像像素的融合,然后使用卷积神经网络对4通道图像数据进行特征提取; 后者则是使用不同的卷积神经网络对RGB图像和Depth图像进行特征提取,然后再对RGB图像特征和Depth图像特征进行融合. 两种不同的融合方式可以分别应用于物体检测与抓取检测等不同的任务.

针对密集遮挡的目标检测难题,学界主要从基于RGB多尺度物体检测、密集物体检测和RGB-D融合3个方面展开研究. 针对RGB图像中密集物体的检测问题,文献[7]提出了Focal Loss损失函数,给予密集物体更大的权重进行有效的特征学习; 文献[8]设计了功能选择性模块,根据物体的形状自适应感受野. 针对多尺度物体检测问题,主要采用设计多尺度特征模块来丰富高维语义特征的策略: 文献[9]提出了FPN (feature pyramid network)特征融合方式,使用自上而下的特征提取路径得到多种尺度的高维语义特征,并与低维特征进行融合; 文献[10]提出了PANet (path aggregation network)模型,在FPN的基础上增加一个自下而上的特征提取路径,用于充分利用包含精确位置信息的低维特征; 文献[11]提出使用sub-stage特征融合算法,将不同层级的特征进行融合. 目前,RGB-D图像融合的模型主要是基于YOLO (you only look once)[12]和Faster-RCNN[13]模型的变体. 文献[14]基于Faster-RCNN检测网络提出了3D Region Proposal Network算法以及joint Object Recognition Network算法来解决物体遮挡的检测问题; 文献[15]认为RGB图像在密集行人检测任务中能提供丰富的颜色细节信息,而Depth图像能提供准确的位置信息,因此,将RGB与Depth图像进行特征融合,并对YOLOv2[16]检测模型进行改进,提高了对密集人群的检测精度; 文献[17]针对YOLOv3[18]进行RGB-D的特征级别的融合,用于密集人群检测. 然而,上述算法的主要研究对象往往是具有较大尺寸的人体或物品,对密集遮挡抓取物的检测依然有待研究.

针对抓取检测问题,根据数据集的类型划分,学界的研究方向主要分为3类: 第1类中训练集、测试集都为单个物体图像,第2类中数据集都为多个物体图像,第3类中训练集为单个物体图像而测试集为多个物体图像. 第1类场景较为单一,因此抓取检测模型也较为精简: 文献[19]提出了基于Depth图像的GG-CNN检测模型,该算法可以同时输出抓取质量、抓取角度和抓取宽度3个参数,用于表征一个抓取框; 文献[20]将残差网络引入到骨干特征提取网络中并提出了基于RGB-D融合的GR-ConvNet模型. 以第2类为主的算法直接在多物体图像中训练抓取检测模型,因此,数据标注的工作量大、人工成本高昂: 文献[21]提出了Graspnet-1billion数据集,用于研究密集遮挡场景的抓取问题; 文献[22]提出了Grasp Pose Refinement Network二阶段网络,用于研究密集场景的抓取姿态问题. 第3类问题比较具有挑战性且对模型的泛化性要求较高: 文献[23]提出了GraspNet抓取点检测网络以及GraspSeg抓取数据集,该算法可以直接输出抓取点的位置. 上述的抓取检测模型基本是端到端进行预测,对于密集遮挡等复杂场景的泛化能力仍有待提高.

综上所述,为了提高对密集遮挡物体的抓取检测精度并减少图像标注的工作量,本文提出分阶段串行检测实现抓取的思路. 具体而言,本文的贡献主要有以下4个方面:

1) 提出基于RGB-D图像融合的物体检测与抓取检测分步骤处理的思路,该思路适用于在单物体图像上训练的抓取检测模型,直接在多物体图像的复杂场

景中进行应用的任务,可以有效减少人工数据标注的工作量,为了表示方便,本文使用YODO (you only detect once)来表示上述方法.

2) 本文使用RGB-D特征级别融合的方法并提出SPA多尺度特征融合模块,用于提高对密集遮挡等多尺度物体的目标检测精度.

3) 本文提出背景填充的图像预处理方法以减少其他物体对目标物体的干扰,用于提高对密集遮挡物体的抓取检测精度.

