Computer Science and Application 计算机科学与应用, 2023, 13(1), 72-82
Published Online January 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.131008

Hans汉斯

基于多层级网络的像素级抓取姿态估计

张志刚[1]，程良伦[2]

[1]广东工业大学自动化学院，广东 广州
[2]广东工业大学计算机学院，广东 广州

收稿日期：2022年12月13日；录用日期：2023年1月9日；发布日期：2023年1月18日

摘 要

为了解决在杂乱场景下从单视角图像中准确估计抓取位姿的问题，本文提出基于多层级特征的像素级端到端抓取检测网络。我们在全卷积神经网络中集成了多层级金字塔池化模块和多分支输出结构形成高精度抓取位姿检测网络，从而有效处理尺寸和位姿各异的未知物体在杂乱场景下的抓取问题。实验表明，我们的方法在Cornell抓取数据集上的理论抓取精度相比现有方法有明显提升；同时，在机械臂实物抓取实验上，我们的方法在多物体杂乱场景中以88.0%的平均抓取成功率实现了100%的抓取完成率。

关键词

机器人抓取，抓取检测，卷积神经网络

Pixel-Wise Grasp Pose Detection Based on Hierarchical Network

Zhigang Zhang[1], Lianglun Cheng[2]

[1]School of Automation, Guangdong University of Technology, Guangzhou Guangdong
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Dec. 13th, 2022; accepted: Jan. 9th, 2023; published: Jan. 18th, 2023

Abstract

In order to address the problem of estimating grasp pose under cluttered scenes using single-view image, we proposed an end-to-end pixel-wise grasp detection network based on hierarchical features. We integrated a pyramid pooling module and multi-head structure into a fully convolutional neural network to form a high-precision grasp pose detection network, effectively handling the problem of predicting grasps for unknown objects with diverse sizes and poses in clutters. The

文章引用: 张志刚, 程良伦. 基于多层级网络的像素级抓取姿态估计[J]. 计算机科学与应用, 2023, 13(1): 72-82.
DOI: 10.12677/csa.2023.131008

experiments showed that our proposed method significantly outperforms existing methods in Cornell grasp dataset. Physical experiments on real robotic arms are also conducted, our method achieved average grasp success rate of 88.0% at 100% completion rate in the challenging grasping task in clutter with multiple unknown objects.

Keywords

Robotic Grasping, Grasp Detection, Convolutional Neural Network

1. 引言

抓取一个物体对于人来说很直观和简单，但是要驱动机器人自主完成一次抓取作业的首要任务是通过视觉信息推断出其抓取位姿。通过机器视觉技术估计抓取位姿的意义在于其能够引导机器人对物体进行精准的操作，主要的应用场景有居家服务机器人和工业自动化机器人。一般来说，机械臂的抓取位姿包含六个维度的信息——末端的三维平移量和三维旋转量。其中三维平移量代表抓取点位置在三维空间中的三维坐标，三维旋转量则表示了机械臂在执行抓取动作时末端的姿态。

因为 SE(3)巨大的搜索空间导致的困难，所以一般会在抓取位姿中添加约束条件，从而简化问题，比如在二维平面上表示抓取位姿[1] [2]。最常用的抓取位姿表达形式是有向抓取矩形框，这种表示方法包含用来代表抓取点的矩形框中心点、矩形框的宽度、矩形框的高度以及矩形框旋转角度。Kumra 等人[3]沿用了 2D 物体检测的思路对抓取位姿进行检测。但是实际的抓取场景中可能存在多个物体，并且物体之间可能存在相互遮挡，此时针对某个单一物体的抓取矩形框就可能会包含其它干扰物体，使得检测结果不准确，这就导致了此方法不能很好地处理多物体的抓取场景。并且基于抓取矩形框的检测算法通常是二阶段算法，这就意味着此类算法需要先在图像上生成大量的抓取候选框，然后再通过神经网络等方法对抓取候选框进行筛选得到目标的抓取位姿。尽管二阶段方法能够进行更细化的处理，但是其计算开销大依然限制了此类算法在实际场景的应用。

除了有向抓取矩形框表示法以外，研究人员们参考了语义分割的解决思路，在输入图像的每个像素点上都预测一个抓取位姿。Morrison 等人[4] [5]提出了生成式的抓取位姿估计策略，对每个像素位置都生成抓取位姿。受到残差卷积网络结构的启发，Kumra 等人[6]使用 RGB-D 图像信息作为输入，提出了生成式残差卷积网络解耦抓取角度，以此来完成实时抓取检测。然而对于弱纹理物体或者是杂乱场景中的物体，这些方法计算的抓取位姿鲁棒性不足，对新物体的表征学习能力不强，无法适应未知物体的抓取检测。

针对上述问题，本文提出一种基于多层级特征的像素级端到端抓取检测网络，该方法使用 RGB 或者深度图作为输入，预测像素级抓取位置图、抓取角度图和抓取宽度图。本文方法遵循编码－解码结构，通过插入多层级金字塔池化模块，可以从有限的训练数据集中学习到更有效的特征，从而可以解决未知物体在杂乱场景下的抓取检测问题。本文的主要工作如下有：1) 提出一个端到端的神经网络结构，通过在神经网络中集成多层级金字塔池化模块，形成高精度的抓取位姿检测网络；2) 在公开的 Cornell 抓取数据集上对本文提出的网络进行训练和验证，效果和正确率超过了现有的抓取检测方法；3) 将本文提出

