DOI:10.11918/202212037

融合注意力机制和多任务学习的机器人抓取检测算法

李钰龙,梁新武

(上海交通大学 航空航天学院,上海 200240)

摘 要:抓取主要分为抓取检测、轨迹规划和执行环节,准确的抓取检测是完成抓取任务的关键。为进行更准确的抓取检测,提高机器人抓取性能表现,本研究以关键点检测算法为基础,提出了一种融合注意力和多任务学习的抓取检测算法。首先,针对任务特点,在特征提取环节引入CA(coordinate attention)注意力模块,显式的学习通道和空间特征,充分利用特征信息。其次,在损失函数环节加入多任务权重学习算法,学习抓取中心坐标、抓手开合宽度及旋转角度信息的最优权重。最后,在Cornell数据集以及更大规模的Jacquard数据集上进行试验。研究结果表明,所提方法相比滑动窗口和锚框类型等经典方法在检测速率上有明显提升,且与单纯的关键点检测方法相比有更高的准确率,所提模型在两个数据集上分别取得98.8%和95.7%的准确率。检测示例体现出所提模型对于非常规物体也有良好的抓取结果,不同Jaccard系数条件下的抓取结果显示模型在精准抓取方面有优秀性能,而对于权重学习算法的不同初始值试验则表明所提模型具有良好的鲁棒性。此外,通过消融实验分析了不同模块对于模型性能表现的影响程度。

关键词:抓取检测;关键点估计;注意力机制;可学习权重;深度学习

中图分类号:TP241 **文献标志码**:A **文章编号**:0367-6234(2023)12-0009-09

Robotic grasp detection algorithm integrating attention mechanism and multi-task learning

LI Yulong, LIANG Xinwu

(School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Grasping is mainly divided into grasping detection, trajectory planning, and execution. Accurate grasping detection is the key to completing grasping tasks. In order to achieve more accurate grasping detection and improve the performance of robot grasping, this paper proposes a grasping detection algorithm that integrates attention and multi-task learning based on key point detection algorithm. Firstly, a coordinate attention (CA) attention module is introduced in the feature extraction process to explicitly learn channel and spatial features and make full use of feature information. Secondly, a multi-task weight learning algorithm is added to the loss function to learn the optimal weights of the grasp center coordinates, gripper opening width, and rotation angle information. Finally, experiments are conducted on the Cornell dataset and the larger-scale Jacquard dataset. The results show that the proposed method has a significant improvement in detection speed compared to classical methods such as sliding windows and anchor box types, and has higher accuracy compared to simple key point detection methods. The proposed model achieves accuracy rates of 98.8% and 95.7% on the two datasets, respectively. Grasping examples show that the proposed model also has good grasping results for unconventional objects, and the model has excellent performance in accurate grasping under different Jaccard coefficient conditions. Moreover, the experiments with different initial values of the weight learning algorithm show that the proposed model has good robustness. In addition, the impact of different modules on the performance of the model is analyzed through ablation experiments.

Keywords: grasp detection; key point estimation; attention module; learnable weights; deep learning

抓取对于机器人而言是一项非常重要的能力。如今,机器人在工业生产、医疗、军事等领域正在扮演越来越重要的角色,其中实现更精准快速的机器人抓取是不可或缺的能力。当人类看到一个未知物体,就可以直觉般的知道如何拿起它。即使已经有很多关于机器人抓取检测领域的工作[1-5],但对于

收稿日期:2022-12-12;**录用日期**:2023-03-03;**网络首发日期**:2023-11-06
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20231106.1140.003
基金项目:国家自然科学基金(62173230);上海市科技计划资助项目(22511101400)
作者简介:李钰龙(1997—),男,硕士研究生
通信作者:梁新武,xinwuliang@sjtu.edu.cn

机器人而言,如何实现鲁棒的目标抓取仍是个挑战。整个抓取流程主要分为抓取检测、轨迹规划和执行。抓取检测是机器人通过光学[3]或者触觉[6-7]等传感器获取周围环境的信息,并从中发现可抓取物体的过程。后续的抓取步骤需要使用抓取检测获得的物体坐标,所以准确的抓取检测结果是顺利实现整个抓取流程的前提和关键。

抓取检测主要分为分析法[8]和数据驱动法[9],分析法要求物体几何、物理模型和力学分析等数据已知,通过这些已知信息进行抓取检测。然而,对于实际需求场景来说,这些信息会经常变化,因此分析法无法成为一种好的通用性方案。数据驱动法利用先前的成功抓取经验数据,可以学习到通用的抓取能力。基于此,本文采用数据驱动法,利用深度学习方法提取物体的抓取信息。有代表性的工作之一是滑动窗口法[10],此方法需要一个分类器来预测每一个小区域是否含有抓取可能,其主要缺点是检测速率过低,很难在现实中使用。另一种代表性的工作是基于锚框的方法[11-13],该方法需要预测很多候选抓取框,通过评分选取最优的候选框。与基于锚框的抓取检测算法不同,本文基于关键点的算法直接通过中心点、宽度图、角度图来预测最优抓取,减少了计算步骤并且可以提高性能。

