

白嫖5个创新点！具身智能最新综述：全面调研！

计算机视觉工坊 2024年08月24日 00:00 江苏

点击下方**卡片**，关注「**3D视觉工坊**」公众号
选择**星标**，干货第一时间送达



计算机视觉工坊
专注于计算机视觉、SLAM、三维重建、自动驾驶、具身智能、Mamba、目标检测、语义...
122篇原创内容

公众号

来源：计算机视觉工坊

添加小助理：cv3d008，备注：方向+学校/公司+昵称，拉你入群。文末附3D视觉行业细分群。

扫描下方二维码，加入「3D视觉从入门到精通」知识星球，星球内凝聚了众多3D视觉实战问题，以及各个模块的学习资料：[近20门秘制视频课程](#)、[最新顶会论文](#)、[计算机视觉书籍](#)、[优质3D视觉算法源码](#)等。想要入门3D视觉、做项目、搞科研，欢迎扫码加入！



3D 视觉从入门到精通

星主：小凡

5800+

成员数量

9500+

内容数量

2329

运营天数

国内最大的 3D 视觉学习平台，加入后，你将可以学习：

- ① 独家秘制课程
- ② 项目对接...

现价：¥11.11

微信扫码加入星球



公众号 · 3D视觉工坊

0. 论文信息

标题：A Survey of Embodied Learning for Object-Centric Robotic Manipulation

作者：Ying Zheng, Lei Yao, Yuejiao Su, Yi Zhang, Yi Wang, Sicheng Zhao, Yiyi Zhang, Lap-Pui Chau

原文链接：<https://arxiv.org/abs/2408.11537>

1. 摘要

以对象为中心的机器人操作中的具身学习是具身人工智能中一个快速发展且充满挑战的领域。它对推动下一代智能机器人至关重要,并且最近引起了广泛关注。与数据驱动的机器学习方法不同,具身学习侧重于机器人通过与环境的物理交互和感知反馈来学习,这使其特别适用于机器人操作。在本文中,我们全面综述了该领域的最新进展,并将现有工作分为三个主要分支:1) 具身感知学习,旨在通过各种数据表示来预测对象的姿态和可负担性;2) 具身策略学习,侧重于使用强化学习和模仿学习等方法生成最优的机器人决策;3) 具身任务导向学习,旨在根据对象抓取和操作中不同任务的特点来优化机器人的性能。此外,我们还概述并讨论了公共数据集、评估指标、代表性应用、当前挑战以及潜在的未来研究方向。与本文调查相关的项目已建立在https://github.com/RayYoh/OCRM_survey上。

2. 引言

在过去十年中,以深度学习为中心的机器学习研究取得了显著进展,彻底改变了包括计算机视觉和自然语言处理在内的各个领域。传统机器学习方法依赖于使用预先构建的数据集进行模式识别和预测来训练模型。然而,这些数据集主要来源于静态资源,如图像、视频和文本,这可能会限制其适用性和有效性。

具身学习作为具身人工智能的基石,与传统机器学习形成鲜明对比。它强调通过物理交互和实践经验来获取知识。其数据源包括广泛的范围,如感官输入、身体动作和即时的环境反馈。这种学习机制高度动态,通过实时交互和反馈循环不断优化行为和操作策略。在机器人技术中,具身学习至关重要,因为它为机器人提供了增强的环境适应能力,使它们能够处理不断变化的条件并承担更复杂和精细的任务。

尽管已经提出了大量的具身学习方法,但本综述主要关注以对象为中心的机器人操作任务。该任务的输入是从传感器收集的数据,输出是机器人执行操作任务的操作策略和控制信号。其目标是使机器人能够高效、自主地执行各种以对象为中心的操作任务,同时提高其在不同环境和任务中的通用性和灵活性。这项任务极具挑战性,因为涉及对象的多样性和操作任务的复杂性、环境的不确定性和复杂性,以及现实应用中的噪声、遮挡和实时性限制等挑战。

图1 (a) 展示了一个典型的机器人操作系统。它包含一个配备有摄像头等传感器和抓手等末端执行器的机械臂,使其能够操作各种对象。该系统的智能围绕三个关键方面展开,对应于图1 (b) 中所示的三种具身学习方法。1) 先进的感知能力,涉及利用不同传感器捕获的数

据来理解目标对象和外部环境；2) 精确的策略生成，涉及分析感知到的信息以做出最优决策；3) 任务导向，确保系统能够通过优化执行过程以适应特定任务，从而实现最大效率。

