第 42 卷 第 4 期
2021 年 4 月

仪 器 仪 表 学 报
Chinese Journal of Scientific Instrument

Vol. 42 No. 4
Apr. 2021

DOI: 10.19650/j.cnki.cjsi.J2107383

基于 Keypoint RCNN 改进模型的物体抓取检测算法*

夏浩宇,索双富,王 洋,安 琪,张妙恬
(清华大学机械工程系 北京 100084)

摘 要:机器人抓取在工业中的应用有两个难点:如何准确地检测可抓取物体,以及如何从检测出的多个物体中选择最优抓取目标。本文在 Keypoint RCNN 模型中引入同方差不确定性学习各损失的权重,并在特征提取器中加入注意力模块,构成了 Keypoint RCNN 改进模型。基于改进模型提出了两阶段物体抓取检测算法,第一阶段用模型预测物体掩码和关键点,第二阶段用掩码和关键点计算物体的抓取描述和重合度,重合度表示抓取时的碰撞程度,根据重合度可以从多个可抓取物体中选择最优抓取目标。对照实验证明,相较原模型,Keypoint RCNN 改进模型在目标检测、实例分割、关键点检测上的性能均有提高,在自建数据集上的平均精度分别为 85.15%、79.66%、86.63%,机器人抓取实验证明抓取检测算法能够准确计算物体的抓取描述、选择最优抓取,引导机器人无碰撞地抓取目标。

关键词: 抓取检测;Keypoint RCNN 改进模型;损失权重;注意力模块;抓取描述;重合度;最优抓取
中图分类号: TP391 TH74 **文献标识码**: A **国家标准学科分类代码**: 520.2060

Object grasp detection algorithm based on improved Keypoint RCNN model

Xia Haoyu, Suo Shuangfu, Wang Yang, An Qi, Zhang Miaotian
(*Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China*)

Abstract: There are two difficulties in the application of robot grasping in industry. How to detect the graspable object accurately and how to select the optimized grasp target among the detected multiple objects. In this paper the homoscedastic uncertainty is introduced into Keypoint RCNN to learn the weights of various losses, the attention modules are integrated into feature extractor, which composes the improved Keypoint RCNN model. A two-stage object grasp detection algorithm is proposed based on the improved Keypoint RCNN model. In the first stage, the improved model is used to predict the masks and keypoints. In the second stage, the masks and keypoints are used to compute the grasp representation and overlap rate of the object, the overlap rate represents the level of collision while grasping. According to the overlap rate, the optimized grasp target can be selected from multiple graspable objects. Comparison experiment indicates that the performances of the improved Keypoint RCNN model are improved in object detection, instance segmentation and keypoint detection compared with those of original model, and the average precisions (AP) on the self-built dataset reach 85.15%, 79.66% and 86.63%, respectively. Robot grasping experiment proves that the proposed grasp detection algorithm can accurately calculate the grasp representation, select the optimized grasp and guide the robot to grasp the target with collision-free grasp.

Keywords: grasp detection; improved Keypoint RCNN model; weight of loss; attention module; grasp representation; overlap rate; optimized grasp

0 引 言

机器人抓取在物流分拣、自动装配等领域有着广泛的应用,是实现自动化、智能化的关键技术,随着深度学习等技术的发展和机器人应用的推广,越发受到研究者们的关注。

根据机器人完成抓取任务的流程,机器人抓取可以细分为抓取检测、抓取规划和机器人控制 3 个子任务,其中最为关键的是抓取检测[1]。在抓取检测问题中,通常

收稿日期:2021-01-14 Received Date: 2021-01-14
*基金项目:国家重点研发计划(2017YYF0108101)项目资助

会利用深度相机等光学仪器感知外界环境,通过计算机视觉检测场景中的可抓取物体,预测目标物体的位置、姿态等信息,由这些信息组成的抓取描述可以引导机器人进行抓取。

基于传统视觉方法的抓取检测多采用点云配准的方法估计物体姿态。这要求已知目标物体的三维模型,将三维模型与实际场景中的目标物体进行配准,从而计算目标物体的空间位置和姿态。其中较为经典的算法包括 Hinterstoisser 等[2]提出的基于模板匹配的 Linemod 算法、Drost 等[3]提出的基于投票的点对特征算法(point pair feature,PPF)及其改进[4]和 Mellado 等[5]提出的基于共面 4 点仿射不变性的超四点一致性集合算法(Super 4 Points Congruent Sets,Super4PCS)。

现在越来越多的研究者开始运用深度学习解决机器人抓取问题。被广泛应用的目标检测算法可分为两阶段和一阶段算法两类。两阶段算法的典型是 He 等[6]提出的掩码区域卷积神经网络(mask region-based convolutional neural network,Mask RCNN),第一阶段生成可能包含物体的候选区域,第二阶段从候选区域中选择出最优结果,这类算法精度较高,但运算速度较慢。一阶段算法中 YOLO[7](You Only Look Once)和 SSD[8](Single Shot MultiBox Detector)较为常用,这类算法不生成候选区域,直接输出目标的检测框位置,运算速度相对较快。

为了计算物体的姿态等信息,目标检测算法还需与其它算法结合,Hernandez 等[9]训练了基于更快速区域卷积神经网络(faster region-based convolutional neural network,Faster RCNN)的网络以检测物体,然后使用 Super4PCS 估计物体姿态。Zeng 等[10]利用全卷积网络实现物体分割,用最近点迭代(iterative closest point,ICP)做物体姿态估计。陈丹等[11]提出了级联式 Faster RCNN 模型,第一级做目标检测,第二级预测物体最优抓取姿态。李秀智等[12]结合 YOLO 和多目标抓取检测网络,计算物体的最优抓取区域。

Mask RCNN 模型是一个多任务模型,在训练时各任务损失的权重对模型性能影响很大。Kendall 等[13]提出任务不确定性表示了不同任务间的相对置信度,通过每个任务的同方差不确定性来设置不同任务损失的权重。Sener 等[14]提出将多任务学习看作多目标优化问题,使总体目标变为寻找帕累托优化,使用基于梯度的多目标优化进行处理。