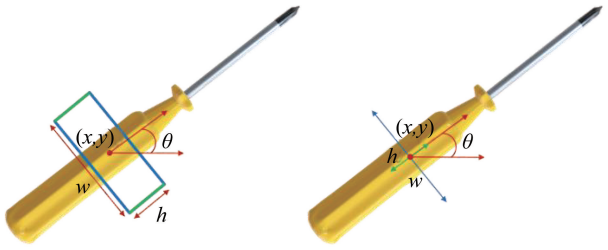
随着计算机视觉领域的发展,为了更好地利用数据,出现了很多注意力机制算法,例如基于通道的 SE(squeeze-and-excitation)注意力[14]和基于空间的 SK(selective kernel)注意力[15]。通道注意力通过显式地学习不同通道的权重信息,动态调整每个通道的权重。而空间注意力机制,则可显式地学习不同空间区域的权重。本文采用基于包含空间和通道注意力的 CA(coordinate attention)注意力[16]作为特征提取模块的补充,能够结合两种注意力机制的优点,并且避免了混合注意力机制 CBAM(convolutional block attention module)注意力[17]在空间上只关注局部位置关系的不足。对于一个抓取检测任务,需要分别学习抓取中心点坐标、角度、抓手开合宽度 3 种信息[10]。为了能够更好的权衡 3 种信息在抓取检测任务中所起的作用,本文引入多任务权重学习算法[18]自动学习这几种信息(即任务)的权重。该算法通过概率建模的思想来学习任务的最佳权重,根据任务不同,利用同方差不确定性[19]刻画多任务学习问题中损失函数的权重。本文设计了一个基于关键点检测算法的模型,其中引入了包含通道和空间注意力的 CA 注意力机制以及多任务权重学习算法。在流行的 Cornell 数据集以及更大规模的 Jacquard 数据集上对所设计的模型进行测试,分别获得了 98.8% 和 95.7% 的抓取检测成功率。对模型模块进行鲁棒性测试和消融实验,分析研究了所设计模型的不同模块对于整体性能的影响程度。

1 问题描述

给定一个物体,本文希望找到一种精确的方法来表示物体的抓取配置,通常用五维抓取来表示[10]。五维表示是七维表示[20]的简化版,在简化维度的同时保留了有效信息,通过抓取点坐标、抓取角度、抓手开合宽度来定义五维抓取表示:

$$\boldsymbol{g}=(x,y,\theta,h,w) \tag{1}$$

式中:(x,y) 对应于抓取矩形的中心点,θ 为抓取矩形关于水平方向的角度,h 为夹持器的高度,w 为抓取手的开合宽度。一个五维表示的例子如图 1(a)所示。因为本文的模型基于像素点,所以本文使用中心点图像、角度图像和宽度图像来定义一个新的抓取表示[21],如图 1(b)所示。



(a) 五维抓取表示 (b) 改进的抓取表示

图 1 抓取表示对比

Fig. 1 Comparison of grasp representations

具体地,中心点图中的像素数值表示此点作为可抓取点的可能性,用 $\boldsymbol{Q}$ 来表示中心点图像,其中每个像素的取值范围为[0,1],该值表示该点的抓取成功率,设数值最大点的坐标为(x,y),其即为抓取中心点。$\boldsymbol{W}$ 为宽度图像,其中的像素数值对应以所在位置为抓取中心时,抓手的开合宽度信息。$\boldsymbol{\Theta}$ 为角度图,其中的像素数值对应以所在位置为抓取中心时,夹具与水平方向之间的夹角。因为抓手是对称的,所以角度只需要取 $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$,文献[12]将角度值回归问题转化为分类问题,忽略了抓取的角度几何属性。本文将角度分为 $\cos(2\theta)$ 和 $\sin(2\theta)$ 分别学习,通过 $\theta=\frac{1}{2}\arctan\left(\frac{\sin(2\theta)}{\cos(2\theta)}\right)$ 计算出唯一角度值。由于特定机械手的硬件限制,h 通常是固定的,为方便与文献[21]比较,本文也将其设为当前抓取框宽度的 1/3。综上所述,抓取表示也可以表示为

$$\boldsymbol{g}=(\boldsymbol{Q},\boldsymbol{W},\boldsymbol{\Theta})\in\mathbf{R}^{3\times h\times w} \tag{2}$$

2　关键点检测模型

2.1　基于关键点的模型

本文使用的是基于关键点的检测算法，其中的代表是 CenterNet[22]。不同于基于锚框的检测算法，此类算法使用物体边界框的中心点来表示物体。该方法是端到端的，不需要非极大值抑制的过程。原 CenterNet 的上采样率为 4，可能会造成信息丢失，为了避免信息丢失，本文使用转置卷积[23]操作来生成与输入图像大小相同的特征图，从而实现更准确的像素级抓取检测。原始 CenterNet 的预测参数为中心点坐标，中心点偏移以及目标尺寸，考虑到在抓取检测任务中需要通过五维抓取表示定义抓取，即回归中心点坐标、抓手的旋转角度以及抓取器的开合宽度，本文将可以表示上述变量的中心图、角度图和宽度图作为新模型的预测参数。

2.2　模型结构概述

模型的总体结构如图 2 所示。先将 3 × 224 × 224 的 RGB 和 1 × 224 × 224 深度图在特征通道的维度拼接得到多模态的 4 × 224 × 224 尺度 RGB-D 图像，再将其输入到模型中。模型先通过 3 个卷积层进行下采样，然后经过带有注意力机制的残差网络组成的特征提取模块，再通过 3 个转置卷积层上采样，得到与输入图片同样大小的 224 特征热力图，随后通过 4 个卷积操作分别得到中心点坐标、以正弦和余弦表示的物体旋转角度以及抓取器的开合宽度。最后，经过后处理模块就能得到该物体的抓取表示。