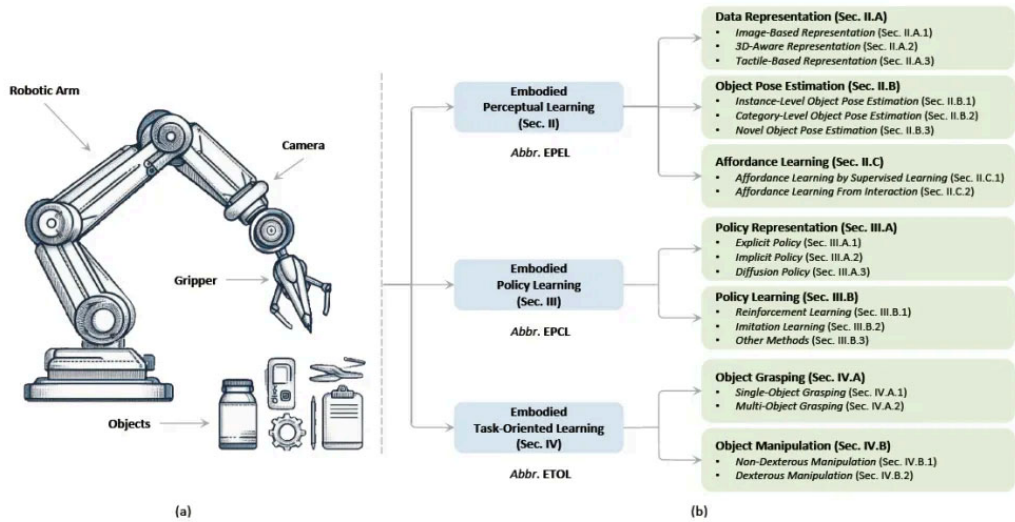


Fig. 1. An illustration of robotic manipulation system (left) and the typology of embodied learning methods for object-centric robotic manipulation (right). EPEL takes the data obtained from sensors such as cameras as input, enhancing the understanding of objects and the environment through interaction. It serves as the basis for EPCL and ETOL. EPCL utilizes the perceptual information provided by EPEL to formulate action strategies for robotic arms and end-effectors like grippers, thereby providing specific operational capabilities for ETOL. ETOL integrates EPEL and EPCL, learning to perform diverse tasks based on the characteristics of different objects. These three closely related learning processes work together to enable robots to accomplish complex tasks.

近年来，围绕上述三个关键方面进行了大量研究，特别是随着大型语言模型（LLMs）、神经辐射场（NeRFs）、扩散模型（Diffusion Models）和3D高斯溅射（3D Gaussian Splatting）的兴起，催生了许多创新解决方案。然而，目前还缺乏一个全面综述这一快速发展领域最新研究的综合性调查报告。这促使我们撰写本综述，系统地回顾前沿进展，总结遇到的挑战，并展望未来的研究方向。

A. 与近期综述的比较

过去几年中，涌现了许多关于具身人工智能和机器人学习的综述文章，涵盖了导航、规划、抓取和操作等不同领域。表I总结了该领域的一些近期相关综述。Cong等人（2021）的综述论文与我们的工作最为相关。他们的研究重点是基于3D视觉的机器人操作，主要回顾了截至2021年的3D视觉感知研究。相比之下，我们的工作不仅限于3D视觉感知方法，还系统地总结和分类了基于图像、3D感知技术和触觉感知的表示方法。此外，我们还对机器人操作的关键方面，如策略和任务导向学习，进行了全面介绍。值得注意的是，我们的综述涵盖了2021年之后发表的大量最新研究成果，提供了更前沿和全面的视角。

TABLE I
SUMMARY OF RECENT SURVEYS RELATED TO EMBODIED AI AND ROBOT
LEARNING. RM: ROBOTIC MANIPULATION; RG: ROBOTIC GRASPING;
RL: REINFORCEMENT LEARNING.

Author	Year	Short Description
Huang [16]	2016	Datasets of RM
Yamanobe [17]	2017	Affordance in RM
Jin [18]	2018	Robot manipulator control using neural networks
Fang [19]	2019	Imitation learning for RM
Billard [20]	2019	Trends and challenges in RM
Kleeberger [21]	2020	Machine learning for vision-based RG
Du [13]	2020	Vision-based RG
Kroemer [22]	2020	Machine learning for RM
Cong [15]	2021	3D vision-based RM
Zhu [23]	2021	Deep learning for embodied visual navigation
Cui [24]	2021	Adaptability of learned RM
Zhu [25]	2021	RM of deformable objects
Mohammed [26]	2022	RL-based RM in cluttered environments
Suomalainen [27]	2022	RM in contact
Duan [28]	2022	Simulators for embodied AI
Francis [29]	2022	Embodied vision-language planning
Zhang [30]	2022	Traditional and recent methods for RG
Xie [31]	2022	Learning-based RG
Gervet [11]	2022	Real-world empirical study for robot navigation
Han [14]	2023	RL for RM
Tian [32]	2023	RG for unknown objects
Newbury [33]	2023	Deep learning approaches to grasp synthesis
Guo [12]	2023	Task and motion planning for robotics
Zare [34]	2023	Imitation learning
Xiao [35]	2023	Foundation models for robot learning
Weinberg [36]	2024	Learning approaches for in-hand RM
Chen [37]	2024	Generative models for offline policy learning
Ma [38]	2024	Vision-language-action models for embodied AI
Xu [39]	2024	Foundation models for robot planning and control
Ours	2024	Embodied learning methods for object-centric RM