除了利用目标检测算法实现抓取检测,有研究者提出了端到端的深度学习模型,输入图像,模型输出物体的姿态或者抓取描述。Wang 等[15]提出了 DenseFusion 算法,用深度学习模型分别提取彩色图像和深度图像的特征图,将两者融合成一个特征,随后预测物体姿态并迭代优化,得到了很好的姿态估计效果,在机器人抓取实验中也验证了其有效性。Mahler 等[16]提出了两阶段算法 DexNet,先生成抓取候选,再用抓取质量卷积网络对抓取候选进行评估,选择最优抓取。马倩倩等[17]基于 SqueezeNet,参考 DenseNet 的网络结构,提出了 SqueezeNet 回归模型,可以直接预测物体的抓取位姿,且检测速度较快。

机器人抓取对深度学习模型的性能提出了较高要求,在图像领域中一个常用的提高模型性能的方法是注意力机制。Woo 等[18]提出了卷积块注意力模块(convolutional block attention module,CBAM),在特征提取器中,加入了通道层面和空间层面的注意力模块,改进了模型在分类和检测方面的性能。

本文提出了基于关键点区域卷积神经网络(keypoint region-based convolutional neural network,Keypoint RCNN)改进模型的物体抓取检测算法,引入了同方差不确定性学习多任务模型各损失的权重,并在特征提取器中加入了 CBAM,改进了 Keypoint RCNN 模型。用该模型进行实例分割和关键点检测,将预测的掩码和关键点用于计算抓取描述,再根据抓取时的重合度对抓取结果进行排序,得到无碰撞的最优抓取。将本文提出的算法应用于机器人抓取的一个具体场景,即纺纱企业络筒工序中管纱的自动抓取和上料,在机器人上进行的抓取实验表明,算法可以准确检测可抓取物体并选择最优抓取。

1　物体关键点检测模型

1.1　Keypoint RCNN 模型

Keypoint RCNN 模型是文献[6]在 Mask RCNN 的基础上增加关键点检测分支所形成的模型,主要包括特征提取器、区域候选网络和感兴趣区域头部网络,其结构如图 1 所示。该算法是一种两阶段模型,可以完成目标检测、实例分割和关键点检测 3 个任务。

1)特征提取器

Keypoint RCNN 采用残差网络(ResNet)作为特征提取的骨干网络,为使模型具有多尺度检测能力,Keypoint RCNN 采用了特征金字塔[19](feature pyramid network,FPN)结构,如图 2 所示。将残差网络分为 5 个阶段:stage1、stage2、stage3、stage4、stage5,每个阶段的特征图尺寸不同,后一阶段特征图长宽是前一阶段的 0.5 倍。stage5 的特征图通过步长为 2 的最大池化,长宽缩小为原来的 0.5,从此开始自上而下上采样,特征图上采样将长宽放大为 2 倍后,与前一阶段通过 1×1 卷积后的特征图加和,再通过 3×3 卷积。因为不使用 stage1 的特征图,上述操作产生 stage2、stage3、stage4、stage5 的新特征图,以及上采样起点的特征图,共 5 个特征图。特征金字塔网络使得特征同时具备强语义和强空间信息。

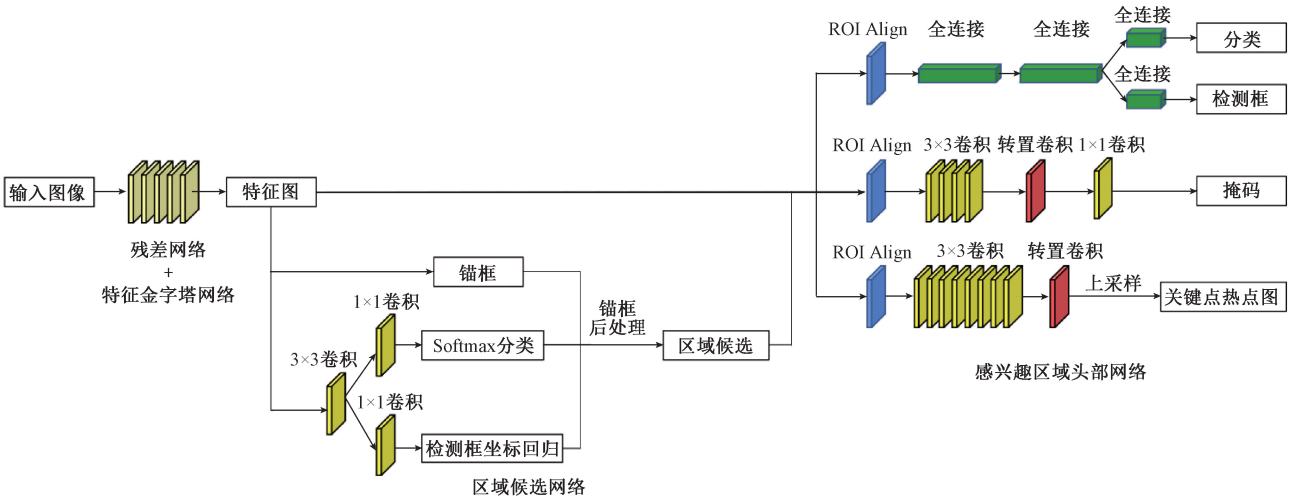


图 1 Keypoint RCNN 结构

Fig. 1 Architecture of Keypoint RCNN

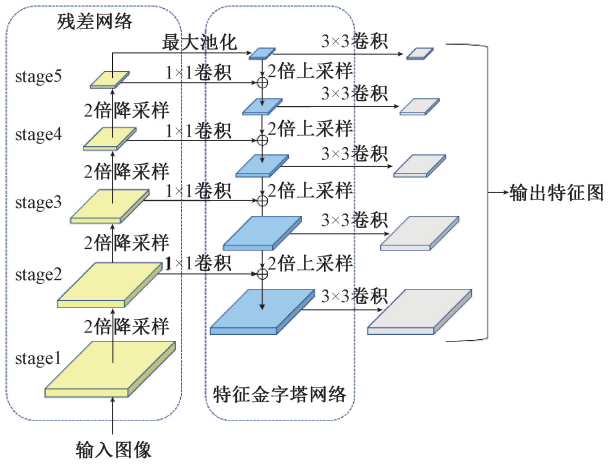


图 2 特征提取器结构

Fig. 2 Architecture of feature extractor

2) 区域候选网络

区域候选网络(region proposal network, RPN)的作用是生成一系列被分类为正例的锚框,并对该锚框的坐标进行修正,以供后续的网络对检测框和掩码等做更为精确的预测,其结构如图 3 所示。

