控制与决策
Control and Decision

基于多层级特征的机械臂单阶段抓取位姿检测

张云洲， 李奇， 曹赫， 王帅， 陈昕

引用本文:
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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1840

基于多层级特征的机械臂单阶段抓取位姿检测

张云洲[1,2†], 李 奇[2], 曹 赫[2], 王 帅[2], 陈 昕[2]

(1. 东北大学 信息科学与工程学院，沈阳 110004；2. 东北大学 机器人科学与工程学院，沈阳 110169)

摘 要: 针对机械臂对尺寸变换、形状各异、任意位姿的未知物体抓取,提出一种基于多层级特征的单阶段抓取位姿检测算法,将物体抓取位姿检测问题视为抓取角度分类和抓取位置回归进行处理,对抓取角度和抓取位置执行单次预测. 首先,利用深度数据替换RGB图像的B通道,生成RGD图像,采用轻量型特征提取器VGG16作为主干网络;其次,针对VGG16特征提取能力较弱的问题,利用Inception模块设计一种特征提取能力更强的网络模型;再次,在不同层级的特征图上,利用先验框的方法进行抓取位置采样,通过浅层特征与深层特征的混合使用提高模型对尺寸多变的物体的适应能力;最后,输出置信度最高的检测结果作为最优抓取位姿. 在image-wise数据集和object-wise数据集上,所提出算法的评估结果分别为95.71 %和94.01 %,检测速度为58.8 FPS,与现有方法相比,在精度和速度上均有明显的提升.

关键词: 机械臂；未知物体；多层级特征；单次预测；最优抓取位姿

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1840 **开放科学(资源服务)标识码(OSID):**

引用格式: 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.

Single-stage grasp pose detection of manipulator based on multi-level features

ZHANG Yun-zhou[1,2†], *LI Qi*[2], *CAO He*[2], *WANG Shuai*[2], *CHEN Xin*[2]

(1. College of Information Science and Engineering，Northeastern University，Shenyang 110004，China；2. Faculty of Robot Science and Engineering，Northeastern University，Shenyang 110169，China)

Abstract: For a manipulator to grasp the novel objects with variable sizes, different shapes, and arbitrary poses, a single-stage grasp pose detection algorithm based on multi-level features is designed by taking the grasp position detection problem of objects as the grasp angle classification and grasp position regression processing, and performing a single prediction for grasp angle and grasp position. The RGD image is generated by replacing the blue channel of RGB image with depth data, and the lightweight feature extractor VGG16 is used as the backbone network. For the problem that the feature extraction ability of VGG16 is weak, the Inception module is used to design a network model with stronger feature extraction capability. Then, grasp position is sampled using the method of priori box on the feature map of different levels, and the adaptability of the model to the objects with variable sizes is improved through the combination of shallow features and deep features. Finally, the detection result with the highest confidence is output as the optimal grasp pose. The evaluation results of the proposed algorithm on the image-wise dataset and the object-wise dataset are respectively 95.71 % and 94.01 %, and the detection speed is 58.8 FPS, and the accuracy and speed are improved compared with the current methods.

Keywords: manipulator；novel objects；multi-level features；single prediction；optimal grasp pose

0 引 言

机械臂抓取是机器人在家庭和工业场景下进行人机合作的重要手段. 在动态变化的场景下,人类能够以准确、稳定和快速的方式进行抓取,但机器人实现准确抓取检测、运动规划以及可靠抓取执行仍是非常具有挑战性的. 机器人在抓取一个物体时,需要先找到物体的抓取位姿. 不适当的抓取位姿将导致在操纵物体期间的不稳定,因此,找到一种更准确的

收稿日期: 2019-12-31；**修回日期:** 2020-04-23.

基金项目: 中央高校基本科研业务费专项资金项目(N172608005, N182608004, N2004022)；装备可靠性重点实验室基金项目(61420030302)；辽宁省高校创新人才支持计划项目(LR2019027).

责任编委: 方勇纯.

†通讯作者. E-mail: zhangyunzhou@ise.neu.edu.cn.

抓取位姿检测方法是非常必要的.

机械臂抓取位姿检测是指,针对给定待抓取对象的情况下由机器人自主生成一个稳定的抓取位姿,并能成功地抓取给定对象的过程. 当前的方法主要分为基于分析的方法和基于经验的方法. 前者也称为硬编码,根据给定任务的需要来手动编程机器人. 这些控制算法仅针对特定任务以及特定环境来控制机器人,并且需要具有相关领域专业知识的人来进行建模[1]. 这种硬编码可以有效地完成任务,但存在很大的局限性,尤其是在模型仅限于少数人有能力构建,却需要频繁更改机器人编程的情况下,该方法显然不能大范围地应用到实际生活中.

非结构化环境仍然是智能机器人的一大挑战[2],需要复杂的分析方法来形成解决方案,然而模型的推导需要大量关于机器人任务的数据和相关专业知识,所以传统分析方法虽然有效但却耗时费力. 在这种情况下,基于经验的方法为机器人增强了认知能力和适应能力,同时减少或完全消除了对手动建模的需要[3]. 早期的经验方法采用经典形式,此时的机器人需要从示范中学习对任务的适应性和认知能力,各种非线性回归技术、高斯混合模型和支持向量机是与此背景相关的一些流行技术[4]. 尽管这些方法为机器人增强了认知能力,但任务的复制仅限于所演示的任务,因而不具有泛化性. 在实际生活场景中,物体往往呈现为尺度多变、种类繁多、形状各异,这些因素决定了机器人抓取检测方法需要具有一定的适应性和泛化性.

近年来,深度学习在目标检测[5]、场景理解[6]和自然语言处理[7]等领域取得了重大进展,展现了很强的特征提取能力. 在机器人抓取检测领域中,基于深度学习的方法逐渐取代了传统的基于经验的方法. Lenz等[8]最先在滑动窗口检测框架中引入神经网络作为分类器,以预测输入图像的图像块中是否存在稳定的抓取,并证明了二维图像中抓取表示可以投影到三维空间. 由于滑动窗口算法的计算成本太大,影响模型的整体实时性,Redmon等[9]通过局部约束的预测机制对图像的每个空间位置进行抓取检测,并利用深度数据替换蓝色通道形成RGD图像,通过AlexNet特征提取器直接回归获得检测结果. 虽然直接回归模型检测速度很快,但是通常会受到平均效应的影响. 例如,对铁板的直接回归模型的抓取位置回归将位于铁板的中心而不是其边缘,因此,直接回归模型很难推广到包含具有多个对象的图像.