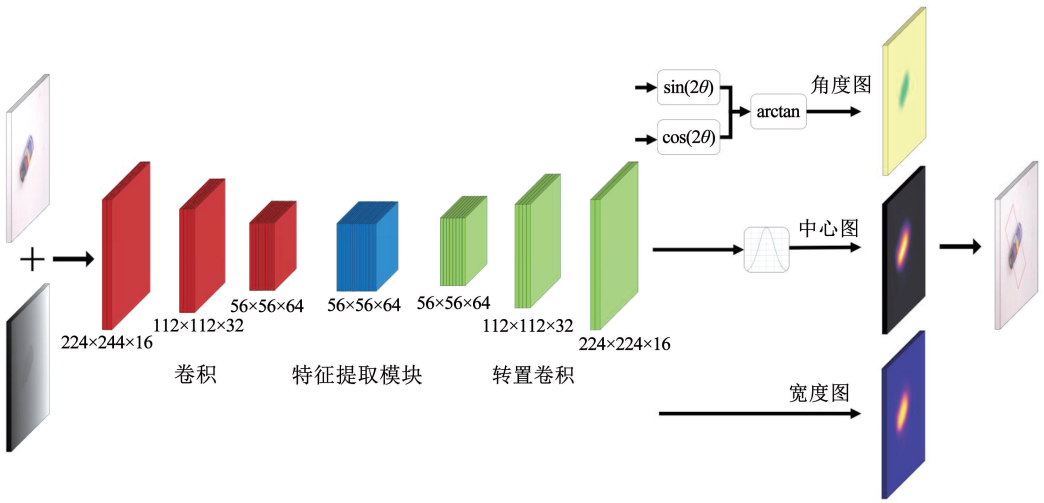


图 2　关键点检测模型结构

Fig. 2　Architecture of key point estimation model

2.3　后处理模块

经过转置的操作后，本文得到了 4 个维度为 $H \times W \times C$ 的特征热力图。对于中心点热力图，本文采用两维的高斯核进行平滑操作，将集中的热力标签分散在其周边。这样可以提高模型的抗随机噪声能力，从而帮助本文提取到最佳的抓取点位置。两维高斯核可以表示为

$$G(x,y) = \frac{1}{\sqrt{2\sigma^2}}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{3}$$

式中 $\sigma = 2$。由于网络输出的角度特征图分别为 $\cos(2\theta)$ 和 $\sin(2\theta)$，所以本文使用下式得到抓取角度为

$$\theta = \frac{1}{2}\arctan\left(\frac{\sin(2\theta)}{\cos(2\theta)}\right) \tag{4}$$

式中，左边的 θ 为通过角度特征图所取的角度值。利用获取的中心点坐标 (x,y)，角度和宽度特征图，本文可以通过下式得到抓取角度和宽度值为

$$[\theta, W] = [\text{feature}_{\text{angle}}(x,y), \text{feature}_{\text{width}}(x,y)] \tag{5}$$

2.4　损失函数

本文选择可以充分利用小误差和控制梯度爆炸的 Smooth L1 函数[24]作为损失函数，Smooth L1 损失函数定义为

$$L(\boldsymbol{G}_l, \hat{\boldsymbol{G}}_l) = \frac{1}{n}\sum_{l=1}^{n} x_l \tag{6}$$

其中 x_l 定义如下：

$$x_l = \begin{cases} 0.5\,(\boldsymbol{G}_l - \hat{\boldsymbol{G}}_l)^2, \text{when} \mid \boldsymbol{G}_l - \hat{\boldsymbol{G}}_l \mid < 1 \\ \mid \boldsymbol{G}_l - \hat{\boldsymbol{G}}_l \mid - 0.5, \text{otherwise} \end{cases}$$

式中：$\boldsymbol{G}_l$ 为模型输出的抓取表示，$\hat{\boldsymbol{G}}_l$ 为真实抓取标签，l 为第 l 个抓取，n 为真实抓取标签的数量。由于需要从中心点、正、余弦角度和宽度图像生成抓取，所以总损失函数需要包含 4 个部分。为了平衡参数之间的分布关系，需要给每个部分加上权重，因此总损失函数定义为

$$L_{\text{total}} = \lambda_{\text{center}} L_{\text{center}} + \lambda_{\cos(2\theta)} L_{\cos(2\theta)} + \lambda_{\sin(2\theta)} L_{\sin(2\theta)} + \lambda_{\text{width}} L_{\text{width}} \tag{7}$$

式中：λ_{center}为中心点的损失权重，$\lambda_{\cos(2\theta)}$、$\lambda_{\sin(2\theta)}$分别为正、余弦角度的损失权重，λ_{width}为末端夹持器宽度的损失权重。

3　模型改进

3.1　引入通道和空间注意力机制

本文使用 RGB-D 多模态信息作为输入，与只有一个通道的神经网络相比，多模态信息会使特征通道上的信息更加丰富。同时，原始的 CenterNet 是针对 RGB 输入设计的，没有考虑 RGB-D 的输入情况。因此，如何充分利用这些高维度信息非常重要。此外，在抓取检测场景中，物体的摆放间距不确定，不同物体的尺寸也可能相差较大，长条形物体笔、圆形物体盘子或者不规则图形门锁等，故对抓取检测识别的精度要求高。根据上述两个问题，可以使用通道注意力机制来充分利用高维信息，以及空间注意力机制来利用整体图像上长程信息。因此，本文引入 CA 注意力模块来提高模型的综合性能。

CA 注意力机制是一种混合注意力机制，包括通道和空间两种注意力机制。SE 注意力模块为仅使用通道注意力的模块，在一定程度上提高了模型性能，但是忽略了空间位置信息，而 CA 注意力机制则可以将其有效利用。虽然 CBAM 这一混合注意力机制可以同时利用通道和空间信息，但其在空间注意力的学习中使用了固定大小的 7 × 7 卷积核来学习空间维度上的特征权重，从而在注意力过程中加入了超参数，并且固定大小的卷积核只能学习到局部相关性，无法学习超过这个范围的长程相关性。

全局池化通常用于通道注意力机制中以编码各个通道的整体信息，但同时也会将该通道下整体的空间信息压缩，从而很难获取空间信息。然而，空间信息对学习空间注意力机制非常重要。为了更好对比，先介绍 SE 模块，其从结构上可以分为压缩和激发两个步骤，全局池化对应压缩步骤。设输入特征图 X 的尺寸为 $C \times H \times W$，其中 C 为特征通道数，H 为高度，W 为宽度。对于第 c 个通道的全局池化可以表示为

$$r_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{8}$$

式中 r_c 为第 c 个通道的全局信息。从式中可以看出其通过压缩空间信息得到的通道信息，无法同时保留两者。CA 注意力机制通过以下两步将通道和空间信息同时编码：坐标信息植入模块和注意力生成模块。CA 注意力模块和残差模块一起组成了模型的特征提取网络，对于其中一层输入 H_k，输出为 Y 的层级展开，其详细组成结构如图 3 所示。

