doi:10.11887/j.issn.1001-2486.26020003 http://journal.nudt.edu.cn

从几何解析到语义推理:机器人抓取感知范式的演进

邹世龙,黄雨行,易任娇,朱晨阳*,徐 凯
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:机器人抓取感知是实现机器人自主操作与具身智能的重要基础,其技术范式正经历从依赖显式几何建模的解析法,向以数据驱动学习与语义推理增强为核心的智能感知体系的深刻变革。围绕机器人抓取感知范式的演进脉络,对相关研究进行了系统综述,阐述了解析几何模型驱动、视觉数据驱动以及语义理解与推理增强三个递进阶段的演变过程,并剖析了各阶段的代表性算法与关键技术路线。通过对不同范式在输入模态、数据需求、泛化能力与任务适应性等方面的对比分析,总结了各类方法在非结构化环境下的优势与局限。此外,系统梳理了抓取数据集从平面基准到大规模综合数据的演进历程,并剖析了由任务可靠性与提议准确度构成的量化评价体系。进一步总结了当前机器人抓取感知在仿真到现实迁移、推理效率、跨模态信息融合以及复杂任务扩展等方面面临的共性挑战,并展望了结合具身基础模型与灵巧操作的发展趋势,旨在为构建高泛化、强理解能力的通用机器人抓取系统提供参考借鉴。

关键词:机器人抓取感知;几何建模;数据驱动学习;语义理解与推理增强
中图分类号:TP242.6 **文献标志码**:A **文章编号**:1001-2486(2026)03-339-18

From geometric analysis to semantic reasoning: the evolution of robotic grasping perception paradigms

ZOU Shilong, *HUANG Yuhang*, *YI Renjiao*, *ZHU Chenyang**, *XU Kai*
(College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Robotic grasping perception is a fundamental prerequisite for autonomous manipulation and embodied intelligence. The technical paradigm is undergoing a profound shift from analytical methods based on explicit geometric modeling to intelligent perception frameworks driven by data-driven learning and enhanced semantic reasoning. Research on robotic grasping perception was systematically reviewed along the lines of paradigm evolution. The evolutionary process was described through three progressive stages: analytical geometry-driven methods, visual data-driven methods, and semantic understanding and reasoning enhancement. Representative algorithms and key technical pathways for each stage were examined and analyzed. Through a comparative analysis of input modalities, data requirements, generalization ability, and task adaptability across different paradigms, the advantages and limitations of various methods in unstructured environments were summarized. Furthermore, the evolution of grasping datasets from planar benchmarks to large-scale comprehensive data was systematically traced, and the quantitative evaluation system composed of task reliability and proposal accuracy was analyzed. Prevailing challenges, including sim-to-real transfer, inference efficiency, cross-modal information fusion, and the extension to complex tasks, were identified. Future development trends that integrate embodied foundation models with dexterous manipulation were discussed to provide references for building general-purpose robotic grasping systems with high generalization performance and robust task comprehension.

Keywords: robotic grasping perception; geometric modeling; data-driven learning; semantic understanding and reasoning enhancement

机器人抓取(robotic grasping)作为机器人与物理世界交互的最基本能力,是实现复杂操作任务(如工业装配[1-4])的核心技术之一。随着机器人应用场景从结构化的工业流水线向非结构化的家庭、物流及开放环境拓展,抓取任务面临的挑战也日益复杂。机器人不仅需要在杂乱、遮挡和

收稿日期:2026-02-02
基金项目:国家自然科学基金资助项目(62522219,62372457,62132021,62572477)
第一作者:邹世龙(2000—),男,山东菏泽人,硕士研究生,E-mail:zoushilong@nudt.edu.cn
***通信作者**:朱晨阳(1991—),男,湖南长沙人,教授,博士,博士生导师,E-mail:zhuchenyang07@nudt.edu.cn
引用格式:邹世龙,黄雨行,易任娇,等. 从几何解析到语义推理:机器人抓取感知范式的演进[J]. 国防科技大学学报, 2026, 48(3): 339-356.
Citation: ZOU S L, HUANG Y H, YI R J, et al. From geometric analysis to semantic reasoning: the evolution of robotic grasping perception paradigms[J]. Journal of National University of Defense Technology, 2026, 48(3): 339-356.

光照变化的场景中准确感知物体的几何属性，还需要理解物体的语义信息以及任务的上下文约束。因此，如何构建稳定、泛化性强且具备鲁棒性的抓取感知模型，是机器人研究领域长期面临的挑战。

尽管近年来关于机器人抓取的综述文献层出不穷，但现有综述在系统性与前瞻性方面存在以下显著局限：①侧重单一视角，缺乏演进脉络的深度串联。现有工作大多局限于对特定技术分支（如解析方法[5-6]或基于深度学习的检测方法[7-11]）进行横向罗列，缺乏从宏观视角梳理不同阶段技术演化的内在逻辑，难以呈现感知范式更迭的必然趋势。②对语义理解与复杂推理能力的关注不足。多数文献仍聚焦于单纯的物理稳定性评估（stable grasping）[11-14]，对近年来兴起的多模态融合、任务驱动操作、语言引导感知以及具身基础模型在抓取中的应用缺乏系统化梳理，难以反映领域最新的认知化趋势。③缺乏跨范式的多维度定量和定性对比。现有综述[15-17]往往忽略了不同范式在数据依赖、适用场景、泛化能力和系统复杂度等维度的本质差异，较少从共性挑战的角度进行归纳总结。

为了填补上述空白，本文从感知范式演进的全新视角切入，构建了全景式的演进分类体系，将抓取技术的发展划分为三个逻辑递进的阶段，对机器人抓取领域的研究成果进行了系统性的梳理与展望。图 1 展示了机器人抓取范式和关键学习算法的演变过程，同时展示了基于数据驱动的感知方法所采用网络架构的演进过程。此外，本文深入调研了机器人抓取领域的公开数据集，梳理了从早期平面抓取基准到大规模 6 自由度（six degrees of freedom，6-DoF）真实与合成数据集的发展脉络。同时，系统剖析了以成功率为核心的任务可靠性指标以及以平均精度为代表的提议准确度评价体系。最后，本文通过多维度的表格对比了三大范式在数据需求、泛化能力及语义理解等方面的优劣，并深入探讨了当前机器人抓取方向面临的共性挑战和未来研究方向。

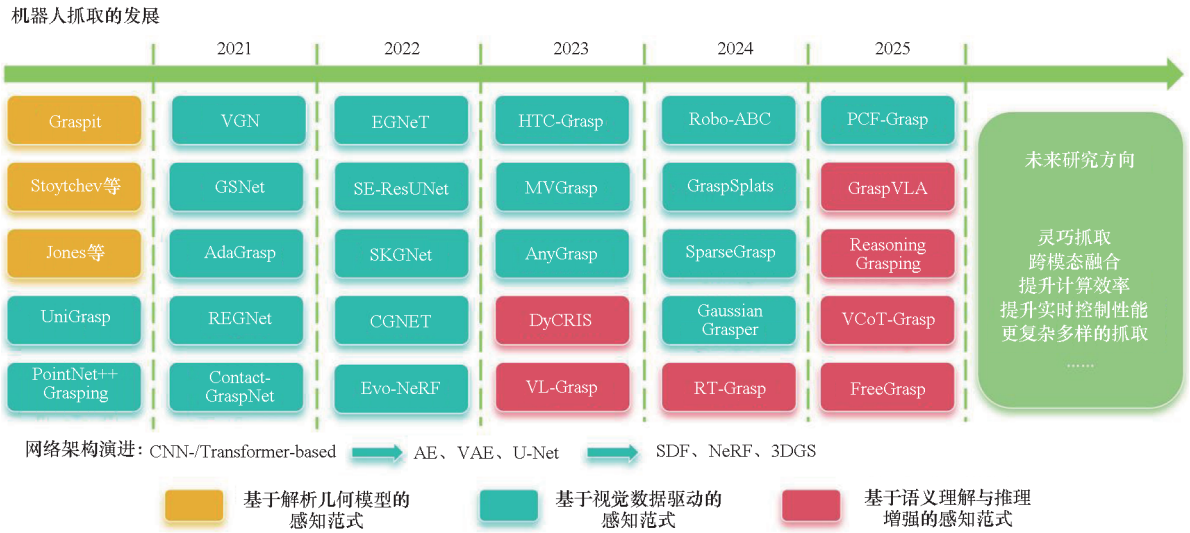


图 1 机器人抓取范式和关键学习算法的演变

Fig. 1 Evolution of robotic grasping paradigms and key learning algorithms

1 问题定义

1.1 总览

机器人抓取的核心本质在于求解机器人末端执行器相对于目标物体的最优配置状态，以确保抓取动作在物理层面的稳定性与任务层面的有效性。从形式化表述来看，抓取问题可定义为一个在多维连续空间中的搜索与优化问题，其目标是根据传感器感知的环境信息（如 RGB-D 图像或三维点云），计算出满足力学平衡、几何约束及避障要求的位姿参数。随着研究的演进，问题定义正从传统的单一力学稳定性评估，向包含任务语义理解、多模态指令对齐及复杂认知推理的智能化感知范式转变。

1.2 抓取表示

抓取表示（grasp representation）是连接视觉感知算法与底层运动控制的逻辑桥梁，其数学描述方式直接决定了算法的计算复杂度和场景适应性。目前主流的抓取表示方法包括以下三种。

1.2.1　基于接触的抓取表示方式

作为解析法的核心理论基础,该方法将抓取建模为作用于物体表面的 N 个接触点的力螺旋集合 $g_j=(w_1,w_2,\cdots,w_N)$,其中,每个力螺旋 $\omega_i=(f_i,\tau_i)$ 包含了作用在接触点 $p=\{p_i\}_{i=1}^{N}$ 上的法向作用力 f_i 和切向力矩 τ_i。其稳定性判据的核心在于:若所有接触点产生的合力螺旋 $w_{g,j}$ 能抵消外部扰动权重 w_e,则物体保持平衡状态。具体数学表示如下:

$$w_{g,j}=\sum_{i=1}^{N}w_i=-w_e \tag{1}$$

依此来判断系统是否达到平衡状态或实现力封闭。该表示方式不依赖特定的夹爪几何形状,具有较高的通用性,因而在多指灵巧手的精细化协同控制与复杂动力学分析中持续发挥作用。

1.2.2　基于有向矩形的抓取表示方式

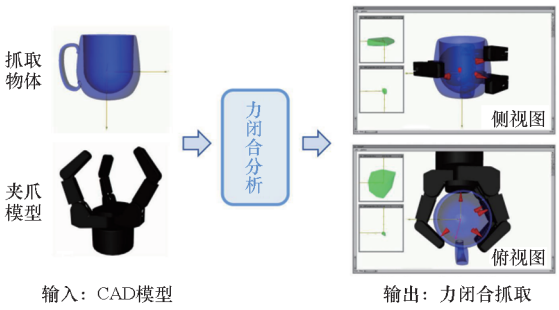
针对工业场景中广泛使用的平行手指抓取器,Jiang 等[18]首次提出了五维有向矩形的抓取表示方法,该表示具体可参数化为 $g=(x,y,w,h,\theta)$,其中 (x,y) 表示图像坐标系下的中心点位置,w 和 h 表示抓取器的开合宽度和夹持高度,θ 表示抓取器相当于水平轴的方向角。这种表示方式将复杂的三维位姿估计问题简化为二维图像层面的检测问题,极大地降低了深度神经网络的学习难度和计算开销,是目前实时视觉抓取检测领域最为广泛采用的范式之一。