Guo等[10]和Redmon等[9]都利用了局部约束预测机制,前者更进一步将每个网格单元与拥有不同尺度和纵横比的默认抓取矩形相关联. 杜学丹等[11]证明了通过引入锚框的不同尺度和纵横比可以提高精度,但该方法没有充分利用底层特征信息,导致在检测小尺寸目标物体的抓取位置时需要重新设计锚框. Chu等[12]在Guo等[10]的基础上将抓取检测过程分成两阶段进行:第1阶段利用区域提议网络提取抓取框所在的图像区域;在第2阶段利用ResNet[13]特征提取器对抓取角度进行细化分类,再将抓取角度与抓取框位置组合成一个输出. 虽然深层网络的应用以及两阶段检测使得抓取检测精度大幅度提升,但是深层网络的计算开销以及两阶段模型的计算开销限制了该方法在实际中的应用. 此外,目前已公开的基于深度学习的抓取位姿检测算法[8-12]都是利用最高层特征图进行最优抓取位姿估计,没有充分利用卷积神经网络浅层特征和深层特征,对尺度多变的待抓取物体的适应性较弱,存在漏检、错检等问题.

针对尺度多变、形状各异、任意位姿的未知物体,本文提出一种基于多层级特征的机械臂抓取位姿检测方法,进一步提升深度学习应用于未知物体抓取位姿检测的速度和准确率. 本文的主要贡献包括以下几个方面:

1) 考虑到抓取位姿检测的特点,将抓取位姿检测任务转化为抓取位置回归与抓取角度分类任务的组合,并且在利用先验框生成抓取位置候选区域的同时进行抓取角度分类和抓取位置回归,实现对抓取角度和抓取位置执行单次预测,从而提升抓取位置的检测速度.

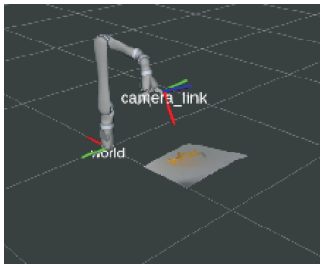
2) 在SSD(single shot multibox detector)[5]网络的基础上,通过引入Inception[14]模块增强网络整体对高判别性抓取特征的提取,并进行底层特征的融合以丰富浅层特征,设计了一种抓取位姿检测网络模型SSGD (single shot grasp detector).

3) 为了增强模型对尺度变换物体的适应能力,采用基于Inception模块的特征金字塔,将多层级特征应用于抓取位姿检测问题. 相比于仅使用最高层特征图的抓取检测方法[8-12],本文方法对不同大小的物体的适应能力更强,抓取检测精度更高.

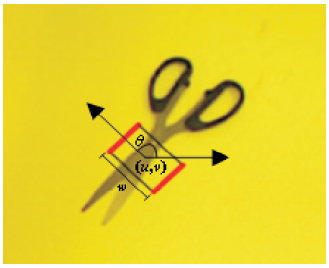
4) 在image-wise数据集和object-wise数据集上,抓取位姿检测模型的评估结果分别为95.71 %和94.01 %,检测速度为58.8 FPS,与现有方法相比在精度和速度上均有明显的提升. 在实际场景下的抓取实验结果表明,本文方法对于任意姿态的未知物体具有较强的适应能力和泛化能力.

1　问题描述

本文考虑在给定抓取场景的RGB图像和深度图像的情况下,在水平工作台上对种类不明、尺寸变换的物体进行抓取位姿检测. 在机械臂平面抓取任务中较为常用的两种抓取策略是顶抓策略和侧抓策略[15],本文采用顶抓策略进行平面抓取任务. 机器人抓取位姿检测任务如图1(a)所示.



(a)　抓取任务示意



(b)　抓取结果示意

图 1　抓取任务和结果示意

为了表示被抓取物体在图像上的抓取位置和姿态,本文采用Jiang等[16]的抓取框定义方法,使用四维向量表示抓取位置检测结果,如图1(b)所示. 四维抓取向量表示如下:

$$\boldsymbol{G}=\{u,v,w,\theta\}. \tag{1}$$

其中:坐标(u,v)表示像素平面坐标系下的抓取中心点;抓取角度θ作为矩形框顺时针旋转时与像素平面坐标系下u轴正方向的夹角,$\theta\in[0,\pi)$;w表示抓取框宽度,对应机器人末端执行器尝试抓取时的两指开度在图像上的投影.

像素坐标系与机器人坐标系之间的关系如下:

$$z_c\begin{bmatrix}u\\v\\1\end{bmatrix}=\boldsymbol{K}\begin{bmatrix}\boldsymbol{R}&\boldsymbol{T}\\0&1\end{bmatrix}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix}. \tag{2}$$

其中:z_c为抓取点在摄像机坐标系下的Z轴坐标值,由抓取点(u,v)在深度图像中对应的深度值转换所得;(X,Y,Z)为抓取点(u,v)在机器人坐标系下的坐标;$\boldsymbol{R}$、$\boldsymbol{T}$为通过手眼标定获得的相机坐标系与机器人坐标系之间的转换矩阵;$\boldsymbol{K}$为相机在针眼相机模型下的内参矩阵.

由式(2)可解得抓取点(u,v)在机器人坐标系下的空间坐标(X,Y,Z). 机器人坐标系下的抓取表示如下:

$$\widetilde{\boldsymbol{G}}=\{X,Y,Z,W,\Phi\}. \tag{3}$$

其中:W为机械臂末端夹具执行抓取时的开度;Φ为机械臂末端夹具在机器人坐标系下围绕Z轴顺时针旋转的角度,$\Phi\in[0,\pi)$.