4) 针对密集遮挡场景的抓取检测任务,提供目标检测数据集以及抓取检测数据集.

1　YODO抓取检测流程

YODO思路如图1所示. 首先,使用基于RGB-D特征级别融合的SPA-YOLO-Fusion对RGB-D图像进行物体检测(如图1(b)所示),得到各个物体的类别及其坐标;其次,将所有物体都进行图像裁剪、背景填充以及图像堆叠,统一物体图像的尺寸;然后,使用GR-ConvNet对批次图像进行抓取检测(如图1(d)所示,蓝色线条表示夹爪的开合方向),该模型在单物体图像上进行训练,且可以直接应用于密集堆叠多物体的场景中;最后,将二维图像坐标转换为机械手臂视角下的三维坐标并进行抓取(如图1(e)所示).

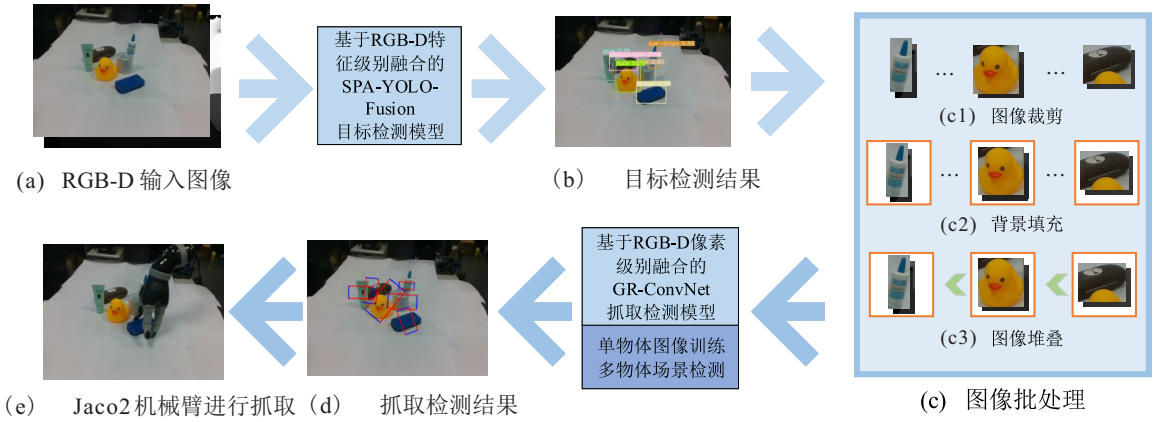


图 1　分阶段抓取检测思路

1.1　SPA-YOLO-Fusion检测网络

目前,很多主流的目标检测模型[24]都是以YOLOv3作为基准模型并进行改进. 因此,本文也以YOLOv3的轻量级网络YOLOv3-tiny为基准模型并进行改进. YOLO模型主要由特征提取器、特征融合器以及检测头3部分组成. 特征提取器主要对输入图像进行特征提取;特征融合器将多分辨率的特征进行融合,以丰富模型的特征信息;检测头用于提取物体的类别以及位置信息. YOLOv3-tiny模型只能处理3通道的图像数据,为了满足RGB和Depth融合检测的任务,本文提出YOLO-Fusion模型,如图2所示. 具体而言,该模型使用两个不同的特征提取网络分别对RGB图像以及Depth图像进行特征提取. 为了提高密集遮挡场景下多尺度物体的检测效率,本文从增加感受野、低维与高维特征融合两方面进行改进,并提出SPA特征融合方式,并将最终的模型命名为SPA-YOLO-Fusion.

感受野大小是指后一个网络层中每个特征能够表征前一个网络层多大的区域,它在卷积神经网络中发挥着非常重要的作用. 如果感受野太小,就会缺失对目标物体的信息,造成召回率降低;如果感受野太大,就会包含较多的干扰信息,造成准确率下降. 增加