1.2.3　基于 6-DoF 的抓取表示方式

为了在更复杂的非结构化环境(如物体堆叠或存在避障需求)中实现灵活抓取,6-DoF 位姿表示法[19-22]成为近年来的研究重心。该方法将抓取描述为三维空间变换矩阵 $\boldsymbol{SE}(3)$,六维姿态可以形式化为 $g=(x,y,z,rx,ry,rz)$,完整定义了抓取器在三维坐标系中的平移 (x,y,z) 与旋转 (rx,ry,rz)。相较于有向矩形表示法,6-DoF 表示法能够与三维点云数据深度耦合,有效捕捉物体的空间拓扑特征,该表征不仅是处理复杂物理交互任务的基础,更是实现通用化具身智能抓取的关键技术桥梁。

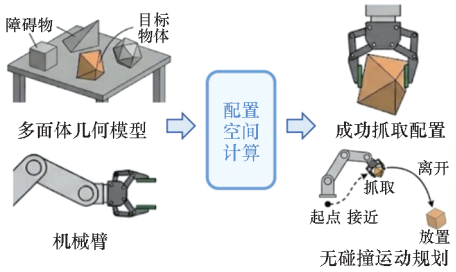
2　基于解析几何模型的感知范式

基于解析几何模型的感知方法是机器人抓取领域的经典研究范式,该方法侧重于通过明确的数学定义、几何特征或人工启发式规则来推导最优抓取构型。早期研究侧重于在结构化环境下,通过显式分析物体的几何拓扑、接触力学约束或特定任务需求来求解闭式解。尽管此类方法在应对非结构化环境与未知物体时存在泛化瓶颈,但其建立的稳定性度量准则与空间表征逻辑不仅构成了多指灵巧操作的理论基石,也为现代数据驱动方法提供了本质的物理约束引导。图 2 展示了基于解析几何模型的感知范式下三种方法的对比示意图。



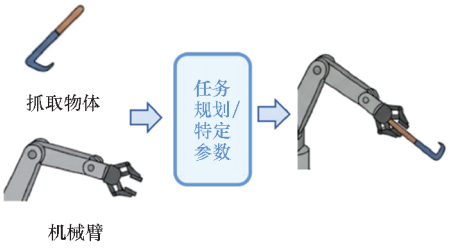
(a) 基于接触力学的稳定性分析

(a) Contact mechanics-based stability analysis



(b) 基于几何特征的显式计算

(b) Geometric feature-based explicit computation



(c) 基于任务规则的抓取

(c) Task-oriented grasping

图 2　解析几何模型的感知范式下三种方法的对比示意图[23-25]

Fig. 2　A comparative diagram of three methods under the analytical geometry perception paradigm[23-25]

2.1　基于接触力学的稳定性分析

力学分析方法主要建立在解析几何模型之上,其核心在于利用物体与手指之间的接触约束来判断抓取在外界扰动下的稳定性。早期研究通常假设物体几何和接触摩擦参数完全已知,以力封闭(force-closure)和形封闭(form-closure)[23]为核心判据,分别对应有摩擦和无摩擦接触模型,强

调抓取是否能够在理论上抵抗任意外部扰动力矩。为了定量描述这种稳定性，Ferrari 等[26]提出了抓取扭矩空间（grasp wrench space，GWS）为代表的连续度量方法，将抓取可施加的所有力/力矩表示为凸空间，并通过最大最小抗扰扭矩[27-30]（largest-minimum resisted wrench，LRW）、GWS 的体积或不同范数下的距离来衡量抓取对外界扰动的鲁棒性。

尽管基于接触力学的方法提供了理论上的最优解，但其严重依赖于精确的物理参数（如摩擦系数、接触位置），难以直接适应非结构化环境，在实际非结构化环境中，微小的感知误差往往会导致分析结果的失效。

2.2 基于几何特征的显式计算

与基于接触力学的分析不同，基于几何特征的方法侧重于从物体的几何属性中直接提取关键抓取特征。这类方法通常不需要完整的物理模型，而是利用局部几何信息来推断抓取构型。

Jones 等[24]提出了一种面向多面体目标的解析式抓取规划框架，其核心思想是将物体、夹爪和障碍物统一建模为多面体，并显式利用物体的面特征来确定抓取平面。具体而言，该方法通过将接近和离开动作限制在抓取平面内，并将复杂的 6-DoF 空间交互约束简化为 3-DoF 构型的空间路径搜索问题，从而使机器人能够在复杂的障碍物环境中快速寻找到几何上可行的抓取方案。

该类方法具有较强的可解释性和较高的几何计算精度，但其局限性同样较为突出：方法高度依赖精确的几何建模与完备的环境先验，在面对高自由度操作、动态变化场景以及形状复杂的非结构化物体时，往往难以兼顾计算效率与泛化能力，实际应用受限。

2.3 基于任务规则的抓取

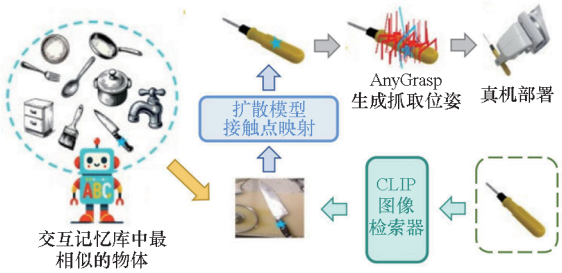
第三类解析方法依赖于预定义的任务规则。这类方法强调将工具功能属性建立在机器人自身的行为与感知能力之上，通常利用预定义的启发式策略或学习机制来应对任务需求。

Stoytchev 等[25]将复杂的抓取过程简化为一种预设位置的夹持动作，即通过程序指定机械臂末端在工具柄部的固定位置执行闭合，从而实现工具与机器人之间的物理耦合。在探索学习阶段，机器人将在选择不同的探索行为时对环境物体的影响记录在功能属性表中。在任务执行阶段，系统根据当前感知状态，利用启发式规则从属性表中动态检索并序列化最优的行为组合，从而驱动机器人完成抓取等操作任务。

这类方法不依赖精确的几何模型或复杂的动力学参数，使其在环境不确定甚至物理模型不准确的情况下，仍能够通过实时经验反馈动态更新功能属性表并完成任务。然而，由于其高度依赖预设的行为库，无法识别超出机器人现有动作能力的潜在功能属性，且随机探索过程效率较低，这也间接推动了后续数据驱动学习方法的发展。

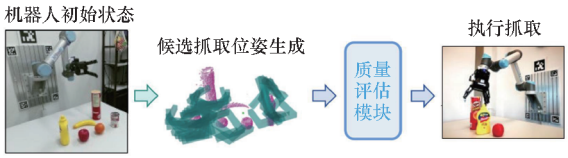
3 基于视觉数据驱动的感知范式

随着机器学习技术的飞速发展，机器人抓取研究正逐渐从依赖精确物理模型的解析范式转向基于数据驱动的学习范式。与解析方法不同，数据驱动范式不依赖于对物体几何、物理参数和预设功能属性表的显式建模，而是受到神经心理学的启发，侧重于从海量“经验”数据中归纳内在规律，从而学习复杂的抓取策略。该范式核心在于利用深度神经网络强大的特征提取能力，直接建立从视觉观测到抓取位姿的映射关系，显著提升了机器人在非结构化环境中面对未知物体的泛化能力。根据学习策略与模型架构的差异，该范式主要分为：模仿学习、生成与评估的两阶段方法、端到端的直接预测方法。为了更直观地理解当前视觉数据驱动抓取技术的方法论差异，图 3 分别展示了三种范式下具有代表性的方法的整体流程。



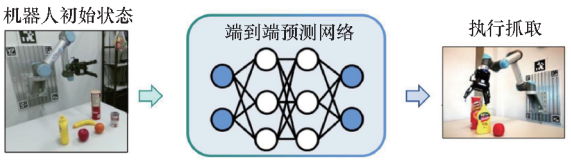
（a）基于模仿学习的范式（模板匹配）

（a）Imitation learning-based paradigm: template matching



（b）基于生成与评估的两阶段范式

（b）Two-stage paradigm based on generation and evaluation



（c）基于端到端的直接预测范式

（c）End-to-end direct prediction paradigm

图 3　视觉数据驱动的感知范式下三种方法的对比示意图[20, 31]

Fig. 3　A comparative diagram of three methods under the visual data-driven perception paradigm[20, 31]

3.1　基于模仿学习的范式

模仿学习的核心思想是通过学习示教数据中的成功行为来习得抓取技能。早期的研究[32－35]主要采用演示编程策略，即利用物理牵引示教或遥操作方式记录专家的抓取轨迹，并在测试时通过轨迹重放或动态调整来执行抓取。然而，单纯的轨迹回放难以应对环境的动态变化。为了提高适应性，文献[36－38]引入了抓取识别模块，将具体抓取姿态映射到预定义的抓取分类体系中。基于识别到的抓取类型，规划器可以通过运动学映射或在受限抓取空间中高效搜索，完成机器人抓取合成。随着深度学习的发展，行为克隆[39]（behavior cloning，BC）和逆强化学习[40－42]（inverse reinforcement learning，IRL）逐渐成为主流。前者将模仿视为监督学习问题，直接建立从输入到动作的映射；后者则试图从示教中推断出隐含的奖励函数，进而通过强化学习优化策略，从而在一定程度上解决了策略的鲁棒性问题。

另一类模仿策略是模板匹配（matching of templates，MoT），其核心假设是相似的物体具有相似的抓取方式。文献[43－44]引入了物体模板匹配（matching of object templates，MOOT），即通过物体识别或位姿估计，在模板库中寻找最相似的对象，并将其预定义的抓取姿态映射到目标物体坐标系中，仅适用于已知物体。为应对几何结构未知的物体，文献[21，45－49]引入了形状模板匹配（matching of shape templates，MOST）。该方法通过将物体分解为基础几何形状基元并预定义抓取，实现了目标物体局部形状与基元的匹配以及抓取策略的有效迁移。这种方法一定程度上扩展了对未知物体的适应能力，但其性能仍然受限于基础几何图元的表达能力，难以处理几何结构极其复杂的物体。与前述关注“形状对齐”或“拓扑匹配”的方法不同，Ju 等[31]提出了一种基于语义对应的可供性泛化框架。该方法不再依赖显式的对象级或形状级模板，而是结合视觉基础模型挖掘语义一致的操作区域，在不同物体之间建立可迁移的操作映射，从而有效实现跨类别的抓取与操作泛化。实验表明，该方法在跨类别抓取任务中表现出极强的鲁棒性。

近年来，为了解决模仿学习对大量示教数据的依赖，元学习（meta-learning）[50]和少样本学习（few-shot learning）[51－53]成为新的研究热点。通过在大规模离线数据上预训练元策略，机器人能够学习到与任务无关的通用特征表示，从而仅凭单次或少量示教快速适应新任务，显著降低了数据采集成本。

