信息与控制　2025 年　第 54 卷　第 3 期：372 ~ 389

**DOI**：10.13976/j.cnki.xk.2024.1731　　**CSTR**：32166.14.xk.2024.1731

# 基于视觉的机器人端到端策略抓取估计综述

苏　康[1]，李嘉良[1]，李俊国[2]，张馨文[3]，刘　闯[1]

1. 石家庄铁道大学机械工程学院，河北 石家庄　050043；
2. 华北理工大学，河北 唐山　063210；
3. 汇通建设集团股份有限公司，河北 保定　074000

通信作者：刘闯，liuc@stdu.edu.can　收稿/录用/修回：2024 - 06 - 11/2024 - 09 - 06/2024 - 10 - 22

**摘要**

人工智能技术的迅速发展和不断突破为智能机器人抓取性能提升创造了很大空间。抓取估计是机器人实现抓取任务的关键，直接影响后续的抓取规划和抓取控制系统。端到端抓取策略不需要分步进行目标定位和目标姿态估计，直接从输入数据中学习并输出抓取信息。本文从平面级抓取和空间级抓取两方面综述基于视觉的端到端策略抓取估计方法，将平面级抓取方法分为估计抓取接触点和估计定向矩形两类，将空间级抓取方法分为面向对象和面向场景两类。此外，本文还对相关的数据集和抓取评估指标进行了简单介绍，指出了基于视觉的机器人端到端策略抓取估计方法面临的挑战及未来的发展方向。

**关键词**

抓取估计
端到端
平面级抓取
空间级抓取

**中图法分类号**：TP393，TP242

**文献标志码**：A

# Review of Vision-based Robot End-to-end Strategic Grasping Estimation

SU Kang[1], LI Jialiang[1], LI Junguo[2], ZHANG Xinwen[3], LIU Chuang[1]

1. *School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang* 050043, *China*;
2. *North China University of Science and Technology, Tangshan* 063210, *China*;
3. *Huitong Construction Group Co., Ltd., Baoding* 074000, *China*

**Abstract**

The rapid development and continuing breakthroughs in artificial intelligence technology have greatly improved the grasping capabilities of intelligent robots. Grasp estimation is critical for robots to perform grasping tasks, which directly affects subsequent grasping planning and control systems. Unlike traditional approaches requiring step-by-step target localization and pose estimation, end-to-end grasping strategies directly learn and output grasping information from input data. We review vision-based end-to-end strategic grasping estimation methods, covering planar-level and spatial-level grasping methods. Planar-level grasping methods are categorized into estimating grasping contact points and estimating oriented rectangles. Spatial-level grasping methods are also divided into two categories: object-oriented and scene-oriented approaches. In addition, we introduce relevant datasets and grasping evaluation metrics and highlights the challenges and future directions in vision-based end-to-end grasping estimation for robots.

**Keywords**

grasp estimation;
end-to-end;
planar-level grasping;
spatial-level grasping

## 0　引言

深度学习技术在计算机视觉领域的迅速发展为智能机器人抓取性能提升创造了很大空间。在基于深度学习的机器人抓取估计方法中，一部分算法通过反向传播和随机梯度下降以端到端的方式训练相当复杂的深度神经网络。这种端到端学习系统覆盖中央学习模块与“外围”模块（如表征学习和记忆形成），确保所有模块对权重参数可微，并将其作为一个整体进行学习[1-3]，以整个系统训练性能提升作为评价基准。

在机器人抓取任务中，端到端抓取策略不需要目标定位和目标姿态估计，从视觉输入（图像或点云信息）中学习完整的抓取过程，直接或间接生成机械手末端执行器的位置和姿态，如图 1 所示。模型复杂度低，且能够有效减少错误累计，这些优点使得端到端策略的深度学习方法在机器人抓取领域逐渐成为研究热点[4]。

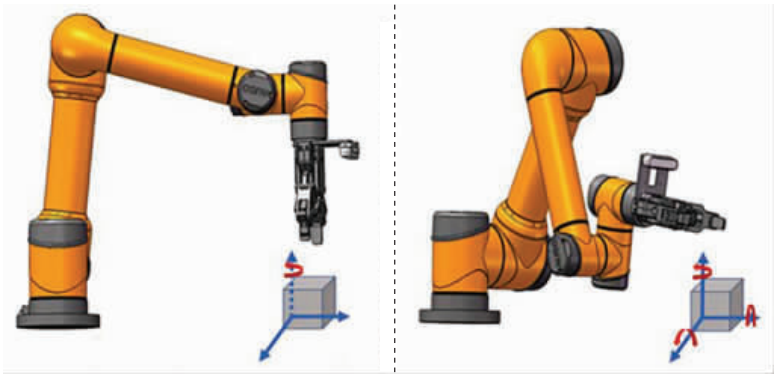


图 1　机器人抓取检测系统

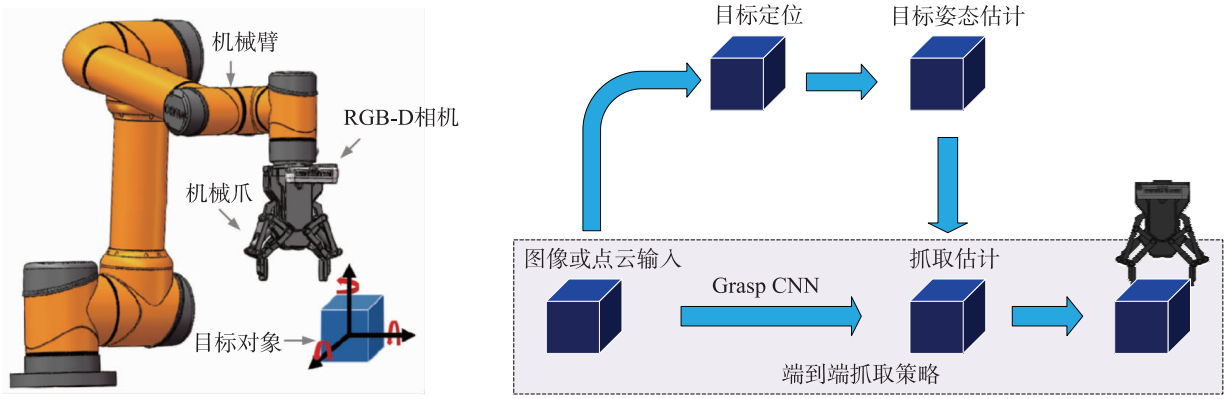
Fig.1　The robotic grasping detection system

近年来，国内外学者对机器人抓取检测方法进行了大量的研究，其中不乏综述类文献报道。曹家乐等针对目标估计中有无锚框，从基于锚点框的方法、无锚点框的方法和端到端预测的方法对单目视觉目标检测进行综述[5]，其中端到端方法属于无锚点框的目标检测方法，但其不需要后处理操作，检测系统架构简洁，文献中将其单独归为一类进行介绍。刘亚欣等在未知物体抓取检测技术研究现状分析一文中，从分析法和经验法两大方面介绍抓取检测技术，重点介绍了经验法中未知物体抓取的两种方法：传统的抓取检测方法（先通过抓取检测生成抓取位姿，再基于轨迹规划生成抓取轨迹，最终实现抓取）和基于视觉运动控制策略的端到端抓取[6]。DUAN 等对基于点云和深度学习的机器人抓取方法提出了新的分类方案：生成-评估框架、学习模式和应用，其中生成-评估框架是分类的核心部分，生成-评估框架外方法包括端到端、强化学习及其它方法[4]。YIN 和 DO 在机器人抓取检测的综述中均按照经典的 2 维和 3 维抓取方法进行分类。前者将 2 维抓取方法按图像模态分为基于彩色（RGB）、深度图及彩色-深度（RGB-D）三类，将 3 维抓取分为已知对象、未知对象及特例[7]。DO 等对基于视觉的机器人抓取进行了全面的综述，对抓取过程中的 3 个关键任务，即物体定位、物体姿态估计和抓取估计进行了详细的分类汇总，并介绍了相关主流数据集与评估指标[8]。

端到端策略的抓取方法网络架构简洁、高效、动态环境下可操作性高，但在基于视觉的机器人抓取检测领域尚缺乏对端到端策略方法进行梳理的综述类文献报道，缺乏对端到端方法进行系统的分析、比较和研究，基于此本文对该类方法进行了详细分类汇总。机器人在真实世界执行抓取任务的应用场景只有两种，即平面场景和空间场景[9-13]，故本文针对末端执行机构为平行夹持器的抓取检测系统，从平面抓取和空间抓取两方面（如图 2 和图 3 所示）对基于视觉的端到端策略抓取估计进行综述，分析了端到端抓取估计方法面临的挑战及未来的发展方向，为后续学者的研究和探索提供借鉴与参考。