的方法部署到真实机械臂进行抓取实验，在杂乱多物体的抓取场景下，取得 88%的平均抓取成功率和100%的抓取完成率，验证了提出方法的有效性。

2. 抓取位姿检测算法

2.1. 抓取位姿描述

本文考虑的抓取工具对象为二指平行夹爪，考虑最典型的二维平面抓取检测问题：给定输入的 RGB 或者 Depth 图像，并从中推断出场景中最优的抓取位姿。得到像素坐标系下的抓取位姿后，通过相机标定和手眼标定操作进行刚体变换，可以将表示在像素坐标系下的抓取位姿变换为表示在机器人基坐标系下的抓取位姿，并最终引导机器人进行抓取操作。

现有的研究[3] [7] [8]都使用了有向矩形框常用来表示抓取位姿，这种表示形式适合基于物体检测的网络框架。但是在多物体的杂乱场景下，这种形式存在局限性。为了适配全卷积神经网络，并且针对多物体的复杂场景，本文提出像素级的抓取位姿表达形式。

给定输入图像 I，我们定义在像素坐标系下的抓取位姿为：

$$g=\left(g_p,g_\theta,g_w\right) \tag{1}$$

其中，$g_p=(u,v)$ 代表在像素坐标系下抓取点的坐标；g_θ 代表以横轴为参照的抓取角度，单位为弧度；g_w 代表在像素坐标系下的抓取宽度，单位为像素。由于二指平行夹爪的对称性，我们限制 $g_\theta\in[0,\pi]$，这样可以减小角度的歧义性，降低神经网络对抓取角度预测的难度。图像坐标系上的抓取位姿表示见图 1 所示。

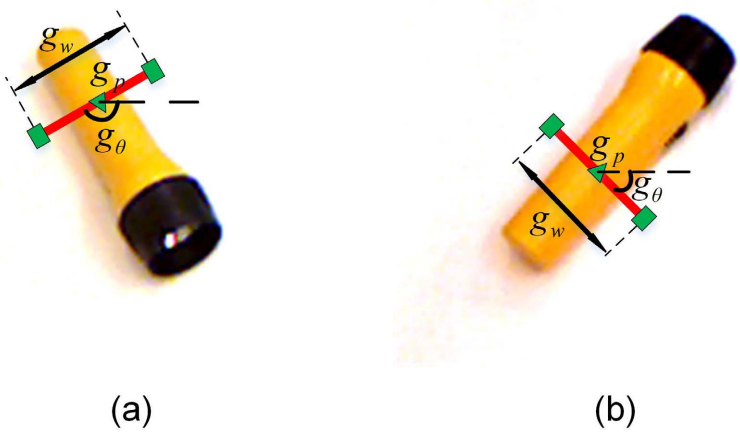

Figure 1. The illustration of grasp pose representation
图 1. 抓取位姿表示

2.2. 网络结构

在本文中，我们提出一种用于在杂乱物体场景中估计抓取位姿的全卷积网络结构。本文的一阶段网络架构由三部分组成，它们分别是主干特征提取器、多层级金字塔池化模块(Pyramid Pooling Module, PPM)和多分支输出网络，如图 2 所示。主干特征提取器由 ResNet [9]网络构成，负责高效地提取特征图；PPM 实现四个尺度的特征图融合，从特征图中计算获得高级特征图，赋予网络处理不同尺寸、不同颜色、不同外形的物体的能力；多分支输出网络对特征进行上采样，并输出三幅和原图尺寸大小一致的抓取图(Grasp Map)。