基于上述抓取框的定义,抓取位置检测问题可描述为:t时刻在抓取候选空间$R^{3\times H\times W}$中,求解使得抓取检测模型置信度$M(G(t))$最大的抓取参数$G^*(t)$,即

$$G^*(t)=\underset{G(t)\in R^{3\times H\times W}}{\operatorname{argmax}}\ (M(G(t))). \tag{4}$$

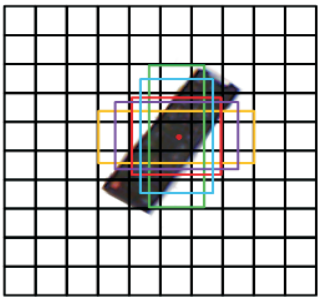
2　抓取位姿检测算法

影响抓取位姿检测算法性能的两个主要因素是抓取位置候选区域的生成和判别性抓取特征的提取. 在进行网络结构设计时,本文利用先验框(prior box)实现抓取位置区域的检测以及提取. 有别于之前的区域提取算法在同一层级上特征图中检测抓取位置区域,本文是从多个卷积层输出的特征图中生成抓取位置区域. 与此同时,为了提升网络检测速度和增强网络模型学习高判别性抓取特征的能力,本文结合Inception模块[14]设计一种基于多层级特征的抓取位姿检测网络.

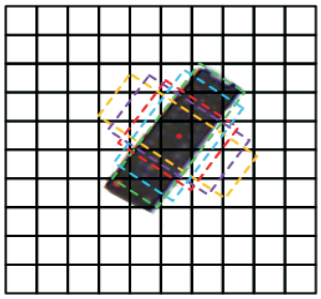
2.1　抓取位置候选区域生成

近年来,抓取位置候选区域的生成方法主要包括滑动窗口法和区域建议网络算法,都是在特征图上进行抓取区域搜索来粗略估计可能含有抓取位置的图像区域,然后利用卷积神经网络重新评估预选抓取框. 尽管区域建议网络比滑动窗口法更快一些,但是这种两阶段检测的方法计算开销大,限制了抓取检测算法的实际应用.

受单阶段目标检测算法SSD和RFBNet[17]的启发,本文通过共享来自主干网络的多层公共特征图,采用先验框的方法生成抓取候选区域,直接进行抓取位置回归和抓取角度分类,从而保证模型的推理速度,最后的输出表示如图2(b)所示. 先验框是指以特征图上的每个单元网格的中点为中心,生成一系列固定大小的同心矩形框. 此外,先验框在不同的特征图上具有不同的尺度,在同一特征图上又有不同的宽高比,因此,能够覆盖输入图像中的各种形状和大小的待抓取物体. 本文采用的先验框设置宽高比为$\{1,2,3,1/2,1/3\}$,如图2(a)所示.



(a) 抓取位置采样



(b) 抓取位置生成

图 2 抓取位置候选区域生成

先验框在不同层级的特征图上的尺寸计算如下:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1). \tag{5}$$

其中: $s_{\min}$ 取为0.2,表示最低层特征图的尺度为0.2; $s_{\max}$ 取为0.9,表示最高层特征图的尺度为0.9; m 取为5,表示共享的特征图数量; k 表示共享的特征图的序列.

为了能够对抓取位置进行更准确的检测,本文采用线性回归方法进行微调,该过程所需的平移量和尺度缩放参数计算公式如下:

$$\begin{aligned} &t_x = (x - x_a)/w_a,\ \hat{t}_x = (\hat{x} - x_a)/w_a;\\ &t_y = (y - y_a)/h_a,\ \hat{t}_y = (\hat{y} - y_a)/h_a;\\ &t_w = \log(w/w_a),\ \hat{t}_w = \log(\hat{w}/w_a);\\ &t_h = \log(h/h_a),\ \hat{t}_h = \log(\hat{h}/h_a). \end{aligned} \tag{6}$$

其中: x、y、w、h 分别表示预测抓取框的中心横纵坐标、宽度和高度, x_a、y_a、w_a、h_a 表示先验框的中心坐标、宽度和高度, t_x、t_y、t_w、t_h 表示预测抓取框的偏移量, $\hat{t}_x$、$\hat{t}_y$、$\hat{t}_w$、$\hat{t}_h$ 表示真值标定抓取框的偏移量.

2.2 抓取检测网络结构

本文的网络模型由主干网络、特征金字塔层和抓取预测层组成,SSGD网络结构如图3所示. 文献[12]使用ResNet-50或ResNet-101作为主干网络,因为ResNet具有从标记图像中学习高判别性抓取特征的能力,但这类深层卷积神经网络计算开销大、耗时严重,难以满足实时检测. 因此,本文采用轻量型网络——VGG16作为模型的主干网络.

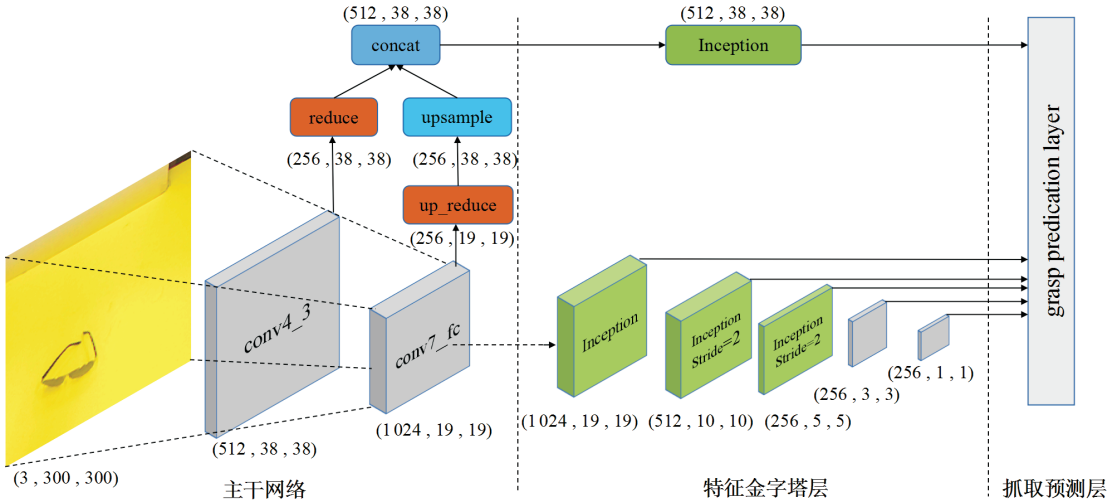


图 3 SSGD网络模型

为了解决轻量型网络存在特征提取能力弱的问题,本文引入了具有多分支卷积层的Inception模块,其结构如图4所示,不同大小的卷积的使用,可以使得网络拥有不同大小的感受野,增加了网络的宽度以及网络对尺度的适应性. 与使用5 × 5的大卷积核相比,使用级联的1 × 1、3 × 3的卷积核拥有相同的感受野,但具有更少的网络参数,shortcuts结构允许提取更深的特征信息. 总体上,Inception模块的引用有利于提升整体网络模型的特征提取能力.