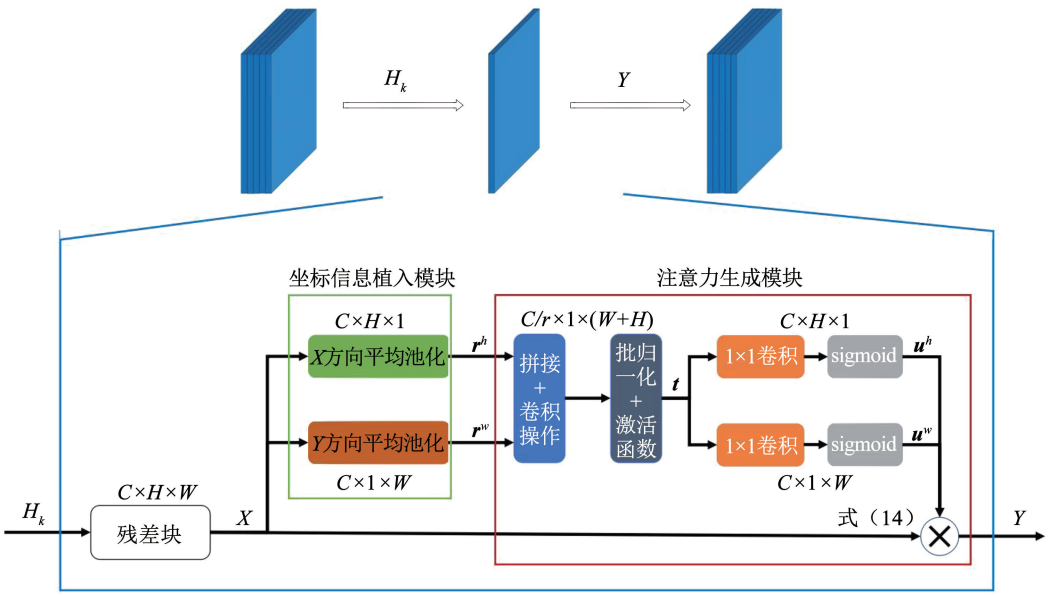


图 3　特征提取模块结构

Fig. 3　Architecture of feature extraction module

3.1.1　坐标信息植入模块

为了强化注意力模块通过精确的位置信息来捕捉长距离的空间关系，此方法分解了式(8)中的二维全局池化，将其变为一对沿着水平方向和垂直方向的一维特征编码操作。对于输入 $\boldsymbol{X}$ 来说，本文使用两个空间延展形池化核 $(H,1)$ 和 $(1,W)$ 来分别沿着水平和垂直方向进行池化操作。对于水平方向来说，第 c 个通道的高度 h 输出信息可以表示为

$$z_c^h(h) = \frac{1}{W}\sum_{0\leqslant i<W} x_c(h,i) \tag{9}$$

同理，第 c 个通道在宽度 w 的输出信息可以表示为

$$z_c^w(w) = \frac{1}{H}\sum_{0\leqslant j<H} x_c(j,w) \tag{10}$$

上述两个转换分别聚合了沿着两个空间方向的特征，产生了一对内含方向信息的特征图。

3.1.2　注意力生成模块

此模块的设计需要尽量简单和低复杂度且能充分利用通道和空间信息。特别地，对于通过式(9)、(10)得到的聚合特征图，首先将其沿着通道维度拼接起来，并且用 C/r 个 1×1 的卷积核对其进行卷积操作，此卷积操作表示为 Conv_1，此过程如下：

$$\boldsymbol{t} = \delta(\mathrm{Conv}_1([\boldsymbol{z}^h,\boldsymbol{z}^w])) \tag{11}$$

式中：$[\cdot,\cdot]$ 为沿着通道维度的拼接操作，δ 为非线性激活函数操作，$\boldsymbol{t}\in\mathbf{R}^{\frac{C}{r}\times(H+W)}$ 编码沿着水平和垂直方向信息的中间特征图，其中 r 为缩减率，可以通过其调整特征图的大小。随后将 $\boldsymbol{t}$ 沿着空间方向分开成两个张量 $\boldsymbol{t}^h\in\mathbf{R}^{\frac{C}{r}\times H}$ 以及 $\boldsymbol{t}^w\in\mathbf{R}^{\frac{C}{r}\times W}$，接着通过同为 C 个 1×1 的卷积核 Conv_h 和 Conv_w 分别进行卷积操作：

$$\boldsymbol{u}^h = \delta(\mathrm{Conv}_h(\boldsymbol{t}^h)) \tag{12}$$

$$\boldsymbol{u}^w = \delta(\mathrm{Conv}_w(\boldsymbol{t}^w)) \tag{13}$$

式中：$\boldsymbol{u}^h\in\mathbf{R}^{C\times H\times1}$，$\boldsymbol{u}^w\in\mathbf{R}^{C\times1\times W}$，$\delta$ 为非线性激活函数。参考文献[16]中讨论的结果，为了在保证性能的同时减低计算复杂度，本文采用 $r=32$。输出 $\boldsymbol{u}^h$ 和 $\boldsymbol{u}^w$ 即为注意力权重，CA 注意力模块输出 $\boldsymbol{Y}$ 可以表示为

$$y_c(i,j) = x_c(i,j)\times u_c^h(i)\times u_c^w(j) \tag{14}$$

CA 注意力机制同时考虑了空间信息的编码，从而可以更精确地定位物体的位置。

3.2　引入多任务学习框架

在抓取检测任务中，模型需要输出抓取中心点坐标、抓手旋转角度以及开合宽度。模型的多种输出可看作不同的子任务，其中子任务间的数值差异较大，会导致总损失函数在每个周期梯度下降时偏向于数值较大的子任务，从而影响抓取检测结果。通常的做法是通过穷举法人工调节参数，然而人为调节参数费时，且最终性能对参数权重的选择敏感。为此，本文参考文献[18]提出的方法，通过引入同方差不确定性，使得不同子任务损失函数的损失权重参数可学习。在贝叶斯模型中，主要有两种不确定性可以建模，第 1 种是由于缺少训练数据导致的认知不确定性，第 2 种是由于噪声导致的偶然不确定性。偶然不确定又可以分为数据相关的不确定性，以及任务相关的不确定性即同方差不确定性。对于多任务学习问题，可以使用同方差不确定性作为加权损失的基础。