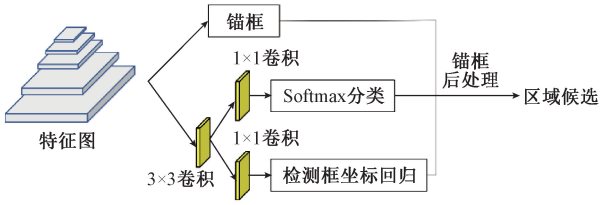


图 3 区域候选网络结构

Fig. 3 Architecture of RPN

该网络的输入是特征提取器输出的 5 个特征图,区域候选网络在特征图上的每个点生成锚框,锚框可以设置多种尺寸和长宽比,文献[6]中实际设置了 3 种尺寸和 3 种长宽比,共 9 种锚框。Softmax 层对锚框进行分类,判断锚框是否包含物体,即是否正例。同时用回归对正例锚框坐标做修正,使其更加接近标注。

得到了大量正例锚框后,为减少数据量,加速训练,将锚框按分类置信度排序,选取置信度高的锚框。然后裁剪超出特征图边界的锚框,删除尺寸小于阈值的锚框。由于网络可能会在同一个物体上生成多个锚框,还需进行非极大值抑制。

3) 感兴趣区域头部网络

RPN 网络输出的锚框尺寸不一,且坐标是对应在输入图像上的坐标,难以直接应用,因此需要感兴趣区域对齐(region of interest align, ROI Align)。首先从 5 个特征图选择一个,选择的依据为

$$level = level_0 + \log_2\left(\frac{\sqrt{area}}{s_0}\right) \tag{1}$$

式中:$level_0 = 4$, $s_0 = 224$, $area$ 为锚框面积,即当 $area = 224\times224$ 时,应当选择 stage4 的特征图。

将锚框坐标转换为对应在特征图上的坐标,将该特征图区域提取出来,重新划分并池化,得到固定尺寸的新特征图,在文献[6]中,尺寸为 7×7。ROI Align 层还用双线性插值解决了锚框坐标在上述过程中直接取整导致的位置偏差问题,提高了检测精度。

感兴趣区域头部网络(region of interest heads, ROI Heads)的输入是 ROI Align 层处理后的新特征图,输出模型在各任务上的损失或预测结果。Keypoint RCNN 中包含了检测框(bounding box)、掩码(mask)和关键点

(keypoint)3 个分支,如图 4 所示,分别对应目标检测、实例分割和关键点检测 3 个任务。

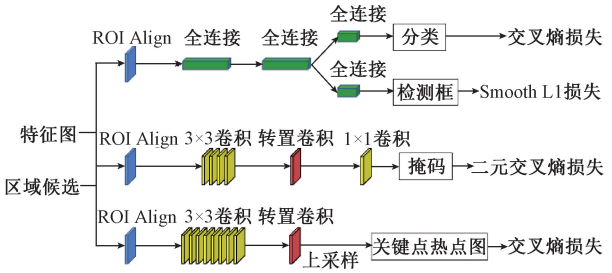


图 4　感兴趣区域头部网络结构

Fig. 4　Architecture of ROI Heads

检测框分支通过多个全连接层输出物体的分类置信度和检测框坐标回归值。用交叉熵损失计算分类损失,记为 $Loss_{cls}$,用 Smooth L1 Loss 计算检测框坐标回归损失,记为 $Loss_{box\_reg}$。

$$Smooth\_L1\_Loss = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & |x| \geqslant 1 \end{cases} \tag{2}$$

掩码分支通过 3×3 卷积和转置卷积得到 N×28×28×256 的特征图,N 为锚框数量,28 为特征图的长与宽,256 为通道数,通过 1 个 1×1 卷积降维,输出尺寸为 N×28×28×C,其中 C 为物体类别数。用输出与标注计算二元交叉熵损失,记为 $Loss_{mask}$。

关键点分支通过 3×3 卷积和转置卷积得到 N×28×28×C 特征图,其中 C 为关键点数量,通过双线性插值放大特征图,将标注转化为热点图,用特征图和热点图计算交叉熵损失,记为 $Loss_{kp}$。

Keypoint RCNN 的损失函数为

$$Loss = Loss_{obj} + Loss_{rpn\_box\_reg} + Loss_{cls} + Loss_{box\_reg} + Loss_{mask} + Loss_{kp} \tag{3}$$

式中:$Loss_{obj}$ 为 RPN 的正负例分类损失,$Loss_{rpn\_box\_reg}$ 为 RPN 的检测框回归损失。

1.2　Keypoint RCNN 改进模型

1)学习损失权重

Keypoint RCNN 是一个多任务模型,多个任务在训练时能否相互促进,提高模型的整体性能是个值得关注的问题。文献[6]中指出 Mask RCNN 的掩码分支可以提升模型目标检测中平均精度(average precision,AP),但关键点分支的加入使得模型在目标检测和实例分割的 AP 降低。

其原因之一是多任务模型的损失不平衡,关键点检测的损失明显高于目标检测和实例分割的损失,以本文训练的基准模型为例,各损失如图 5 所示,训练终止时,关键点检测的损失比其他损失大一个数量级。因此,关键点检测主导了梯度下降的方向,影响了其他任务的训练和优化。

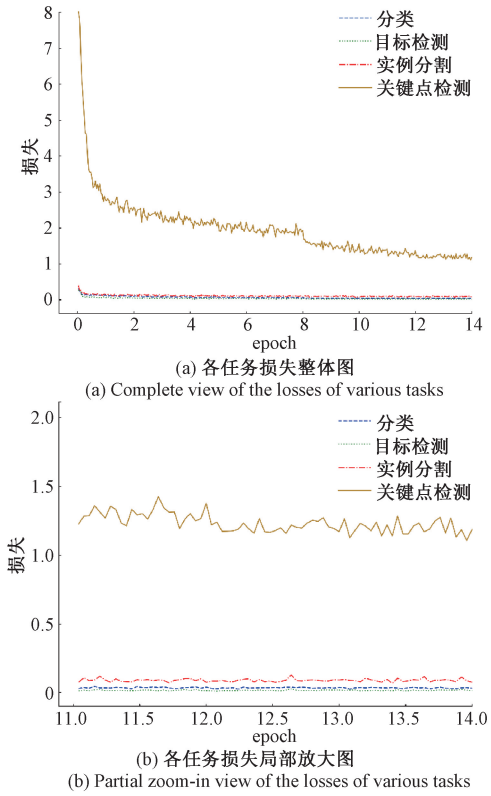


图 5　各任务损失

Fig. 5　Losses of various tasks

因此,参考 Kendall 等[13]的方法,通过引入同方差不确定性学习不同任务损失的最优权重。同方差不确定性是贝叶斯模型中偶然不确定的一种,对所有输入数据保持不变,在不同任务中的取值不同,因此同方差不确定性可以看作任务不确定性,表示不同任务间的相对置信度。