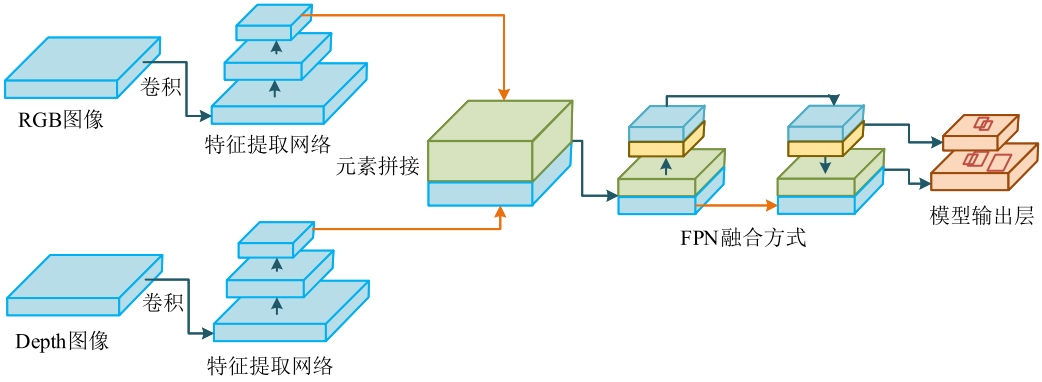


图 2　YOLO-Fusion图像融合检测网络

网络感受野主要有3个手段: 增加网络的层数, 使用下采样进行信息提取以及使用空洞卷积. 本文主要针对前两种方法进行改进: 通过添加自下而上的卷积提取路径来增加网络层数; 并使用sub-stage融合方式提高下采样次数, 从而增加模型的感受野.

在卷积神经网络中, 低维特征图主要包含边缘和线条等物体局部位置特征信息, 而高维特征图主要包含细节和轮廓等全局语义信息, 通常丰富的高维语义特征信息能够获得更好的检测效果. 图3(a)的FPN使用自上而下(up-down)的上采样方式获得多尺度特征信息, 并进行高、低维特征融合, 但与图3(c)中的sub-stage融合方式相比, 该融合方式所带来的感受野增益较低; 图3(b)的path aggregation中额外添加了自下而上(bottom-up)的特征提取路径, 主要是为了在高维决策特征中加入低维特征, 从而获得物体更准确的位置信息. 基于此, 针对密集遮挡环境下存在多尺度检测与高维特征表征较差的问题, 本文提出SPA特征融合方式: 使用sub-stage代替FPN特征融合方式, 用于进一步扩大模型的感受野并获得更大范围的特征信息, 从而有效解决密集遮挡所带来的多尺度检测问题; 增加一条自上而下的特征提取路径并进行高、低维特征融合, 从而丰富高维特征的表征能力.

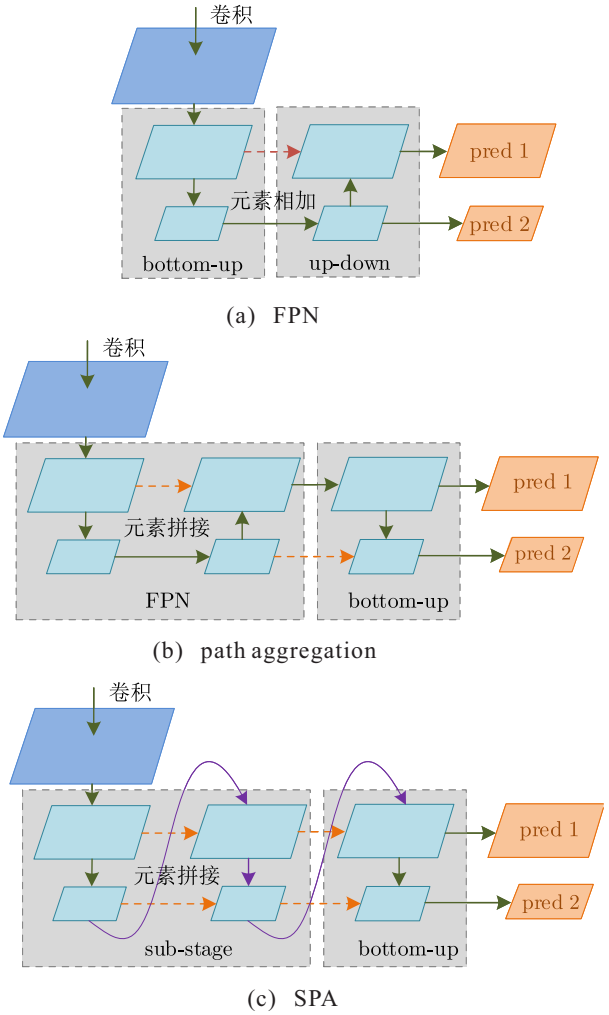
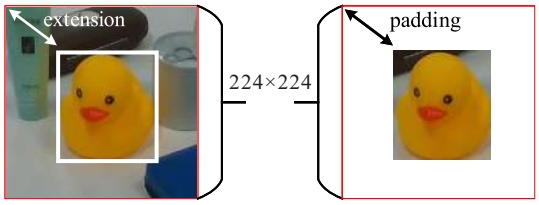