在文献[8-10,12]的模型中,抓取位置都是直接在主干网络生成的最高层特征图上进行预测. 对于这种单一的特征图而言,感受野是非常有限的,而且没有充分利用前几级网络的特征信息. 同时,Cornell数据集中目标物体的尺寸大小变换很大,要求检测

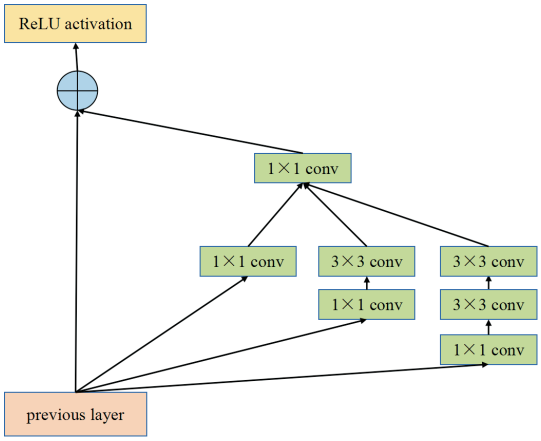


图 4　Inception 模块

模型对物体的尺寸大小具有良好的适应性. 因此,本文在抓取检测中引入了基于 Inception 模块的特征金字塔,进一步利用不同层级的特征信息,增强网络模型对尺度多变的物体的适应能力. 不同层级上的特征信息用途不同,网络中浅层特征有利于检测小尺度目标,深层特征适用于检测大尺度目标. 本文分别从不同的卷积层上提取出尺寸为(512, 38, 38)、(1 024, 19, 19)、(215, 10, 10)、(256, 5, 5)、(256, 3, 3)、(256, 1,1) 的 6 个特征图来构建特征金字塔层,以此提取浅层特征和深层特征,如图 3 所示. 为了缓解模型对小尺度物体的漏检与错检,本文将底层卷积 conv4_3 和 conv7_fc 输出的特征图进行融合,从而丰富浅层特征.

抓取预测层如图 5 所示,主要由两部分组成: 抓取角度分类层和抓取位置回归层. 全连接层是将整幅图像特征进行分类,因此,这两层都是利用 3×3 的卷积代替全卷积层进行分类. 特征图上每生成 k 个抓取候选区域时,抓取位置回归层便产生 $4\times k$ 个坐标,这 $4\times k$ 个坐标的含义已在式(6)中给出了解释. 同时,抓取角度分类层产生 $20\times k$ 个置信度,20 代表包括背景标签和 19 个角度标签在内的语义标签数量. 角度分类公式如下:

$$\theta_i=\left[(i-1)\times\frac{180^\circ}{19},i\times\frac{180^\circ}{19}\right],\ i\in[1,19]. \tag{7}$$

其中: θ_i 为抓取角度的分类标签, i 为抓取角度分类的类别数量, $\frac{180^\circ}{19}$ 表示抓取角度分类的区间长度.

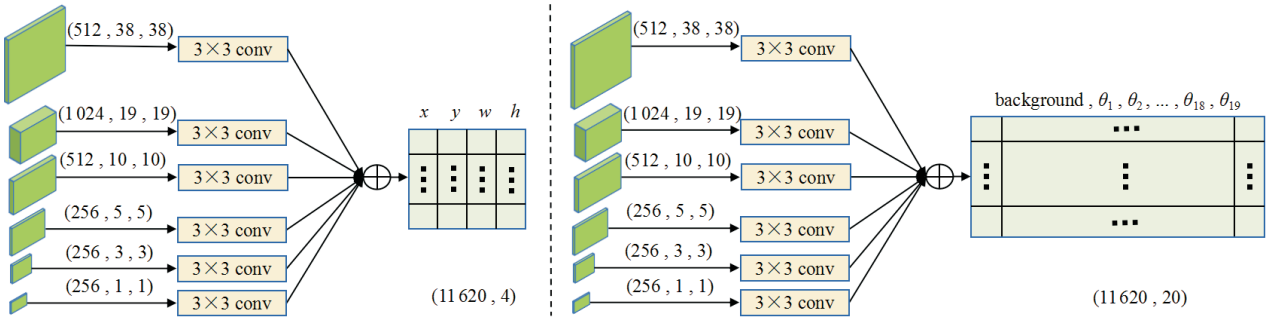


图 5　抓取预测层

关于最优抓取框的选择方法,为了测试模型的性能,本文仅输出置信度最高的抓取角度和相应的抓取位置. 文献[18-19]利用最优抓取位姿搜索算法提升检测精度,但是在搜索过程中花费了很多时间; 文献[12]利用阈值控制表示最优抓取位姿,但需要额外的非极大值抑制算法来搜索局部极大值,抑制非极大值元素,在提升检测精度的同时为算法增加了额外的计算开销. 因此,本文在检测最优抓取位姿时,直接输出置信度最高的结果,从而避免算法额外的计算开销.

2.3　损失函数定义

本文采用 smooth L_1 函数作为抓取位置回归损失函数. 通过式(6)得到平移量和尺度缩放参数,本文将抓取位置回归的损失函数 $L_{\rm greg}$ (the loss of grasp regression)定义如下:

$$L_{\rm greg}(t_m,\hat{t}_m)=\sum_{i\in P}^{N}\sum_{m\in\{x,y,w,h\}}\text{smooth}_{L_1}(t_m^{(i)}-\hat{t}_m^{(i)}). \tag{8}$$

其中: N 是与真值标定的抓取矩形相匹配的含有抓取位置区域的先验框的数量, t_m 是网络预测的抓取位置偏移量, $\hat{t}_m$ 是真值标定的抓取位置偏移量. 当先验框中不含有与真值标定相匹配的抓取位置区域时,抓取位置损失函数为零,不再考虑定位损失.