设模型的输入为 x,神经网络的参数为 W,网络输出为 $\boldsymbol{f}^W(x)$,回归问题的输出符合高斯分布,似然函数为:

$$p(y \mid \boldsymbol{f}^W(x)) = N(\boldsymbol{f}^W(x), \sigma^2) \tag{4}$$

其中,σ^2 是观测噪声。对数似然估计为:

$$\log p(y \mid \boldsymbol{f}^W(x)) \propto -\frac{1}{2\sigma^2} \| y - \boldsymbol{f}^W(x) \|^2 - \log \sigma \tag{5}$$

分类问题通常以 Softmax 函数为输出,似然函数为:

$$p(y \mid \boldsymbol{f}^W(x)) = Softmax(\boldsymbol{f}^W(x)) \tag{6}$$

符合玻尔兹曼分布,σ 为该分布的温度参数,对数似然估计为:

$$\log p(y = c \mid \boldsymbol{f}^W(x), \sigma) = \log Softmax\left(\frac{1}{\sigma^2}\boldsymbol{f}^W(x)\right) = \frac{1}{\sigma^2}\boldsymbol{f}_c{}^W(x) - \log \sum_{c'} \exp\left(\frac{1}{\sigma^2}\boldsymbol{f}_{c'}{}^W(x)\right) \tag{7}$$

其中,$\boldsymbol{f}_{c'}{}^W(x)$ 为特征向量 $\boldsymbol{f}^W(x)$ 的第 c' 项。

考虑两个输出的情况,设 y_1 为回归问题的输出,y_2

为分类问题的输出,则对数似然估计为:

$$L(\boldsymbol{W},\sigma_1,\sigma_2)=-\log p(y_1,y_2=c\mid \boldsymbol{f}^{\boldsymbol{W}}(x))=-\log N(y_1;\boldsymbol{f}^{\boldsymbol{W}}(x),\sigma_1^2)\cdot Softmax(y_2=c;\boldsymbol{f}^{\boldsymbol{W}}(x),\sigma_2)=\frac{1}{2\sigma_1^2}\|y_1-\boldsymbol{f}^{\boldsymbol{W}}(x)\|^2+\log\sigma_1-\log p(y_2=c\mid \boldsymbol{f}^{\boldsymbol{W}}(x),\sigma_2)=\frac{1}{2\sigma_1^2}L_1(\boldsymbol{W})+\frac{1}{2\sigma_2^2}L_2(\boldsymbol{W})+\log\sigma_1+\log\frac{\sum_{c'}\exp\left(\frac{1}{\sigma_2^2}f_{c'}^{\boldsymbol{W}}(x)\right)}{\sum_{c'}\exp\left(\frac{1}{\sigma_2^2}f_{c'}^{\boldsymbol{W}}(x)\right)^{\frac{1}{\sigma_2^2}}}\approx\frac{1}{2\sigma_1^2}L_1(\boldsymbol{W})+\frac{1}{2\sigma_2^2}L_2(\boldsymbol{W})+\log\sigma_1+\log\sigma_2 \tag{8}$$

式中:$L_1(\boldsymbol{W})=\|y_1-f^{\boldsymbol{W}}(x)\|^2$ 是 L2 损失,$L_2(\boldsymbol{W})=-\log Softmax(y_2,f^{\boldsymbol{W}}(x))$ 是交叉熵损失。

$L(\boldsymbol{W},\sigma_1,\sigma_2)$即为多任务模型的损失函数,通过上式可以学习各任务损失的权重。增大 σ 会使对应损失的权重减小,同时有惩罚项 $\log\sigma$ 控制,使其不会过度增大。

Keypoint RCNN 的损失为交叉熵损失和 Smooth L1 损失等之和,因此可以利用上式学习各损失的权重。在实际操作中,用数值更稳定的 $s=\log\sigma^2$ 代替 σ^2,因此 Keypoint RCNN 的损失函数改为:

$$Loss=Loss_{obj}+Loss_{rpn\_box\_reg}+\exp(-s_{cls})\cdot Loss_{cls}+s_{cls}+\exp(-s_{box\_reg})\cdot Loss_{box\_reg}+s_{box\_reg}+\exp(-s_{mask})\cdot Loss_{mask}+s_{mask}+\exp(-s_{kp})\cdot Loss_{kp}+s_{kp} \tag{9}$$

由式(9)可知,用该方法学习损失权重仅需在网络中增加 4 个可学习的参数,和网络的其他参数一起训练,且 4 个参数的运算、求导简单,相较于原 Keypoint RCNN 模型,引入了损失权重学习的新模型参数量和计算量增加很少。

2)加入 CBAM 注意力模块

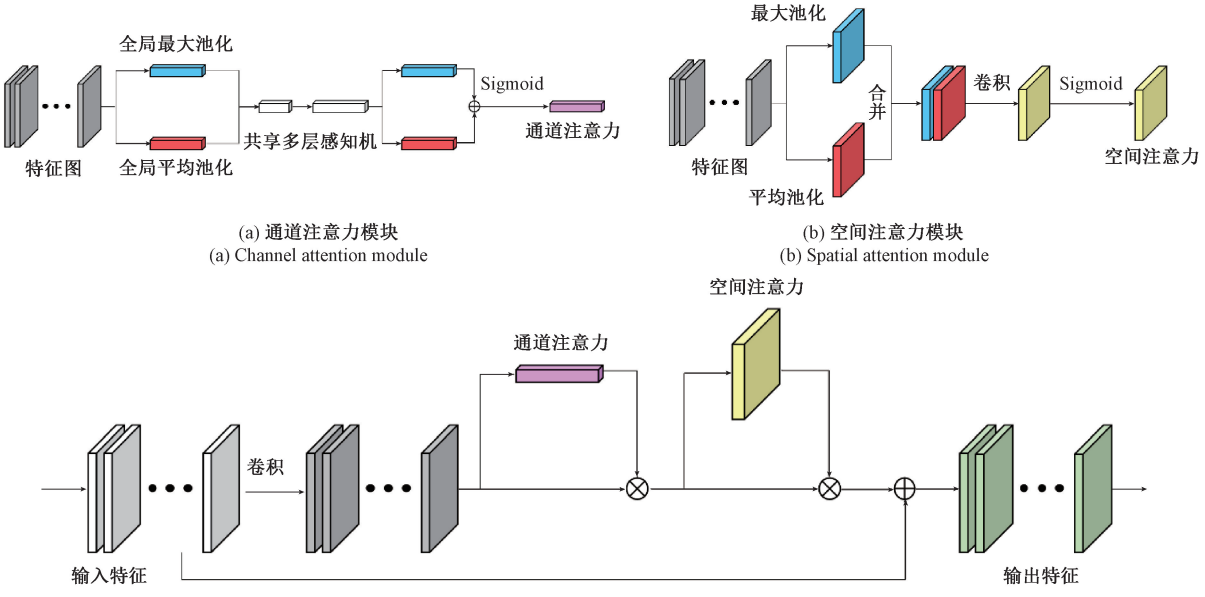
复杂抓取场景中物体间距离小、相互堆叠、机器人抓取时不应与物体发生碰撞,对抓取检测的精度要求高,因此加入 CBAM 注意力模块来提高模型性能。