3.2　基于生成与评估的两阶段范式

基于生成与评估的两阶段范式将抓取合成解耦为“候选生成”和“质量评估”两个独立的子问题[54]。其核心思想是将复杂的决策过程拆解：首先在状态空间中生成一组潜在的抓取构型，随后利用判别模型对这些候选进行评分与筛选，最终输出最优解。这种解耦设计不仅降低了单一步骤的学习难度，还使得系统具有更好的模块化解释性和算法可扩展性。然而，它在性能和运行速度方面严重依赖于更好的抓取采样器。

在抓取位姿检测的候选生成阶段，全空间的随机采样往往面临“维度灾难”且搜索效率低下。为解决这一问题，相关研究通常将抓取感知视为类似于传统计算机视觉中的物体检测任务，结合几何启发式规则或学习到的先验知识来约束搜索空间，从而确定抓取感兴趣区域并引导采样过程。针对 2D 图像输入，采样器通常用于生成基于有向矩形表示的抓取候选。主流方法引入学习机制，利用深度网络预测物体的抓取可供性（affordance）密度[55－60]作为先验概率分布，从而引导采样集中在物体的高置信度区域。为增强表示能力，2D 采样点常被映射到 3D 空间[53, 56, 59－60]，或通过启发式方法将简单的矩形配置拓展为完整抓取构型[58]。针对 3D 点云输入，采样器主要用于生成更具通用性的 6-DoF 抓取位姿。为了生成高质量的 6-DoF 抓取候选，相关研究采用体素网格搜索[19－20, 22, 61]或交叉熵方法[62－64]（cross entropy method，CEM），通过迭代优化的方式动态调整采样分布，使其逐步收敛至物体的最佳抓取位姿分布。

在质量评估阶段，判别器的任务是构建一个鲁棒的评分函数，以从大量采样候选集中筛选出最优位姿。早期的研究[18, 53, 55, 59, 65－68]多依赖支

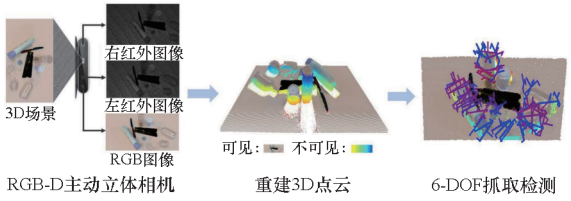
持向量机[69]（support vector machine，SVM）或概率模型[70]结合人工设计的几何特征进行分类，而现代方法则全面转向以深度卷积神经网络（convolutional neural network，CNN）为核心的判别架构，能够自动从原始数据中提取高阶特征。为了训练高精度的判别器，海量的多样化数据尤为重要。例如，Dex-Net 系列工作[62-63]通过物理仿真引擎批量生成了数百万级的合成抓取数据及对应的鲁棒性标签，极大地提升了深度网络在现实世界中面对噪声和不确定性时的泛化能力。此外，随着深度学习的发展，文献[19-20, 22, 57]将更复杂的输入模态点云也引入判别器设计中，使得模型能够直接处理无序、非结构化的 3D 点云数据，从而更精准地捕捉局部几何特征与抓取稳定性之间的非线性关系。

3.3 基于端到端的直接预测范式

基于感知输入的端到端抓取范式已成为当前机器人学领域的主流研究方向。其核心思想是将特征提取、决策与执行逻辑完整集成于统一的神经网络模型中，通过端到端训练直接建立从多模态原始观测数据（如 RGB 图像、深度图或 3D 点云）到最优抓取配置的映射关系。在这种架构下，包括抓取位姿采样与质量评估在内的所有关键中间步骤，均通过模型内部可训练参数的迭代更新进行全局自适应调整。相较于模仿学习的方法或生成与评估的两阶段方法，端到端范式能够更充分地挖掘大规模数据集中的隐含先验，且在推理阶段表现出更快的运行速度与实时响应能力。在具体实现上，端到端抓取学习主要包括三种代表性形式。

第一类方法以图像信息作为输入，借鉴目标检测的思想，将抓取视为一种特殊的目标，通过检测网络直接回归二维抓取有向矩形及置信度。文献[71-72]引入深度卷积神经网络将抓取候选采样与质量评估继承于单一模型中，显著提升了推理速度。然而，与传统视觉任务不同，抓取检测对物体的局部几何特征及旋转方向极度敏感。为解决这一挑战，研究者引入了有向锚框[73-74]、空间变换网络[75-76]以及旋转集成模块[77]，以实现更精准的位姿估计，并开发出反应式策略，能够实现抓取过程的实时调整与纠错[78-82]。

第二类方法以 3D 点云作为输入，通过主干网络提取点的层级化特征，直接回归每个采样点的 $SE(3)$ 抓取位姿及其置信度得分。目前，该方向的主流研究广泛探索了多种网络架构，包括卷积网络[13, 83]、Transformer[84-86]、生成式自编码器[87]、变分自编码器[88]以及 U-Net[89]等。其中，ASGrasp[83]、6-DOF GraspNet[87]和 ZeroGrasp[88]三种方法的输入与输出结果分别如图 4（a）~（c）所示。从最初的单阶段无锚框检测器[90-92]，到引入 $SE(3)$ 抓取锚框及多阶段精细化预测模型[93-94]，该方法在处理杂乱场景时的鲁棒性得到了极大增强。为了克服模型对特定执行器的依赖，相关工作尝试在模型输入中编码夹爪特征，使网络能够学习并适应不同类型的夹爪模型，从而提高了算法的通用性[95-96]。此外，针对真实场景中常见的点云畸变与噪声问题，文献[97-98]等通过引入点云补全或去噪模块来保证输入的干净一致性，从而提高了位姿估计的精度。



（a）ASGrasp 方法[83]

（a）ASGrasp method[83]

（b）6-DoF GraspNet 方法[87]

（b）6-DoF GraspNet method[87]

（c）ZeroGrasp 方法[88]

（c）ZeroGrasp method[88]

图 4 ASGrasp、6-DOF GraspNet 和 ZeroGrasp 三种方法的输入与输出效果图

Fig. 4 Visualizations of input and output for three methods: ASGrasp, 6-DOF GraspNet, and ZeroGrasp

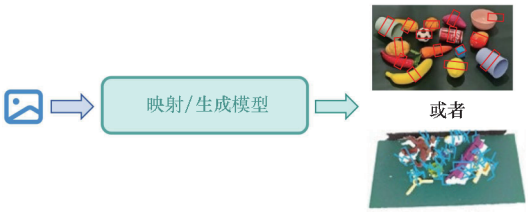
第三类方法是受图像语义分割启发而发展的像素级抓取图合成技术，为端到端范式提供了另一种高效路径。该方法通常采用 Encoder-Decoder 架构（如 U-Net[99]）生成抓取可供性图，不仅能指示最优抓取位置，还能表征全局范围内连续、一致的抓取可行性表示[100-103]。这种表征方式支持

机器人在未发现合适抓取位姿时，通过主动交互动作（如推挤或视角切换）来改变环境状态[82, 101, 104]或观测视角[12, 105]，从而在复杂堆叠场景中实现更智能的自主作业。

除了上面的三种主要方法，符号距离场[106－108]（signed distance field，SDF）、神经辐射场[109－110]（neural radiance field，NeRF）以及 3D 高斯泼溅[111－113]（3D Gaussian splatting，3DGS）等新兴 3D 表征正成为研究热点。这些方法通过隐式或显式地建模场景的连续几何与辐射场信息，能够提取比原始点云更高阶的几何特征，并有效辅助抓取点的精细采样，为下游任务提供更为丰富的底层信息支撑。

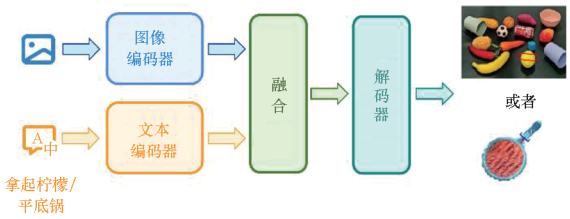
4　基于语义理解与推理增强的感知范式

随着大语言模型（large language model，LLM）与视觉－语言模型的快速发展，机器人抓取技术正经历从单纯的几何适应向语义驱动的感知范式转变。这一范式，即语言驱动的抓取，不再局限于物理稳定性的计算，而是旨在建立机器人对物理世界的深层理解。它要求机器人不仅具备处理开放词汇物体的泛化能力，还需理解复杂的任务意图与逻辑关系。图 5 展示了视觉数据驱动的感知范式和语义理解与推理增强的感知范式之间的差异。



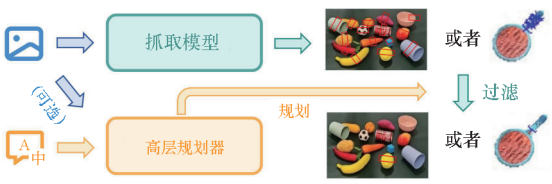
（a）纯视觉输入

（a）Vision-only input



（b）多模态融合

（b）Multimodal fusion



（c）分层规划引导

（c）Hierarchical planning-guided paradigm

（d）端到端基础模型

（d）End-to-end foundation models

图 5　视觉数据驱动的感知范式和语义理解与推理增强的感知范式的对比[15]

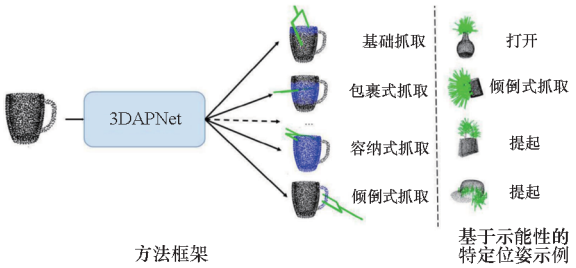
Fig. 5　Comparison between visual data-driven perception paradigm and perception paradigms enhanced by semantic understanding and reasoning[15]

4.1　基于多模态融合的范式

基于多模态融合的范式主要聚焦于视觉－语言联合建模下的目标定位与抓取决策问题，其目标在于建立抽象自然语言指令与具体物理场景中物体实例之间的映射关系。其核心思想是通过融合自然语言理解和视觉感知能力，使机器人能够解析人类的高层指令与操作意图，并据此生成与任务语义一致的抓取策略。与传统仅依赖视觉或几何信息的抓取方法不同，语言驱动抓取在决策过程中引入了显式的语义约束与任务引导，例如“抓取杯子的把手”“拿起红色的物体”或“为倒水抓取容器”等指令。这类语言信息不仅为目标实例的选择提供了重要线索，还隐含地编码了抓取部位、抓取方式以及潜在的后续操作意图，从而为实现面向任务的抓取与操作奠定了基础。