实际上,含有抓取位置区域的先验框仅占整个先验框的小部分,大多数先验框中并不含有抓取位置区域. 这种不平衡的比例会影响梯度更新的正确方向,导致抓取检测模型无法学习到有用的可抓取特征. 为了避免上述问题,本文采用启发式采样方法,对不包含抓取位置区域的先验框进行抽样. 抽样时按

照可抓取的置信度误差进行降序排列,选取误差较大的样本作为训练的负样本,保证正负样本比例接近1:3. 对于抓取分类的损失函数$L_{\rm gcls}$ (the loss of grasp classification),本文采用Softmax Loss函数,定义为

$$L_{\rm gcls}(\theta_{\rm P},\theta_{\rm N})=-\sum_{i\in P}^{N}\log(\hat{\theta}_{\rm P}^{(i)})-\sum_{i\in N}^{3N}\log(\hat{\theta}_{\rm N}^{(i)}),$$

$$\hat{\theta}_{\rm P}^{(i)}=\frac{\exp(\theta_{\rm P}^{(i)})}{\sum\limits_{\rm P}\exp(\theta_{\rm P}^{(i)})},\ \hat{\theta}_{\rm N}^{(i)}=\frac{\exp(\theta_{\rm N}^{(i)})}{\sum\limits_{\rm N}\exp(\theta_{\rm N}^{(i)})}. \tag{9}$$

其中:$\theta_{\rm P}$表示网络预测可抓取的置信度,$\theta_{\rm N}$表示网络预测不可抓取的置信度.

最后,本文的多任务损失函数L定义为

$$L(t_m,\hat{t}_m,\theta_{\rm P},\theta_{\rm N})=$$
$$\frac{1}{N}(L_{\rm greg}(t_m,\hat{t}_m)+L_{\rm gcls}(\theta_{\rm P},\theta_{\rm N})). \tag{10}$$

3 实验结果与分析

本文利用PyTorch深度学习框架构建网络模型以及相关代码. 在模型训练与评估过程中,所采用的实验平台为Nvidia GTX Titan Xp服务器. 在训练抓取位姿检测网络时,使用文献[17]提供的预训练模型,可以加快模型的收敛速度. 在训练过程中,使用基于SGD (stochastic gradient descent)的参数优化算法,批量(batchsize)设定为32,动量(momentum)设定为0.9,权重下降(weight decay)设定为0.000 5. 训练时采用“warmup”策略:前5个epochs,将学习率从10^{-6}逐步提升至0.004,然后保持不变. epoch = 75,每个epoch的迭代次数为2 200. 在进行模型评估时,使用单个GPU (Nvidia GTX Titan Xp)进行测试.

3.1 数据预处理

Cornell数据集包含了208类物体,合计885张图片数据和全局坐标系下的千万个点云数据. 图片与点云数据是对齐的,一共标注了8 019个抓取矩形. 本文采用文献[9]中的数据模态,即利用深度数据替换RGB图像中的蓝色通道,文献[9]的实验结果表明,多模态数据的应用有利于提升检测精度. 由于RGB数据位于0~255之间,深度信息被归一化到相同的范围,而在没有信息的深度图像上的像素被替换为0,在深度信息替换之后形成了黄色背景的图像,预处理生成的RGD图像如图6所示. 对于用来评估网络的图像,本文仅在RGD图像基础上对其进行300 × 300的中心裁剪,不再进行其他任何的数据增强. 与训练时图像批量输入的形式不同,测试阶段将图像逐一地输入到网络中.

图6 图像预处理

此外,与其他深度学习中的数据集相比,Cornell数据集是非常小的,因此在读入网络之前需要进行扩充. 首先,将图像中心裁剪为351 × 351;然后,裁剪后的图像在0° ~ 360°之间进行随机旋转,旋转的图像在x和y方向上随机平移最多50个像素;最后,再次进行中心裁剪,以获得300 × 300的图像. 预处理为每个图像生成100个增强数据,最终将图像以300 × 300的分辨率批量读入抓取位姿检测网络模型.

文献[8-10, 15-16, 20-25]在进行Cornell数据集训练时,采用五重交叉验证方法. 为了有效地与其他方法进行实验结果对比,本文将训练集与测试集的比例设置为4:1,将数据集分别采用以下两种方式进行分割:

1) image-wise分割:将图片随机分成训练集和测试集,这主要是为了测试网络模型在检测同一物体在不同位置、不同角度时的适应性;

2) object-wise分割:将数据集按照对象实例分割,利用之前没有出现在训练集中的数据测试模型对未知物体的泛化性.

3.2 评估度量方法

为了与现有的研究工作,如文献[8-10, 15-16, 20-25]进行比较,本文使用矩形度量来评估抓取位姿检测结果. 如果同时满足以下两个条件,则预测的结果被认为是一个很好的抓取参数:

1) 预测的抓取角度与真值标定的抓取角度之间的角度差在30°以内.

2) 真值标定的抓取框与预测的抓取框的Jaccard指数大于25 %. Jaccard指数的定义为

$$J(S_G,S_{\hat{G}})=\frac{|S_G\bigcap S_{\hat{G}}|}{|S_G\bigcup S_{\hat{G}}|}. \tag{11}$$

其中:S_G为预测的抓取框面积,$S_{\hat{G}}$为真值标定的抓取框面积. $S_G\bigcap S_{\hat{G}}$是这两个抓取矩形面积的交集,$S_G\bigcup S_{\hat{G}}$是这两个抓取矩形面积的并集. 在本文

中, Jaccard 指数仅用作模型的评估, 并不是抓取框选择的度量方法.

3.3　模型评估结果

为了对比 SSGD 模型性能改进的程度, 采用 SSD300 模型进行本文的基线实验, 在相同条件下利用 Cornell 抓取数据集进行测试, 得到的对比结果如表 1 所示. 表 1 结果显示, 相比于 SSD300 模型, SSGD 模型在 image-wise 数据集和 object-wise 数据集上的准确率分别提升了 15.28 % 和 12.88 %, 检测时间增加了 4 ms, 所以 SSGD 模型在处理抓取位置检测问题时比 SSD300 更加有效, 通过牺牲较小的检测时间的代价, 可以改善模型对高判别性抓取特征的提取能力, 最终提升模型的检测精度.