CBAM 注意力模块是一种混合注意力机制,由通道注意力和空间注意力两个部分组成。通道注意力计算模型应该注意什么特征,空间注意力计算模型应该注意特征图什么位置。

设输入的特征图尺寸为 $N\times W\times H\times C$,$N$ 为批尺寸(batch size),W 为宽,H 为高,C 为通道数。

通道注意力机制的过程如图 6(a)所示,分别对特征图做全局平均池化和全局最大池化,得到两个尺寸为 $N\times1\times1\times C$ 的特征图。通过共享多层感知器得到 $N\times1\times1\times C$ 的特征图,将两特征图相加,得到通道注意力特征图。

空间注意力机制的过程如图 6(b)所示,分别对特征图做通道维度上的平均池化和全局池化,得到两个尺寸为 $N\times W\times H\times1$ 的特征图,将两者拼接形成 $N\times W\times H\times2$ 的特征图,再通过一个 7×7 卷积将通道数压缩为 1,得到的 $N\times W\times H\times1$ 的空间注意力特征图。CBAM 模块可以添加到残差网络中,先通道注意力再空间注意力,添加后的残差网络如图 6(c)所示。



(a) 通道注意力模块
(a) Channel attention module

(b) 空间注意力模块
(b) Spatial attention module

(c) CBAM在残差网络中的用法
(c) The usage of CBAM in ResNet

图 6 CBAM 的结构及其在残差网络中的用法

Fig. 6 Architecture of CBAM and its usage in ResNet

在本文中,由于输入图像尺寸较大,使用的显卡显存有限,因此批尺寸 $N=2$。Keypoint RCNN 改进模型的特征金字塔网络中,stage1 和 stage2 的特征图长宽分别为输入图像的 0.5 和 0.25 倍,如果在 stage1 和 stage2 中加入 CBAM 注意力模块,模型的参数量会明显增加,减慢训练和推理速度,因此仅在 stage3、stage4、stage5 中加入 CBAM 注意力模块,对应的通道数 C 分别为 512、1 024、2 048。

2　抓取描述生成与排序方法

2.1　抓取描述生成

Jiang 等[20]提出用包含抓取位置、朝向和夹爪宽度的 7 维向量描述抓取,本文将 $\boldsymbol{G}=(x, y, z, n_x, n_y, n_z)$ 作为抓取描述,其中(x, y, z)为抓取点坐标,(n_x, n_y, n_z)为抓取点法向量,本文将物体的一个特定关键点作为抓取点,具体在 3.2 节中说明。

假设使用的深度相机的焦距为 f_x, f_y,主点偏移为 p_x, p_y,,深度图上任意像素点 P_i 的坐标为(u_i, v_i),距离值为 d_i。则像素点 P_i 在相机坐标系下的坐标为:

$$\begin{cases} z_i = d_i \\ x_i = z_i \cdot (u_i + 1 - p_x)/f_x \\ y_i = z_i \cdot (v_i + 1 - p_y)/f_{yx} \end{cases} \tag{10}$$

因此在已知关键点图像坐标的情况下,可以计算得到其三维坐标。

需要计算关键点法向量,以确定机械臂的抓取时夹爪的姿态。求取关键点法向量的问题,可以转化为求关键点及其邻域构成的平面的法向量,用最小二乘法拟合平面求该法向量。

关键点及其邻域构成点集 $\boldsymbol{P}$:$\{x_1, x_2, \cdots, x_n\}$,设平面方程为:

$$ax + by + cz + d = 1 \tag{11}$$

其中 $d \geqslant 0$, $a^2 + b^2 + c^2 + d^2 = 1$

点集 $\boldsymbol{P}$ 中任意点到平面的距离为:

$$d_i = | ax_i + by_i + cz_i - d | \tag{12}$$

根据最小二乘法,使 d_i 之和最小,利用拉格朗日乘子法,将带约束的优化问题转化为无约束的优化问题。

$$f = \sum_{i=1}^{n} d_i^2 - \lambda (a^2 + b^2 + c^2 - 1) \tag{13}$$

对 d 求偏导

$$\frac{\partial f}{\partial d} = -2 \sum_{i=1}^{n} (ax_i + by_i + cz_i - d) = 0 \tag{14}$$

$$d = \frac{\sum_{i=1}^{n} x_i}{n} a + \frac{\sum_{i=1}^{n} y_i}{n} b + \frac{\sum_{i=1}^{n} z_i}{n} c \tag{15}$$

将 d 带入式(13)后,对 a 求偏导

$$\frac{\partial f}{\partial a} = 2 \sum_{i=1}^{n} [a(x_i - \bar{x}) + b(y_i - \bar{y}) + c(z_i - \bar{z})] \times (x_i - \bar{x}) - 2\lambda a \tag{16}$$

对 b、c 同理,令 $\Delta x_i = x_i - \bar{x}$,$\Delta y_i = y_i - \bar{y}$,$\Delta z_i = z_i - \bar{z}$。