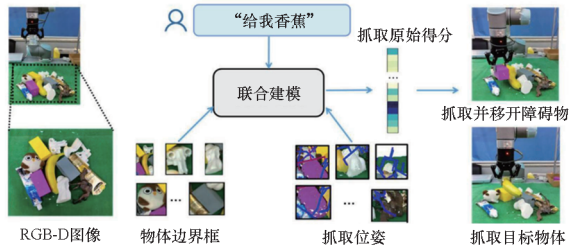
现有研究通常采用多模态表示学习框架，将自然语言指令编码为语义嵌入，并与视觉特征（如 RGB/RGB-D 图像或点云）进行对齐，从而在共享语义空间中完成目标实例的定位与抓取策略的生成。一类方法通过语言引导视觉注意力机制，使模型在语言约束下聚焦于与任务相关的目标区域，并进一步生成符合操作意图的抓取位姿[108]。另一类方法则采用联合建模策略，直接学习从“语言－视觉”联合表征到抓取构型的映射关系[109]。上面两种方法的典型代表 3DAPNet[114] 和 Vision-Language-Grasping[115] 具体的概述图分别如图 6（a）和图 6（b）所示。在这两类方法奠定的对齐基础上，为了进一步克服 Transformer 架构在处理长序列多模态数据时的计算开销与推理效率瓶颈，Nguyen 等[116] 将具备线性复杂度的状态空间模型引入语言驱动的抓取感知中。该框架通过在视觉主干网络的多个阶段同步注入文本特征，实现了视觉与语义的深度对齐，

从而更精准地捕捉目标物体的细粒度局部可供性。



（a）3DAPNet 方法[114]

（a）3DAPNet method[114]



（b）Vision-Language-Grasping 方法[115]

（b）Vision-Language-Grasping method[115]

图 6　3DAPNet 和 Vision-Language-Grasping 两种方法的概述

Fig. 6　Overview of 3DAPNet and Vision-Language-Grasping

此外，随着大规模视觉－语言模型的快速发展，语言驱动抓取方法逐渐展示出具备零样本或弱监督条件下的泛化能力，能够在未见过的物体类别及指令组合中生成合理的抓取行为。这一发展趋势显著增强了机器人在开放世界环境中的通用操作能力。

4.2　基于分层规划引导的抓取范式

在语义驱动的抓取研究中，另一类重要的方向是基于分层规划引导的抓取范式，其主要思想是将抓取任务解耦为底层候选动作生成与高层语义规划引导两个阶段。具体而言，系统首先利用现有的成熟抓取模型在物理空间中生成大规模的候选抓取位姿，随后利用高层规划器，如 LLM 或视觉语言模型（vision language model，VLM）对这些候选点进行筛选、排序或评分，从而在保证物理可实现性的同时，精确匹配任务的高层语义需求。

诸多代表性工作进一步验证了这一范式的有效性。Lu 等[117]通过 VLM 增强对目标物体的注意力，从而引导生成更具语义感知的抓取位姿，该方法的概述如图 7 所示；Tziafas 等[118]则利用基于 VLM 的语义先验，在物体发生遮挡的复杂环境下实现稳健的规划；Shi 等[119]提出了一种视觉语言驱动的主动感知与抓取分层框架，通过主动视点规划与不确定度融合，实现了在目标完全被遮挡的极端情况下的鲁棒抓取。除此之外，在认知与推理能力方面，Jin 等[120]进一步整合了视觉－语言推理能力，显著提升了系统对物体属性及操作逻辑的理解；Jiao 等[121]则专注于解析自由形式的语言指令，通过为场景中所有物体标注关键点并辅以视觉提示，增强 VLM 对复杂空间关系的推理能力，从而在存在多个同类物体的模糊场景中准确识别目标并规划抓取顺序；禹鑫燚等[122]提出了一种任务自适应的多模态融合框架，该方法创新性地引入了基于思维链的语言－视觉联合提示，使机器人能够根据场景复杂度动态分解任务链；Tang 等[123]将面向任务的抓取解耦为“知识生成—特征表征—位姿评估”三个阶段，即首先利用 LLM 生成目标物体的语义与几何描述，随后通过 Transformer 架构对抓取候选进行评分。还有研究尝试利用多模态大模型进行环境感知的错误纠正[124]，或通过学习以物体为中心的属性来实现跨任务的快速抓取自适应[125]。与此同时，基于可供性驱动的方法也开始将语言和视觉线索融入抓取生成过程中，使机器人的操作更符合人类的常识逻辑[126－128]。

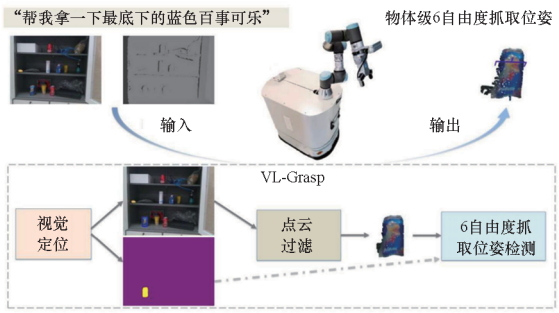


图 7　VL-Grasp 方法的概述[117]

Fig. 7　Overview of VL-Grasp[117]

然而，这类基于分层规划引导的方法也面临显著的挑战：一方面，多组件叠加使系统日益复杂，大幅增加了计算与内存开销；另一方面，该范式存在显著的信息传递瓶颈与累积误差问题，系统的最终性能高度受限于底层模型生成的候选位姿质量。

4.3　基于端到端的基础模型范式

基础模型（foundation models）的兴起促使机器人抓取从“位姿预测”向“动作生成”范式演进。不同于以往侧重于静态抓取位姿预测且脱离执行控制的方案，该范式致力于构建感知—决策—控制一体化的架构，直接建立从多模态输入（如视

觉、语言、本体状态)到抓取动作或连续操作序列的映射。基础抓取模型通常基于大规模跨任务数据预训练,具备更强的泛化能力和知识迁移能力。这使得模型能够在不同物体、不同任务乃至不同机器人平台之间实现零样本或少样本迁移。

Deng 等[129]提出了一个依托大规模合成动作数据预训练的视觉 - 语言 - 动作抓取基础模型。该研究首先构建了首个 10 亿量级的合成抓取数据集 SynGrasp-1B,通过引入高保真物理仿真、光线追踪渲染及大规模域随机化技术,实现了对多元化物体与复杂环境表征的深度覆盖。在模型架构层面,该方法将目标检测与抓取位姿预测作为中间"推理环境",并将其融入后续的动作生成过程,实现从多模态输入到操作序列的端到端统一映射。Zhang 等[130]创新性地引入了视觉思维链(visual chain-of-thought)推理机制,通过一种两回合的进阶处理范式显著提升了模型对视觉信息的深层理解能力。在首回合推理中,模型首先预测目标对象的粗粒度边界框以实现初步定位;随后,系统通过对目标区域进行动态裁剪与放大,引导模型在第二回合中聚焦于更为精细的视觉特征,从而生成高精度的最终抓取位姿。

尽管目前端到端的基础模型范式在极度杂乱场景下的执行效率与精确度仍是学术界关注的焦点,但其展现出的通用性已成为迈向通用机器人智能的重要基石。

5　公开数据集和评价指标

从早期的基于解析几何的分析方法,到目前基于语义理解与推理增强的方法,抓取技术的演进历程深度依赖于标准化数据集的规模化发展。数据集不仅为深度神经网络提供了学习特征表示的数据基础,更定义了任务的复杂边界,而评价指标则构建了衡量算法鲁棒性、泛化性及作业效率的客观尺度。本节旨在对机器人抓取领域的常用数据集进行多维度的分类梳理,并对涵盖任务可靠性和提议准确度在内的量化评价体系进行剖析,以期为该方向的研究提供系统的参考框架。

5.1　公开数据集

抓取数据集的演进历程体现了从简单到复杂、从单一模态到多模态语义的转变趋势。表 1 对当前机器人抓取研究中具有代表性的公开数据集进行对比,从物体与场景规模、抓取表示形式以及数据来源等维度进行归纳。根据抓取自由度、场景复杂度及数据生成方式,可将抓取数据集的发展划分为如下几个核心阶段。

表 1　典型机器人抓取数据集对比

Tab.1　Comparision of representative robotic grasping datasets

数据集	物体/场景规模	抓取表示形式	数据来源
Cornell	240 个物体/885 张图像	平面矩形框	真实世界
Jacquard	11 000 个物体/54 000 张图像	平面矩形框	物理仿真
ACRONYM	8 000 个物体	6-DoF 抓取位姿	虚拟环境
GraspNet-1Billion	144 类物体/190 个场景	6-DoF 抓取位姿	真实世界
SynGrasp-1B	10 000 个物体/10 亿帧数据	6-DoF 抓取位姿	光追仿真
ZeroGrasp-11B	12 000 个物体/100 万张图像	6-DoF 抓取位姿	仿真合成
GraspClutter6D	200 个物体/1 000 个场景	6-DoF 密集标注	真实世界

5.1.1　早期平面抓取基准:从矩形框到 *SE*(2)简化

在深度学习介入抓取研究的初期,研究者通常将 3D 空间的抓取任务简化为图像平面上的检测问题。这类数据集的核心逻辑是在 RGB 或 RGB-D 图像中寻找最优的抓取位置和旋转角度,通常被定义为 3-DoF 或 4-DoF 任务。

Cornell[131](康奈尔)数据集是该阶段最具代表性的资源。它包含约 240 个常见物体的 885 张图像及对应的点云,通过人工标注的抓取矩形框来定义作业区域。尽管规模较小且仅支持单视角观察,但它为后续抓取检测网络的算法验证奠定了基础。然而,由于视角的局限性,康奈尔数据集难以支撑复杂环境下(如物体堆叠或侧向抓取)的算法需求。

为解决数据规模瓶颈,Depierre 等[132]利用物理仿真引擎在大规模虚拟环境下生成 Jacquard 数据集。它包含了超过 11 000 个物体的 54 000 张图像,标注数量达到了 110 万个。与 Cornell 数据集的人工标注不同,Jacquard 的标注通过仿真环境中的抓取尝试和物理验证产生,保证了标注的物理有效性。然而,这类数据集本质上仍属于 *SE*(2)简化范畴,忽略了夹爪在 3D 空间中的趋近矢对抓取稳定性的影响。

5.1.2　6-DoF 真实世界感知基准

随着工业与家庭服务场景对抓取精度要求的提升,研究重心从平面抓取转向全 6-DoF 抓取。

这意味着算法不仅需要预测抓取中心,还需给出夹爪在 $SE(3)$ 空间中的三维旋转 。

GraspNet-1Billion[133] 作为目前该领域较权威的大规模真实世界抓取数据集,通过精细设计的标注流程突破了数据规模限制。该数据集包含144 类真实物体,覆盖了 190 个场景,每个场景均从 256 个不同视角进行拍摄。其核心创新在于通过高精度 3D 扫描与 6D 位姿标注,将解析生成的10 亿级抓取标签准确映射至真实点云观测空间,使模型能够有效学习应对真实深度相带来的传感噪声,从而显著提升 Sim-to-Real 场景下的泛化能力。

5.1.3 面向通用大模型的大规模合成数据集

在视觉 - 语言 - 动作模型和具身基础模型兴起的背景下,抓取数据集的规模达到了亿级甚至10 亿级。这类数据集不局限于简单的位姿回归,而是强调感知与动作的端到端对齐。