## 1　平面级抓取

平面内抓取通常直接使用 RGB 图像、深度图像或 RGB-D 图像进行抓取估计。抓取目标位于水平工作台上，机械臂竖直向下从单个角度进行抓




(a) 平面级抓取示意图（机器人抓取可实现 $X$、$Y$ 轴方向的平移及围绕 $Z$ 轴的旋转）

(b) 空间级抓取示意图（机器人抓取可实现 $X$、$Y$ 和 $Z$ 轴方向的平移及围绕 $X$、$Y$ 和 $Z$ 轴的旋转）

图 2　2 维平面抓取和 6 自由度抓取

Fig.2　2D plane grasping and 6-DOF grasping

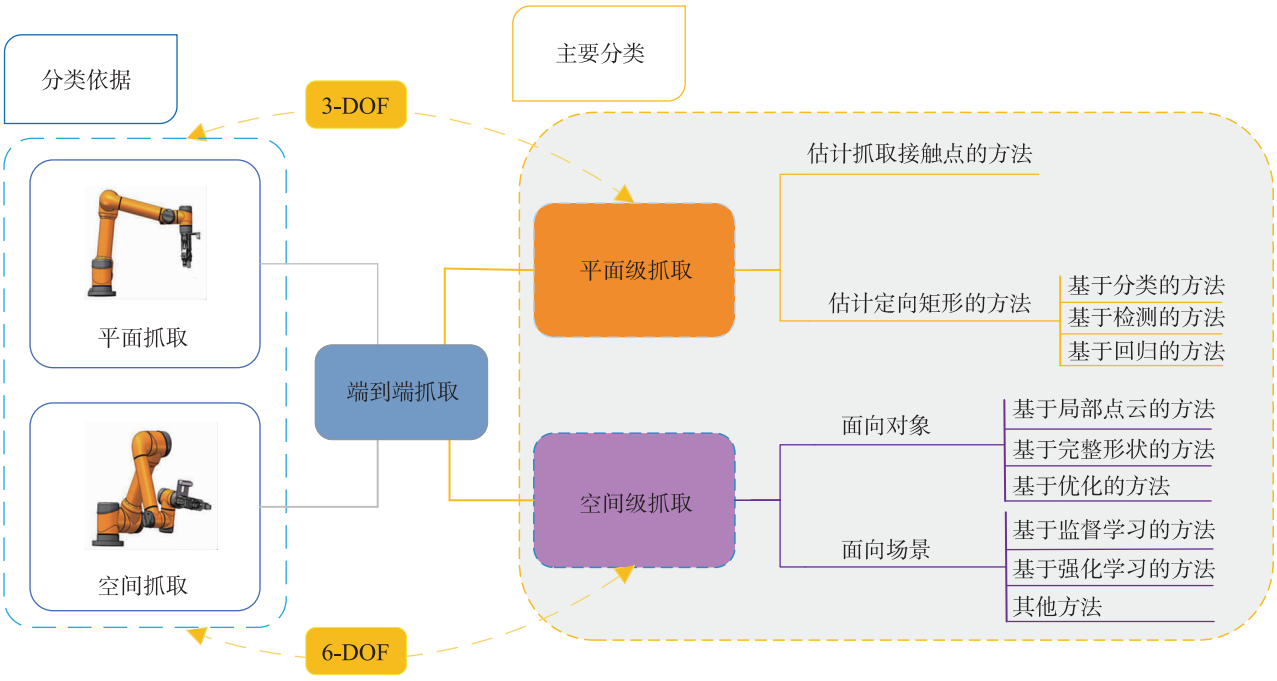



图 3　端到端抓取研究框架

Fig.3　A research framework for end-to-end grasping

取，其姿态能够简化为 3 个自由度，即 2 维平面内位置和 1 维旋转角度，抓取接触点和 2 维定向矩形能够唯一定义抓取姿态，这两种方法代表了两种候选抓取表示模型。图 4 绘制近年来平面级端到端抓取估计方法发展时间表，表 1 对其中的代表性方法进行了详细汇总。

### 1.1　估计抓取接触点的方法

估计抓取接触点的方法本质上属于分类问题，首先对候选抓取接触点进行采样，然后使用分析法（经验法）或基于深度学习的方法来评估抓取成功的可能性。分析法在机器人抓取的某些先验知识（对象几何学、物理模型或力分析）已知的前提下进行。

关于估计抓取接触点的方法，文献报道的国内外代表性研究成果如下。2018 年，ZENG 团队提出了一种能够处理广泛对象类别的抓取方法，该方法无需为新对象提供任何针对特定任务的训练数据，而能够在杂乱的环境中抓取和识别已知和新颖的物体[14]。在前者的基础上，CAI 提出了一种高精度抓取网络——Affordance interpreter network，将抓取检

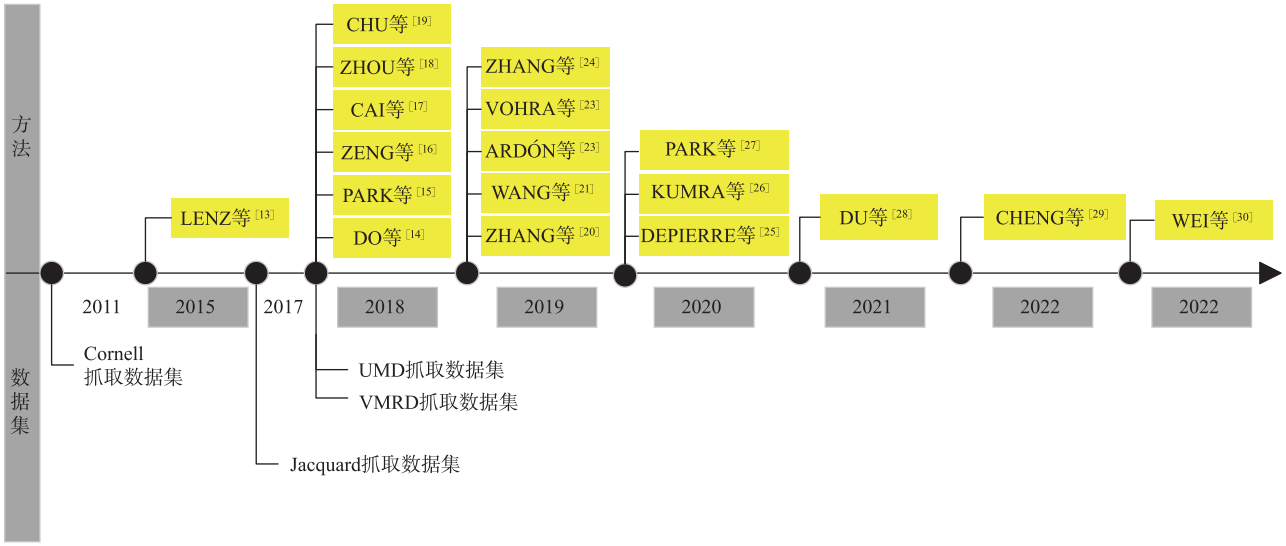



图 4　平面级端到端抓取估计方法发展时间表

Fig.4　Timeline for development of planar-level estimation methods based on end-to-end strategy

表 1　平面级抓取估计方法汇总与比较

Tab.1　Summary and comparison of planar-level grasping estimation methods

| 分类 | 方法 | 数据集 | 准确率 | | | 输入大小 |
|---|---|---|---|---|---|---|
| 评估抓取接触点的方法 | ZENG 等[14] | 自收集数据集 | 96.7% | | | 640×480 |
| | CAI 等[15] | V-REP 模拟生成 | 94% | | | – |
| | DO 等[16] | UMD 抓取数据集 | 73% | | | 244×244 |
| | | ImageNet 抓取数据集(IIT-AFF) | 79% | | | |
| | ARDÓN 等[21] | 自收集数据集 | 84% | | | – |
| | | | 图像级 | | 对象级 | |
| 评估定向矩形的方法 | LENZ 等[13] | Cornell 抓取数据集 | 92.0% | | | 640×480 |
| | PARK 等[17] | VMRD 抓取数据集 | 98.6% | | | 320×320 |
| | | Cornell 抓取数据集 | 74.2% | | | |
| | Zhou 等[18] | Cornell 抓取数据集 | 97.74% | | 96.61% | 320×320 |
| | CHU 等[19] | Cornell 抓取数据集 | 89.0% | | 96.1% | 227×227 |
| | WANG 等[20] | Cornell 抓取数据集 | 94.42% | | 91.02% | 400×400 |
| | ZHANG 等[22] | VMRD 抓取数据集 | 杂乱场景 | 熟悉堆叠场景 | 复杂堆叠场景 | – |
| | | | 90.6% | 71.4% | 59.4% | |
| | VOHRA 等[23] | 真实环境 | 88.16% | | 77.03% | – |
| | ZHANG 等[24] | VMRD 抓取数据集 | 86.0% | 67.1% | 88.8% | – |
| | PARK 等[25] | VMRD 抓取数据集 | | 76.4% | | 360×360 |
| | | Cornell 抓取数据集 | 98.6% | | 97.2% | |
| | DEPIERRE 等[26] | Jacquard 抓取数据集 | | 85.74% | | 320×320 |
| | KUMRA 等[27] | Cornell 抓取数据集 | 97.7% | | 94.6% | 640×480 |
| | | Jacquard 抓取数据集 | | 94.6% | | |
| | CHENG 等[29] | Cornell 抓取数据集 | | 95.4% | | 448×448 |
| | | Jacquard 抓取数据集 | | 91.8% | | |

注：– 无法获取。

测精确到像素级，该模型仅用少量抓取样本训练，使用模糊神经网络对 RGB 图像的每个像素集进行抓取启示图预测，这种基于视觉的数据驱动方法在遮罩目标和背景上获得显著性能[14]。深度学习的方法可以通过像素集评估候选抓取接触点的抓取效果，也可通过像素级掩码估计最佳抓取接触点。DO 提出一种深度学习网络——AffordanceNet，能够从 RGB 图像中同时检测多个对象及其启示，每幅图像的检测时间为 150 ms[16]。ARDÓN 在 AffordanceNet 网络的基础上，使用马尔可夫逻辑网络构建涉及不同语义属性的知识库图谱，这使得该抓取检测方法的泛化能力更优秀[21]。

上述方法中，ZENG 的模型特点是其不需要进行针对性训练便可以得到较高的抓取成功率，即使是面向杂乱场景中的新物体；CAI 的模型在有限量抓取样本条件下有优势；DO 和 ARDÓN 的模型更倾向于解决模型的泛化问题。现有基于深度学习的估计抓取接触点的方法大都趋向于设计高性能的特征提取器，但代价是更重的计算负担和特定训练数据过拟合[20]。端到端模型对不同层的特征进行上采样和融合构建，实现了更高的抓取检测率，并对输入区间内的变化具有鲁棒性[29]。

### 1.2 估计定向矩形的方法

早在 2011 年 JIANG 就首次提出使用定向矩形来表示抓取器配置(gripper configuration)，该方法通过两步程序进行，先利用快速计算出的特征有效修剪搜索空间，再使用高级特征来精确选择抓取[3]。针对基于定向矩形的抓取模式，深度学习方法可分为 3 类，即基于分类、回归和检测。这些方法采用机器人抓取的 5 个参数来对抓取姿态进行编码，即对象位置坐标、旋转角度和矩形框尺寸，表示为($x$, $y$, $\theta$, $h$, $w$)。由于这种表示不能区分每个抓取矩形的质量，且对象检测框长度 $h$ 也可视作夹持器的宽度，因此最新方法将参数 $h$ 省略并引入 $q$(抓取质量分数)，新表示为($x$, $y$, $\theta$, $h$, $w$)。