令偏导为 0,将三式整理后可得:

$$\begin{bmatrix} \sum \Delta x_i \Delta x_i & \sum \Delta x_i \Delta y_i & \sum \Delta x_i \Delta z_i \\ \sum \Delta y_i \Delta x_i & \sum \Delta y_i \Delta y_i & \sum \Delta y_i \Delta z_i \\ \sum \Delta z_i \Delta x_i & \sum \Delta z_i \Delta y_i & \sum \Delta z_i \Delta z_i \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{17}$$

即 $\boldsymbol{Ax} = \lambda \boldsymbol{x}$,矩阵 $\boldsymbol{A}$ 为点集 $\boldsymbol{P}$ 的协方差矩阵。因此问题转化为求协方差矩阵的特征值和特征向量问题。

协方差矩阵 $\boldsymbol{A}$ 最小特征值对应的特征向量即为待求平面的法向量。

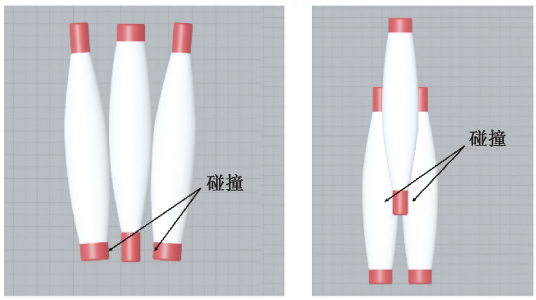
2.2　抓取描述排序

本文的待抓取物体管纱类似圆柱体,在实际场景中,大量管纱堆放在筐中,间距小、互相堆叠、场景复杂,如图 7 所示。因此抓取时夹爪容易与物体碰撞。

图 7　复杂抓取场景

Fig. 7　Complex grasp scene

常见的碰撞有两种形式,一种是物体紧密排列,抓取时夹爪与相邻物体碰撞,如图 8(a)所示;一种是物体堆叠在其他物体上,在特定堆叠方式下,夹爪与下方物体碰撞,如图 8(b)所示。



(a) 物体紧密排列
(a) Objects are arranged closely

(b) 物体互相堆叠
(b) Objectsare pileed up

图 8　抓取时发生碰撞的场景

Fig. 8　Collision scene while grasping

因此本文提出了通过计算夹爪与场景点云中特定区域的重合度来判断是否会发生碰撞。

根据 Keypoint RCNN 改进模型预测的两个关键点，可以计算管纱的朝向。已知模型预测的掩码轮廓，过直径较小端的关键点做与管纱朝向垂直的直线，与轮廓形成两个交点。根据上述两交点，可以得到抓取时夹爪在场景点云中所处区域，设区域中有 N 个点，在相机坐标系下，抓取点距离值为 p，区域中任意点距离值为 p_i，抓取时夹爪伸入长度为 l，如图 9 所示，重合度为：

$$overlap\_rate = \frac{\sum_{i=1}^{N} \min\left(\frac{p + l - p_i}{l}, 0\right)}{N} \tag{18}$$

若重合度小于阈值，则认为抓取时不会发生碰撞。本文实验中重合度阈值取 0.3。

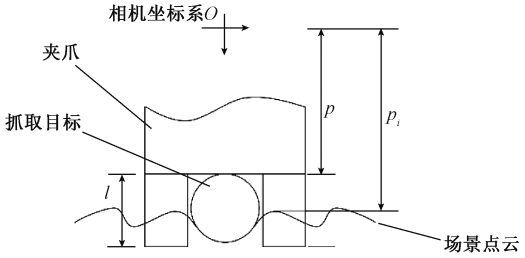


图 9 重合度计算示意图

Fig. 9 Schematic diagram of overlap rate computation

图像中一般包含多个物体，因此模型通常会检测出多个可抓取物体，根据重合度排序，仅抓取重合度小于阈值的物体，重合度最小的物体为最优抓取。

3 实验结果及分析

3.1 实验平台

实验使用的机器人为 Aubo i3 协作机器人，所用夹爪为二指平行夹爪，如图 10 所示。

使用的深度相机为 Intel Realsense D415，彩色图像最高分辨率为 1 920×1 080，深度图像最高分辨率为 1 280×720，为了匹配彩色和深度图像，将彩色和深度图像分辨率均设置为 1 280×720。

使用的 CPU 是 Intel Core i7-8750H@ 2.20 GHz×12，显卡是 Nvidia GeForce GTX 1070，在 Pytorch1.5.1 中运行深度学习模型，用 CUDA10.2 和 cuDNN8.0.2 加速，其他算法用 Python3.7 运行。

3.2 数据集与训练方案

实验使用的数据集为自建数据集，用 Intel Realsense D415 采集图像。数据集仅包含管纱一种物体。管纱是在塑料管上缠绕纱线而成的物体，类似圆柱体，标注如

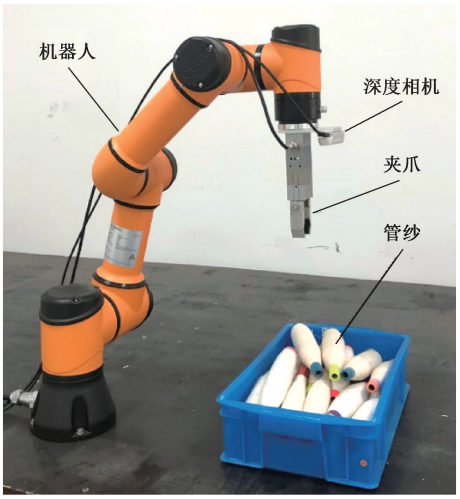


图 10 机器人平台

Fig. 10 Robot platform

图 11 所示，关键点定义在管纱的两头，其中直径小的、塑料的一头的关键点即为抓取点。管纱摆放在筐中，布置在不同背景中。数据集涵盖了在实际抓取场景中管纱可能出现的多种摆放情况：少量管纱松散摆放、没有堆叠的简单场景；多个管纱紧密摆放、存在堆叠的复杂场景。数据集中部分图像如图 12 所示。

图 11 掩码和关键点标注

Fig. 11 Mask and keypoint annotations

(a) 场景1 (b) 场景2

(a) Scene 1 (b) Scene 2

(c) 场景3 (d) 数据增强后的场景3

(c) Scene 3 (d) Scene 3 after data augmentation

图 12 自建数据集部分图像

Fig. 12 Partial images in self-built dataset

采集图像后对数据集做离线数据增强,随机旋转 45°、90°或 135°,再随机缩放为 0.75 ~ 1.2 倍,使数据增多到原来的 2 倍,按照 5∶1的比例划分训练集和测试集。