表 1　抓取位姿检测方法的基线实验

方法	准确率 / %		时间 / s
	image-wise	object-wise	
SSD300[5]	80.43	81.13	**0.013**
ours	**95.71**	**94.01**	0.017

本文将 SSGD 模型与仅使用最高层级特征的文献 [8-10, 15-16, 20-25] 在 Cornell 抓取数据集上进行了比较. 对比结果由表 2 可见, 相比于其他 10 种方法, 本文所提出的方法在两种类型的数据集测试中识别率最高, 在 image-wise 数据集和 object-wise 数据集上的准确率分别达到了 95.71 % 和 94.01 %. 与文献 [8-10, 15-16, 20-25] 相比, 本文模型所提取的特征信息更加丰富多样, 浅层特征与深层特征混合使用, 提高了模型对不同尺寸物体的适应能力, Inception 模块的引入有效地改善了抓取位姿检测精度. 与此同时, image-wise 的评估结果表明, 本文模型对于不同摆放位置和角度的物体具有较强的适应性; object-wise 的评估结果表明, 本文模型即使在处理未曾出现在训练集中的待抓取物体时, 也能够进行准确有效的抓取位姿检测, 对未曾参与训练的物体具有很好的泛化性.

表 2　不同抓取位姿检测方法的结果

方法	准确率 / %	
	image-wise	object-wise
Jiang[16]	60.50	58.30
Lenz[8]	73.90	75.60
Redmon[9]	88.00	87.10
Zhang[20]	88.90	88.20
Kumra[21]	89.21	88.96
Guo[10]	93.20	89.10
Asif[22]	90.60	90.20
Xia[15]	93.80	91.30
Yu[25]	94.10	93.30
Wang[23]	94.42	91.02
ours	**95.71**	**94.01**

表 3 所示为在更严格的 Jaccard 指数条件下的评估结果. 虽然 SSGD 模型性能在更严格的评估度量条件下有所下降, 但即使在 IoU (intersection over union) = 35 % 条件下, 本文方法与其他方法[8-9, 16, 20] 相比也非常具有竞争力.

表 3　本文方法在不同 Jaccard 阈值下的准确率

数据集	Jaccard 阈值		
	25 %	30 %	35 %
image-wise	95.71	94.24	91.53
object-wise	94.01	90.86	87.25

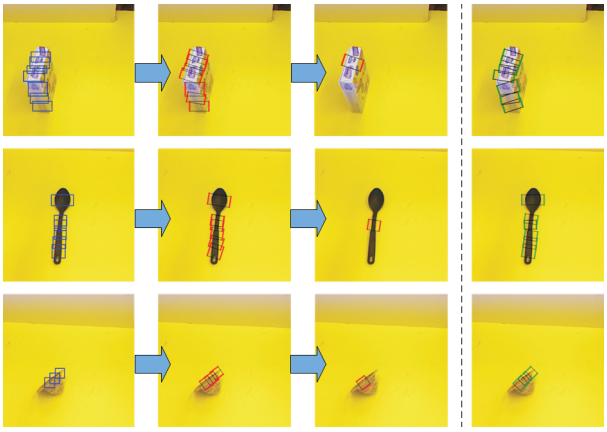
SSGD 模型与其他 6 种已公开检测方法的最快检测速度对比结果如表 4 所示. SSGD 模型测试一张图像的平均速度可达到 17 ms, 与表 4 结果中的最快检测速度相比, 本文模型的检测时间下降了 40 ms, 达到了实时检测的需求. 表 4 实验结果表明, 对抓取角度和抓取位置执行单次预测, 可有效地提升抓取位置检测速度.

表 4　检测速度对比结果

方法	时间 / s
Jiang[16]	50.000
Lenz[8]	13.500
Redmon[9]	0.076
Zhang[20]	0.117
Kumra[21]	0.103
Xia[15]	0.057
ours	**0.017**

3.4　模型测试结果

在 image-wise 测试集下, 可视化 Cornell 抓取数据集中的一些物体的真值标定和检测结果如图 7 所示. 其中: 图 7(a) 所示的蓝色矩形为抓取搜索过程回归的抓取候选区域, 它们与真值标定的抓取框之间的 IoU > 0.25, 抓取角度与真值标定的抓取角度之间的角度差在 30° 以内; 图 7(b) 中的矩形为抓取角度分类后的抓取框, 矩形的黑色边长表示机器人末端夹持器的两指张开尺寸, 矩形的红色边长与图像 X 轴的正方向夹角为抓取角度分类结果; 图 7(c) 中的抓取框为置信度最高的候选抓取框; 图 7(d) 所示抓取框为 Cornell 数据集的标定. 从图 7(b) 可以看出, 本文模型预测的抓取覆盖了大多数的真值标定, 图 7(c) 中的最优抓取结果也非常具有代表性, 完全符合模型评估度量标准.



(a) 区域搜索 (b) 角度分类 (c) 检测结果 (d) 真值标定

图 7 抓取位姿检测结果

在模型测试过程中,本文发现虽然某些预测的抓取矩形框不满足矩形度量的评估条件,但这些预测仍然可行,如图8所示,本文将这类预测结果称为假阴性抓取检测结果. 这是因为Cornell抓取数据集中的抓取位姿标定并不全面,难免会出现假阴性结果的产生,从而导致模型评估结果下降. 因此,本文模型的准确率远远高于目前所报告的精确度.

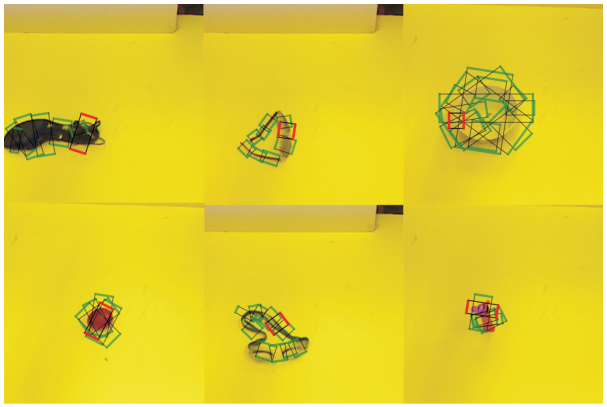


图 8 抓取位姿检测假阴性结果

在待抓取物体尺度多变的情况下,SSGD模型的检测效果如图9所示. 浅层输出特征可以检测较小的待抓取目标,深层输出的特征可以检测较大的待抓取目标,因此,SSGD模型通过结合多层级特征图检测结果,使得抓取检测模型对待抓取物体的尺度变换具有适应性.