图3 3种不同的特征融合方式

1.2 GR-ConvNet抓取检测模型

GR-ConvNet[20]抓取检测模型主要用于单个物体图像的抓取检测任务. 对于输入的RGB-D图像(224×224×4), 首先进行编码操作和解码操作; 然后输出每个像素点处的抓取质量(224×224×1)、抓取角度(224×224×1)以及抓取宽度(224×224×1); 最后, 由这3种抓取特征确定最优的抓取框. 由于GR-ConvNet抓取检测模型只选择一个最优的抓取点, 对于密集遮挡堆叠的多物体的复杂场景需要进行多次抓取检测, 检测效率低且误检率高.

为了减少密集遮挡场景下物体之间的影响, 本文提出图4(b)所示的背景填充的方法来代替图4(a)中心点扩展的原始方法. 原始方法的思路是根据目标物体的中心点坐标从原图中裁剪固定尺寸大小的图像像素, 这样会导致输入图像中包含较多的其他物体图像像素的干扰, 从而影响GR-ConvNet检测模型的输出. 为了降低干扰, 本文提出给目标物体填充白色背景的方法.



(a) 原始方法 (b) 本文方法

图4 3种不同的特征融合方式

2 SPA-YOLO-Fusion检测网络

2.1 抓取检测实验平台

本文的实验平台配置为: 使用RealSense D435i深度相机进行图像数据的采集, 使用Jaco2机械臂对目标物体进行抓取. 尽管抓取研究中存在GraspSeg、LineMOD等数据集, 但这些数据集或不开源, 或是在特定场景使用特定物品得到的数据集, 难以获得相关物品并在本平台上复现其场景.

为此, 本文选择6类形状尺寸各异的生活物品作为目标检测任务以及抓取检测任务的实验对象, 并自制了部分数据集. 为了描述方便, 将从实际实验环境下采集的目标检测数据集以及抓取检测数据集分别命名为YODO_Det与YODO_Grasp. YODO_Det数据集的特点是密集摆放的物体之间存在不同程度的堆叠与遮挡; YODO_Grasp数据集的特点是训练集只包含单个物体的图像而测试集是只包含密集遮挡堆叠的多物体图像.

2.2 密集遮挡场景的目标检测实验

本文的所有目标检测模型均使用交叉熵损失函数来降低模型的分类误差,使用GIOU[25]损失函数来提高模型的位置精度,并使用锚框机制辅助卷积神经网络进行位置信息的学习,以提高模型的学习效果. 在训练集中通过聚类的方法得到6种宽、高尺寸不同的锚框,其尺寸分别为: [31, 48], [36, 70], [46, 55], [53, 82], [71, 95], [83, 132].

本文的实验平台配置为Intel Xeon(R) Bronze 3106 CPU@1.70 Hz x16处理器,64 GB内存和12 GB GeForce RTX 2080Ti GPU,在Ubuntu 18.04环境下使用PyTorch构建检测模型,并进行模型的训练以及推理. 所使用的超参数为: 训练的学习率为0.001,并采用随机梯度下降法进行学习率的调整,其momentum参数为0.937;权重衰减的参数设置为0.000 484,整个训练过程的epoch为100,训练时间大约为200 min. 在测试中,置信度阈值为0.5,交并比阈值为0.6. 为了加快模型的训练速度和效果,在场景较为单一的lifelong[26]数据集上进行模型预训练.

本文选用目标检测任务中常用的平均精度(average precision, AP)和均值平均精度(mean average precision, mAP)作为评价指标. 平均精度表示检测模型对某一类物体的平均检测精度,而均值平均精度表示检测模型对所有物体的平均检测精度.