训练时,用在 COCO 数据集[21]上训练的模型作为预训练模型,用预训练模型的权重初始化 Keypoint RCNN 改进模型,预训练模型没有的部分采用 He 等[22]提出的 Kaiming 初始化。

用自建数据集对模型进行微调,学习率为 0.002 5,共训练 12 个 epochs,并在第 6 个 epoch 和第 10 个 epoch 时做学习率衰减,将学习率变为原来的 0.1。为保证网络在训练开始时损失不会过大,采用学习率热身策略,以 0.002 5/3 的学习率开始训练,逐渐线性增长到 0.002 5。

训练时的图像增强为随机水平翻转,概率为 0.5。

自建数据集按照 COCO 数据集的评估方式,给出模型在不同任务上的 AP 作为性能指标。

3.3　对照试验

为了验证本文提出的 Keypoint RCNN 改进模型的有效性,进行对照实验,对比 Mask RCNN、Keypoint RCNN、引入了权重的 Keypoint RCNN(Keypoint RCNN + Weight)、本文提出的 Keypoint RCNN 改进模型(Keypoint RCNN + Weight + CBAM),共 4 个模型,每个模型均生成 10 个随机种子,训练 10 次,将 10 次训练的 AP 平均值作为该模型的 AP,对照实验结果如表 1 所示。

表 1　对照实验数据

Table 1　Comparison experiment data

模型	检测框平均精度/%	掩码平均精度/%	关键点平均精度/%	每批图像训练时间/s	每张图像推理时间/s
Mask RCNN	81.13	77.38		0.693 9	0.130 2
Keypoint RCNN	79.98	76.41	86.16	1.084 4	0.139 8
Keypoint RCNN+Weight	84.52	78.91	85.53	1.086 2	0.138 3
Keypoint RCNN+Weight+CBAM	**85.15**	**79.66**	**86.63**	1.243 4	0.191 6

从对照实验可以看出,模型增加了关键点分支后,在目标检测和实例分割上的 AP 分别下降了 1.15% 和 0.97%,说明关键点损失过大影响了这两个任务的优化;此外还增大了模型的参数量,使每批图像的训练时间增加了约 0.4 s,每张图像的推理时间增加了约 0.01 s。

引入权重平衡损失后,两者都有不同程度的上升,分别比 Mask RCNN 高 3.39% 和 1.53%,说明权重不仅消除了关键点损失过大的影响,还使得目标检测损失和实例分割损失的权重比 Mask RCNN 模型更合理。而关键点检测的的 AP 有约 0.63% 左右的下降,这是由于加权后,关键点损失的权重小于 1,其他损失的权重大于 1,与加权前相比,各损失的数值无明显差距,关键点损失不再主导总损失的数值,因此可能不再得到更多的优化。加权后各损失如图 13 所示,虽然分类、目标检测和实例分割的损失变大,但由于式(9)中存在惩罚项,总损失逐渐下降。训练时间与原 Keypoint RCNN 模型基本相等,说明学习损失权重的开销很小。

加入 CBAM 注意力模块后,模型在 3 个任务上的 AP 均有提升,比加入前分别提升了 0.63%、0.72%、1.1%,说明加入 CBAM 模块能够有效提升模型的性能,但也增大了模型的参数量,使得模型的训练时间和推理时间高于其他模型。部分图像的检测结果如图 14 所示,图中管纱边缘实线表示掩码轮廓,实心点表示关键点,两点连线表示关键点连线。

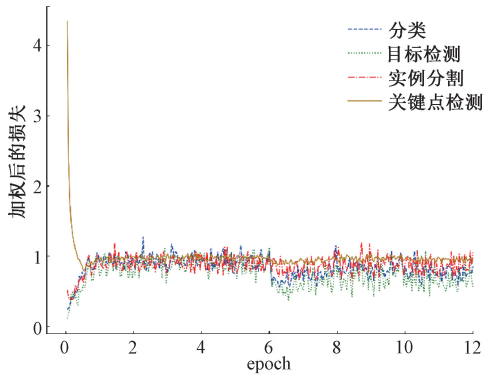
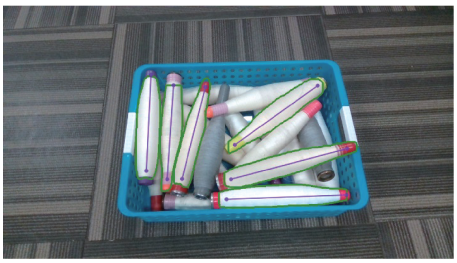


图 13　加权后的损失

Fig. 13　The losses after weighting

对照实验说明,与原模型相比,加入了权重和 CBAM 模块的改进模型虽然推理时间有所增长,但具备更好的检测性能,预测检测框、掩码和关键点的精度更高。而为了避免碰撞,复杂场景的抓取对检测精度要求高,因此本文提出的 Keypoint RCNN 改进模型更适合用于复杂场景的抓取检测。

(a) 简单场景检测结果
(a) Detection results of simple scene



(b) 复杂场景检测结果
(b) Detection results of complex scene

(c) 增强图像检测结果
(c) Detection results of augmented image

图 14 Keypoint RCNN 改进模型检测结果

Fig. 14 Detection results of the improved Keypoint RCNN model

3.4 抓取实验

为了验证本文提出的抓取检测算法的有效性,将算法部署在机器人上,按照数据集的图像构建实际抓取场景,每次将 25 个管纱随机摆放在筐中,机器人连续抓取管纱直至将筐中的管纱全部取出。共抓取了两筐、50 个管纱,两次抓取的成功率分别为 80%、84%,平均成功率为 82%,检测出的每个管纱的平均抓取描述生成和排序时间为 0.026 3 s。实验中某次抓取的检测结果如图 15 所示,其中物体的重合度如表 2 所示,其中上、下重合度分别表示管纱两侧的重合度。图 15 中,0、1、2 号管纱的重合度均小于阈值,抓取时无碰撞,其中 0 号管纱的平均重合度最小,因此是最优抓取,3 号管纱有一侧的重合度大于阈值,4 号管纱两侧重合度均大于阈值,抓取时可能发生碰撞,与实际情况相符。