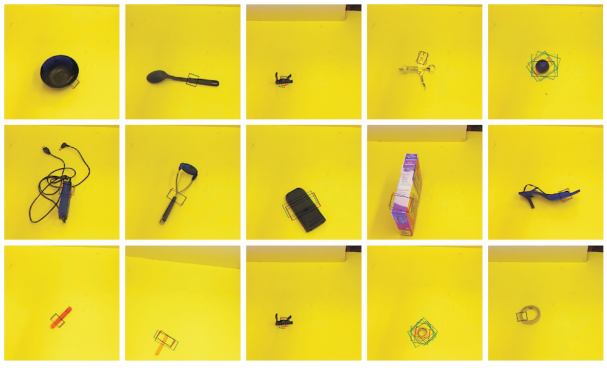


图 9 尺度多变物体的检测效果

3.5 机械臂抓取测试

为了在实际环境中确定并测试抓取位姿检测算法的效果,本文建立了图10所示的机器人抓取系统. 抓取执行机构为Jaco机械臂,基座是机械臂的固定装置,图10中标注的数字为机械臂的6个自由度关节,其中机械臂的手腕与末端夹具相连. 在抓取实验过程中,将Intel® RealSense Camera SR300相机固定在机械臂上面,用于采集目标的彩色图像和深度图像.

图 10 实际抓取位姿检测场景

由式(2)可得,在给定检测的抓取位置像素坐标以及相应深度值时,可求得抓取位置的空间坐标. 然后根据检测得到的抓取角度θ,通过采用顶抓策略控制机械臂末端夹具围绕空间坐标系Z轴顺时针旋转Φ,获得物体在空间坐标下的抓取位姿,具体流程如下:

1) 初始化机械臂位姿及夹具开度. 设机械臂在笛卡尔空间下的初始位姿为$(x_0, y_0, z_0, rx_0, ry_0, rz_0)$.

2) 根据抓取位姿检测结果,执行相应位姿. 控制机械臂移动至目标位置(x, y, z),夹具围绕Z轴顺时针旋转角度Φ.

3) 闭合末端夹具,将待抓取物体拾起.

4) 控制机械臂将目标物体放置在指定位置后返回初始位姿.

5) 若需要进行连续抓取,则返回步骤1),否则结束抓取任务.

抓取实验一共选取10种物体,都是训练数据集中所不存在的,对于机器人均为未知物体. 在实验过程中,将物体按照不同方向随机放置在不同位置,然后对每个物体进行10次抓取,抓取过程如图11所示,图像从左到右分别表示抓取位姿检测、抓取执行、抓取拾起过程. 抓取位姿检测结果如图12所示.

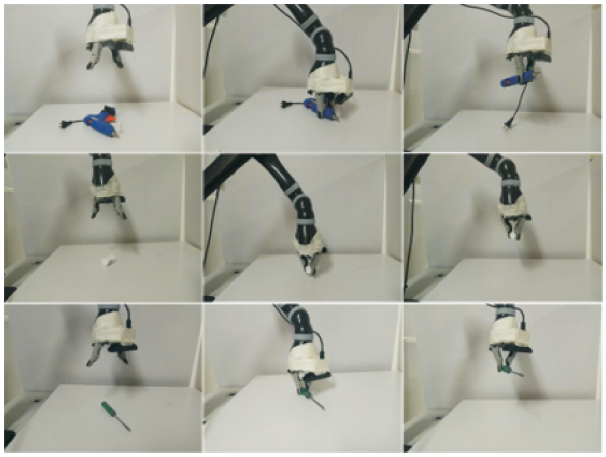


图 11　机器人抓取过程

抓取实验结果如表 5 所示, 从有限的代表性实验中得出, 机器人抓取平均成功率为 94 %, 测试一张图像的平均速度可达到 17 ms. 实验结果表明, 单阶段的抓取检测算法能够快速地计算出抓取位姿, 并验证了基于多层级特征的 SSGD 模型的泛化能力, 能够检测未知物体的最优抓取位姿, 从而对未知物体进行抓取. 本文算法能够有效地检测未知物体的抓取位姿, 但是在热胶枪和药瓶上分别出现了两次抓取失败的结果. 这是因为热胶枪的形状影响到了物体质心位置, 并且热胶枪本身重量也影响着抓取过程; 而在抓取尺寸较小的药瓶时出现了两次抓取失败的结果, 表明本文算法对小尺度物体的检测精度还有待于加强, 需进一步丰富浅层特征.

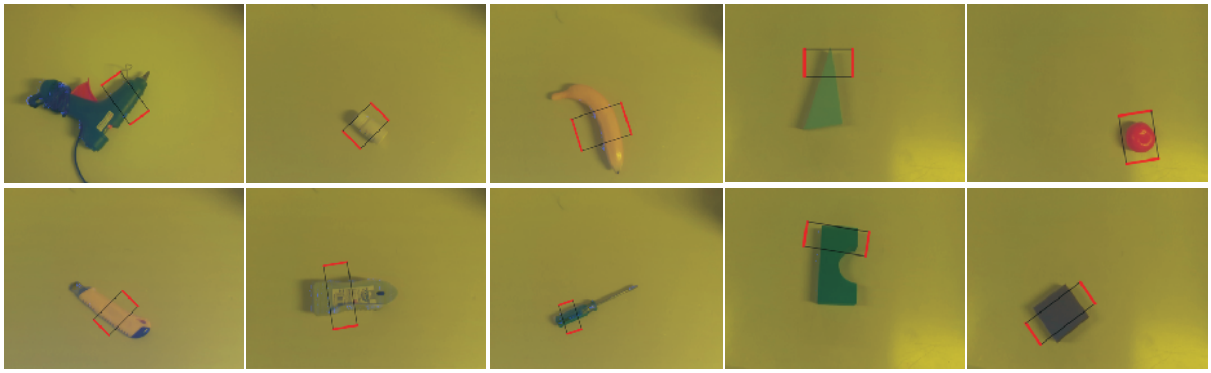


图 12　实际抓取位姿检测结果

表 5　实际抓取统计结果

物体	抓取次数	成功次数	成功率 / %
香蕉	10	10	100
刷子	10	9	90
热胶枪	10	8	80
美工刀	10	10	100
药瓶	10	8	80
螺丝刀	10	10	100
橘子	10	10	100
拱形积木	10	10	100
角形积木	10	9	90
方形积木	10	10	100
总计	100	94	94