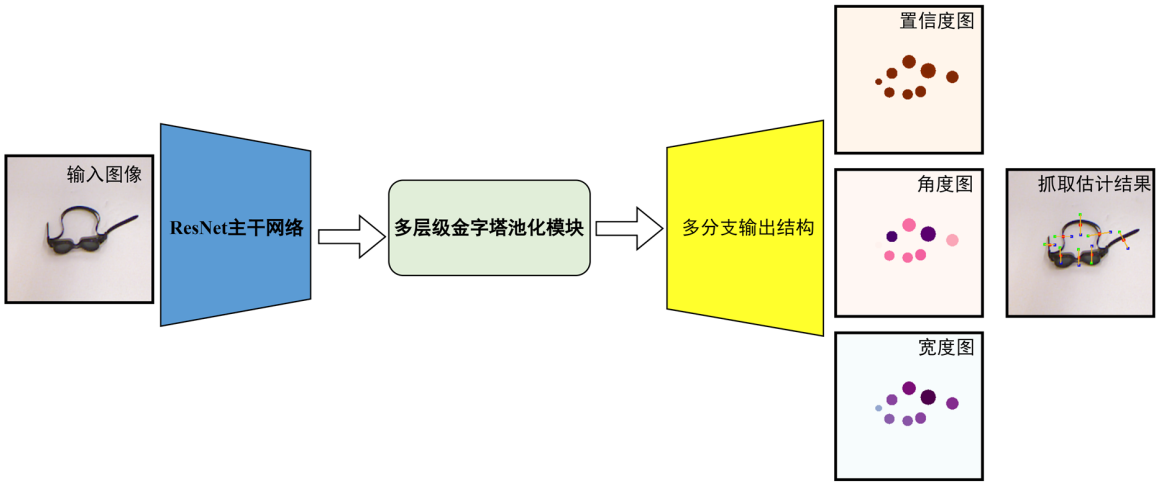

Figure 2. The diagram of our proposed network architecture
图 2. 本文提出的网络结构的整体模块图

2.3. 抓取图

在本文中，我们使用像素级的抓取位姿表示方法，即在输入图像 I 的每个像素位置 p 都使用一个三元组 $t=(c,\theta,w)$ 来表示一个抓取，其中 c 表示当前像素位置的存在合理抓取位姿的置信度，$c\in[0,1]$，置信度越高表明该位置存在合理抓取位姿的概率越高；θ 表示抓取角度；w 表示抓取宽度。对于给定的一张图像，网络模型需要生成三幅尺寸相同的抓取图——抓取置信度图(Confidence Map)、抓取角度图(Angle Map)和抓取宽度图(Width Map)来表征三元组 t。

抓取置信度图：抓取点 $g_p=(u,v)$ 表示输入图像上的一个像素点坐标。尽管输入图像 I 中可能包含多个物体，并且每个物体都可能存在多个合理的抓取位姿，但是抓取点的数量还是远少于非抓取点的数量，这种情况会造成训练数据中的正样本和负样本的极度不平衡，进而导致网络训练难以收敛。为了解决这个问题，我们提出使用抓取点扩散策略。我们将每个抓取点 g_p 映射到一个处于 I 上的抓取区域 g_p^R，位于区域内所有像素都可以接受其为可抓取点。具体来说，g_p^R 设计为一个圆心位于 g_p，半径为 $\frac{g_w}{4}$ 的圆。位于圆心处的置信度设置为 1，圆内其它像素的抓取置信度使用类截断二维正态分布计算得到：

$$c(x,y)\big|_{(x,y)\in g_p^R}=\mathrm{e}^{-\frac{(x-u)^2+(y-v)^2}{2\sigma^2}} \tag{2}$$

其中，$(x,y)\in g_p^R$ 表示抓取区域内的像素点，e 为自然常数，σ 为超参数，σ 设置越大则抓取区域边缘的置信度越大。

抓取角度图：由于角度值具有对称性，使用连续的角度值不利于神经网络直接进行回归学习，这将导致网络训练的收敛难度增大。本文将抓取角度的预测问题视作一个分类问题，可以缓解连续角度值带来的问题。本文不预测连续的抓取角度，而是将 $[0,\pi]$ 范围的角度分成 N_A 等份，每个抓取角度都对应与其距离最接近的角度标签 $\theta_i\left(i=\{1,2,\cdots,N_A\}\right)$。需要注意的是，图像上的非抓取区域不应该分配任何的抓取角度标签，因此我们使用 $\theta_i=0$ 来描述这种情况。所以，角度类别数目为 $C=N_A+1$。在同一个抓取区域 g_p^R 内的抓取角度相同。

抓取宽度图：本文将抓取宽度表示为图像上以像素为单位的长度。同一抓取区域内的抓取宽度取相同的值。为了尺度一致性和输出稳定性，我们将抓取宽度归一化到 $[0,1]$ 之间。

2.4. 多层级金字塔池化模块

为了充分利用复杂场景下输入图像不同层级的特征信息，并且为了进一步挖掘不同层级特征信息之间的关系，我们在全卷积神经网络中引入了 PPM，该技术在文献[10]中首先被提出，在室外场景的语义分割任务上表现出了出色的性能。PPM 技术能够充分提取输入图像的全局信息，提升感受野大小，增强网络模型对不同尺寸，不同类型的物体的适应能力，这对于在复杂场景下执行抓取检测任务具有很大的帮助。

输入图像经过主干网络处理得到初步特征后，将初步特征图送入 PPM 处理，如图 3 所示。在 PPM 中，使用四个子区域大小分别为 1 × 1、2 × 2、3 × 3 和 6 × 6 的自适应平均池化层(Adaptive Average Pooling)将初步特征图划分成四个不同层级的特征图。接着，通过四个平行的 1 × 1 卷积层和双线性插值上采样操作将四个层级的特征图大小恢复为初步特征图的尺寸大小。最后，将四幅特征图和初步特征图在通道维度上进行拼接，得到高级特征图。PPM 的输出结果中即包含了输入图像场景中的低级特征信息，也包含了细化处理后的高级特征。高级特征图后续将被送入多分支输出网络进行抓取图的预测。