设模型的输入为 $\boldsymbol{x}$，网络参数 $\boldsymbol{W}$，输出为 $f^{\boldsymbol{W}}(\boldsymbol{x})$。对于回归任务，本文将似然定义为以模型输出作为平均值的高斯函数：

$$p(\boldsymbol{y}|f^{\boldsymbol{W}}(\boldsymbol{x})) = N(f^{\boldsymbol{W}}(\boldsymbol{x}),\sigma^2) \tag{15}$$

式中 σ 为噪声参数。对于多输出模型，本文定义 $f^{\boldsymbol{W}}(\boldsymbol{x})$ 为充分统计量，根据独立性假设，故可通过因式分解得到多任务可能性：

$$p(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_K|f^{\boldsymbol{W}}(\boldsymbol{x})) = p(\boldsymbol{y}_1|f^{\boldsymbol{W}}(\boldsymbol{x}))\cdots p(\boldsymbol{y}_k|f^{\boldsymbol{W}}(\boldsymbol{x})) \tag{16}$$

式中，$\boldsymbol{y}_1,\cdots,\boldsymbol{y}_K$ 为不同任务的输出。对数似然估计为

$$p(\boldsymbol{y}|f^{\boldsymbol{W}}(\boldsymbol{x})) \propto -\frac{1}{2\sigma^2}\boldsymbol{y}-f^{\boldsymbol{W}}(\boldsymbol{x})^2 - \log\sigma \tag{17}$$

假设模型输出由两个向量 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 组成：

$$p(\boldsymbol{y}_1,\boldsymbol{y}_2|f^{\boldsymbol{W}}(\boldsymbol{x})) = p(\boldsymbol{y}_1|f^{\boldsymbol{W}}(\boldsymbol{x}))\cdots p(\boldsymbol{y}_2|f^{\boldsymbol{W}}(\boldsymbol{x})) = N(\boldsymbol{y}_1;f^{\boldsymbol{W}}(\boldsymbol{x}),\sigma_1^2)N(\boldsymbol{y}_2;f^{\boldsymbol{W}}(\boldsymbol{x}),\sigma_2^2) \tag{18}$$

在上述条件下，为了最小化损失函数 $L(\boldsymbol{W},\sigma_1,\sigma_2)$，有下列推导：

$$\begin{aligned}L(\boldsymbol{W},\sigma_1,\sigma_2) &= -\log p(\boldsymbol{y}_1,\boldsymbol{y}_2|f^{\boldsymbol{W}}(\boldsymbol{x})) \propto \\ &\frac{1}{2\sigma_1^2}\boldsymbol{y}_1 - f_1^{\boldsymbol{W}}(\boldsymbol{x})^2 + \frac{1}{2\sigma_2^2}\boldsymbol{y}_2 - \\ &f_2^{\boldsymbol{W}}(\boldsymbol{x})^2 + \log\sigma_1\sigma_2 = \frac{1}{2\sigma_1^2}L_1(\boldsymbol{W}) + \\ &\frac{1}{2\sigma_2^2}L_2(\boldsymbol{W}) + \log\sigma_1\sigma_2\end{aligned} \tag{19}$$

式中：$L_1(\boldsymbol{W})$、$L_2(\boldsymbol{W})$ 分别为两个任务的损失函数，σ_1、σ_2 可以基于数据自适应的分别学习到 $L_1(\boldsymbol{W})$ 和 $L_2(\boldsymbol{W})$ 的相对权重。σ_1 作为 $\boldsymbol{y}_1$ 变量的噪声参数，当其增加时，$L_1(\boldsymbol{W})$ 的权重会减少。从另一个角度来说，当噪声减少的时候，相对应的损失权重就会增加。由于式(19)中 $\log\sigma_1\sigma_2$ 正则项的存在，也会抑制噪声数值相对过大。在实践中，本文采用 $s=\log\sigma^2$ 作为学习参数，避免了直接学习 σ 的值，因 σ 在训练过程中可能会出现 0 值，导致式(19)中出现被除数为 0 的问题，从而可获得更强的数值稳定性。针对本文的任务，结合式(7)，增加可学习噪声参数后的损失函数为

$$\begin{aligned}L_{\mathrm{total}} = &\exp(-s_{\mathrm{center}})L_{\mathrm{center}} + \exp(-s_{\cos})L_{\cos(2\theta)} + \\ &\exp(-s_{\sin})L_{\sin(2\theta)} + \exp(-s_{\mathrm{width}})L_{\mathrm{width}} + \\ &s_{\mathrm{center}} + s_{\cos} + s_{\sin} + s_{\mathrm{width}}\end{aligned} \tag{20}$$

由式(20)可知，该方法通过引入 4 个可学习权重参数，可以在反向传播中自动更新权重，与其他模

型参数同时训练。相较于原 CenterNet 模型,增加的参数数量和计算量低。

4 结果和分析

4.1 实验条件

本文实验使用的 GPU 为 NVIDIA RTX2060, NVIDIA CUDA 版本为 11.6.134,处理器型号为 AMD Ryzen™ 7 4800HS,操作系统是 Window 11,版本号 21H2。

4.2 数据集

具有代表性的公开数据集基本情况见表 1。

表 1 抓取检测数据集比较

Tab. 1 Comparison of grasp detection datasets

数据集	数据类型	物体数量	图片量	抓取标签
Dexnet	Depth	1 500	6.7×10^6	6.7×10^6
Cornell	RGB-D	240	885	8 019
Jacquard	RGB-D	1 100	5 400	1.1×10^6