2.2.1 LineMOD数据集

LineMOD[27]数据集是研究目标识别与抓取常用的数据集,它包括13类物体,共13 000余张图像,但数据集的每张图像中只标注单个物体. 为了研究密集、遮挡场景下模型的效果,本文选取常见的10类物体,共5 000余张图像进行实验对比. 使用LableImg工具对10类物体进行标注. 将整个数据集按照8:1:1的比例划分成训练集、测试集和验证集,即训练集共4 000张图像,测试集和验证集分别是500张图像.

如表1所示,在基于RGB图像检测的检测器中: MobileNet使用深度可分离卷积等技巧获得了较快的检测速度,但检测精度较低; YOLOX-s模型具有最高的检测精度,但却带来较大的推理损耗; YOLOX-tiny是YOLOX-s的轻量级版本,与YOLOX-s相比检测速度有所提高,但检测精度有了大幅度的降低,性价比不高. 对于使用RGB图像和Depth图像融合的检测网络,其多模态融合的检测网络与改进之前的模型都有较大性能的提升,而且其推理速度的性能损耗也较小: SPA-YOLO相比于改进之前的YOLOv3-tiny检测网络,在检测速度没有显著增长的情况下,其检测精度有了4%的性能提升,尤其是增加了对小物体(如eggbox、driller和glue等)的检测精度. 基于RGB-D融合的SPA-YOLO-Fusion与YOLOv3-tiny相比,有1.3 ms的推理时间损耗,但是检测性能分别提高4%,表明了RGB-D图像融合策略与SPA模块的有效性.

2.2.2 YODO_Det目标检测数据集

为了测试在实际场景中模型的性能,从实际场景中采集生活中常见的尺寸形状各异的6类物体作为YODO_Det数据集,共5 000张图像,并按照8:1:1的比例将其划分成训练集、验证集以及测试集.

各类模型的检测性能如表2所示. 可见YOLOv4-tiny与YOLOv3-tiny的检测速度很接近,但YOLOv3-tiny的检测精度更高. 因此,本文在YOLOv3-tiny模型基础上提出了SPA-YOLO-Fusion, 该模型虽然增加

表 1 LineMOD检测精度对比

检测模型	ape	benchvise	cam	can	cat	driller	duck	eggbox	glue	holepuncher	mAP	推理速度/ms
MobileNet[28]	0.75	0.1	0.31	0.48	0.67	0.1	0.82	0.61	0.26	0.66	0.47	**3.2**
YOLOX-tiny[24]	0.91	0.69	0.71	0.90	0.80	0.32	0.91	0.44	0.27	0.91	0.69	6.9
YOLOX-s[24]	0.91	**0.90**	**0.99**	0.91	0.91	**0.91**	0.91	0.90	0.90	0.91	**0.92**	7.3
YOLOv4-tiny[29]	0.83	0.79	0.92	0.90	0.85	0.63	0.88	0.80	0.67	0.83	0.81	3.7
YOLOv3-tiny[18]	0.85	0.82	0.93	0.92	0.86	0.67	0.90	0.82	0.67	0.87	0.83	3.5
YOLO-Fusion	0.94	0.69	0.77	0.88	0.92	0.72	0.93	0.88	0.85	0.91	0.85	4.7
SPA-YOLO-Fusion	**0.95**	0.80	0.90	**0.94**	**0.93**	0.84	**0.97**	**0.96**	**0.91**	**0.93**	0.91	4.8

表 2 YODO_Det检测精度对比

检测模型	can	duck	eraser	eye-drops	glasses-box	glue	mAP	推理速度/ms
MobileNet	0.36	0.82	0.90	0.56	0.63	0.45	0.62	**3.0**
YOLOX-s	**0.96**	**0.99**	**0.99**	**0.91**	0.90	**0.99**	**0.96**	7.1
YOLOv3-tiny	0.34	0.78	0.77	0.56	0.92	0.64	0.67	3.4
YOLOv4-tiny	0.35	0.66	0.59	0.50	0.90	0.55	0.59	3.5
SPA-YOLO-Fusion	0.87	**0.99**	**0.99**	0.88	**0.99**	**0.99**	0.95	4.6

了1.2 ms的推理损耗,但提高了28 %的推理精度,尤其增加了对can、eye-drops和glue等与背景相近物体的检测精度.