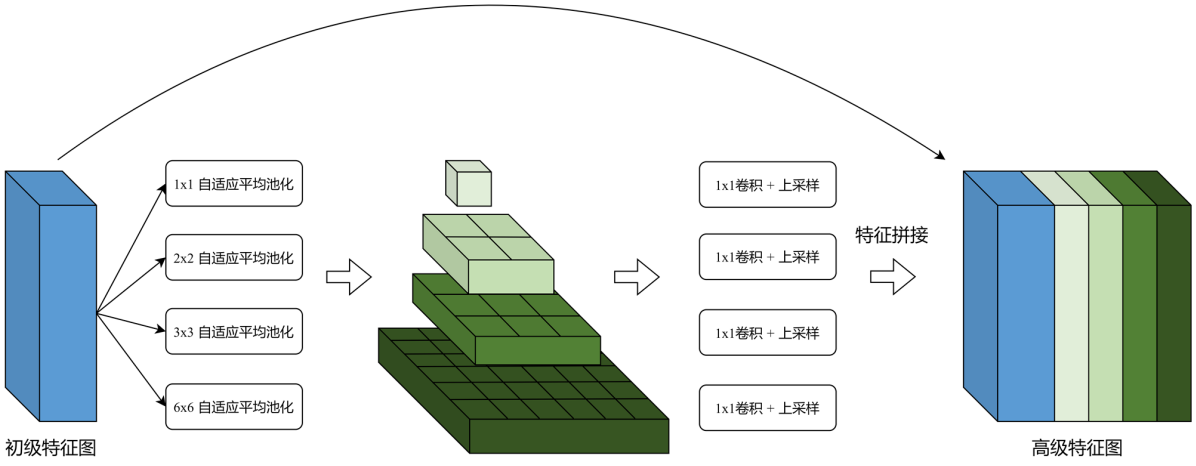

Figure 3. The diagram of hierarchical pyramid pooling module
图 3. 多层级金字塔池化模块流程结构图

2.5. 多分支输出网络

多分支输出网络输出三幅抓取图——抓取置信度图、抓取角度图和抓取宽度图，图 4 展示了多分支输出网络结构图。对应地，输出网络包含了三个分支，置信度估计分支、角度估计分支和宽度估计分支，每个分支都内置了连续的卷积层、批归一化层、ReLU 激活函数层和 Dropout 层。具体来说，首先将来自 PPM 的高级特征图输入到置信度估计分支，随后经过 Sigmoid 激活函数的计算得到抓取置信度图。

由于抓取角度图和抓取宽度图都依赖于抓取置信度图，参考文献[11] [12]的做法，我们为角度估计分支和宽度估计分支引入空间注意力机制，具体做法为：将抓取置信度图与高级特征图进行相乘操作，将相乘后的结果分别输入用以计算抓取角度图和抓取宽度图。

三幅不同的抓取图分别计算损失，然后联合加权计算抓取图全局损失，用以反向传播更新神经网络的参数。三者的损失函数具有相同的形式，这种损失函数的设计能够使得网络更加关注感兴趣区域，也就是场景中的抓取区域，从而降低无关的区域(比如背景、干扰物)对训练的干扰。统一形式的损失函数如下公式(3)所示：

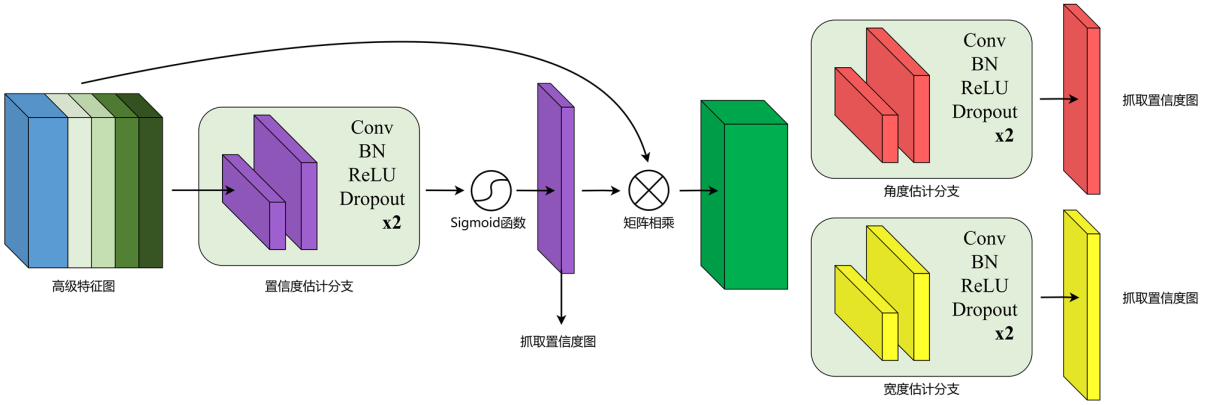

Figure 4. The diagram of multi-head output structure
图 4. 多分支输出网络结构图

$$\mathcal{L}\left(R_{pos},R_{neg},gt,pred,l\right)=\frac{\alpha}{N_{pos}}\sum_{R_{pos}}l\left(gt,pred\right)+\frac{\beta}{N_{neg}}\sum_{R_{neg}}l\left(gt,pred\right) \tag{3}$$

其中，R_{pos} 和 R_{neg} 指明了该如何定义正负样本，N_{pos} 和 N_{neg} 分别是正负样本的数量，分别表示正样本区域和负样本区域；gt 和 $pred$ 分别表示了真实值和预测值；函数 l 计算 gt 和 $pred$ 之间损失值；α 和 β 为超参数，用来调节正负样本在损失函数中的占比，防止过拟合。