#### 1.2.1 基于分类的方法

使用基于分类的方法训练分类器来评估候选抓取矩形时，在分类器中得分最高的定向矩形会被选中。LENZ 等[13]提出了一个两级联模型，第一个网络具有较少的特征，运行速度快，且能够有效剔除不可能的候选项；第二个网络具有较多的特征，运行速度慢，只需在少数特定输入上运行。尽管该网络达到了很高的精度，但迭代计算过程十分缓慢。针对抓取检测，通常回归方法要优于分类方法，但 PARK 采用了一种多级空间变换器网络(STN)的分类方法能够为掌握抓取候选的中间步骤(抓取位置、方向)提供可观察性[17]。在物体堆叠场景下，考虑抓取对象与场景物体之间隶属关系的机器人抓取任务一直具有挑战性。ZHANG 提出了基于目标区域的机器人抓取检测(ROI-GD)，使用感兴趣区域而非整个场景中的特征来检测抓取，在物体堆叠场景下验证了算法的有效性[22]。该团队通过标记视觉操作关系数据集，贡献了一个多对象抓取数据集，数据规模远大于 Cornell 抓取数据集。上述方法中，LENZ 的模型结构复杂，计算精度高，但运行速度慢；PARK 方法中采用的 STN 网络能够对高分辨图像进行实时检测，并保证较高检测率；ZHANG 提出的 ROI-GD 算法[22]在单个对象抓取上与最先进抓取检测算法相当。这类方法都以包含更多信息的 RGB-D 图像作为输入，程序简单明了、准确度较高，但总体运行速度相对较慢。

#### 1.2.2 基于检测的方法

基于检测的方法使用锚框(先验框)帮助生成和评估候选抓取矩形。通过使用锚框，这类方法不直接回归抓取配置，而是先预测抓取框的变形，通过引入预期抓取尺寸的先验知识来简化回归问题[25]。针对多个对象的抓取：ZHOU 通过引入定向锚框来消除方向分类，定向锚框机制模型采用一个锚框与多个方向，而非采用多个尺度或宽高比的参考框，根据参考方向预测抓取的角度，该网络预测了特征图中每个定向参考框的 5 个回归值和一个抓取质量得分[26]。CHU 提出了一个保留所有候选矩形同时输出所有候选矩形排名的网络，将回归问题转换为区域检测和方向分类问题的组合，方向分类包含抓取质量得分，因此该网络同时预测抓取回归值和离散方向的分类得分[19]。DEPIERRE 在 ZHOU 的基础上引入一个新的损失函数将抓取参数的回归与可抓取性评分相关联，进一步说明了对于抓取精度而言，方向比多尺寸锚框更重要[26]。上述方法中，ZHOU 所采用的定向锚框和匹配策略在当时对提高机器人抓取精度很有启发性，模型各方面性能都优于 CHU；DEPIERRE 提出的网络架构通过新引入的损失来改进抓取回归，但模型的各项性能指标都较低。

#### 1.2.3 基于回归的方法

基于回归的端到端方法是一种单阶段方法，它不过滤候选抓取，而是直接预测抓取姿态。这类方法通过模型训练，直接生成目标对象位置与方向的

抓取参数。KUMRA 提出的生成残差卷积神经网络模型——GR-ConvNet 能够从通道的输入图像对未知对象生成鲁棒的逆抓取(antipodal grasps)[27]。对于复杂形状的物体，从物体中心点或沿物体主轴抓取，往往会失败。VOHRA 提出了一种抓取估计策略，通过估计点云中的对象轮廓，得到图像平面中的物体骨架，用每个骨架点的抓取矩形和物体中心点相对应的点云数决定最终的抓取矩形。该策略对复杂形状物体的表现非常出色，并能预测出有效的抓取配置，且不需要繁琐的抓取配置和采样步骤，使得输出更加稳定[23]。针对平面杂乱场景下的复杂多任务问题，机器人需对抓取对象是否存在堆叠问题进行分析[28]。ZHANG 基于感兴趣区域，使用 ROI-GD 抓取检测算法从输入特征回归抓取参数。该网络由多个深度神经网络组成，分别负责生成局部特征图、抓取估计、物体检测和关系推理，优化堆叠物体抓取并解决物体检测与视觉操控关系推理的组合问题[24]。相比之下，PARK 使用单一多任务深度神经网络，通过后处理功能同样实现了目标对象的抓取估计、物体检测与物体之间关系推理信息的提取[25]。这类基于回归方法能够缩减网络模型，同时可以减少一些重复的计算，所以其在速度上有了较大的提升。上述方法中，KUMRA 提出模块化抓取系统在 Cornell 数据集上，抓取成功率远高于其它方法，但其图像分割和对象分割准确率低于 PARK 方法；在 VMRD 抓取数据集上，PARK 网络的抓取准确率要高于 ZHANG 所提出的 ROI-GD 抓取检测算法。

表 2 分别从优势、抓取成功率和局限性对每一类平面级端到端抓取估计方法进行了对比。

表 2　平面级抓取方法对比分析

Tab.2　Comparative analysis of planar-level grasping methods

| 分类 | | 优势 | 成功率 /% | 局限性 | 参考文献 |
|---|---|---|---|---|---|
| 评估抓取接触点的方法 | | 应用广泛，对于实际抓取任务，所需执行抓取动作信息少 | 73.0 ~ 96.7 | 忽略实际几何形状可能导致真实手无法到达的接触位置 | 文[14 – 16, 21] |
| 评估抓取矩形的方法 | 基于分类的方法 | 基于分类的方法训练分类器来评估候选掌握，并将选择得分最高的一个，这类方法模型简单，准确度高 | 74.2 ~ 98.6 | 效率低，迭代缓慢 | 文[13, 17, 22] |
| | 基于检测的方法 | 这类方法借助参考锚框来帮助生成和评估候选人掌握的信息，这使模型对复杂场景的特征信息检测更加准确 | 85.74 ~ 89.0 | 启发式引导的特征选择，基于覆盖锚取样 | 文[18 – 19, 26] |
| | 基于回归的方法 | 训练模型以直接产生位置和方向的抓取参数，其模型容易理解解释，且统一模型提高了性能指标 | 67.1 ~ 97.7 | 有较大偏差，模型更易欠拟合 | 文[23 – 24, 27 – 28, 30] |

### 1.3　数据集和评估指标

表 3 对公开可用的 2 维平面抓取数据集进行了汇总。2 维平面抓取数据集有 CMU 数据集[31]、Jacquard 数据集[32]和 Cornell 抓取数据集[3]，其中 Cornell 抓取数据集使用最为广泛。Cornell 数据集包含有图像分割和对象拆分数据，图像分割指将图像随机分割，用来测试算法对已知物体在新位置的泛化能力；对象拆分是将同一物体的所有图像放入相同的交叉验证拆分中，用以测试算法对未知物体的泛化能力。与 Cornell 抓取数据集相比，Jacquard 数据集拥有更庞大的数据内容，抓取方式也更多样化，近期研究者的工作开始在 Jacquard 抓取数据集上进行实验。

表 3　公开可用的平面级抓取数据集

Tab.3　Summaries of public available planar-level grasping datasets

| 数据集 | 模态 | 物体 | 图像 | 抓取 |
|---|---|---|---|---|
| Cornell 抓取数据集[31] | RGB-D | 240 | 1 035 | 8 019 |
| VMRD 抓取数据集[25] | RGB | 17 688 | 4 683 | $1\times10^5$ |
| CMU 抓取数据集[31] | RGB-D | >150 | 50 567 | – |
| Jacquard 抓取数据集[32] | RGB-D | 1 500 | $6.7\times10^7$ | $6.7\times10^7$ |

目前端到端策略的抓取估计性能评价指标主要有点指标和矩形指标[8]，前者需要设定一个阈值来评估预测中心和实际抓取中心之间的距离，但它在确定距离阈值方面有困难，且无法考虑到抓取角度；后者需要确保预测角度与真实抓取角度相差范

围在30°以内，且预测抓取的Jaccard系数大于25%。此外，还存在一些指标用来评估预测抓取点的性能，如抓取成功率、鲁棒抓取率、计划时间等。2维平面抓取方法适用于从单一角度抓取的固定场合，如果改变抓取角度，网络将无法学习到合适抓取的姿态，因此，平面级抓取方法不适合任意角度的抓取任务。

此外，公开的平面级抓取数据集真实场景中的数据规模较小，多在模拟环境中评估测试，很难进行拓展迁移。这些问题在一定程度上限制了模型的发展和应用，6-DOF抓取可以在3维空间中完成从不同角度的抓取任务[33]。随着深度相机技术的进步、计算机计算能力的提高及传感技术的发展，科研人员逐渐将目光转向6-DOF抓取。

## 2 空间级抓取

随着现实生活中机器人3维应用场景的增多，机器人空间内抓取成为当下研究的热点[34]。空间内抓取则需要考虑物体在3维空间中的位置、形状和姿态，RGB-D摄像头或双目视觉系统能够提供更多的物体信息，帮助机器人判断物体在3维空间中的准确位置和表面特征，从而进行更加复杂和精准的抓取操作并帮助抓取系统动态地推理机器人、目标对象以及环境三者之间的交互关系。本文将空间内抓取分为面向对象和面向场景两类[35]，是否考虑环境中的约束是区分面向对象方法和面向场景方法的标志。面向对象方法的典型形式即试图将物体的预定义刚性3维模型与感知场景中对象的几何图形匹配叠加，利用该匹配得出相应的抓取姿态。面向场景的方法追求对整个场景的理解，这种方法在采样过程中动态生成抓取候选对象。图5和表4对近期文献报道的空间内抓取方法进行了汇总与比较。