2.3 密集遮挡的抓取检测实验

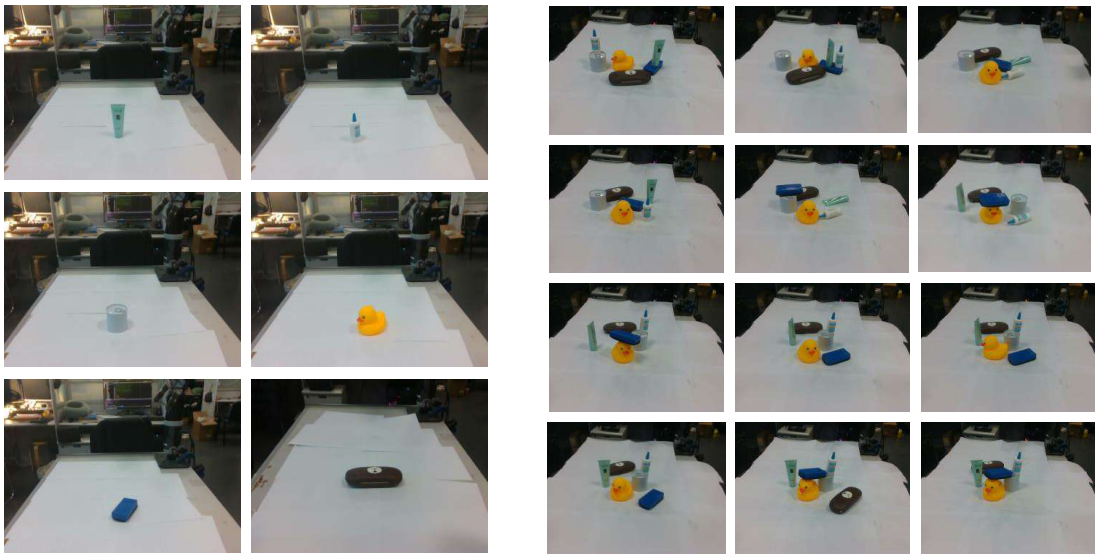
2.3.1 YODO_Grasp抓取检测数据集

GraspSeg[23]数据集中缺少堆叠场景图像,且该数据集暂未开源,因此,本文在实验室场景下对YODO_Grasp数据集进行采集. 选择RGB-D单物体图像各3 000张作为训练集(如图5(a)所示),并在多物体图像上(如图5(b)所示)进行测试,测试集的图像共500张,且图像中各个物体之间存在不同程度的遮挡与堆叠情况.

2.3.2 实验结果与分析

本文使用模型的smooth-L1损失函数进行训练,训练策略与GR-ConvNet论文保持一致. 由于GG-CNN以及GR-ConvNet模型无法一次性检测多物体图像中所有物体的抓取点,为了公平比较,所有的抓取检测模型都使用SPA-YOLO-Fusion先进行目标识别,然后再进行抓取检测处理.

表3是抓取检测模型的实验结果,可见:GG-CNN检测模型推理速度较快但准确率较低;本文提出的背景填充方法优于原始图像裁剪的方法,改进前后的抓取检测模型的准确率提高了23 %,而且并不会增加模型的推理损耗.



(a) 训练集(单物体图像) (b) 测试集(多物体图像)

图5 YODO_Grasp抓取检测数据集

表3 YODO_Grasp抓取检测结果

算法	准确率/%	推理速度/ms
GG-CNN (extension)	56	5
GG-CNN (padding)	73	5
GR-ConvNet (extension)	62	14
GR-ConvNet (padding)	85	14