置信度损失函数如公式(4)所示，其中 $c_{x,y}$ 为真实置信度，$\hat{c}_{x,y}$ 是估计置信度，l_{smooth} 表示 Smooth-L1 损失函数，δ 为划分正负样本的阈值，正样本区域记作 R_{pos}，负样本记作 R_{neg}。

$$\begin{cases}\mathcal{L}_c=\mathcal{L}\left(R_{pos},R_{neg},c_{x,y},\hat{c}_{x,y},l_{smooth}\right)\\ R_{pos}:x,y\in I,c_{x,y}\geq\delta\\ R_{neg}:x,y\in I,c_{x,y}<\delta\end{cases} \tag{4}$$

抓取角度损失函数如公式(5)所示，其中 $\theta_{x,y}$ 是真实抓取角度，$\hat{\theta}_{x,y}$ 是估计抓取角度，l_{focal} 表示 Focal 损失函数。

$$\mathcal{L}_\theta=\mathcal{L}\left(\theta_{x,y}\in\{1,2,\cdots,N_A\},\theta_{x,y}\in\{0\},\theta_{x,y},\hat{\theta}_{x,y},l_{focal}\right) \tag{5}$$

抓取宽度损失函数如公式(6)所示，其中 w 是真实抓取宽度，$\hat{w}$ 是估计抓取宽度。

$$\mathcal{L}_w=\mathcal{L}\left(R_{pos},R_{neg},w,\hat{w},l_{smooth}\right) \tag{6}$$

联合损失函数定义为三个损失函数的加权和，如公式(7)所示，λ_c、λ_θ、λ_w 为三个权重超参数。

$$\mathcal{L}_g=\lambda_c\mathcal{L}_c+\lambda_\theta\mathcal{L}_\theta+\lambda_w\mathcal{L}_w \tag{7}$$

3. 理论实验及结果

3.1. 数据集介绍

本文用于训练神经网络的数据基于 Cornell 抓取数据集[13]。该数据集包含 885 张 RGB 和 Depth 图像，每张图像中都包含一个日常生活常见的物体，每个物体标注了多个正样本抓取矩形框和负样本抓取矩形框。为了防止过拟合，提升算法的鲁棒性，我们在数据集中引入了不同的图像对比度、亮度、随机缩放、随机旋转平移变换等在线数据增强策略来扩展数据集的规模。

3.2. 评估指标

为了方便与其它方法进行对比，本文从两个维度考虑一个估计抓取位姿是否正确，为与之前的文献[6] [7] [8]使用的评估指标保持一致性，当二维平面抓取位姿同时满足下列条件的时候，那么将其视为正确：

1) 预测抓取角度和真实抓取角度之间的差值小于阈值 δ_A，本文中 $\delta_A = \frac{\pi}{6}$：

$$\left|g_\theta - \hat{g}_\theta\right| < \delta_A \tag{8}$$

2) 当预测抓取位置和真实抓取位姿之间的 Jaccard 系数大于阈值 δ_J，本文中 $\delta_J = 0.25$：

$$J\left(g,\hat{g}\right) = \frac{\left|g \cap \hat{g}\right|}{\left|g \cup \hat{g}\right|} > \delta_J \tag{9}$$

尽管本文使用的像素级的抓取位姿表示方式，但是仍然可以将其转化为抓取矩形框的表示方式，从而可以使用上述的评估指标对抓取位姿质量进行评估。

3.3. 实验参数细节

为了训练提出的神经网络，本文使用深度学习框架 PyTorch。网络训练使用 Adam 优化器，初始学习率设置为 0.001，每 20 个回合学习率减半；权重衰减系数设为 0.0001；整个网络以批大小为 4 一共训练 100 个回合。本文使用的训练平台为 Ubuntu 操作系统，硬件方面采用英伟达 RTX2080 显卡，配备 Intel i5 处理器和 16GB 的内存。

3.4. 实验结果

我们与现有的抓取位姿估计算法进行实验比较。我们同样将数据集按照两种依据进行划分——按照图片划分(Image-Wise, IW)和按照物体划分(Object-Wise, OW)。IW 划分方式主要评估网络模型对于不同位置，不同姿态的物体的抓取位姿检测能力；OW 则更加关注神经网络模型在新物体下的适应性。主要实验结果如表 1 所示。我们分别使用 RGB 和 Depth 作为网络的输入，分别执行前向传播。从表 1 中可以看出，和其它现有的算法进行对比，我们的方法在准确率上有一定的提升。无论输入时 RGB 还是 Depth 图像，无论在 IW 划分规则还是在 OW 划分规则下，我们的方法所获得的准确率都优于现有的方法。本文方法的部分实验可视化结果如图 5 所示。除此之外，我们还设置了不使用 PPM 的基准网络，他们分别使用最大池化模块(Max)和平均池化(Avg)模块作为 PPM 的对比模型，从而验证 PPM 模块在提升网络的性能上有关键作用。两个基准网络在所有的实验设置下均仅有接近 90%的准确率，远远低于嵌入了 PPM 模块的网络。

Table 1. Experiment results on cornell dataset
表 1. 在 Cornell 数据集上的实验结果

作者	方法	准确率(%)		输入格式
		IW	OW	
Morrison *et al.* [5]	GG-CNN	73.0	69.0	Depth
Kumra *et al.* [3]	ResNet50x2	88.9	89.2	RGB
Chu *et al.* [7]	ResNet50	94.4	95.5	RGB
Zhang *et al.* [14]	ROI-GD, ResNet101	93.6	93.5	RGB