实验中抓取失败的主要原因是检测出的所有管纱的重合度均大于阈值,没有最优抓取,此时需人工干预,重新摆放管纱。

实验结果表明 Keypoint RCNN 改进模型预测的掩码和关键点是准确的;抓取描述计算准确;按照重合度选择最优抓取合理且有效。本文提出的基于 Keypoint RCNN 改进模型的物体抓取检测算法可以在物体数量多、有堆叠的复杂场景中实现抓取检测,确定抓取顺序,更适合工业应用。

图 15 抓取描述排序示例

Fig. 15 An example of grasp representation sorting

表 2 图 15 对应管纱重合度数据

Table 2 Overlap rate data of the corresponding bobbins in Fig. 15

管纱编号	上重合度	下重合度	平均重合度	是否大于阈值
0	0.197 7	0	0.098 9	否
1	0.001 1	0.239 0	0.120 1	否
2	0.202 9	0.113 7	0.158 3	否
3	0.217 2	0.484 5	0.350 9	是
4	0.562 4	0.570 5	0.566 5	是

4 结 论

本文基于 Keypoint RCNN 模型,通过引入同方差不确定性学习模型各损失的权重,以及在特征提取器中加入 CBAM 注意力机制,改进了 Keypoint RCNN 模型,并根据模型预测的掩码和关键点,计算物体的抓取描述,根据场景点云和抓取时机器人夹爪在点云中的对应位置,计算两者的重合度,依据重合度排序选择无碰撞的最优抓取作为机器人的抓取目标。

将本文提出的基于 Keypoint RCNN 改进模型的物体抓取检测算法应用于具体的工业问题中,即纺纱企业络筒工序的管纱自动抓取和上料。在自建管纱数据集上的

对照实验中,相对于原 Keypoint RCNN 模型,Keypoint RCNN 改进模型在目标检测、实例分割和关键点检测任务上的 AP 都更高,证明对模型的改进提高了检测性能。将算法部署在机器人上后进行的抓取实验表明,本文提出的算法可以准确检测可抓取物体,并从多个抓取中选择无碰撞的最优抓取,能够满足复杂场景的检测需求,可以应用于实际的机器人抓取工作。

参考文献

[1] DU G, WANG K, LIAN S, 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: 1677-1734.

[2] HINTERSTOISSER S, LEPETIT V, ILIC S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[C]. Asia Conference on Computer Vision, 2012: 548-562.

[3] DROST B, ULRICH M, NAVAB N, et al. Model globally, match locally: Efficient and robust 3D object recognition [C]. Computer Vision and Pattern Recognition, 2010: 998-1005.

[4] HINTERSTOISSER S, LEPETIT V, RAJKUMAR N, et al. Going further with point pair features [C]. European Conference on Computer Vision, 2016: 834-848.

[5] MELLADO N, AIGER D, MITRA N. Super 4PCS fast global pointcloud registration via smart indexing[J]. Computer Graphics Forum, 2014, 33(5): 205-215.

[6] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. International Conference on Computer Vision, 2017: 2980-2988.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. Computer Vision and Pattern Recognition, 2016: 779-788.

[8] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]. European Conference on Computer Vision, 2016: 21-37.

[9] HERNANDEZ C, BHARATHEESHA M, KO W, et al. Team delft's robot winner of the amazon picking challenge 2016[C]. RobotCup2016: Robot World Cup XX, 2017: 613-624.

[10] ZENG A, YU K, SONG S, et al. Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge[C]. International Conference on Robotics and Automation, 2017: 1386-1393.

[11] 陈丹,林清泉. 基于级联式 Faster RCNN 的三维目标最优抓取方法研究[J]. 仪器仪表学报, 2019, 40(4): 229-237.
CHEN D, LIN Q Q. Research on 3D object optimal grasping method based on cascaded Faster RCNN[J]. Chinese Journal of Scientific Instrument, 2019, 40(4): 229-237.

[12] 李秀智,李家豪,张祥银,等. 基于深度学习的机器人最优抓取姿态检测方法[J]. 仪器仪表学报, 2020, 41(5): 108-117.
LI X ZH, LI J H, ZHANG X Y, et al. Detection method of robot optimal grasp posture based on deep learning[J]. Chinese Journal of Scientific Instrument, 2020, 41(5): 108-117.

[13] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]. Computer Vision and Pattern Recognition, 2018: 7482-7491.

[14] SENER O, KOLTUN V. Multi-task learning as multi-objective optimization[C]. Neural Information Processing Systems, 2018: 525-536.

[15] WANG C, XU D, ZHU Y, et al. DenseFusion: 6D object pose estimation by iterative dense fusion[C]. Computer Vision and Pattern Recognition, 2019: 3343-3352.

[16] 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[C]. Robotics: Science and Systems, 2017.

[17] 马倩倩,李晓娟,施智平. 轻量级卷积神经网络的机器人抓取检测研究[J]. 计算机工程与应用, 2020, 56(10): 141-148.
MA Q Q, LI X J, SHI ZH P. Research on light-weight convolutional neural network for robotic grasp detection[J]. Computer Engineering and Applications, 2020, 56(10): 141-148.

[18] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module [C]. European Conference on Computer Vision, 2018: 3-19.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Computer Vision and Pattern Recognition, 2017: 936-944.

[20] JIANG Y, MOSESON S, SAXENA A. Efficient grasping from RGBD images: Learning using a new rectangle representation[C]. International Conference on Robotics and Automation, 2011: 3304-3311.

[21] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context [C]. European Conference on Computer Vision, 2014: 740-755.

[22] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[C]. International Conference on Computer Vision, 2015: 1026-1034.

作者简介

夏浩宇(通信作者),2018 年于清华大学获得学士学位,现为清华大学硕士研究生,主要研究方向为深度学习与机器人感知。

E-mail:hzxiahaoyu@ foxmail. com

Xia Haoyu (Corresponding author) received his B. Sc. degree in 2018 from Tsinghua University. Now, he is an M. Sc. candidate in Tsinghua University. His main research interest includes deep learning and robot perception.

索双富,1984 年于中国矿业大学获得学士学位,1991 年于中国矿业大学获得硕士学位,1995 年于中国矿业大学获得博士学位,现为清华大学副教授,主要研究方向为图像处理与机器人控制。

E-mail:sfsuo@ tsinghua. edu. cn

Suo Shuangfu received his B. Sc. degree in 1984, M. Sc. degree in 1991 and Ph. D. degree in 1995 all from China University of Mining and Technology. Now, he is an associate professor in Tsinghua University. His main research interest includes image processing and robot control.