由于本文需要使用 RGB-D 数据训练模型,故选用其中两种包含 RGB-D 信息的抓取检测数据集。

4.2.1 Cornell 数据集,

Cornell 数据集部分样例如图 4(a)所示。该数据集是抓取检测领域最流行的数据集,包含 240 个不同物体的 885 张 RGB-D 图片,分辨率为 640 × 480。每张图片都有正向和负向抓取标签,一共有 8 019 个标签,其中 5 110 个正标签和 2 909 个负标签,在训练过程中,本文只使用正标签。考虑到图片数量相对较少,为了能更好训练,本文利用了随机裁剪、放缩和旋转等图片增强手段。与其他工作一样,本文通过以下两种方式对数据集进行分割。

1)图像分割。图像分割随机打乱数据集中的图片进行训练,可检测模型对于不同位姿下的之前见过物体的抓取生成能力。

2)对象分割。对象分割根据不同物体的种类进行分割。有助于检测模型对于未知物体的抓取能力,具有更强的现实意义,更能体现模型的鲁棒性。

4.2.2 Jacquard 数据集

Jacquard 数据集部分样例如图 4(b)所示。该数据集是通过 CAD 模型建立的数据集,包含 11 619 个不同物体的 54 485 张 RGB-D 图片,分辨率为 1 024 × 1 024。图片的正标签是通过在仿真环境中的成功抓取获得的,约含有一百万个抓取标签。由于此数据集已经足够大,训练时并不需要采用数据增强。

图 4 Cornell 和 Jacquard 数据集样例

Fig. 4 Examples from Cornell and Jacquard dataset

4.3 模型训练评估标准

为了便于与之前的研究工作比较,本文采用文献[20]提出的矩形抓取标准作为判定条件。一个成功的抓取检测需要满足以下两个判定条件:

1)预测抓取矩形和真实标签抓取矩形之间的夹角应小于 30°;

2)预测抓取矩形和真实标签抓取矩形的 Jaccard 相似系数应该大于 25%。

Jaccard 相似系数定义如下:

$$\text{Jaccard}(\boldsymbol{g},\hat{\boldsymbol{g}}) = \frac{|\boldsymbol{g}\cap\hat{\boldsymbol{g}}|}{|\boldsymbol{g}\cup\hat{\boldsymbol{g}}|} \tag{21}$$

式中:$\boldsymbol{g}$ 为模型预测的抓取矩形区域,$\hat{\boldsymbol{g}}$ 为真实标签抓取矩形区域。

4.4 模型检测结果及性能分析

模型在 Cornell 数据集上的检测结果示例如图 5 所示。第 1、2 列分别为 RGB 和深度图,第 3 列为模型生成的热力图,第 4 列为角度图,第 5 列为宽度图。通过热力图定位抓取中心并结合对应坐标下的角度图和宽度图信息,得到最后 1 列的抓取表示。从图 5 中例子可以看出模型在面对锁型等不规则物体时也有良好表现。

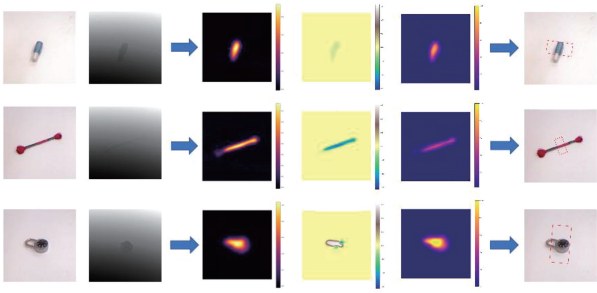


图 5 Cornell 数据集的抓取检测结果

Fig. 5 Detection results on Cornell dataset

表 2 展现了本文在康奈尔数据集下与其他模型的性能对比。本文采用图像和对象分割方法分别对数据集分割,主要检测准确率和检测速率。在图像和对象分割条件下,本文获得了 98.8% 的准确率和 25.30 fps 的检测速率。可以发现,与其他工作相

比，本文的检测准确率和检测速率都有一定提升。其中与基于锚框的检测算法相比，见表 2 中的 Fast-RCNN，本文改进的 CenterNet 模型在取得相近抓取准确率的同时，由于不需要生成候选锚框和非极大值抑制等操作，在检测速率上有明显优势。

为了进一步验证本模型的性能和鲁棒性，本文在更加严苛的条件下与其他模型进行对比。如表 3 所示，本文测试了本模型在 Jaccard 相似系数以0.05为梯度、从 0.25 ~0.45 等梯度变化时的对象分割准确率。在系数小于 0.45 时，随着 Jaccard 相似系数的提高，本模型可以保持在超过 90% 的准确率。在 0.45 时，本模型的准确率和其他工作相比也能获得更好的结果。表 3 的结果对比表明本模型在更严苛条件下具有较强的性能和鲁棒性。

表 2　康奈尔数据集下模型精度对比

Tab. 2　Comparison of model prediction accuracy on Cornell dataset

文献	算法	准确率/%		检测速率/fps
		图像分割	对象分割	
文献[20]	Fast Search	60.5	58.3	0.02
文献[10]	SAE, struct. reg. two-stage	75.9	75.6	0.07
文献[13]	AlexNet, MultiGrasp	88.0	87.1	13.16
文献[25]	STEM-CaRFs (Selective Grasp)	88.2	87.5	
文献[26]	ZF-net, 3 scales	93.2	89.1	
文献[12]	Faster RCNN, dual resnet-50	96.1	96.0	8.33
文献[22]	FCGN, ResNet-101	97.7	96.6	5.51
本文	基于 CA 和 Multi-task 的改进 CenterNet	**98.8**	**98.8**	**25.30**