Continued

Song *et al.* [15]	ResNet50	96.2	95.6	RGB
Kumra *et al.* [6]	GR-ConvNet-RGB	96.6	95.5	RGB
Ours	Baseline-Avg-RGB	89.2	88.1	RGB
	Baseline-Avg-Depth	88.6	88.1	Depth
	Baseline-Max-RGB	90.9	90.3	RGB
	Baseline-Max-Depth	90.9	89.8	Depth
	PPM-RGB	**97.2**	96.6	RGB
	PPM-Depth	96.0	**97.2**	Depth

从公式(8)和公式(9)可知，两个可以改变的阈值参数 δ_A 和 δ_J 与准确率相关，上述实验结果都是在 $\delta_A=\frac{\pi}{6}$ 和 $\delta_J=0.25$ 下得到的。为了进一步验证本文网络模型性能，我们改变这两个阈值，测试网络模型在不同的阈值下准确率表现。如果在更加严格的阈值条件(δ_A 越小或 δ_J 越大)下准确率依旧维持在较高数值的话，那么可以认为网络模型具有更强的鲁棒性。我们设置从 $\delta_A=\frac{\pi}{6}$ 开始，按照固定步长不断缩小 δ_A；从 $\delta_J=0.25$ 开始，以固定步长不断增大 δ_J，从图 6 可以看出，我们的算法在角度阈值变小的时候保持了较高的准确率；在 δ_J 变大的过程中，准确率下降缓慢。而 GR-ConvNet 算法则无法很好地应对更加严格的阈值条件。这充分表明了本文的方法能够预测精度更高的抓取位姿，结果更加稳定，并且鲁棒性也更强。

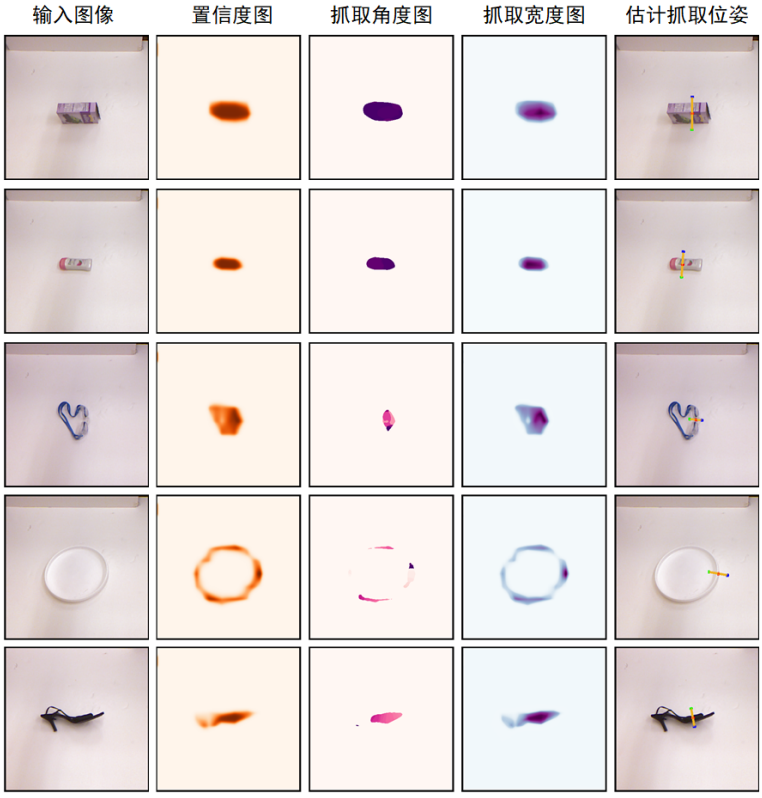

Figure 5. Visualization of grasp detection results on cornell dataset
图 5. 在 Cornell 数据集上的检测结果可视化