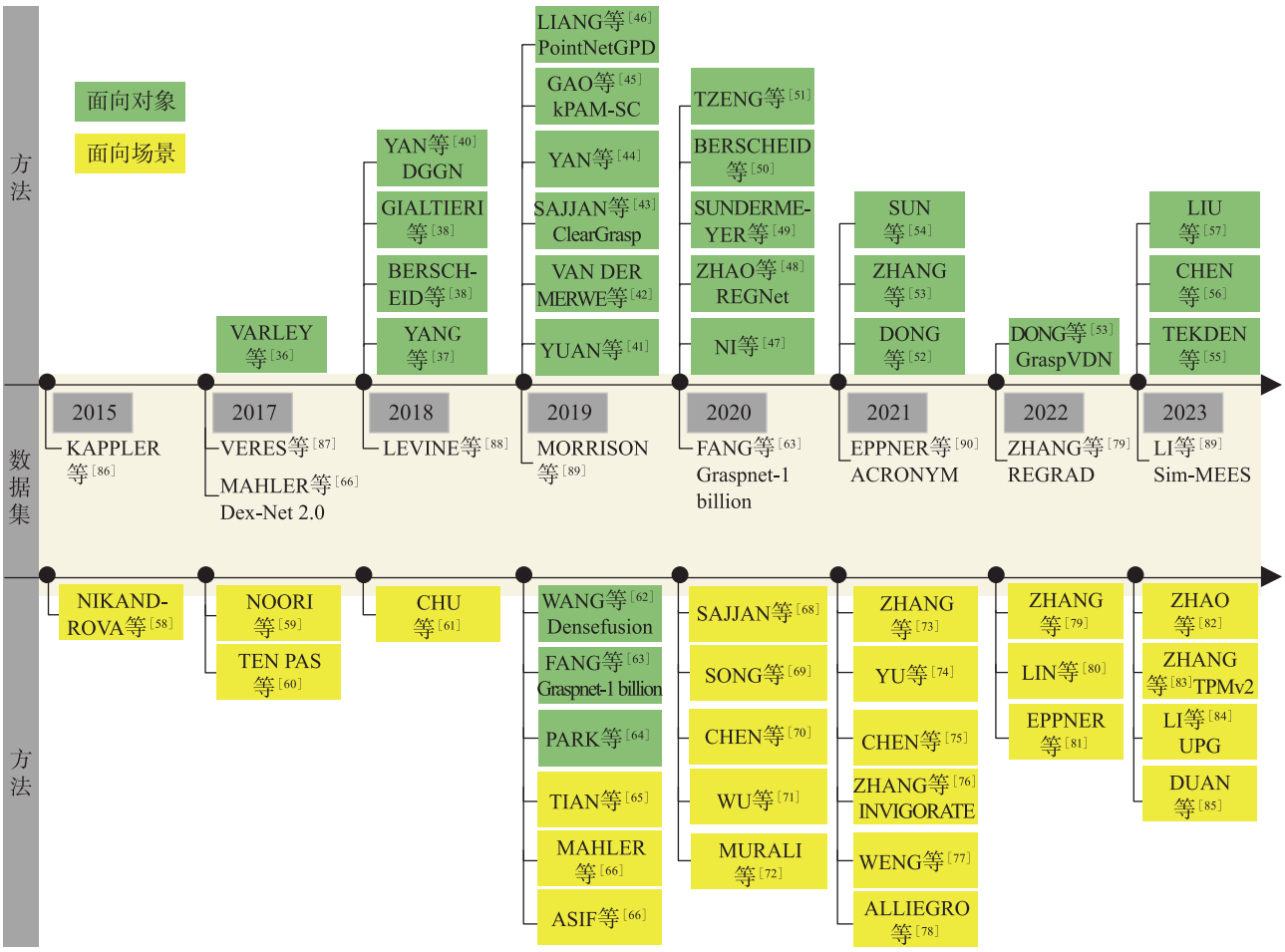



图5 空间级端到端抓取估计方法发展时间表

Fig.5 Timeline for development of spatial-level grasping estimation methods based on end-to-end strategy

表 4　面向对象的抓取方法汇总与比较

Tab.4　Summary and comparison of object-oriented grasping approaches

| 分类 | 工作 | 骨干网络 | 成功率（完成率）/% | 环境 | | | | 模拟/真实（S/R） |
|---|---|---|---|---|---|---|---|---|
| | | | | 对象排列 | 对象数量 | 对象形状 | 新环境测试 | |
| 基于局部点云的方法 | LIANG 等[46] | Pointnet | 89.33<br>100 | 杂乱 | 47 | 不规则 | 是 | R |
| | NI 等[47] | Pointnet++ | 82.95<br>91.50 | 杂乱 | 79 | 不规则 | 是 | R |
| | ZHAO 等[48] | Pointnet++ | 79.34<br>96.00 | 杂乱 | – | 不规则 | 是 | R |
| | FANG 等[63] | Pointnet++ | 96.00 | 杂乱 | 10 | 不规则 | 是 | R |
| 基于完整形状的方法 | VARLEY 等[36] | CNN | 93.33 | 杂乱 | – | 不规则 | 是 | R |
| | YANG 等[37] | DMPs | 85.71 | 单一 | – | 规则 | 否 | R |
| | YUAN 等[41] | Q-learning | 78.00 | 单一 | – | 规则 | 否 | R |
| | VAN DER MERWE 等[42] | PointSDF | 78.66 | 单一 | – | 规则 | 是 | S/R |
| | WANG 等[45] | CNN | 94.00 | 杂乱 | 21 | 规则 | 否 | R |
| | SUNDERMEYER 等[49] | ResNet-50 | ADD | 单一 | – | – | 否 | R |
| | BERSCHEID 等[50] | FCNN | 86.00 | 杂乱 | – | 规则 | 是 | R |
| | TZENG 等[51] | CNN | 71.4 ± 14 | 单一 | – | 规则 | 否 | S/R |
| | DONG 等[52] | VGG-16 | 98.38 | 单一 | – | 不规则 | 否 | R |
| | ZHANG 等[53] | Darknet19 | 图像分割 88.7　对象分割 87.2 | 单一 | – | 规则 | 否 | R |
| | SUN 等[54] | PointNet++ | 87.40 | 单一 | – | 不规则 | 否 | R |
| 基于优化的方法 | SAJJAN 等[43] | DRN-D-54 | 72 | 杂乱 | – | 不规则 | 是 | S/R |
| | YAN 等[44] | CNN | 75.76 | 单一 | 7 – 39 | 不规则 | 是 | S |
| | GAO 等[45] | ShapeHD | 74.00 | 单一 | 33 | 不规则 | 否 | R |

### 2.1　面向对象

面向对象的方法利用目标对象的预定义刚性 3 维模型与感知场景中对象的几何形状之间的匹配叠加，生成机械手对应的抓取姿态，即将已知对象的抓取姿态转移到检测对象。当新对象与已知抓取对象相似时，新对象能够用类似的姿态成功抓取。本文将面向对象的抓取方法分为基于局部点云的方法、基于完整形状的方法和基于优化的方法三种。

#### 2.1.1　基于局部点云的方法

基于局部点云的方法适用于未知对象，目前存在两种形式，一是先对大量抓取候选进行采样，然后利用多种方法评估抓取质量，这是一种基于分类的方式；二是隐式地估计抓取质量，并以单个视图输入的方式直接预测 6-DOF 抓取姿势，这是一种基于回归的方式。LIANG 提出的 PointNetGPD 模型属于分类方法，它通过直接处理抓取空间内的 3 维稀疏点云对目标对象进行抓取估计[46]。在数据集训练生成过程中，该方法结合力闭合和无摩擦抓取扳手空间（GWS）分析来评估抓取质量，相比同类别方法更具轻量化，且对新物体的泛化性能更优。在 PointNetGPD 模型的启发，NI 提出的 Pointnet++ 抓取模型不需要抓取采样或搜索过程，而是结合全局数据信息直接预测所有抓取姿态、类别和抓取质量得分[47]。随后，ZHAO 使用 PointNet++ 模型将一个单视图点云作为输入，提出 REGNet 抓取检测网络，其包括 3 个部分：用于对选定的抓取正点生成一组抓取建议的分数网络（SN），用于选择具有最高置信度抓取正点的抓取区域网络（GRN）和用于根据本地抓取特征细化抓取配置的细化网络（RN）[48]。REGNet 是先进的空间抓取检测方法，其性能优于 PointnetGPD。为了解决复杂场景训练数据不足和缺乏评价基准的问题，FANG 团队[63]建立了一个具有 10 亿数据规模的抓取检测数据集 GraspNet，其包含 97 280 张 RGB-D 图像和超过 10 亿的抓取配置。基于几何推理和上下文编码，该团队提出一种采用 PointNet++ 骨干网络的端到端抓取姿态预测模型，以解耦的方式学习接近方向和操作参数，能够有效减少错误产生。上述方法中，LI-

ANG 提出的 PointNetGPD 模型在点云输入不精确和传感信息不足的情况下仍然保持了较高抓取成功率，在 YCB 抓取数据集，抓取准确率明显高于 NI[47] 和 ZHAO[48] 所采用的方法；NI、ZHAO 和 FANG[63] 都是基于 PointNet++模型，其中 FANG 的方法抓取准确度远高于其余两种方法，原因在于该团队建立了规模庞大的抓取数据集。

### 2.1.2 基于完整形状的方法

基于完整形状的方法主要有两种形式：一是估计已知物体完整 3 维形状上的 6-DOF 抓取姿态，并将其从图像坐标转换到摄像机坐标；二是直接从输入的单视角点云数据生成摄像机坐标下的完整 3 维形状，再进行 6-DOF 抓取姿态估计。基于完整形状的方法在机器人空间内抓取检测领域成为主流。