表 3　在不同 Jaccard 系数下的检测准确率

Tab. 3　Prediction accuracy at different Jaccard thresholds

Jaccard 系数	0.25	0.30	0.35	0.40	0.45
文献[4]	85.4	93.1	77.5	75.3	70.8
文献[12]	96.0	94.9	92.1	84.7	
本文	98.8	97.7	95.5	93.2	86.5

除此之外，本文希望验证本模型在更大规模数据集上的效果，故对其在 Jacquard 数据集上进行训练，相应的检测结果示例如图 6 所示。第 1 行显示的陀螺物体，其在纵向的空间上有一定高度，在倾斜放置时会给抓取检测带来一定挑战。Jacquard 数据集是通过 CAD 模型建立的，其中的物体有一定的随机性，相对于现实情况更加苛刻，如第 2 行中的盆栽，但本文的模型仍获得了准确的抓取结果。

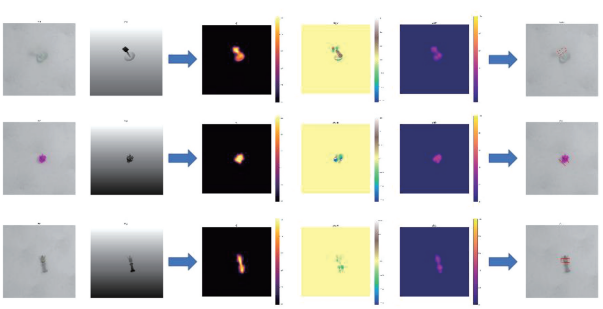


图 6　Jacquard 数据集的抓取检测结果

Fig. 6　Detection results on Jacquard dataset

表 4 表明本模型不仅在康奈尔数据集上取得出色性能，并且在规模更大的 Jacquard 数据集上也能取得优秀结果。

表 4　Jacquard 数据集下模型精度对比

Tab. 4　Comparison of model prediction accuracy in Jacquard grasp dataset

文献	算法	准确率/%
文献[4]	GG-CNN	84.0
文献[22]	FCGN, ResNet-101	91.8
文献[5]	GR-ConvNet	94.6
本文	基于 CA 和 Multi-task 的改进 CenterNet	**95.7**

4.5　可学习权重的鲁棒性实验

关于可学习权重的实验，通常仅考虑初始权重为 0 的情况[18]。为了更好地验证本模型的鲁棒性和适用性，本文将每个子任务的初始权重设定为相差很大的初值，这样中心点、正、余弦角度、抓手开合宽度一共包含 16 种初始权重组合情况。同时本文也随机交换了不同任务间的学习权重作为初始权重。权重参数随着训练进行的变化情况如图 7 所示，可以看出其收敛效果对于初始值不敏感，反应了本算法的鲁棒性。

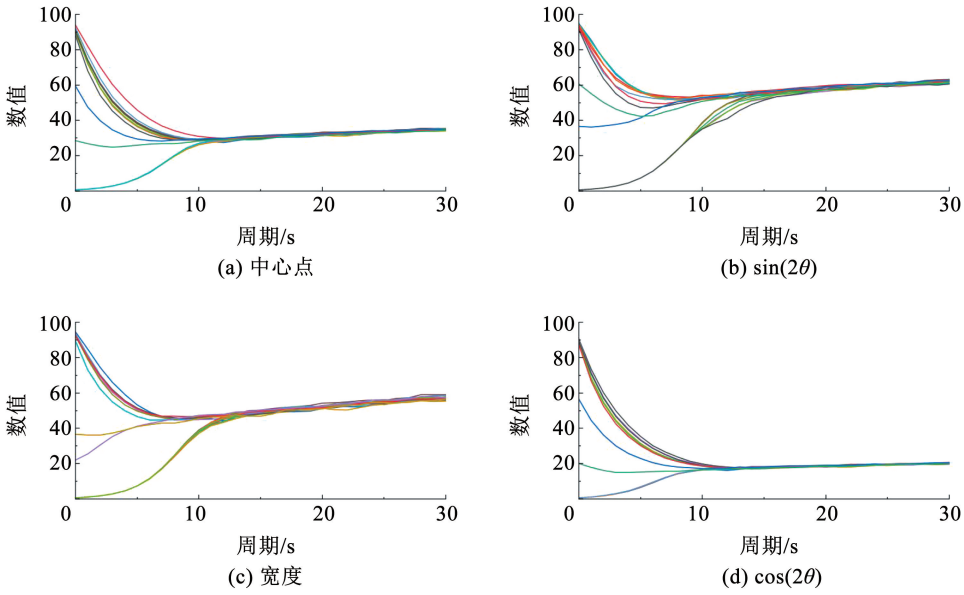


图 7 不同初始权重下参数变化情况

Fig. 7 Detection results on Jacquard dataset

4.6 注意力机制和多任务学习的消融实验

为了进一步验证本方法各模块的有效性,以及其对于整体性能的影响,本文对模型进行消融实验,实验结果见表 5。

表 5 消融实验结果

Tab. 5 Results from ablation experiment

算法	准确率/%		检测速率/fps
	图像分割	对象分割	
改进的 CenterNet	95.5	94.3	26.70
基于 CA 的改进 CenterNet	97.7	96.6	25.90
基于 Multi-task 的改进 CenterNet	96.6	95.5	26.40
基于 CA 和 Multi-task 的改进 CenterNet	**98.8**	**98.8**	**25.30**

从表 5 实验结果可以看出,CA 注意力机制和权重学习机制的引入,可有效提升抓取检测的准确率,且由于实现过程中需要的参数较少,对检测速率的影响较小。

5 结 论

1)针对抓取检测任务,本文结合 CA 注意力机制和多任务学习提出了一种新的抓取检测模型。模型利用 RGB-D 图像作为输入信息,根据输出的中心图、宽度图、角度图,生成像素级别的抓取表示,在 Cornell 和 Jacquard 数据集上分别实现了 98.8% 和 95.7% 的准确率。