SynGrasp-1B[129] 展示了大规模合成数据在基础模型训练中的显著优势。该数据集基于来自Objaverse-LVIS[134] 的万级物体网格,包含 10 亿帧抓取数据,并通过光线追踪渲染与大规模领域随机实现高真实感与强多样性,从而使模型具备优异的零样本泛化能力,可直接由仿真迁移部署至真实机器人平台。

ZeroGrasp-11B[88] 则通过将 3D 形状重建与抓取预测耦合,强调了几何完整性对操作安全性的重要性。该数据集提供了 113 亿个物理有效的抓取标注。

5.2 评价指标

评估一个抓取系统的优劣,需要构建一套完整的评价体系。目前主流的评价指标可以分为两大类:任务可靠性指标和提议准确度指标。

5.2.1 任务可靠性指标

成功率(success rate,SR)是衡量抓取系统最核心的指标。通常定义为:在给定的尝试次数内,机器人成功将目标物体提升至指定高度并保持稳定的次数比例。

对于杂乱场景,清空率(declutter rate,DR)或清除率(completion rate)用于衡量算法在复杂环境下清理所有目标的能力。在多轮次的杂乱场景实验中,清空率能够体现算法在面对最后几个极难抓取、被重度遮挡物体时的韧性。

5.2.2 提议准确度指标

在 6-DoF 位姿预测任务中,平均精度(average precision,AP)逐渐成为重要的性能评价指标之一。为了更细致地分析物体材质对抓取的影响,GraspNet 基准引入了不同摩擦系数下的 AP 计算。较低的摩擦系数代表更光滑的物体表面,对抓取的稳定性提出了更高要求。最终的 AP 通常是多个摩擦阈值下精度的均值,从而综合反映模型在复杂接触环境中的鲁棒性与泛化能力。

6 对比、挑战与未来展望

6.1 范式对比

机器人抓取感知经历了从“解析几何模型”到“视觉数据驱动”再到“语义理解与推理增强”的深刻演变。为了系统性地梳理这三类感知路径的本质差异,表 2 从输入模态、理论基础、泛化能力及局限性等多个核心维度进行了深刻对比分析。表 3 列出了三类感知范式下典型算法的定量分析结果及所用数据集。

6.2 挑战和未来展望

尽管基于学习的机器人抓取技术已取得显著进展,但在仿真到现实迁移、推理效率、跨模态信息融合以及复杂任务扩展等方面仍面临诸多挑战,当前机器人抓取领域面临以下共性技术瓶颈。

6.2.1 弥补 Sim-to-Real 迁移差距

域偏移(domain shift)是数据驱动抓取算法从实验室走向开放环境的核心障碍。由于深度学习模型在训练阶段高度依赖合成数据或特定的仿真场景,当其部署于分布存在偏差的真实环境时,感知精度与决策稳健性往往大幅下降。尽管物理引擎为抓取交互提供了低成本的仿真环境,但“仿真与现实的鸿沟”(Sim-to-Real gap)依然显著:一方面是视觉层面的表观差异,包括环境光照、物体纹理及传感器噪声的模拟失真;另一方面则是深层的动力学差异(dynamics gap),仿真环境难以精确建模物体间的摩擦特性、非刚性形变以及复杂的力反馈交互。这种双重偏移导致在仿真中表现优异的策略在真实机器人上难以直接泛化。

域随机化(domain randomization, DR)是目前提升模型鲁棒性的主流技术路线。Huber 等[135-136] 的研究证明,通过在仿真中引入光照、纹理及物理参数的随机化,可以使生成的抓取策略在迁移到真实世界时表现出更强的鲁棒性。同时,预训练视觉语言模型通过在互联网规模数据上的学习,积累了丰富的真实世界常识,能够为解决域偏移提供新思路[137] 。通过采用轻量级的提示学习(prompt learning)策略[138] ,可以在保留大规模模型通用认知能力的同时,通过少量真实交

表 2　机器人抓取感知范式的演进对比:从解析几何到语义推理

Tab. 2　Evolution of robotic grasping perception paradigms: from geometric constraints to semantic reasoning

对比维度	基于解析几何模型的感知	基于视觉数据驱动的感知	基于语义理解与推理增强的感知
代表工作	Jones 等[24]、Stoytchev 等[25]	Robo-ABC[31]、AnyGrasp[13]	GraspVLA[129]、ReasoningGrasping[120]
输入模态	已知几何模型(CAD)、力反馈	RGB/RGB-D、点云	RGB/RGB-D、点云、语言指令
任务目标	物理稳定性与力/形闭合条件满足	抓取鲁棒性与成功率最大化	任务适应性与人类意图对齐
理论基础	接触力学、几何约束理论	统计学习理论、深度神经网络	大规模预训练、多模态对齐、常识推理
数据需求	无需训练数据,依赖精确几何模型	较高,依赖大规模标注或仿真	极高,依赖大规模多模态对齐数据
泛化能力	较弱,受限于已知几何模型	较强,可适应未知几何物体	极强,具备零样本或少样本迁移能力
语义理解	无,仅关注几何可行性	有限,缺乏高层任务语义理解	强,支持自然语言指令与任务推理
主要局限	对感知噪声敏感,模型构建成本高	缺乏显式逻辑推理机制	计算与推理开销较大,存在累计误差
目前应用	结构化工业环境下的精密装配、基于已知 CAD 模型的精细化协同控制	非结构化物流仓储分拣、家庭环境中的通用物体抓取、复杂背景下的实时位姿检测	开放词汇物体交互、基于自然语言指令的任务驱动抓取、语义受限场景下的目标定位
未来应用	为现代混合感知模型提供本质物理约束引导、高精度微纳操作与复杂动力学分析	跨场景大规模通用抓取模型、具备实时纠错能力的反应式策略、低质量感知下的鲁棒操作	通用具身基础模型、复杂长时程任务的逻辑推理与自主决策、多机器人/移动平台协同操作

表 3　典型机器人抓取感知算法性能对比

Tab. 3　Performance comparison of typical robotic grasping perception algorithms

感知范式	代表性算法	核心数据集	成功率/%
基于解析几何模型的感知	Stoytchev 等[25]	自建“可供性表”	75.0
基于视觉数据驱动的感知	AnyGrasp[13]	GraspNet-1Billion	93.3
	ASGrasp[83]	GraspNet-1Billion	95.2
	6-DOF GraspNet[87]	ShapeNet	90.0
基于语义理解与推理增强的感知	VL-Grasp[117]	RoboRefIt	78.7
	GraspVLA[129]	SynGrasp-1B	93.3
	VCoT-Grasp[130]	VCoT-GraspSet	76.0

互数据,以较低的计算成本实现向机器人操作领域的知识迁移。

6.2.2　提升计算效率与实时控制性能

在工业制造与物流仓储等实际应用场景中,计算效率是限制抓取技术落地的关键因素。复杂模型的生成速度慢且存储需求大,严重阻碍了系统的实时运行效率。为了应对这一挑战,可以借鉴计算机视觉领域的模型优化方案:

1)知识蒸馏:通过将复杂“教师模型”的视觉表征能力迁移至轻量化“学生模型”,在显著降低模型尺寸的同时保持模型性能[139-140]。

2)模型压缩:减少参数量与模型复杂度,使其适配资源受限的嵌入式部署环境[141]。

6.2.3　跨模态异构信息融合

触觉信息在机器人抓取任务中扮演着不可替代的角色,它能通过实时的力反馈实现精确的接触检测。如何高效整合视觉与触觉传感器的输入信息,以提升在面对复杂形状和多元材质物体时的抓取精度,是当前的重大技术挑战。触觉感知作为近距离感知的核心,能够有效补偿视觉信息在物体遮挡或动态变化时的缺失。

目前已有研究构建了新型传感器,实现了视觉与触觉数据的融合[142],或通过 PoseFusion[143]等方法利用 LSTM 网络[144]筛选最优目标位姿。这种多模态协同感知不仅能突破单一模态的局限,还显著提升了机器人在操纵精密物体时的灵巧度与稳健性。

6.2.4　拓展到更多样化的抓取任务

当前大多数研究仍局限于单目标的刚性物体抓取。然而,真实世界中的任务远比实验室环境复杂,未来的研究需要重点突破以下领域:

1)复杂目标抓取:包括超大尺寸物体[145]、透明物体[146]、柔软可变形物体[147-149]以及具有不

可预测行为的生命体[150]。

2)杂乱场景下目标导向灵巧抓取:在杂乱场景下实现闭环目标导向灵巧抓取[151-152]。

3)多物体协同抓取:在杂乱场景下实现多目标抓取,并处理物体间的复杂物理交互[153-155]。

4)移动操作抓取:机器人在导航移动的同时,利用机械臂或灵巧手对物体进行抓取[156-161]。

目前面向此类复杂操作任务的基准数据集和系统化方法论仍十分匮乏,已成为机器人迈向通用化发展的关键瓶颈。因此,加强对复杂抓取与操作场景的研究投入,并构建高质量、可复现的数据集体系,对提升机器人的操作柔性和任务通用性具有重要的学术意义与工程应用价值。

7 总结

作为目前具身智能与机器人学研究的前沿交叉,机器人抓取感知旨在解决机器人与物理世界交互过程中的最优抓取位姿的预测问题,尤其关注非结构化环境下由感知噪声、遮挡与不确定性所带来的挑战。该技术经历了从以几何模型为核心的解析方法,向数据驱动学习方法,再到融合高层语义推理能力的感知范式演进,使机器人逐步具备了从深层语义层面理解物理场景与任务意图的能力。在显著提升系统泛化性能与任务理解能力的同时,机器人抓取感知技术也不断拓展其在家庭服务、物流分拣等复杂实际应用场景中的适用范围。

综合来看,机器人抓取感知在迈向通用人工智能与通用机器人系统的过程中具有重要的研究意义与应用潜力。围绕其感知范式演进开展系统性研究,不仅有助于提升机器人自主操作与环境适应能力,也对推动智能机器人技术的工程化落地与规模化应用具有关键支撑作用。基于此,本文系统梳理了机器人抓取感知领域的研究进展,将现有方法归纳为三类逐级演进的感知范式,并从输入模态、泛化能力、任务适应性等多个维度对不同范式的优势与适用场景进行了对比分析。同时,本文总结了当前机器人抓取感知在仿真到现实迁移、跨模态信息融合及语义一致性建模等方面面临的共性挑战,并探讨了未来可能的研究方向。

展望未来,机器人抓取感知研究亟须进一步加强具身基础模型与高自由度灵巧手抓取动作生成之间的深度融合,将大模型中蕴含的常识推理与任务理解能力有效注入底层抓取与控制策略中,以推动系统在实时性、稳定性与操作灵巧度等方面取得实质性突破。同时,有必要将现有抓取感知算法扩展至更具挑战性的柔性物体操作、复杂接触建模及移动操作等任务场景中。期待随着相关技术的不断发展与完善,机器人抓取感知能够为通用机器人走向真实世界、融入日常生活提供更加坚实的技术基础与广阔的发展空间。

参考文献(References)

[1] WANG Z F, ZHAO H, XU J Z, et al. RoboBPP: benchmarking robotic online bin packing with physics-based simulation[EB/OL]. (2025-12-04)[2026-01-25]. https://arxiv.org/abs/2512.04415.