针对算法的精度及高效性问题，以下研究提出的方法在测试中表现优异[64]。VARLEY 提出了一个通过完整形状来实现机器人抓取规划的深度学习框架，卷积神经网络(CNN)通过单视角点云数据在离线训练下进行快速 3 维形状补全，最后在 GraspIt! 环境中重建网格并执行了高质量抓取[36]。WANG 提出了一种异构体系结构的通用框架 DenseFusion，用于从 RGB-D 图像估计已知对象的抓取姿态，该框架集成了一个端到端迭代姿态优化过程，在 YCB Video 和 LineMOD 这两个数据集上，精确姿态估计和快速推理性能表现优异[62]。SUNDERMEYER 提出了一种基于 RGB 图像的实时目标检测和 6-DOF 姿态估计系统，这是一种基于自动去噪编码器的变体，它不需要真实的、带姿势注释的训练数据，直接在渲染的 3 维模型视图中学习隐式表示，在码本中找到最佳匹配作为预测抓取姿态[49]。以上方法简化了部分抓取流程如大量样本训练、后处理步骤，同时结果实现了对遮挡、杂乱背景的鲁棒性，并能够推广到不同的环境[92]。

针对多层 CNN 过拟合问题，DONG 提出了一种识别物体的类别并利用物体本身的特征来预测抓取配置的网络模型，其丢弃了网络中的冗余特征提取结构，设计了一种目标特征关注机制，引导模型根据语义信息关注目标对象本体的特征进行抓取估计。这种方法有效地减少了与目标对象弱相关的背景特征，使目标特征更加独特，提高了抓取估计的准确性和效率[52]。此外，ZHANG 联合采用角度和位置的拟合，引入数据集预处理和迁移学习的方法来避免网络过拟合[53]。

针对重叠对象检测精度较低的问题，SUN 借助结构光对场景内的点云信息进行高精度重建，利用模板与场景目标之间的转换关系得到场景中多个堆叠目标对象的抓取姿态[54]。BERSCHEID 和 YANG 团队把运动原语和全卷积神经网络(FCNN)进行奖励估算的学习方法扩展到抓取动作，将机器人对运动原语的学习与一次性模仿学习相结合，实现了完整的抓取技能转移，并很好地优化了抓取过程的刚度泛化和运动泛化问题[37, 50]。

在真实世界中验证抓取算法，积累抓取经验需要大量成本，但在虚拟仪器环境或模拟中很容易实现[41]。为了减少使用真实世界收集的训练数据量，YUAN 提出了一种从模拟到现实的迁移方法，使仿真模拟学习到的重排策略适应现实世界的输入数据，并借助 Baxter Robot 在模拟和真实环境中进行验证，结果表明该方法可以有效地改进模拟训练过程，并能使所学策略适应真实世界的抓取任务[41]。TZENG 团队的研究与其相似，他们通过在源域和目标域中使用弱对齐的图像对来消除对昂贵注释的需求，这种弱成对匹配技术能够有效地补偿域偏移，从而在现实世界中实现更佳的机器人性能[51]。VAN DER MERWE 提出使用 PointSDF 隐式曲面重建算法直接从 3 维点云回归检测对象的符号距离函数，并提出一种抓取预测的深度学习框架，隐式学习几何感知的点云编码[42]。

当精确 3 维模型可用时，上述方法均能实现目标抓取姿态估计，并生成目标物体的抓取姿态。当现有的 3 维模型与目标模型有差异时，6-DOF 姿态估计将产生较大偏差，进而导致抓取失败。在这种情况下，通过补全 3 维点云并对其进行三角剖分以获得完整形状，在重建的完整 3 维模型上进行 6-DOF 抓取姿态估计会更精确，且使用者能够自主开发各种抓取模拟工具包，以方便结合触觉信息更好地进行抓取估计[57]。

### 2.1.3 基于优化的方法

基于优化的方法在深度学习网络中增加了优化器对模型参数进行优化。针对 3 维传感器难以对透明物体进行精确深度估计的问题，SAJJAN 提出了一个深度学习框架 ClearGrasp，给定透明物体的单张 RGB-D 图像，ClearGrasp 使用深度卷积网络推断曲面法线、透明表面的遮罩和遮挡边界，这优化了场景中所有透明曲面的初始深度估计[68]。从现有抓取姿态中转移抓取在高级机器人操作任务中有潜在用途，YAN 等提出一种深度几何感知抓取网络(DGGN)，它包括形状生成网络和结果预测网络，

从 RGB-D 图像输入中学习 6-DOF 抓取，然后使用简单的无导数优化技术寻找最优抓取姿势[44]。GAO 等[45]结合关键点和密集几何(点云或网格)的组合优势提出一种新的混合对象表示法，利用基于关键点检测和形状生成的学习方法，直接从传感器输入的未处理数据中感知密集几何体和关键点信息[60]。上述 3 种方法抓取成功率都在 75% 左右，远低于其余两类方法，原因在于基于优化的方法大多针对一些特殊对象，如透明物体、巨大对象等。这类方法在问题规模较小时，算法能在合理的时间内找到问题的最优解；但当问题规模较大时算法的计算，复杂度高，求解时间呈指数级增长。

面向对象的方法在目标形状和纹理重建完整时能够产生精确的抓取，但它依赖于一个包含丰富对象 3 维模型的数据库，帮助算法转移到新对象或提供关键的语义和几何坐标信息。目前面向对象的方法主要有几个缺点：一是这类方法高度依赖于对象分割的准确性，且训练支持多种类对象的网络并不容易；二是构建大型 3 维数据库是一项非常繁重且艰巨的任务；三是要求数据库中存在与抓取对象相似的物体形状和抓取姿态；四是对于遮挡对象，关键坐标特征丢失，计算高质量的抓取点也是一项挑战。因此，依赖于静态数据库的面向对象方法很难扩展到动态场景和可变形对象，它的优点和缺点都来自于对数据库的严重依赖。

### 2.2　面向场景

面向场景的方法起源于 SAXENA 的工作，这种方法不需要也不尝试构建或补全对象的 3 维模型，而是给定一个物体的两个(或更多)图像，然后识别几个抓取点，用一个有监督学习模型来定位对象，最后对这组稀疏点云进行三角剖分以获得一个 3 维位置，并在该位置尝试抓取(只关注混乱场景)。面向场景的方法添加了对整个场景的理解[74]，故能够推广到新的对象和环境，并对环境作出动态反应[58, 65, 69, 70, 83, 100-102]。本文将面向场景的方法分为 3 类，分别是基于监督学习、基于强化学习及其它方法，表 5 对近年来空间级面向对象的端到端抓取方法进行了汇总与比较。

表 5　面向场景抓取方法汇总与比较

Tab.5　Summary and comparison of scene-oriented grasping approaches

| 分类 | 工作 | 骨干网络 | 成功率(完成率)/% | 环境 | | | | 模拟/真实(S/R) |
|---|---|---|---|---|---|---|---|---|
| | | | | 对象排列 | 对象数量 | 对象形状 | 新环境测试 | |
| 基于监督学习的方法 | LIN 等[80] | ResNet | 81.67 | 杂乱 | 8 | 不规则 | 否 | R |
| | ZHANG 等[94] | CNN | 98.54 | 杂乱 | 1 | 规则 | 否 | R |
| 基于强化学习的方法 | WU 等[71] | PointNet++ | 93.20 | 单一 | 196 | 规则 | 是 | S/R |
| | WANG 等[95] | PointNet&PointNet++ | 81.20 ± 8.9 | 杂乱 | 9 | 不规则 | 否 | S/R |
| | LI 等[97] | PointNet++ | 91.30 | 杂乱 | 9 | 规则 | 否 | S/R |
| | ZHAO 等[98] | PoseCNN | 85.80 | 杂乱 | 30 | 规则 | 是 | S |
| 其他方法 | DONG 等[35] | ResNet18 | 60.00 | 杂乱 | 20 | 不规则 | 是 | S |
| | TEN PAS 等[60] | CNN | 93.00 | 杂乱 | 55 | 不规则 | 否 | R |
| | CHEN 等[75] | ResNet-50&ImageNet | 93.20 | 杂乱 | 31 | 不规则 | 否 | R |
| | ZHANG 等[76] | NN modules, O-Net, R-Net, G-Net, and Q-Net | 83.00 | 杂乱 | - | 不规则 | 否 | R |

#### 2.2.1　基于监督学习的方法

在某种特殊环境，自由手绘草图也能成为机器人生成抓取配置的一种交互方式。LIN 提出了一种通过理解自由手绘草图内容进行条件抓取的方法，该模型以端到端的方式进行训练和测试，在 VMRD 和 GraspNet-1Billion 数据集上验证了方法在杂乱场景中的通用性和有效性[80]。在无序抓取技术中，对象的无序叠加和自我遮挡下造成样本难以标记。ZHANG 提出了一种无序场景下基于自动标注数据集的机器人抓取检测方法，该方法结合端到端和采样评价抓取策略的优势，构建了基于机器人抓取检测的两阶段训练和端到端无序抓取模型，在虚拟环境中验证了方法的有效性和泛化性能[94]。上述两种方法，ZHANG 的抓取成功率为 98.5%，远高于 LIN 的 81.6%，原因在于对于存在遮挡和重叠的复杂环境，ZHANG 从抓取角度方面展开研究，建立

了基于自监督学习的抓取角度分类模型。监督学习最突出的特点是数据集内每个样本具有唯一标签，这给模型带来了许多局限性，但其易于实现、易于训练且适用范围广。端到端的学习方法在一定程度上是一种基于监督学习思想的变体。

#### 2.2.2 基于强化学习的方法

对于6-DOF机械手抓取设置，现有大多数方法采取的策略在采样效率(启发式采样)和最佳抓取覆盖率之间存在冲突。针对这个问题，WU提出了一种端到端抓取建议网络GPNet(Grasp Proposal Networks)，从单一的未知目标物体图像中进行6-DOF抓取姿态预测。GPNet对抓取提案模块进行了重点设计，使其在离散但规则的3维网格边界顶点处定义抓取中心的锚，可灵活支持精确多样的抓取预测[71]。