4　结　论

本文在 SSD 网络的基础上进行优化, 设计了一种特征提取能力更强的抓取位姿检测网络模型 SSGD, 并提出了一种基于多尺度特征的单阶段抓取位姿检测算法. 算法利用先验框从多级卷积层输出的特征图中生成抓取位置候选区域, 并通过 SSGD 网络模型对抓取位置候选区域进行抓取角度分类以及抓取位置回归, 最终得到稳定的平面抓取位姿. 实验表明: 本文设计的单阶段抓取检测算法速度为 58.8 FPS, 在 image-wise 数据集和 object-wise 数据集上的评估结果分别为 95.71 % 和 94.01 %, 检测精度和实时性较以前的方法均有明显的提升; 并且在实际场景中能够对训练中未出现的未知物体进行抓取位姿检测, 从而实现对未知物体的抓取; 对于尺度多变的物体也具有较强的适应能力.

参考文献 (References)

[1] Kober J, Peters J. Imitation and reinforcement learning[J]. IEEE Robotics & Automation Magazine, 2010, 17(2): 55-62.

[2] Ju Z F, Yang C G, Li Z J, et al. Teleoperation of humanoid baxter robot using haptic feedback[C]. 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems. Beijing: IEEE, 2014: 1-6.

[3] Konidaris G, Kuindersma S, Grupen R, et al. Robot learning from demonstration by constructing skill trees[J]. The International Journal of Robotics Research, 2012, 31(3): 360-375.

[4] Billard A G, Calinon S, Dillmann R. Learning from Humans[C]. Springer Handbook of Robotics. Cham:

Springer International Publishing, 2016: 1995-2014.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[6] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]. IEEE International Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6230-6239.

[7] Yang M C, Lee D G, Park S Y, et al. Knowledge-based question answering using the semantic embedding space[J]. Expert Systems with Applications, 2015, 42(23): 9086-9104.

[8] 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.

[9] Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks[C]. IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2015: 1316-1322.

[10] Guo D, Sun F C, Liu H P, et al. A hybrid deep architecture for robotic grasp detection[C]. IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2017: 1609-1614.

[11] 杜学丹, 蔡莹皓, 鲁涛, 等. 一种基于深度学习的机械臂抓取方法[J]. 机器人, 2017, 39(6): 820-828.
(Du X D, Cai Y H, Lu T, et al. A robotic grasping method based on deep learning[J]. Robot, 2017, 39(6): 820-828.)

[12] Chu F J, Vela P A. Deep grasp: Detection and localization of grasps with deep neural networks[J]. 2017, arXiv: 1802.00520.

[13] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. LasVegas: IEEE, 2016: 770-778.

[14] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-ResNet and the impact of residual connections on learning[J]. 2016, arXiv: 1602.07261.

[15] 夏晶, 钱堃, 马旭东, 等. 基于级联卷积神经网络的机器人平面抓取位姿快速检测[J]. 机器人, 2018, 40(6): 794-802.
(Xia J, Qian K, Ma X D, et al. Fast planar grasp pose detection for robot based on cascaded deep convolutional neural networks[J]. Robot, 2018, 40(6): 794-802.)

[16] Jiang Y, Moseson S, Saxena A. Efficient grasping from RGBD images: Learning using a new rectangle representation[C]. IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2011: 3304-3311.

[17] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[C]. Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 404-419.

[18] 仲训昱, 徐敏, 仲训昱, 等. 基于多模特征深度学习的机器人抓取判别方法[J]. 自动化学报, 2016, 42(7): 1022-1029.
(Zhong X G, Xu M, Zhong X Y, et al. Multimodal features deep learning for robotic potential grasp recognition[J]. Acta Automatica Sinica, 2016, 42(7): 1022-1029.)

[19] Mahler J, Liang J, Niyaz S, et al. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics[J]. 2017, arXiv:1703.09312.

[20] Zhang Q, Qu D K, Xu F, et al. Robust robot grasp detection in multimodal fusion[C]. MATEC Web of Conferences. France: EDP Sciences, 2017, 139: 00060.

[21] Kumra S, Kanan C. Robotic grasp detection using deep convolutional neural networks[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2017: 769-776.

[22] Asif U, Tang J B, Harrer S. Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices[C]. International Joint Conferences on Artificial Intelligence. Stockholm: IJCAI, 2018: 4875-4882.

[23] Wang S F, Jiang X, Zhao J, et al. Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images[C]. IEEE International Conference on Robotics and Biomimetics (ROBIO). Piscataway: IEEE, 2019: 474-480.

[24] Park D, Chun S Y. Classification based grasp detection using spatial transformer network[J]. 2018, arXiv: 1803.01356.

[25] 喻群超, 尚伟伟, 张驰. 基于三级卷积神经网络的物体抓取检测[J]. 机器人, 2018, 40(5): 762-768.
(Yu Q C, Shang W W, Zhang C. Object grasp detecting based on three-level convolution neural network[J]. Robot, 2018, 40(5): 762-768.)

作者简介

张云洲(1974−), 男, 教授, 博士生导师, 从事智能机器人、计算机视觉等研究, E-mail: zhangyunzhou@ise.neu.edu.cn;

李奇(1996−), 男, 硕士生, 从事机器人抓取位姿检测的研究, E-mail: liqi008@foxmail.com;

曹赫(1995−), 男, 博士生, 从事机械臂抓取的研究, E-mail: caohe17@outlook.com;

王帅(1995−), 男, 硕士生, 从事基于单目视觉的跟随机器人算法的研究, E-mail: shuaiwang@stumail.neu.edu.cn;

陈昕(1996−), 男, 硕士生, 从事机械臂抓取的研究, E-mail: 1901936@stu.neu.edu.cn.

(责任编辑：李君玲)