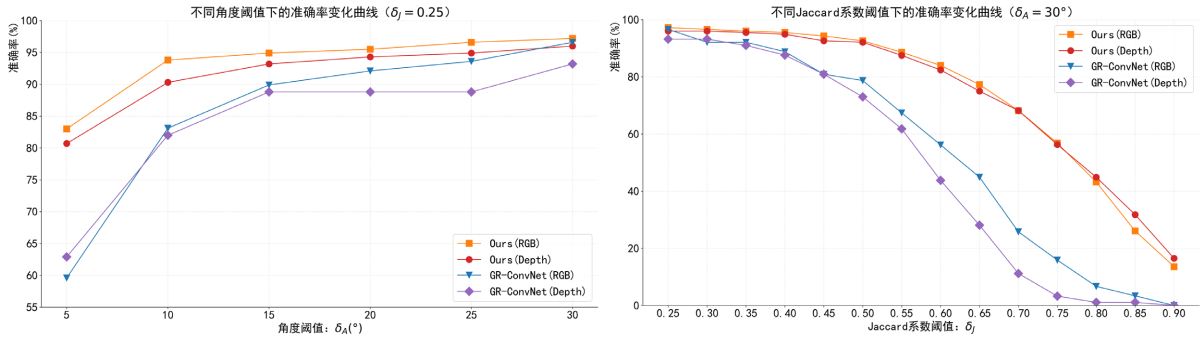

Figure 6. Grasp detection of accuracy at different thresholds
图 6. 不同阈值下的准确率变化曲线

4. 机械臂实物抓取实验

本文的方法在抓取数据集上取得了优秀的理论抓取位姿估计性能，但是仍有必要将本文算法部署到实际机器人系统上进行抓取测试验证本文提出方法的有效性。本文使用安川 MA2010 机械臂配合 Robotiq 二指夹爪进行操作，使用 RealSense D435i 相机进行场景信息的捕获，深度相机固定在机械臂的手腕处。在实验之前相机标定和手眼标定均已经完成。我们使用 ROS 配合 Moveit!运动控制框架驱动机械臂。本文选取了日常生活中常见的 15 中物体作为候选物体集，并将它们大致分成四大类——圆柱体、金属类、环状物体类、其它。如图 7 所示，这些物体不存在于用来训练网络的数据集中，可以用来评估网络对陌生物体检测抓取位姿的性能。

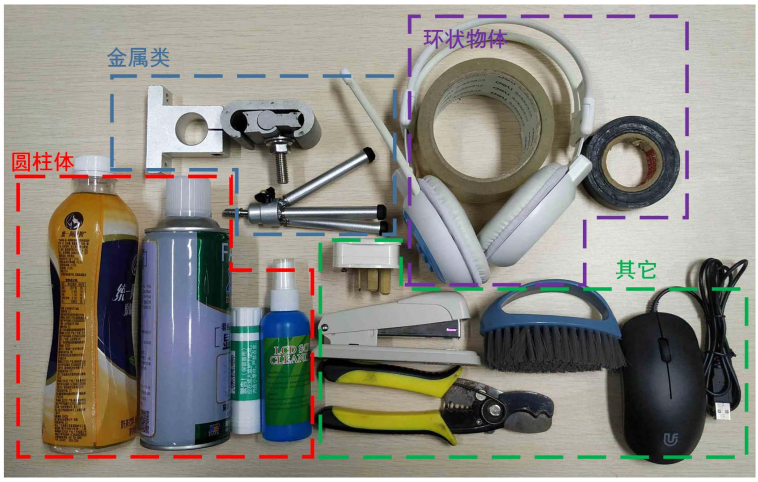

Figure 7. Object candidates for robotic grasping experiments
图 7. 实物抓取实验的候选物体集

4.1. 实验流程

我们从候选物体集中选取 n 个物体并将它们以随机的姿态放在操作台上。当机械臂抓住一个物体并且移动到指定的目标地点松开夹爪后，这次抓取尝试就视作成功，如果在这个过程中物体掉落这此次抓取尝试失败。对于每个场景，机械臂自主执行最多 $1.5n$ 次抓取操作，记录抓取成功的次数为 n_{ok}，抓取尝试的次数记作 $n_{attempt}$，我们使用抓取成功率(Success Rate，$SR = n_{ok}/n_{attempt}$)和抓取完成率(Completion Rate，$CR = n_{ok}/n$)来量化评估系统的性能。

4.2. 实验结果

我们用文献[6]的开源实现在相同场景下的表现作为对比。实验的结果如表 2 所示。我们设置了三个不同的 n 参数，用来表示场景中的物体数量，场景中的物体数量越多，任务的难度越高。从表格中的数据可看出，本文的方法可以在最大允许抓取次数的限制下以高成功率和高完成率完成复杂场景下的抓取作业。文献[6]在 $n=4$ 和 $n=6$ 时以 86.7%和 75.8%的平均抓取成功率完成物体的清理，而我们的系统的平均抓取成功率为 93.3%和 95.2%，我们的方法表现出更加优秀的性能。在 $n=8$ 的困难场景，文献[6]的平均抓取成功率仅为 52.8%，而我们的系统可以以 75.6%的平均抓取成功率达到 100%的抓取完成率。这表明我们的网络在有限的数据集中得到了更加充分的训练，在无需额外数据的情况下就有潜力处理复杂的物体抓取场景，即使待抓取的物体在数据中从未出现过。部分抓取实验场景展示见图 8，在图中我们只展示了置信度最高的 5 个抓取位姿。

Table 2. Results of robotic grasping experiment
表 2. 机械臂抓取实验结果

		Kumra *et al.* [6]			**Ours**		
设置	场景	尝试	*SR*	*CR*	尝试	*SR*	*CR*
$n=4$	1#	4/5	80	100	4/5	80	100
	2#	4/5	80	100	4/4	100	100
	3#	4/4	100	100	4/4	100	100
$n=6$	1#	6/7	85.7	100	6/6	100	100
	2#	6/8	75	100	6/7	85.7	100
	3#	6/9	66.7	100	6/6	100	100
$n=8$	1#	7/12	58.3	87.5	8/12	66.7	100
	2#	6/12	50	75	8/10	80	100
	3#	6/12	50	75	8/10	80	100
		49/74	71.7 (Avg.)	93.1	54/64	**88.0** (Avg.)	**100**