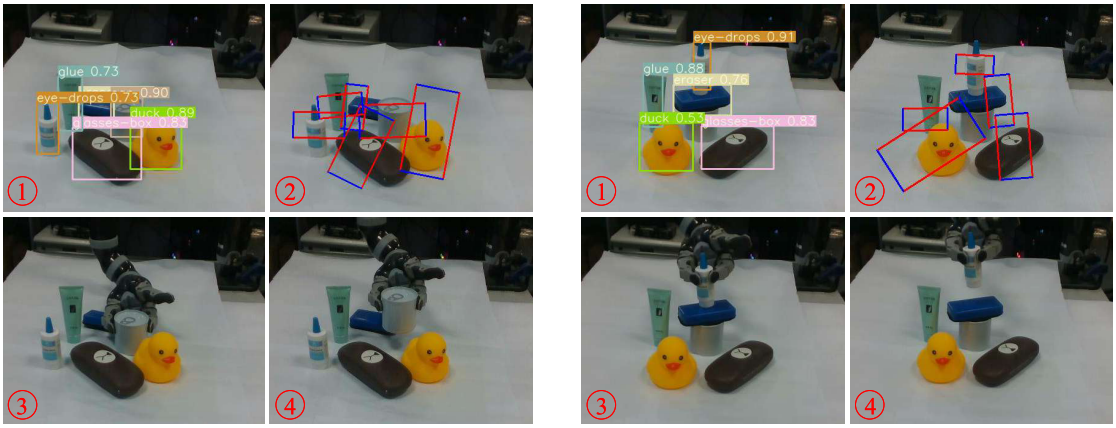
2.4 抓取实验与分析

本文主要研究抓取检测,因此,机械手臂的运动规划、密集物体的抓取策略等不是本文的研究重点. 针对密集遮挡物体的抓取,本文采用较为简单的策略:首先抓取目标检测物体置信度较大的物体,如果置信度相同,则优先抓取距离相机较近的物体. 表4列出使用了Jaco2机械手臂对检测出正确抓取框的物体进行50次抓取的成功率. 不难发现,机械臂对配备白色背景填充的预处理算法的GG-CNN以及GR-ConvNet模型的抓取成功率较高,较改进之前的抓取检测模型而言,分别有2 %和3 %的准确率提升.

表4 实际抓取结果

算法	成功率/%
GG-CNN (extension)	90
GG-CNN (padding)	92
GR-ConvNet (extension)	90
GR-ConvNet (padding)	93

图6是使用SPA-YOLO-Fusion模型进行目标检测、GR-ConvNet模型进行抓取检测,并使用Jaco2机械手臂进行抓取的分阶段流程图. 实际过程中依然存在抓取失败的案例,比如对于glue等物体,由于该类物体表面过于光滑且重心过高,夹爪与该物体之间的摩擦阻力不够,因而未能成功夹起物体.



(a) can抓取示意图　　(b) eye-drops抓取示意图

图 6　实际抓取示意图

3　结　论

针对密集遮挡场景的抓取检测以及人工标注图像费时且昂贵的问题,本文提出了基于RGB-D图像融合的目标检测与抓取检测的分步处理的思路,提供了YODO_Det目标检测数据集以及YODO_Grasp抓取检测数据集,并且在单物体图像中训练的抓取检测模型可以在密集遮挡场景中进行应用. 首先使用本文所提出的SPA-YOLO-Fusion检测模型对密集遮挡场景图像进行目标检测,其次对每个物体都使用GR-ConvNet模型进行抓取检测,最后从每个物体中选择一个最优的抓取框即可. 其中:本文提出的SPA多尺度融合方式可以有效地提高检测器的检测精度并降低漏检率;所提出的背景填充的方式可以有效降低密集遮挡场景下其他物体对目标物体进行抓取估计的干扰,从而提高抓取检测的准确率. 未来工作将继续研究使用RGB-D图像融合方式提高抓取检测模型的成功率.

参考文献(References)

[1] 刘亚欣, 王斯瑶, 姚玉峰, 等. 机器人抓取检测技术的研究现状[J]. 控制与决策, 2020, 35(12): 2817-2828.
(Liu Y X, Wang S Y, Yao Y F, et al. Recent researches on robot autonomous grasp technology[J]. Control and Decision, 2020, 35(12): 2817-2828.)

[2] 张云洲, 李奇, 曹赫, 等. 基于多层级特征的机械臂单阶段抓取位姿检测[J]. 控制与决策, 2021, 36(8): 1815-1824.
(Zhang Y Z, Li Q, Cao H, et al. Single-stage grasp pose detection of manipulator based on multi-level features[J]. Control and Decision, 2021, 36(8): 1815-1824.)

[3] Du G G, 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.