WANG将端到端抓取策略扩展到有障碍物的复杂场景[95]，该研究采用分层框架基于局部点云观测学习无碰撞目标驱动抓取，使用累积的分割场景点云来表示目标和障碍物，减少了模拟与真实的感知差距，使用二元分类器来选择预训练的抓取策略，以提高抓取成功率，使用潜在条件闭环策略执行避障和精确抓取[96]。上述2种方法都是基于Pointnet++框架，WU方法的抓取成功率为93.2%，略高于WANG，但是WANG主要解决的是有障碍物的复杂场景下的抓取，这个不确定因素使得其抓取成功率为72%~90%。这些基于强化学习的方法增强了机器人在试错中自我探索的能力，训练成功的模型在规划抓取上更具灵活性。然而，算法的搜索空间会变得非常大或者需要试错的抓取姿势趋于无穷多，从而使学习效率变得非常缓慢[97]。

#### 2.2.3 其他方法

基于监督学习、强化学习的方法不一定采用端到端策略，反之亦然。机器人在真实环境下执行抓取任务时，可能会面临以下问题：机器人有限的工作空间无法囊括理想的抓取姿势；场景中的碰撞阻碍特定轨迹的执行；夹持器改变或环境改变时抓取策略失效。应对这个问题，DONG考虑环境因素对抓取姿势的约束，提出了一种面向场景的抓取估计方法GraspVDN(Grasp Vector Detection Network)，以场景的RGB图像、深度图像或RGB-D图像作为输入，其估计值通过相应的策略转换为6-DOF抓取姿态[35]。该方法在包含十亿抓取姿势的GraspNet数据集中表现优异，但其与拥有复杂网络和丰富输入数据的部分方法仍有较大差距，当场景中添加新对象时，泛化性能会随着添加数量的增大而下降。

许多抓取检测方法对孤立或杂乱场景下的新对象抓取成功率较低，且所评估的轻度杂乱场景往往不能反映真实世界的抓取现实。TEN PAS提出了一种不需要精确分割对象的抓取假设生成算法。该研究通过结合关于对象类别的先验知识来提高抓取分类的准确性，同时由于该算法不分割对象，故能够将多个对象视为整体进行抓取检测[60]。考虑基于自然语言命令查询的抓取任务，在给定查询对象的定位上需要单独的抓取检测模块。两个深度学习管道的级联应用在重叠的多目标情况下会由于单个输出的模糊性而产生错误。CHEN提出了一种命令抓取网络CGNet(Command Grasping Network)，直接从RGB图像和文本命令输入中输出满足抓取的命令[75]。为了更好地实现机器人与人之间的交互，ZHANG开发了一种能够与人类进行互动的机器人系统INVIGORATE(Interactive visual grourding and grasping)，这是一种通过理解语言命令并结合部分可观察的马尔可夫决策过程(POMDP)的方法，它集成学习到的神经网络模块，通过近似的POMDP规划，来帮助机器人实现自主目标物体的识别和抓取[79]。上述方法中，TEN PAS方法和CHEN方法的抓取成功率基本相等，但TEN PAS方法可以在任何可见表面上生成抓取假设，适应能力更强；ZHANG开发的系统结合了基于模型的POMDP规划和数据驱动的深度学习二者的优势，为机器人抓取检测贡献了一种新的思路。

对于空间杂乱场景中的单个或多个物体，抓取孤立对象的主要挑战来自对象外部的场景信息；抓取堆叠对象的困难主要是由对象外部的场景信息和对象之间的位置关系(堆叠和阻挡)造成的[85]。面向场景的方法引入环境因素，能够更好地完成对象分割、抓取估计，提高抓取方法的各项性能并避免碰撞。以上研究工作针对面向场景的端对端机器人抓取方法存在的各种问题，提出了自己的解决办法，这为今后的研究提供了有效参考，表6对空间级抓取方法进行了对比。

### 2.3 数据集和评估指标

本文将抓取数据集从收集方式上分为两类，一是使用传感器收集的真实世界数据集，其获取输入信息的传感器与平面抓取相同，但在结构上增加获取深度信息的双目相机(两个2维摄像头)，二是利用模拟引擎生成的标注数据集，常见的公开空间抓取数据集见表7。Cornell抓取数据集作为通用的真

表 6　空间级抓取方法对比分析

Fig.6　Comparative analysis of spatial-level grasping methods

| 数据集 | 评价方法 | 标签 | 抓取 | 抓取量(Cat.) | 抓取量(物体/场景) | 场景 |
|---|---|---|---|---|---|---|
| GraspNet-1billion[63] | 模拟 + 真实 | 模型分析 | $1.1\times10^7$ | 88 | $1.25\times10^8$ | 多个 |
| Dex-Net[66] | 模拟 | 模型分析 | $6.7\times10^7$ | 1 500(50) | 100 | – |
| Eppnerl[81] | – | 物理模拟 | $1.1\times10^7$ | 88 | $1.25\times10^8$ | 多个 |
| Sim-MEES[84] | 模拟 | 模型分析 + 物理模拟 | $1.1\times10^8$ | $1.5\times10^3$ | 1 – 20 | 多个 |
| Kappler[86] | 模拟 | 手动 + 真实 | | | | |
| 世界实验标注 | $3\times10^5$ | 700(80) | ≈430 | 单一 | | |
| Veres[87] | 模拟 | 物理模拟 | $5\times10^4$ | – | – | 单一 |
| EGAD[89] | 模拟 | 模型分析 | $2.33\times10^5$ | 2 331 | 100 | 单一 |
| 6-DOF GraspNet[99] | 模拟 | 物理模拟 | $7.07\times10^7$ | 206(6) | $3.4\times10^4$ | 单一 |
| ACRONYM[90] | 模拟 | 物理模拟 | $1.77\times10^8$ | 8 872(262) | 2 000 | 多个 |

注：– 无法获取。

表 7　公开可用的空间级抓取数据集

Tab.7　Public available spatial-level grasping datasets

| 分类 | 方法 | 优势 | 成功率 /% | 局限性 | 参考文献 |
|---|---|---|---|---|---|
| 面向对象 | 基于局部点云的方法 | 适用于未知对象 | 82.95 ~ 96.00 | 点云配准时间久，空间复杂度较大，收敛缓慢，对应点匹配易错 | 文[46 – 48, 63] |
| | 基于完整形状的方法 | 适用于已知对象 | 71.4 ~ 98.38 | 当现有的 3 维模型与目标模型不同时，抓取姿态会有很大的偏差，将导致抓取失败 | 文[36 – 37, 41 – 42, 45, 49 – 54] |
| | 基于优化的方法 | 更新和计算影响模型训练和模型输出的网络参数，使其逼近或达到最优值，从而最小化(或最大化)损失函数，使用优化器来提高模型性能，其适用范围广 | 72.00 ~ 75.76 | 很难选择出合适的学习率；相同的学习率并不适用于所有的参数更新 | 文[43, 45] |
| 面向场景 | 基于监督学习的方法 | 架构简单，易于实施，易于训练，适合大多数任务 | 81.67 ~ 98.54 | 其数据集通常每个样本只有一个标签，这给模型增加了很多限制 | 文[80, 94] |
| | 基于强化学习的方法 | 学习自我探索能力，使机器人具有更高的灵活性 | 81.20 ~ 93.20 | 算法的探索空间会变得非常大，或者需要反复尝试的抓取姿势数不胜数，使学习效率变得非常慢，并且由于探索空间太大，正负样本将极不平衡 | 文[71, 95, 97 – 98] |
| | 其他方法 | – | 60.00 ~ 93.00 | – | 文[35, 60, 75 – 76] |

实世界数据集在很多抓取方法研究中应用广泛[86]，但该数据集只有 885 幅图像，对于机器人抓取的研究现状来说，数据量过小，一些研究团队在其数据集上进行扩展，如 ImageNet 数据集(拥有 14 197 122 幅图像)是将相同协议构建的数据集扩展到多个对象场景[99]。早在 2016 年，PINTO 和 LEVINE 就尝试开发专用设备来大规模收集真实世界数据，但庞大数据的标记工作需要大量的人工成本和强大

的硬件支持[91-95]。为了解决这种问题，许多学者研究使用仿真环境来模拟抓取，如 MURALI[72]、ALLIEGRO[78]、ZHANG[79]等。这种方法能够低成本生成大规模的标注数据集，但其与真实场景的差距始终是一个障碍。对于机器人抓取算法性能提升来说，建立一个规模更大、对象类别更丰富的数据集是有效且必要的。

针对空间内抓取，2 面抓取的评价指标不适用。目前使用的指标主要有：

1）有效抓取率(VGR)，成功完成抓取对象的概率；

2）模型点的平均距离(Average Distance of model Points，ADP)，抓取姿态主要包括旋转量 $R$ 和位移量 $T$。给定 3 维模型 $M$，设真实姿态为 $R$ 和 $T$，预测姿态为 $R'$和 $T'$，则：

$$e_{\mathrm{ADP}} = \mathrm{avg}(Rx + T) - (R'x + T'),\ x \in M$$

其中，$e_{\mathrm{ADP}}$是模型点的平均距离的数值化，$x$ 是模型 $M$ 上的任意点。

3)倒角距离(chamfer distance，CD)，计算生成点云与真实点云之间的平均最短点距离。

通常用来评估这些指标的是 YCB Video 抓取数据集[97]。

## 3 结论

本文从平面内抓取和空间内抓取两个方面对基于视觉的机器人端到端策略抓取估计进行综述，介绍其细分方法、相关数据集、评价指标和对应方法的优势和缺陷。目前，学者们虽然提出了许多基于视觉的机器人端到端策略抓取算法、框架和多功能方法来辅助机器人抓取任务，但在实际任务场景的应用中仍然存在挑战。

### 3.1 研究挑战

1）算法的通用性不足。

数据问题：现有的大多数方法都局限于特定对象或特定场景范围，大多方法只在单个场景中模拟开发和验证，其泛化性能不理想。究其原因，根本上是数据和算法本身的问题。对于数据，例如采集数据的功能性不足、对象数据类型的完整性不够、场景数据的多样性有限以及训练数据量不理想等。

算法本身问题：端到端策略在学习过程中必然会忽略问题分解中包含的有价值信息，这并不适用于所有抓取任务。随着网络复杂性或任务难度的增加，端到端策略可能会变得低效或失败。

2）模型的轻量化有限。一些研究通过构建大型复杂的网络架构、多模态特征聚合等实现更高性能的机器人抓取，但相应代价是高计算成本和大量参数，不适用于机器人实时操作。

3）任务评价机制缺乏多元化。现有的机器人抓取任务的评价机制主要依托于精度。然而，现有精度的评价机制主要是“抓住”而不是“抓好”。在实际的机器人操作场景，只是抓起目标物，是难以实现替代人类工作的角色。

### 3.2 未来展望

1）数据和算法方面。

可采用多源数据融合，如多视图数据(更广泛的透视数据)、多传感器数据(如深度数据、触觉信息)及利用模拟环境生成虚拟数据建立大规模数据集或半监督方法直接生成标注数据等策略来提高抓取任务性能。算法开发需考虑结构适用性问题以提高算法在真实场景的表现性能，未来端到端学习的发展需要新的结构化学习范式。

2）模型轻量化方面。

通常的方法是进行模型压缩，使网络携带更少的网络参数，解决内存和速度问题；算法开发要综合考虑模型的实时性、轻量化与高性能融合问题，这才能让卷积神经网络走出实验室，走入应用更广泛的移动端。

3）模型评估方面。

建立统一的机器人抓取模型评价指标体系和评估方法，实现机器人抓取相关任务以及评价机制的多元化，推动机器人拟人化高质量的完成任务会更有研究价值。

## 参考文献

[1] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61: 85-117.