2)从充分利用高维信息角度出发,本文引入 CA 注意力机制,其使用通道注意力机制来充分利用高维信息,以及空间注意力机制来利用整体图像上长程信息,从而可以更精确地定位物体的位置。消融实验证明其对于图像分割和对象分割可以达到 2% 以上的性能提升。

3)从损失函数优化角度出发,本文引入多任务学习框架,能根据任务特点自动获取不同子损失函数的权重,无需对权重进行人为选定。鲁棒性实验进一步证明此方法可以稳定学习到任务的最优权重,对于初始值的干扰不敏感。

4)本文通过在不同数据集、不同精度要求以及干扰下进行的实验,验证了所提出的抓取检测模型在各项指标下的良好表现。同时,通过消融实验,从模型组成方面证明了各模块在提升抓取检测性能方面的有效性。在未来的工作中,本文将尝试搭建闭环抓取系统,通过对真实世界中物体的抓取,进一步验证模型的性能表现。

参考文献

[1] AINETTER S, FRAUNDORFER F. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE, 2021: 13452. DOI: 10.1109/ICRA48506.2021.9561398

[2] BICCHI A, KUMAR V. Robotic grasping and contact: A review[C]//Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). San Francisco, CA: IEEE, 2002: 348. DOI: 10.1109/ROBOT.2000.844081

[3] KUMRA S, KANAN C. Robotic grasp detection using deep convolutional neural networks[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, BC, Canada: IEEE, 2017: 769. DOI: 10.1109/IROS.2017.8202237

[4] MORRISON D, CORKE P, LEITNER J. Closing the loop for robotic

grasping: A real-time, generative grasp synthesis approach[EB/OL]. 2018: arXiv: 1804.05172. https://arxiv.org/abs/1804.05172.pdf

[5] 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 (IROS). Las Vegas, NV: IEEE, 2021: 9626. DOI: 10.1109/IROS45743.2020.9340777

[6] KOLAMURI R, SI Zilin, ZHANG Yufan, et al. Improving grasp stability with rotation measurement from tactile sensing[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021: 6809. DOI: 10.1109/IROS51168.2021.9636488

[7] 张建军, 刘卫东, 张溢文, 等. 基于微机电系统的水下灵巧手触觉力测量传感器[J]. 上海交通大学学报, 2018, 52(1): 76
ZHANG Jianjun, LIU Weidong, ZHANG Yiwen, et al. Tactile force sensor of underwater dexterous hand based on micro electromechanical system [J]. Journal of Shanghai Jiao Tong University, 2018, 52(1): 76. DOI: 10.16183/j.cnki.jsjtu.2018.01.012

[8] BOHG J, MORALES A, ASFOUR T, et al. Data-driven grasp synthesis—a survey[J]. IEEE Transactions on Robotics, 2014, 30(2): 289. DOI: 10.1109/TRO.2013.2289018

[9] CALDERA S, RASSAU A, CHAI D. Review of deep learning methods in robotic grasp detection[J]. Multimodal Technologies and Interaction, 2018, 2(3): 57. DOI: 10.3390/mti2030057

[10] LENZ I, LEE H, SAXENA A. Deep learning for detecting robotic grasps[J]. International Journal of Robotics Research, 2015, 34(4/5): 705. DOI: 10.1177/0278364914549607

[11] ZHOU Xinwen, LAN Xuguang, ZHANG Hanbo, et al. Fully convolutional grasp detection network with oriented anchor box [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2019: 7223. DOI: 10.1109/IROS.2018.8594116

[12] CHU F J, XU Ruinian, VELA P A. Real-world multiobject, multigrasp detection[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3355. DOI: 10.1109/LRA.2018.2852777

[13] REDMON J, ANGELOVA A. Real-time grasp detection using convolutional neural networks [C]//2015 IEEE International Conference on Robotics and Automation (ICRA). Seattle, WA: IEEE, 2015: 1316. DOI:10.1109/ICRA.2015.7139361

[14] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 7132. DOI: 10.1109/CVPR.2018.00745

[15] LI Xiang, WANG Wenhai, HU Xiaolin, et al. Selective kernel networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA: IEEE, 2020: 510. DOI: 10.1109/CVPR.2019.00060

[16] HOU Qibin, ZHOU Daquan, FENG Jiashi. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE, 2021: 13708. DOI: 10.1109/CVPR46437.2021.01350

[17] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018: 3. DOI: 10.1007/978-3-030-01234-2_1

[18] CIPOLLA R, GAL Y, KENDALL A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 7482. DOI: 10.1109/CVPR.2018.00781

[19] KENDALL A, GAL Y. What uncertainties do we need in Bayesian deep learning for computer vision? [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5580. DOI: 10.5555/3295222.3295309

[20] JIANG Yun, MOSESON S, SAXENA A. Efficient grasping from RGBD images: Learning using a new rectangle representation[C]//2011 IEEE International Conference on Robotics and Automation. Shanghai, China: IEEE, 2011: 3304. DOI: 10.1109/ICRA.2011.5980145

[21] 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. DOI: 10.1177/0278364919859066

[22] ZHOU Xingyi, WANG Dequan, KRÄHENBÜHL P. Objects as points[EB/OL]. 2019: arXiv: 1904.07850. https://arxiv.org/abs/1904.07850.pdf

[23] DUMOULIN V, VISIN F. A guide to convolution arithmetic for deep learning[EB/OL]. 2016: arXiv: 1603.07285. https://arxiv.org/abs/1603.07285.pdf

[24] GIRSHICK R. Fast r-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2016: 1440. DOI: 10.1109/ICCV.2015.169

[25] ASIF U, BENNAMOUN M, SOHEL F A. RGB-D object recognition and grasp detection using hierarchical cascaded forests [J]. IEEE Transactions on Robotics, 2017, 33(3): 547. DOI: 10.1109/TRO.2016.2638453

[26] GUO Di, SUN Fuchun, LIU Huaping, et al. A hybrid deep architecture for robotic grasp detection [C]//2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017: 1609. DOI: 10.1109/ICRA.2017.7989191

（编辑　张　红）