[4] Kleeberger K, Bormann R, Kraus W, et al. A survey on learning-based robotic grasping[J]. Current Robotics Reports, 2020, 1(4): 239-249.

[5] Wang H, Sridhar S, Huang J W, et al. Normalized object coordinate space for category-level 6D object pose and size estimation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 2637-2646.

[6] 刘政怡, 段群涛, 石松, 等. 基于多模态特征融合监督的RGB-D图像显著性检测[J]. 电子与信息学报, 2020, 42(4): 997-1004.
(Liu Z Y, Duan Q T, Shi S, et al. RGB-D image saliency detection based on multi-modal feature-fused supervision[J]. Journal of Electronics & Information Technology, 2020, 42(4): 997-1004.)

[7] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. IEEE International Conference on Computer Vision. Venice, 2017: 2999-3007.

[8] Pan X J, Ren Y Q, Sheng K K, et al. Dynamic refinement network for oriented and densely packed object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 11204-11213.

[9] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 936-944.

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

[11] Li H C, Xiong P F, Fan H Q, et al. DFANet: deep feature aggregation for real-time semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 9514-9523.

[12] Redmon J, Divvala S, Girshick R, et al. You only

look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016 : 779-788.

[13] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE, 2015: 1137-1149.

[14] Song S R, Xiao J X. Deep sliding shapes for amodal 3D object detection in RGB-D images[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 808-816.

[15] Ophoff T, Beeck K V, Goedemé T. Improving real-time pedestrian detectors with RGB+Depth fusion[C]. The 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, 2018: 1-6.

[16] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 6517-6525.

[17] Linder T, Pfeiffer K Y, Vaskevicius N, et al. Accurate detection and 3D localization of humans using a novel YOLO-based RGB-D fusion approach and synthetic training data[C]. IEEE International Conference on Robotics and Automation. Paris, 2020: 1000-1006.

[18] Farhadi A, Redmon J. Yolov3: An incremental improvement[J/OL]. 2018, arXiv: 1804.02767.

[19] 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.

[20] Kumra S, Joshi S, Sahin F. Antipodal robotic grasping using generative residual convolutional neural network[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, 2020: 9626-9633.

[21] Fang H S, Wang C X, Gou M H, et al. GraspNet-1Billion: A large-scale benchmark for general object grasping[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 11441-11450.

[22] Wei W, Luo Y K, Li F Y, et al. GPR: grasp pose refinement network for cluttered scenes[C]. IEEE International Conference on Robotics and Automation. Xi'an, 2021: 4295-4302.

[23] Asif U, Tang J B, Harrer S. GraspNet: An efficient convolutional neural network for real-time grasp detection for low-powered devices[C]. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, 2018: 4875-4882.

[24] Ge Z, Liu S T, Wang F, et al. Yolox: Exceeding yolo series in 2021[J/OL]. 2021, arXiv: 2107.08430.

[25] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 658-666.

[26] She Q, Feng F, Hao X Y, et al. OpenLORIS-object: A robotic vision dataset and benchmark for lifelong deep learning[C]. IEEE International Conference on Robotics and Automation. Paris, 2020: 4767-4773.

[27] Hinterstoisser S, Holzer S, Cagniart C, et al. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes[C]. International Conference on Computer Vision. Barcelona, 2011: 858-865.

[28] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018 : 4510-4520.

[29] Bochkovskiy A, Wang C Y, Liao Hongyuanmark. Yolov4: Optimal speed and accuracy of object detection[J/OL]. 2020, arXiv: 2004.10934.

作者简介

李明 (1962−), 男, 教授, 博士生导师, 从事智能控制与机器视觉等研究, E-mail: liming@cumt.edu.cn;

鹿朋 (1996−), 男, 硕士生, 从事目标检测与抓取检测的研究, E-mail: lp_oreo@163.com;

朱龙 (1996−), 男, 硕士生, 从事目标检测与抓取检测的研究, E-mail: lzhu2020@qq.com;

朱美强 (1979−), 男, 副教授, 博士, 从事机器学习、机器人与计算机视觉等研究, E-mail: zhumeiqiang@cumt.edu.cn;

邹亮 (1987−), 男, 副教授, 博士, 从事机器视觉的研究, E-mail: liangzou@ece.ubc.ca.