[2] GLASMACHERS T. Limits of end-to-end learning[C]//Asian Conference on Machine Learning. New York, USA: PMLR, 2017, 77: 17-32.

[3] 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, USA: IEEE, 2011: 3304-3311.

[4] DUAN H, WANG P, HUANG Y, et al. Robotics dexterous grasping: The methods based on point cloud and deep learning[J].

Frontiers in Neurorobotics, 2021, 15: 65 – 82.

[5] 曹家乐，李亚利，孙汉卿，等. 基于深度学习的视觉目标检测技术综述[J]. 中国图象图形学报，2022，27(6)：1697 – 1722.
CAO J L, LI Y L, SUN H Q, et al. A review of visual target detection techniques based on deep learning[J]. Chinese Journal of Image Graphics, 2022, 27(6): 1697 – 1722.

[6] 刘亚欣，王斯瑶，姚玉峰，等. 机器人抓取检测技术的研究现状[J]. 控制与决策，2020，35(12)：2817 – 2828.
LIU Y X, WANG S Y, YAO Y F, et al. Research status of robot grasping detection technology[J]. Control and Decision Making, 2020, 35(12): 2817 – 2828.

[7] YIN Z, LI Y. Overview of robotic grasp detection from 2D to 3D[J]. Cognitive Robotics, 2022, 2: 73 – 82.

[8] 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(3): 1677 – 1734.

[9] XIE Z, LIANG X, ROBERTO C. Learning-based robotic grasping: A review[J]. Frontiers in Robotics and AI, 2023, 10: 10 – 38.

[10] PLATT R. Grasp learning: Models, methods, and performance[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2023, 6: 363 – 389.

[11] YAO K, BILLARD A. Exploiting kinematic redundancy for robotic grasping of multiple objects[J]. IEEE Transactions on Robotics, 2023, 39(3): 1982 – 2002.

[12] WU Y, FU Y, WANG S. Information-theoretic exploration for adaptive robotic grasping in clutter based on real-time pixel-level grasp detection[J]. IEEE Transactions on Industrial Electronics, 2023, 71(3): 2683 – 2693.

[13] 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.

[14] ZENG A, SONG S, YU K T, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching[J]. The International Journal of Robotics Research, 2022, 41(7): 690 – 705.

[15] CAI J, CHENG H, ZHANG Z, et al. Metagrasp: Data efficient grasping by affordance interpreter network[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 4960 – 4966.

[16] DO T T, NGUYEN A, REID I. Affordancenet: An end-to-end deep learning approach for object affordance detection[C]// IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 5882 – 5889.

[17] PARK D, CHUN S Y. Classification based grasp detection using spatial transformer network [EB/OL]. (2016 – 03 – 04) [2024 – 10 – 17]. https://arxiv.org/abs/1803.01356.

[18] ZHOU X, LAN X, ZHANG H, et al. Fully convolutional grasp detection network with oriented anchor box[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 7223 – 7230.

[19] CHU F J, XU R, VELA P A. Real-world multiobject, multigrasp detection[J]. IEEE Robot Autom Letters, 3(4): 3355 – 3362.

[20] WANG S, 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. Piscataway, USA: IEEE, 2019: 474 – 480.

[21] ARDÓN P, PAIRET È, PETRICK R P, et al. Learning grasp affordance reasoning through semantic relations[J]. IEEE Robot Automation Letters , 2019, 4(4): 4571 – 4578.

[22] ZHANG H, LAN X, BAI S, et al. Roi-based robotic grasp detection for object overlapping scenes[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscatway, USA: IEEE, 2019: 4768 – 4775.

[23] VOHRA M, Prakash R, Behera L. Real-time grasp pose estimation for novel objects in densely cluttered environment[C]// IEEE International Conference on Robot and Human Interactive Communication. Piscataway, USA: IEEE, 2019: 1 – 6.

[24] ZHANG H, LAN X, BAI S, et al. A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2019: 6435 – 6442.

[25] PARK D, SEO Y, SHIN D, et al. A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 7300 – 7306.

[26] DEPIERRE A, DELLANDRéA E, CHEN L. Optimizing correlated grasp ability score and grasp regression for better grasp prediction [EB/OL]. (2020 – 02 – 03) [2024 – 10 – 17]. https://arxiv.org/abs/2002.00872.

[27] KUMRA S, JOSHI S, SAHIN F. Antipodal robotic grasping using generative residual convolutional neural network[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2020: 9626 – 9633.

[28] DUAN S, TIAN G, WANG Z, et al. A semantic robotic grasping framework based on multi-task learning in stacking scenes[J]. Engineering Applications of Artificial Intelligence, 2023, 121: 10 – 60.

[29] CHENG H, WANG Y, MENG M Q H. A robot grasping system with single-stage anchor-free deep grasp detector[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1 – 12.

[30] WEI B, YE X, LONG C, et al. Discriminative active learning for robotic grasping in cluttered scene[J]. IEEE Robotics and Automation Letters, 2023, 8(3): 1858 – 1865.

[31] CALLI B, SINGH A, BRUCE J, et al. Yale-CMU-Berkeley dataset for robotic manipulation research[J]. The International Journal of Robotics Research, 2017, 36(3): 261 – 268.

[32] DEPIERRE A, DELLANDRéA E, CHEN L. Jacquard: A large scale dataset for robotic grasp detection[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 3511 – 3516.

[33] 葛俊彦, 史金龙, 周志强, 等. 基于三维检测网络的机器人抓取方法[J]. 仪器仪表学报, 2023(8): 146 – 153.
GE J Y, SHI J L, ZHOU Z Q, et al. Robot grasping method based on three-dimensional detection network[J]. Journal of Instrumentation, 2023(8): 146 – 153.

[34] ZHANG C, LIN C, LENG Y, et al. An effective head-based HRI for 6D robotic grasping using mixed reality[J]. IEEE Robotics and Automation Letters, 2023, 8(5): 2796 – 2803.

[35] DONG Z, TIAN H, BAO X, et al. GraspVDN: Scene-oriented grasp estimation by learning vector representations of grasps[J]. Complex & Intelligent Systems, 2022: 1 – 12.

[36] VARLEY J, DECHANT C, RICHARDSON A, et al. Shape completion enabled robotic grasping[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 2442 – 2447.

[37] YANG C, ZENG C, FANG C, et al. A DMPs-based framework for robot learning and generalization of humanlike variable impedance skills[J]. IEEE/ASME Transactions on Mechatronics, 2018, 23(3): 1193 – 1203.

[38] BERSCHEID L, MEIβNER P, KRÖGER T. Self-supervised learning for precise pick-and-place without object model[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4828 – 4835.

[39] GUALTIERI M, TEN PAS A, PLATT R. Pick and place without geometric object models[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 7433 – 7440.

[40] YAN X, HSU J, KHANSARI M, et al. Learning 6-DoF grasping interaction via deep geometry-aware 3d representations[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 3766 – 3773.

[41] YUAN W, HANG K, KRAGIC D, et al. End-to-end nonprehensile rearrangement with deep reinforcement learning and simulation-to-reality transfer[J]. Robotics and Autonomous Systems, 2019, 119: 119 – 134.

[42] VAN DER MERWE M, LU Q, SUNDARALINGAM B, et al. Learning continuous 3D reconstructions for geometrically aware grasping[EB/OL]// (2019 – 10 – 02) [2024 – 10 – 17]. https://arxiv.org/abs/1910.00983.

[43] SAJJAN S, MOORE M, PAN M, et al. Clear grasp: 3D shape estimation of transparent objects for manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 3634 – 3642.

[44] YAN X, HSU J, KHANSARI M, et al. Learning 6-DoF grasping interaction via deep geometry-aware 3d representations[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 3766 – 3773.

[45] GAO W, TEDRAKE R. KPAM-SC: Generalizable manipulation planning using keypoint affordance and shape completion[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2021: 6527 – 6533.

[46] LIANG H, MA X, LI S, et al. Pointnetgpd: Detecting grasp configurations from point sets[C]//International Conference on

Robotics and Automation. Piscataway, USA: IEEE, 2019: 3629 – 3635.

[47] NI P, ZHANG W, ZHU X, et al. Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 3619 – 3625.

[48] ZHAO B, ZHANG H, LAN X, et al. Regnet: Region-based grasp network for end-to-end grasp detection in point clouds[C]// IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2021: 13474 – 13480.

[49] SUNDERMEYER M, MARTON Z C, DURNER M, et al. Augmented autoencoders: Implicit 3D orientation learning for 6D object detection[J]. International Journal of Computer Vision, 2020, 128(3): 714 – 729.

[50] BERSCHEID L, MEIβNER P, KRÖGER T. Self-supervised learning for precise pick-and-place without object model[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4828 – 4835.