[2] ZHAO H, XU J Z, YU K X, et al. Deliberate planning of 3D bin packing on packing configuration trees[EB/OL]. (2025-09-04)[2026-01-25]. https://arxiv.org/abs/2504.04421.

[3] LI W H, YU Z Y, SHE Q J, et al. LLM-enhanced scene graph learning for household rearrangement[C]//Proceedings of SIGGRAPH Asia 2024 Conference Papers, 2024: 32.

[4] LIU P Q, ORRU Y, VAKIL J, et al. OK-Robot: what really matters in integrating open-knowledge models for robotics[EB/OL]. (2024-02-29)[2026-01-25]. https://arxiv.org/abs/2401.12202.

[5] BICCHI A, KUMAR V. Robotic grasping and contact: a review[C]//Proceedings of IEEE International Conference on Robotics and Automation, 2000: 348-353.

[6] SHIMOGA K B. Robot grasp synthesis algorithms: a survey[J]. International Journal of Robotics Research, 1996, 15(3): 230-266.

[7] 苏康, 李嘉良, 李俊国, 等. 基于视觉的机器人端到端策略抓取估计综述[J]. 信息与控制, 2025, 54(3): 372-389.
SU K, LI J L, LI J G, et al. Review of vision-based robot end-to-end strategic grasping estimation[J]. Information and Control, 2025, 54(3): 372-389. (in Chinese)

[8] KLEEBERGER K, BORMANN R, KRAUS W, et al. A survey on learning-based robotic grasping[J]. Current Robotics Reports, 2020, 1(4): 239-249.

[9] BAI Q, LI S B, YANG J, et al. Object detection recognition and robot grasping based on machine learning: a survey[J]. IEEE Access, 2020, 8: 181855-181879.

[10] WOLF R, SHI Y T, LIU S, et al. Diffusion models for robotic manipulation: a survey[J]. Frontiers in Robotics and AI, 2025, 12: 1606247.

[11] ZHENG Y, YAO L, SU Y J, et al. A survey of embodied learning for object-centric robotic manipulation[J]. Machine Intelligence Research, 2025, 22(4): 588-626.

[12] WANG C X, FANG H S, GOU M H, et al. Graspness discovery in clutters for fast and accurate grasp detection[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 15944-15953.

[13] FANG H S, WANG C X, FANG H J, et al. AnyGrasp: robust and efficient grasp perception in spatial and temporal domains[J]. IEEE Transactions on Robotics, 2023, 39(5):

3929 – 3945.
[14] SONG X, LI Y Y, ZHANG Y F, et al. An overview of learning-based dexterous grasping: recent advances and future directions[J]. Artificial Intelligence Review, 2025, 58(10): 300.
[15] BAI S H, SONG W X, CHEN J Y, et al. Towards a unified understanding of robot manipulation: a comprehensive survey[EB/OL]. (2025 – 10 – 13) [2026 – 01 – 25]. https://arxiv.org/abs/2510.10903.
[16] ZHANG H B, TANG J, SUN S G, et al. Robotic grasping from classical to modern: a survey[EB/OL]. (2022 – 02 – 08) [2026 – 01 – 25]. https://arxiv.org/abs/2202.03631.
[17] SUN J H, MAO P J, KONG L J, et al. A review of embodied grasping[J]. Sensors, 2025, 25(3): 852.
[18] JIANG Y, MOSESON S, SAXENA A. Efficient grasping from RGBD images: learning using a new rectangle representation[C]//Proceedings of 2011 IEEE International Conference on Robotics and Automation, 2011: 3304 – 3311.
[19] GUALTIERI M, TEN PAS A, SAENKO K, et al. High precision grasp pose detection in dense clutter [C]// Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016: 598 – 605.
[20] LIANG H Z, MA X J, LI S, et al. PointNetGPD: detecting grasp configurations from point sets[C]//Proceedings of 2019 International Conference on Robotics and Automation (ICRA), 2019: 3629 – 3635.
[21] MILLER A T, KNOOP S, CHRISTENSEN H I, et al. Automatic grasp planning using shape primitives [C]// Proceedings of 2003 IEEE International Conference on Robotics and Automation, 2003: 1824 – 1829.
[22] TEN PAS A, GUALTIERI M, SAENKO K, et al. Grasp pose detection in point clouds[J]. International Journal of Robotics Research, 2017, 36(13/14): 1455 – 1473.
[23] BICCHI A. On the closure properties of robotic grasping[J]. The International Journal of Robotics Research, 1995, 14(4): 319 – 334.
[24] JONES J L, LOZANO-PEREZ T. Planning two-fingered grasps for pick-and-place operations on polyhedra [C]// Proceedings of IEEE International Conference on Robotics and Automation, 1990: 683 – 688.
[25] STOYTCHEV A. Behavior-grounded representation of tool affordances [C]//Proceedings of 2005 IEEE International Conference on Robotics and Automation, 2005: 3060 – 3065.
[26] FERRARI C, CANNY J. Planning optimal grasps [C]// Proceedings of 1992 IEEE International Conference on Robotics and Automation, 1992: 2290 – 2295.
[27] MILLER A T, ALLEN P K. Examples of 3D grasp quality computations [C]//Proceedings of 1999 IEEE International Conference on Robotics and Automation, 1999: 1240 – 1246.
[28] MIRTICH B, CANNY J. Easily computable optimum grasps in 2-D and 3-D[C]//Proceedings of 1994 IEEE International Conference on Robotics and Automation, 1994: 739 – 747.
[29] MISHRA B. Grasp metrics: optimality and complexity[R]. New York: New York University, 1995.
[30] TEICHMANN M. A grasp metric invariant under rigid motions[C]//Proceedings of IEEE International Conference on Robotics and Automation, 1996: 2143 – 2148.
[31] JU Y C, HU K Z, ZHANG G W, et al. Robo-ABC: affordance generalization beyond categories via semantic correspondence for robot manipulation[C]//Computer Vision-ECCV 2024, 2025: 222 – 239.
[32] ALEOTTI J, CASELLI S. Grasp recognition in virtual reality for robot pregrasp planning by demonstration [C]// Proceedings of 2006 IEEE International Conference on Robotics and Automation, 2006: 2801 – 2806.
[33] DE GRANVILLE C, SOUTHERLAND J, FAGG A H. Learning grasp affordances through human demonstration[EB/OL]. [2026 – 01 – 25]. https://www.cs.ou.edu/~fagg/papers/2006/degranville_etal_2006_affordance.pdf.
[34] EKVALL S, KRAGIC D. Grasp recognition for programming by demonstration [C]//Proceedings of 2005 IEEE International Conference on Robotics and Automation, 2005: 748 – 753.
[35] KANG S B, IKEUCHI K. Toward automatic robot instruction from perception-recognizing a grasp from observation [J]. IEEE Transactions on Robotics and Automation, 1993, 9(4): 432 – 443.
[36] ZOLLNER R, ROGALLA O, DILLMANN R, et al. Dynamic grasp recognition within the framework of programming by demonstration [C]//Proceedings of the 10th IEEE International Workshop on Robot and Human Interactive Communication, 2001: 418 – 423.
[37] CUTKOSKY M R. On grasp choice, grasp models, and the design of hands for manufacturing tasks [J]. IEEE Transactions on Robotics and Automation, 1989, 5(3): 269 – 279.
[38] FEIX T, PAWLIK R, SCHMIEDMAYER H B, et al. A comprehensive grasp taxonomy[EB/OL]. [2026 – 01 – 25]. https://www.csc.kth.se/grasp/taxonomyGRASP.pdf.
[39] ZHANG T H, MCCARTHY Z, JOW O, et al. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation[C]//Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018: 5628 – 5635.
[40] ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning[C]//Proceedings of the Twenty-First International Conference on Machine Learning, 2004: 1.
[41] HORN M W. Quantifying grasp quality using an inverse reinforcement learning algorithm[D]. Austin: University of Texas at Austin, 2017.
[42] XIE X, LI C Y, ZHANG C, et al. Learning virtual grasp with failed demonstrations via Bayesian inverse reinforcement learning[C]//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 1812 – 1817.
[43] DO M, ROMERO J, KJELLSTRÖM H, et al. Grasp recognition and mapping on humanoid robots [C]// Proceedings of 2009 9th IEEE-RAS International Conference on Humanoid Robots, 2009: 465 – 471.
[44] MORALES A, AZAD P, ASFOUR T, et al. An anthropomorphic grasping approach for an assistant humanoid robot[J]. VDI Berichte, 2006, 1956: 149.
[45] CURTIS N, XIAO J. Efficient and effective grasping of novel objects through learning and adapting a knowledge base[C]//

Proceedings of 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008: 2252 – 2257.

[46] EKVALL S, KRAGIC D. Learning and evaluation of the approach vector for automatic grasp generation and planning[C]// Proceedings of 2007 IEEE International Conference on Robotics and Automation, 2007: 4715 – 4720.

[47] HERZOG A, PASTOR P, KALAKRISHNAN M, et al. Template-based learning of grasp selection[C]//Proceedings of 2012 IEEE International Conference on Robotics and Automation, 2012: 2379 – 2384.

[48] HERZOG A, PASTOR P, KALAKRISHNAN M, et al. Learning of grasp selection based on shape-templates[J]. Autonomous Robots, 2014, 36(1): 51 – 65.

[49] HSIAO K, LOZANO-PEREZ T. Imitation learning of whole-body grasps [C]//Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006: 5657 – 5662.

[50] VANSCHOREN J. Meta-learning: a survey[EB/OL]. (2018 – 10 – 08) [2026 – 01 – 25]. https://arxiv.org/abs/1810.03548.

[51] SERMANET P, LYNCH C, CHEBOTAR Y, et al. Time-contrastive networks: self-supervised learning from video[C]// Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018: 1134 – 1141.

[52] ZHOU A, JANG E, KAPPLER D, et al. Watch, try, learn: meta-learning from demonstrations and reward [EB/OL]. (2020 – 01 – 30) [2026 – 01 – 25]. https://arxiv.org/abs/1906.03352.

[53] BOHG J, KRAGIC D. Grasping familiar objects using shape context[C]//Proceedings of 2009 International Conference on Advanced Robotics, 2009: 1 – 6.

[54] 李小晗, 张哲戬, 徐胜军, 等. 几何先验引导的堆叠点云抓取位姿联合预测方法[J/OL]. 控制与决策. (2025 – 12 – 04) [2026 – 01 – 25]. https://doi.org/10.13195/j.kzyjc.2025.0696.
LI X H, ZHANG Z J, XU S J, et al. A joint grasp pose prediction method for stacked point clouds guided by geometric priors[J/OL]. Control and Decision. (2025 – 12 – 04) [2026 – 01 – 25]. https://doi.org/10.13195/j.kzyjc.2025.0696. (in Chinese)

[55] BOHG J, KRAGIC D. Learning grasping points with shape context [J]. Robotics and Autonomous Systems, 2010, 58(4): 362 – 377.

[56] GOU M H, FANG H S, ZHU Z D, et al. RGB matters: learning 7-DoF grasp poses on monocular RGBD images[C]// Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 13459 – 13466.

[57] LI Y K, SCHOMAKER L, KASAEI S H. Learning to grasp 3D objects using deep residual U-Nets[C]//Proceedings of 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2020: 781 – 787.