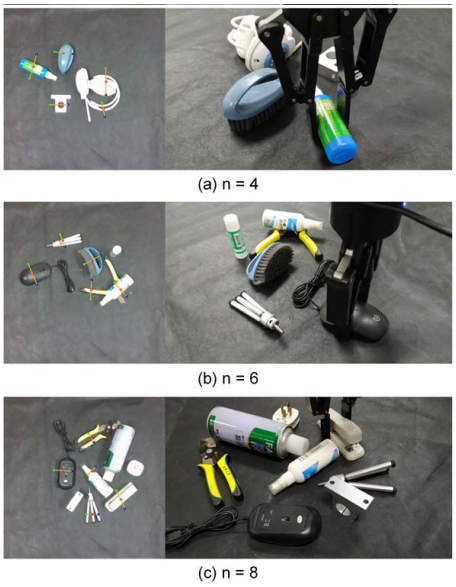

Figure 8. Illustration of robotic grasping in clutter scenarios
图 8. 杂乱场景下的机械臂抓取实验图

5. 结论

本文提出了一个有效的基于多层级特征的像素级端到端抓取检测网络。通过在卷积神经网络中嵌入多层级金字塔池化模块和多分支输出结构，我们的方法可以从有限的数据集中充分挖掘特征，在多物体杂乱场景下处理尺寸和姿态各异的未知物体的抓取位姿检测问题。我们使用 Cornell 数据集评估了提出的方法的理论性能，结果表明我们的方法的理论抓取性能优于现有方法；在机械臂实物抓取实验上，我们在多物体杂乱场景中以 88.0%的平均抓取成功率实现了 100%的抓取完成率。本文的算法并未针对运动物体的抓取位姿检测问题做出深入的研究，这需要调整算法网络模型和抓取系统架构来适应动态场景，这将是后续需要着重关注的问题。

参考文献

[1] Redmon, J. and Angelova, A. (2015) Real-Time Grasp Detection Using Convolutional Neural Networks. *Proceedings of the* 2015 *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, 26-30 May 2015, 1316-1322. https://doi.org/10.1109/ICRA.2015.7139361

[2] Lenz, I., Lee, H. and Saxena, A. (2015) Deep Learning for Detecting Robotic Grasps. *The International Journal of Robotics Research*, **34**, 705-724. https://doi.org/10.1177/0278364914549607

[3] Kumra, S. and Kanan, C. (2017) Robotic Grasp Detection Using Deep Convolutional Neural Networks. *Proceedings of the* 2017 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Vancouver, 24-28 September 2017, 769-776. https://doi.org/10.1109/IROS.2017.8202237

[4] Morrison, D., Corke, P. and Leitner, J. (2018) Closing the Loop for Robotic Grasping: A Real-Time, Generative Grasp Synthesis Approach. https://doi.org/10.15607/RSS.2018.XIV.021

[5] Morrison, D., Corke, P. and Leitner, J. (2019) Learning Robust, Real-Time, Reactive Robotic Grasping. *The International Journal of Robotics Research*, **39**, 183-201. https://doi.org/10.1177/0278364919859066

[6] Kumra, S., Joshi, S. and Sahin, F. (2020) Antipodal Robotic Grasping Using Generative Residual Convolutional Neural Network. *Proceedings of the* 2020 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Las Vegas, 24 October-24 January 2021, 9626-9633. https://doi.org/10.1109/IROS45743.2020.9340777

[7] Chu, F.J., Xu, R. and Vela, P.A. (2018) Real-World Multiobject, Multigrasp Detection. *IEEE Robotics and Automation Letters*, **3**, 3355-3362. https://doi.org/10.1109/LRA.2018.2852777

[8] Zhou, X., Lan, X., Zhang, H., Tian, Z., *et al*. (2018) Fully Convolutional Grasp Detection Network with Oriented Anchor Box. *Proceedings of the* 2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Madrid, 1-5 October 2018, 7223-7230. https://doi.org/10.1109/IROS.2018.8594116

[9] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the* 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[10] Zhao, H., Shi, J., Qi, X., Wang, X., *et al*. (2017) Pyramid Scene Parsing Network. *Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6230-6239. https://doi.org/10.1109/CVPR.2017.660

[11] Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. 15*th European Conference on Computer Vision*, Munich, 8-14 September 2018, 1-17. https://doi.org/10.1007/978-3-030-01234-2_1

[12] Hu, J., Shen, L., Albanie, S., Sun, G., *et al*. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[13] Yun, J., Moseson, S. and Saxena, A. (2011) Efficient Grasping from RGBD Images: Learning Using a New Rectangle Representation. *Proceedings of the* 2011 *IEEE International Conference on Robotics and Automation*, Shanghai, 9-13 May 2011, 3304-3311.

[14] Zhang, H., Lan, X., Bai, S., Zhou, X., *et al*. (2019) ROI-Based Robotic Grasp Detection for Object Overlapping Scenes. *Proceedings of the* 2019 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Macau, 3-8 November 2019, 4768-4775. https://doi.org/10.1109/IROS40897.2019.8967869

[15] Song, Y., Gao, L., Li, X. and Shen, W. (2020) A Novel Robotic Grasp Detection Method Based on Region Proposal Networks. *Robotics and Computer-Integrated Manufacturing*, **65**, Article ID: 101963. https://doi.org/10.1016/j.rcim.2020.101963