[51] TZENG E, DEVIN C, HOFFMAN J, et al. Adapting deep visuomotor representations with weak pairwise constraints[M]//Algorithmic Foundations of Robotics XII. Berlin, Germany: Springer, 2020: 688 – 703.

[52] DONG M, WEI S, YIN J, et al. Real-world semantic grasp detection based on attention mechanism [EB/OL]. (2021 – 11 – 20) [2024 – 10 – 17]. https://arxiv.org/abs/2111.10522.

[53] ZHANG L, WU D. A single target grasp detection network based on convolutional neural network[J]. Computational Intelligence and Neuroscience, 2021, 2021(1): 11 – 55.

[54] SUN H, CUI X, SONG Z, et al. Precise grabbing of overlapping objects system based on end-to-end deep neural network[J]. Computer Communications, 2021, 176: 138 – 145.

[55] TEKDEN A, DEISENROTH M P, BEKIROGLU Y. Grasp transfer based on self-aligning implicit representations of local surfaces[J]. IEEE Robotics and Automation Letters, 2023, 8(10): 6315 – 6322.

[56] CHEN S, TANG W, XIE P, et al. Efficient heatmap-guided 6-DoF grasp detection in cluttered scenes[J]. IEEE Robotics and Automation Letters, 2023, 8(8): 4895 – 4902.

[57] LIU J, SUN W, LIU C, et al. Robotic continuous grasping system by shape transformer-guided multi-object category-level 6D pose estimation[J]. IEEE Transactions on Industrial Informatics, 2023, 19(11): 11171 – 11181.

[58] NIKANDROVA E, KYRKI V. Category-based task specific grasping[J]. Robotics and Autonomous Systems, 2015, 70: 25 – 35.

[59] NOORI N S, WANG Y, COMES T, et al. Behind the scenes of scenario-based training: Understanding scenario design and requirements in high-risk and uncertain environments[C]//ISCRAM. Piscataway, USA: 2017: 948 – 959.

[60] TENPAS A, GUALTIERI M, SAENKO K, et al. Grasp pose detection in point clouds[J]. The International Journal of Robotics Research, 2017, 36(13/14): 1455 – 1473.

[61] HU F J, XU R N, PATRICIO A V, et al. Real-world multi object, multigrasp detection[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3355 – 3362.

[62] WANG C, XU D, ZHU Y, et al. Densefusion: 6D object pose estimation by iterative dense fusion[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscatway, USA: IEEE, 2019: 3343 – 3352.

[63] FANG H S, WANG C, GOU M, et al. Graspnet-1billion: A large-scale benchmark for general object grasping[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 11444 – 11453.

[64] PARK K, PATTEN T, PRANKL J, et al. Multi-task template matching for object detection, segmentation and pose estimation using depth images[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 7207 – 7213.

[65] TIAN H, WANG C, MANOCHA D, et al. Transferring grasp configurations using active learning and local replanning[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 1622 – 1628.

[66] 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[EB/OL]. (2019 – 03 – 27) [2024 – 10 – 17]. https://arxiv.org/abs/1703.09312.

[67] ASIF U, TANG J, HARRER S. Densely supervised grasp detector (DSGD)[C]//The AAAI Conference on Artificial Intelligence. Keystone, USA: AIAA, 2019: 8085 – 8093.

[68] SAJJAN S, MOORE M, PAN M, et al. Clear grasp: 3D shape estimation of transparent objects for manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 3634 – 3642.

[69] SONG S, ZENG A, LEE J, et al. Grasping in the wild: Learning 6-DOF closed-loop grasping from low-cost demonstrations[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4978 – 4985.

[70] CHEN Y, KEE H, LEE K, et al. Hierarchical 6-DoF grasping with approaching direction selection[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 1553 – 1559.

[71] WU C, CHEN J, CAO Q, et al. Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps[J]. Advances in Neural Information Processing Systems, 2020, 33: 13174 – 13184.

[72] MURALI A, MOUSAVIAN A, EPPNER C, et al. 6-DoF grasping for target-driven object manipulation in clutter[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 6232 – 6238.

[73] ZHANG L, WU D. A single target grasp detection network based on convolutional neural network[J]. Computational Intelligence and Neuroscience, 2021, 2021(1): 11 – 55.

[74] YU H, LAI Q, LIANG Y, et al. A cascaded deep learning framework for real-time and robust grasp planning[C]//IEEE International Conference on Robotics and Biomimetics. Piscataway, USA: IEEE, 2021: 1380 – 1386.

[75] CHEN Y, XU R, LIN Y, et al. A Joint network for grasp detection conditioned on natural language commands [EB/OL]. (2021 – 04 – 01) [2024 – 10 – 17]. https://arxiv.org/abs/2104.00492.

[76] ZHANG H, LU Y, YU C, et al. Invigorate: Interactive visual grounding and grasping in clutter [EB/OL]. (2021 – 08 – 25) [2024 – 10 – 17]. https://arxiv.org/abs/2108.11092.

[77] WENG Y, SUN Y, JIANG D, et al. Enhancement of real-time grasp detection by cascaded deep convolutional neural networks [J]. Concurrency and Computation: Practice and Experience, 2021, 33(5): 59 – 76.

[78] ALLIEGRO A, VALSESIA D, FRACASTORO G, et al. Denoise and contrast for category agnostic shape completion[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2021: 4629 – 4638.

[79] ZHANG H, YANG D, WANG H, et al. Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2929 – 2936.

[80] LIN H, CHEANG C, FU Y, et al. I Know what you draw: Learning grasp detection conditioned on a few freehand sketches [C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2022: 8417 – 8423.

[81] EPPNER C, MOUSAVIAN A, FOX D. A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set[C]//International Symposium on Robotics Research. Berlin, Germnay: Springer, 2022: 890 – 905.

[82] ZHAO M, ZUO G, YU S, et al. Position-aware pushing and grasping synergy with deep reinforcement learning in clutter[J]. CAAI Transactions on Intelligence Technology, 2024, 9(3): 738 – 755.

[83] ZHANG F, GAO J, SONG C, et al. TPMv2: An end-to-end tomato pose method based on 3D key points detection[J]. Computers and Electronics in Agriculture, 2023, 210: 10 – 78.

[84] LI J, CAPPELLERI D J. Sim-MEES: Modular end-effector system grasping dataset for mobile manipulators in cluttered environments [EB/OL]. (2023 – 05 – 17) [2024 – 10 – 17]. https://arxiv.org/abs/2305.10580.

[85] DUAN S, TIAN G, WANG Z, et al. A semantic robotic grasping framework based on multi-task learning in stacking scenes[J]. Engineering Applications of Artificial Intelligence, 2023, 121: 10 – 60.

[86] KAPPLER D, BOHG J, SCHAAL S. Leveraging big data for grasp planning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2015: 4304 – 4311.

[87] VERES M, MOUSSA M, TAYLOR G W. An integrated simulator and dataset that combines grasping and vision for deep learning [EB/OL]. (2017 – 02 – 07) [2024 – 10 – 17]. https://arxiv.org/abs/1702.02103.

[88] LEVINE S, PASTOR P, KRIZHEVSKY A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection[J]. The International Journal of Robotics Research, 2018, 37(4/5): 421 – 436.

[89] MORRISON D, CORKE P, LEITNER J. Egad: An evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4368 – 4375.

[90] EPPNER C, MOUSAVIAN A, FOX D. Acronym: A large-scale grasp dataset based on simulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2021: 6222 – 6227.

[91] MILLER A T, ALLEN P K. Graspit! a versatile simulator for robotic grasping[J]. IEEE Robotics & Automation Magazine,

2004，11(4)：110 – 122.

[92] ZHAI D H，YU S，XIA Y. FANet：Fast and accurate robotic grasp detection based on keypoints[J]. IEEE Transactions on Automation Science and Engineering，2023：2974 – 2986.

[93] SAXENA A，DRIEMEYER J，NG A Y. Robotic grasping of novel objects using vision[J]. The International Journal of Robotics Research，2008，27(2)：157 – 173.

[94] ZHANG T，ZHANG C，HU T. A robotic grasp detection method based on auto-annotated dataset in disordered manufacturing scenarios[J]. Robotics and Computer-Integrated Manufacturing，2022，76：10 – 23.

[95] WANG L，MENG X，XIANG Y，et al. Hierarchical policies for cluttered-scene grasping with latent plans[J]. IEEE Robotics and Automation Letters，2022，7(2)：2883 – 2890.

[96] WANG L，XIANG Y，YANG W，et al. Goal-auxiliary actor-critic for 6d robotic grasping with point clouds[C]//Conference on Robot Learning. New York，USA：PMLR，2022.

[97] LI Z，XU B，WU D，et al. A YOLO-GGCNN based grasping framework for mobile robots in unknown environments[J]. Expert Systems with Applications，2023，225：11 – 99.

[98] ZHAO M，ZUO G，YU S，et al. Position-aware pushing and grasping synergy with deep reinforcement learning in clutter[J]. CAAI Transactions on Intelligence Technology，2024，9(3)：738 – 55.

[99] MOUSAVIAN A，EPPNER C. Fox D. 6-DoF graspnet：Variational grasp generation for object manipulation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway，USA：IEEE，2019：2901 – 2910.

[100] ZHANG H，LAN X，BAI S，et al. Roi-based robotic grasp detection for object overlapping scenes[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway，USA：IEEE，2019：4768 – 4775.

[101] LERREL P，ABHINAV G. Supersizing self-supervision：Learning to grasp from 50k tries and 700 robot hours[C]//IEEE International Conference on Robotics and Automation. Piscataway，USA：IEEE，2016：3406 – 3413.

[102] LI X，ZHANG X，ZHOU X，et al. UPG：3D vision-based prediction framework for robotic grasping in multi-object scenes[J]. Knowledge-Based Systems，2023，270：11 – 104.

## 作者简介

苏　康(1984)，男，博士，讲师，硕士生导师。研究领域为摩擦学，机械工程，人工智能，3D 打印。

李嘉良(1998)，男，硕士生。研究领域为视觉检测。