[58] RAO D, LE Q V, PHOKA T, et al. Grasping novel objects with depth segmentation[C]//Proceedings of 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010: 2578 – 2585.

[59] SAXENA A, DRIEMEYER J, KEARNS J, et al. Robotic grasping of novel objects [C]//Proceedings of the 20th International Conference on Neural Information Processing Systems, 2006: 1209 – 1216.

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

[61] TEN PAS A, PLATT R. Using geometry to detect grasp poses in 3D point clouds [M]//BICCHI A, BURGARD W. Robotics Research. Cham: Springer, 2018: 307 – 324.

[62] 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]. (2017 – 08 – 08) [2026 – 01 – 25]. https://arxiv.org/abs/1703.09312.

[63] MAHLER J, MATL M, LIU X Y, et al. Dex-Net 3.0: computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning [C]// Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018: 5620 – 5627.

[64] YAN X C, KHANSARI M, HSU J, et al. Data-efficient learning for sim-to-real robotic grasping using deep point cloud prediction networks[EB/OL]. (2019 – 06 – 21) [2026 – 01 – 25]. https://arxiv.org/abs/1906.08989.

[65] DANG H, ALLEN P K. Learning grasp stability [C]// Proceedings of 2012 IEEE International Conference on Robotics and Automation, 2012: 2392 – 2397.

[66] LE Q V, KAMM D, KARA A F, et al. Learning to grasp objects with multiple contact points[C]//Proceedings of 2010 IEEE International Conference on Robotics and Automation, 2010: 5062 – 5069.

[67] PELOSSOF R, MILLER A, ALLEN P, et al. An SVM learning approach to robotic grasping[C]//Proceedings of IEEE International Conference on Robotics and Automation, 2004: 3512 – 3518.

[68] SCHILL J, LAAKSONEN J, PRZYBYLSKI M, et al. Learning continuous grasp stability for a humanoid robot hand based on tactile sensing[C]//Proceedings of 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), 2012: 1901 – 1906.

[69] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273 – 297.

[70] KOLLER D, FRIEDMAN N. Probabilistic graphical models: principles and techniques [M]. Cambridge: MIT Press, 2009.

[71] CHU F J, XU R N, VELA P A. Real-world multiobject, multigrasp detection [J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3355 – 3362.

[72] GUO D, SUN F C, KONG T, et al. Deep vision networks for real-time robotic grasp detection [J]. International Journal of Advanced Robotic Systems, 2016, 14(1): 1729881416682706.

[73] ZHANG H B, ZHOU X W, LAN X G, et al. A real-time robotic grasping approach with oriented anchor box[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(5): 3014 – 3025.

[74] ZHOU X W, LAN X G, ZHANG H B, et al. Fully

convolutional grasp detection network with oriented anchor box[C]// Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 7223 – 7230.

[75] GARIÉPY A, RUEL J C, CHAIB-DRAA B, et al. GQ-STN: optimizing one-shot grasp detection based on robustness classifier[C]//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 3996 – 4003.

[76] PARK D, CHUN S Y. Classification based grasp detection using spatial transformer network[EB/OL]. (2018 – 03 – 04)[2026 – 01 – 25]. https://arxiv.org/abs/1803.01356.

[77] PARK D, SEO Y, CHUN S Y. Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module[C]//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020: 9397 – 9403.

[78] BAIER-LOWENSTEIN T, ZHANG J W. Learning to grasp everyday objects using reinforcement-learning with automatic value cut-off [C]//Proceedings of 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007: 1551 – 1556.

[79] LAMPE T, RIEDMILLER M. Acquiring visual servoing reaching and grasping skills using neural reinforcement learning [C]//Proceedings of 2013 International Joint Conference on Neural Networks (IJCNN), 2013: 1 – 8.

[80] 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.

[81] MAHLER J, GOLDBERG K. Learning deep policies for robot bin picking by simulating robust grasping sequences[C]// Proceedings of the 1st Annual Conference on Robot Learning, 2017: 515 – 524.

[82] ZENG A, SONG S R, WELKER S, et al. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning[C]//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 4238 – 4245.

[83] SHI J, A Y, JIN Y X, et al. ASGrasp: generalizable transparent object reconstruction and 6-DoF grasp detection from RGB-D active stereo camera[C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024: 5441 – 5447.

[84] FAN Q Y, CAI Y H, LI C, et al. MISCGrasp: leveraging multiple integrated scales and contrastive learning for enhanced volumetric grasping[EB/OL]. (2025 – 07 – 03)[2026 – 01 – 25]. https://arxiv.org/abs/2507.02672.

[85] 陈鹏, 白勇, 陈旭, 等. 融合点云 Transformer 的多尺度抓取检测模型[J]. 计算机工程与应用, 2025, 61(22): 196 – 204.
CHEN P, BAI Y, CHEN X, et al. Multiscale grasping detection model integrating point cloud Transformer[J]. Computer Engineering and Applications, 2025, 61(22): 196 – 204. (in Chinese)

[86] 仓欣. 基于模态融合 Transformer 的轻量化网络设计及其应用[D]. 扬州: 扬州大学, 2025.
CANG X. Lightweight network design and its application based on modal fusion transformer[D]. Yangzhou: Yangzhou University, 2025. (in Chinese)

[87] MOUSAVIAN A, EPPNER C, FOX D. 6-DOF GraspNet: variational grasp generation for object manipulation[C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 2901 – 2910.

[88] IWASE S, IRSHAD M Z, LIU K, et al. ZeroGrasp: zero-shot shape reconstruction enabled robotic grasping[C]// Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 17405 – 17415.

[89] XIE P W, CHEN S A, TANG W, et al. Rethinking 6-DoF grasp detection: a flexible framework for high-quality grasping[J]. Pattern Recognition, 2026, 170: 112088.

[90] NI P Y, ZHANG W G, ZHU X X, et al. PointNet + + grasping: learning an end-to-end spatial grasp generation algorithm from sparse point clouds[C]//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020: 3619 – 3625.

[91] QIN Y Z, CHEN R, ZHU H, et al. S4G: amodal single-view single-shot $\mathbb{SE}(3)$ grasp detection in cluttered scenes[C]// Proceedings of the Conference on Robot Learning, 2020: 53 – 65.

[92] SUNDERMEYER M, MOUSAVIAN A, TRIEBEL R, et al. Contact-GraspNet: efficient 6-DoF grasp generation in cluttered scenes[C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 13438 – 13444.

[93] ZHAO B L, ZHANG H B, LAN X G, et al. REGNet: region-based grasp network for end-to-end grasp detection in point clouds[C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 13474 – 13480.

[94] WEI W, LUO Y K, LI F Y, et al. GPR: grasp pose refinement network for cluttered scenes[C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 4295 – 4302.

[95] SHAO L, FERREIRA F, JORDA M, et al. UniGrasp: learning a unified model to grasp with multifingered robotic hands[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2286 – 2293.

[96] XU Z J, QI B C, AGRAWAL S, et al. AdaGrasp: learning an adaptive gripper-aware grasping policy[C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 4620 – 4626.

[97] CHENG Y F, ZHA F S, GUO W, et al. PCF-Grasp: converting point completion to geometry feature to enhance 6-DoF grasp[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2026, 56(1): 617 – 628.

[98] CAI J F, CHEN Z B, WU X M, et al. Real-to-sim grasp: rethinking the gap between simulation and real world in grasp detection[EB/OL]. (2024 – 10 – 09)[2026 – 01 – 25]. https://arxiv.org/abs/2410.06521.

[99] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted

Intervention-MICCAI 2015, 2015: 234 – 241.

[100] CHALVATZAKI G, GKANATSIOS N, MARAGOS P, et al. Orientation attentive robotic grasp synthesis with augmented grasp map representation [EB/OL]. (2021 – 02 – 02) [2026 – 01 – 25]. https://arxiv.org/abs/2006.05123.

[101] LIU H P, YUAN Y, DENG Y H, et al. Active affordance exploration for robot grasping[C]//Intelligent Robotics and Applications, 2019: 426 – 438.

[102] SHAO Q Q, HU J. Combining RGB and points to predict grasping region for robotic bin-picking[EB/OL]. (2019 – 04 – 24) [2026 – 01 – 25]. https://arxiv.org/abs/1904.07394.

[103] SHAO Q Q, HU J, WANG W M, et al. Suction grasp region prediction using self-supervised learning for object picking in dense clutter[C]//Proceedings of 2019 IEEE 5th International Conference on Mechatronics System and Robots (ICMSR), 2019: 7 – 12.

[104] DENG Y H, GUO X F, WEI Y X, et al. Deep reinforcement learning for robotic pushing and picking in cluttered environment[C]//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 619 – 626.

[105] KASAEI H, KASAEI M. MVGrasp: real-time multi-view 3D object grasping in highly cluttered environments [J]. Robotics and Autonomous Systems, 2023, 160: 104313.

[106] BREYER M, CHUNG J J, OTT L, et al. Volumetric grasping network: real-time 6 DOF grasp detection in clutter[C]//Proceedings of the 2020 Conference on Robot Learning, 2021: 1602 – 1611.

[107] SONG P, LI P, DETRY R. Implicit grasp diffusion: bridging the gap between dense prediction and sampling-based grasping [C]//Proceedings of Machine Learning Research 270, 2024: 2948 – 2964.

[108] JAUHRI S, LUNAWAT I, CHALVATZAKI G. Learning any-view 6DoF robotic grasping in cluttered scenes via neural surface rendering[EB/OL]. (2024 – 05 – 29) [2026 – 01 – 25]. https://arxiv.org/abs/2306.07392.

[109] RASHID A, SHARMA S, KIM C M, et al. Language embedded radiance fields for zero-shot task-oriented grasping[C]//Proceedings of The 7th Conference on Robot Learning, 2023: 178 – 200.

[110] DAI Q Y, ZHU Y, GENG Y R, et al. GraspNeRF: multiview-based 6-DoF grasp detection for transparent and specular objects using generalizable NeRF[EB/OL]. (2023 – 03 – 15) [2026 – 01 – 25]. https://arxiv.org/abs/2210.06575.

[111] ZHENG Y H, CHEN X Y, ZHENG Y P, et al. GaussianGrasper: 3D language Gaussian splatting for open-vocabulary robotic grasping [J]. IEEE Robotics and Automation Letters, 2024, 9(9): 7827 – 7834.

[112] JI M, QIU R Z, ZOU X Y, et al. GraspSplats: efficient manipulation with 3D feature splatting[EB/OL]. (2024 – 09 – 03) [2026 – 01 – 25]. https://arxiv.org/abs/2409.02084.

[113] YU J Q, REN X L, GU Y C, et al. SparseGrasp: robotic grasping via 3D semantic Gaussian splatting from sparse multi-view RGB images[EB/OL]. (2024 – 12 – 03) [2026 – 01 – 25]. https://arxiv.org/abs/2412.02140.

[114] NGUYEN T, VU M N, HUANG B R, et al. Language-conditioned affordance-pose detection in 3D point clouds[C]// Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024: 3071 – 3078.

[115] XU K C, ZHAO S Q, ZHOU Z X, et al. A joint modeling of vision-language-action for target-oriented grasping in clutter[C]// Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023: 11597 – 11604.

[116] NGUYEN H H, VUONG A, NGUYEN A, et al. GraspMamba: a mamba-based language-driven grasp detection framework with hierarchical feature learning[C]// Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2025: 15808 – 15815.

[117] LU Y H, FAN Y X, DENG B X, et al. VL-Grasp: a 6-Dof interactive grasp policy for language-oriented objects in cluttered indoor scenes[C]//Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023: 976 – 983.

[118] TZIAFAS G, KASAEI H. Towards open-world grasping with large vision-language models[C]//Proceedings of the 8th Conference on Robot Learning, 2025: 3304 – 3332.

[119] SHI Y T, WEN D, CHEN G Q, et al. VISO-Grasp: vision-language informed spatial object-centric 6-DoF active view planning and grasping in clutter and invisibility[EB/OL]. (2025 – 08 – 06) [2026 – 01 – 25]. https://arxiv.org/abs/2503.12609.

[120] JIN S Y, XU J X, LEI Y T, et al. Reasoning Grasping via multimodal large language model[C]//Proceedings of the 8th Conference on Robot Learning, 2025: 3809 – 3827.

[121] JIAO R Y, FASOLI A, GIULIARI F, et al. Free-form language-based robotic reasoning and grasping[EB/OL]. (2025 – 07 – 28) [2026 – 01 – 25]. https://arxiv.org/abs/2503.13082.

[122] 禹鑫燚, 何威, 欧林林. 融合多模态感知的机器人抓取策略研究[J/OL]. 小型微型计算机系统. (2025 – 05 – 21) [2026 – 01 – 25]. https://link.cnki.net/urlid/21.1106.TP.20250521.0938.002.

YU X Y, HE W, OU L L. Research on robot grasping strategy integrating multimodal perception[J/OL]. Journal of Chinese Computer Systems. (2025 – 05 – 21) [2026 – 01 – 25]. https://link.cnki.net/urlid/21.1106.TP.20250521.0938.002.(in Chinese)

[123] TANG C, HUANG D H, DONG W L, et al. FoundationGrasp: generalizable task-oriented grasping with foundation models[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 12418 – 12435.

[124] LUO Z, YANG Y X, ZHANG Y F, et al. RoboReflect: a robotic reflective reasoning framework for grasping ambiguous-condition objects[EB/OL]. (2025 – 03 – 10) [2026 – 01 – 25]. https://arxiv.org/abs/2501.09307.

[125] YANG Y, YU H J, LOU X B, et al. Attribute-based robotic grasping with data-efficient adaptation [J]. IEEE Transactions on Robotics, 2024, 40: 1566 – 1579.

[126] DONG W L, HUANG D H, LIU J S, et al. RTAGrasp: learning task-oriented grasping from human videos via retrieval, transfer, and alignment[C]//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA), 2025: 1 – 7.

[127] SONG Y X, SUN P L, JIN P P, et al. Learning 6-DoF fine-grained grasp detection based on part affordance grounding[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 15200 – 15214.

[128] DESHPANDE A, DENG Y Q, RAY A, et al. GraspMolmo: generalizable task-oriented grasping via large-scale synthetic data generation[EB/OL]. (2025 – 09 – 12) [2026 – 01 – 25]. https://arxiv.org/abs/2505.13441.

[129] DENG S L, YAN M, WEI S L, et al. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data[EB/OL]. (2025 – 08 – 27) [2026 – 01 – 25]. https://arxiv.org/abs/2505.03233.

[130] ZHANG H R, BAI S H, ZHOU W Q, et al. VCoT-Grasp: grasp foundation models with visual chain-of-thought reasoning for language-driven grasp generation[EB/OL]. (2025 – 10 – 07) [2026 – 01 – 25]. https://arxiv.org/abs/2510.05827.

[131] Robot Learning Lab. Cornell grasping dataset[EB/OL]. (2017 – 01 – 09)[2026 – 01 – 25]. https://www.kaggle.com/datasets/oneoneliu/cornell-grasp.

[132] DEPIERRE A, DELLANDRÉA E, CHEN L M. Jacquard: a large scale dataset for robotic grasp detection[C]// Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 3511 – 3516.

[133] FANG H S, WANG C X, GOU M H, et al. GraspNet-1Billion: a large-scale benchmark for general object grasping[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 11441 – 11450.

[134] DEITKE M, SCHWENK D, SALVADOR J, et al. Objaverse: a universe of annotated 3D objects[C]// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 13142 – 13153.

[135] HUBER J, HELENON F, CONINX M, et al. Quality diversity under sparse interaction and sparse reward: application to grasping in robotics[J/OL]. Evolutionary Computation, 2025: 1 – 30 (2025 – 01 – 14) [2026 – 01 – 25]. https://pubmed.ncbi.nlm.nih.gov/39823378/.

[136] HUBER J, HÉLÉNON F, WATRELOT H, et al. Domain randomization for Sim2real transfer of automatically generated grasping datasets[C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024: 4112 – 4118.

[137] WANG X, CHEN G Y, QIAN G W, et al. Large-scale multi-modal pre-trained models: a comprehensive survey[J]. Machine Intelligence Research, 2023, 20(4): 447 – 482.

[138] BAIDOO-ANU D, ANSAH L O. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning[EB/OL]. [2026 – 01 – 25]. https://scale.stanford.edu/ai/repository/education-era-generative-artificial-intelligence-ai-understanding-potential-benefits.

[139] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789 – 1819.

[140] WANG L, YOON K J. Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3048 – 3068.

[141] CHOUDHARY T, MISHRA V, GOSWAMI A, et al. A comprehensive survey on model compression and acceleration[J]. Artificial Intelligence Review, 2020, 53(7): 5113 – 5155.

[142] HUANG B H, WANG Y X, YANG X Y, et al. 3D-ViTac: learning fine-grained manipulation with visuo-tactile sensing[EB/OL]. (2025 – 01 – 06) [2026 – 01 – 25]. https://arxiv.org/abs/2410.24091.

[143] TU Y Y, JIANG J N, LI S, et al. PoseFusion: robust object-in-hand pose estimation with SelectLSTM[C]// Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023: 6839 – 6846.

[144] GRAVES A. Long short-term memory[M]//GRAVES A. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012: 37 – 45.

[145] SHAO Y M, XIAO C X. Bimanual grasp synthesis for dexterous robot hands[J]. IEEE Robotics and Automation Letters, 2024, 9(12): 11377 – 11384.

[146] 吴超达, 王珍, 刘雪飞, 等. 基于ArUco码的透明物体识别与抓取系统设计[J]. 现代电子技术, 2025, 48(22): 172 – 178.
WU C D, WANG Z, LIU X F, et al. Design of transparent object recognition and grasping system based on ArUco marker[J]. Modern Electronics Technique, 2025, 48(22): 172 – 178. (in Chinese)

[147] ZHAOLE S, ZHU J H, FISHER R B. DexDLO: learning goal-conditioned dexterous policy for dynamic manipulation of deformable linear objects[C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024: 16009 – 16015.

[148] 赵洲, 耿明强, 何秋实, 等. 基于“形态 – 感知 – 动作”仿生机理的机器人自适应力控抓取方法[J/OL]. 自动化学报. (2026 – 02 – 03) [2026 – 01 – 25]. https://doi.org/10.16383/j.aas.c250453.
ZHAO Z, GENG M Q, HE Q S, et al. Robot adaptive force control grasping method based on bionic mechanism of "shape-perception-action"[J/OL]. Acta Automatica Sinica. (2026 – 02 – 03) [2026 – 01 – 25]. https://doi.org/10.16383/j.aas.c250453. (in Chinese)

[149] 高茂源, 王廷, 王好臣, 等. 堆叠场景下零件视觉识别与定位抓取系统研究[J]. 机床与液压, 2025, 53(21): 46 – 52.
GAO M Y, WANG T, WANG H C, et al. Research on visual recognition and positioning gripping system of parts in stacked scenes[J]. Machine Tool & Hydraulics, 2025, 53(21): 46 – 52. (in Chinese)

[150] HU Z, ZHENG Y, PAN J. Grasping living objects with adversarial behaviors using inverse reinforcement learning[J]. IEEE Transactions on Robotics, 2023, 39(2): 1151 – 1163.

[151] CHEN Z Y, YAN Q Y, CHEN Y P, et al. ClutterDexGrasp: a sim-to-real system for general dexterous grasping in cluttered scenes[EB/OL]. (2025 – 09 – 04)[2026 – 01 – 25]. https://arxiv.org/abs/2506.14317.

[152] 贾军营, 艾良, 卢鑫. 融合跨模态特征与颜色引导的遮挡目标6D位姿估计算法[J/OL]. 计算机应用研究. (2025 – 12 – 16)[2026 – 01 – 25]. https://doi.org/10.19734/j.issn.1001 – 3695.2025.07.0288.
JIA J Y, AI L, LU X. 6D pose estimation algorithm for occluded objects fusion of cross-modal features and color guidance[J/OL]. Application Research of Computers. (2025 – 12 – 16)[2026 – 01 – 25]. https://doi.org/10.19734/j.issn.1001 – 3695.2025.07.0288. (in Chinese)

[153] HE S C, SHANGGUAN Z Y, WANG K N, et al. Sequential multi-object grasping with one dexterous hand[EB/OL]. (2025 – 08 – 02)[2026 – 01 – 25]. https://arxiv.org/abs/2503.09078.

[154] LI Y Y, LIU B, GENG Y R, et al. Grasp multiple objects with one hand[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4027 – 4034.

[155] CHEN T Z, SUN Y. Multi-object grasping—experience forest for robotic finger movement strategies[J]. IEEE Robotics and Automation Letters, 2024, 9(6): 5222 – 5229.

[156] ZHANG J Z, GIREESH N, WANG J L, et al. GAMMA: graspability-aware mobile manipulation policy learning based on online grasping pose fusion[C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024: 1399 – 1405.

[157] YAN S X, ZHANG Z Y, HAN M Z, et al. M2 diffuser: diffusion-based trajectory optimization for mobile manipulation in 3D scenes[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2025 – 03 – 21)[2026 – 01 – 25]. https://ieeexplore.ieee.org/document/10937276.

[158] POHL C, REISTER F, PELLER-KONRAD F, et al. MAkEable: memory-centered and affordance-based task execution framework for transferable mobile manipulation skills[C]//Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024: 3674 – 3681.

[159] QIU R Z, HU Y F, SONG Y C, et al. Learning generalizable feature fields for mobile manipulation[EB/OL]. (2024 – 11 – 26)[2026 – 01 – 25]. https://arxiv.org/abs/2403.07563.

[160] JAUHRI S, LUETH S, CHALVATZAKI G. Active-perceptive motion generation for mobile manipulation[C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024: 1413 – 1419.

[161] 罗国庆, 袁庆霓, 曲鹏举, 等. 结构化环境中UR3机械臂对于移动物体的抓取研究[J]. 计算机工程与应用, 2025, 61(16): 106 – 115.
LUO G Q, YUAN Q N, QU P J, et al. Grasp of moving objects by UR3 manipulator in structured environment[J]. Computer Engineering and Applications, 2025, 61(16): 106 – 115. (in Chinese)