B. 文本组织

本文全面综述了以对象为中心的机器人操作中的具身学习方法，涵盖三个主要领域和七个子方向。三个领域分别是具身感知学习（第二节）、具身策略学习（第三节）和具身任务导向学习（第四节）。七个子方向包括数据表示（第二节A）、对象姿态估计（第二节B）、可负担性学习（第二节C）、策略表示（第三节A）、策略学习（第三节B）、对象抓取（第四节A）和对象操作（第四节B）。我们还广泛涵盖了该领域常用的数据集和评估指标（第五节），以及几个代表性应用（第六节）。此外，我们深入探讨了主要挑战，并提供了对未来研究方向的见解（第七节）。

TABLE II
SUMMARY OF EMBODIED POLICY LEARNING. RL: REINFORCEMENT LEARNING; IL: IMITATION LEARNING.

Task	Type	Subfields & references
Policy Representation	Explicit Policy	Deterministic policy [118], Stochastic policy [119]
	Implicit Policy	EBMs [120], Implicit behavioral cloning [121], IDAC [122], EBIL [123]
	Diffusion Policy	Diffusion Policy [124], Decision Diffuser [125], Diffusion-QL [126], HDP [127], UniDexFPM [128], BESO [129]
Policy Learning	Incorporating language instructions: MDT [130], Lan-o3dp [131]	
	RL	ViSkill [132], RMA ² [119], SAM-RL [133], Offline RL [134], [135], Demonstration-guided RL [136]
		Rewards function learning: Text2reward [137] and EUREKA [138]
		DMPs [139], DAgger [140], SpawnNet [141], ACT [142]
	IL	Scaling up demonstration data: MimicGen [143], Bridge Data [144], Open X-embodiment [145]
		Learning from human videos: Vid2Robot [146], Ag2Manip [147], MPI [148]
		Equivariant models: NDFs [149], L-NDF [150], EDFs [151], EDGI [152], Diffusion-EDFs [153], SE(3)-DiffusionFields [154]
Other Methods	Combination of RL & IL: UniDexGrasp [155], UniDexGrasp++ [156]	
	LLM- or VLM-driven: VILA [157], Grounding-RL [158], OpenVLA [159], 3D-VLA [160]	

3. 潜在研究方向

在过去的几年中，以对象为中心的机器人操作任务的具身学习方法研究显著增加，推动了该领域的快速发展。然而，当前技术仍面临一些极具挑战性的问题。进一步探索这些问题对于促进智能机器人在各个领域的广泛应用至关重要。本节将讨论几个挑战和潜在的未来研究方向。

A. 从模拟到现实的泛化

收集现实世界中的机器人操作数据是困难的，因此创建大规模数据集面临挑战。为了解决这个问题，当前研究主要集中在在模拟环境中训练模型，这些环境提供了安全、可控且成本效益高的学习场景，并能够生成几乎无限量的模拟训练数据。然而，现实环境往往存在模拟环境无法准确复制的意外挑战和变化。这种差异可能会显著降低在模拟环境中训练的模型在现实世界中的性能。具体来说，虚拟世界与现实世界之间的差距源于多种因素，如感知差距、控制器不准确和模拟偏差。近期研究已经关注于通过使用领域随机化、物理约束正则化和迭代自训练等方法来缩小这一差距。对这一问题的进一步研究将有助于提升机器人操作方法对现实环境的适应性和在实际场景中的表现。

B. 多模态具身大语言模型（LLMs）

人类拥有丰富的感知能力，如视觉、听觉和触觉，这些能力帮助他们收集周围环境的详细信息。此外，人类还能利用学习到的经验来执行各种任务。这种多功能性也是通用智能机器人的最终目标。为了实现这一目标，机器人必须配备多个传感器来感知环境并收集多模态数据。此外，机器人还必须快速学习和适应新环境和新任务以执行有效操作。然而，这对智能机器人来说是一个重大挑战。

近期研究已经关注于使用多模态LLMs来增强机器人的感知、推理和动作生成能力。例如，Xu等人介绍了一种调整推理的方法，该方法利用LLMs的广泛先验知识为机器人抓取生成准确的数值输出。Huang等人将可负担性和物理概念整合到LLMs中，超越了常规的图像和文

本模态，从而在机器人操作中取得了更好的性能。这些工作推动了多模态具身LLMs的发展，但总体而言，该领域仍处于起步阶段，需要进一步广泛和深入的研究。

C. 人机协作

智能机器人有潜力彻底改变制造业、医疗保健和服务等行业。为了充分发挥这一潜力，人机协作至关重要。通过协同工作，机器人可以辅助人类，提高效率并减少人类的工作量和安全风险。同时，人类可以指导和监控机器人操作以提高准确性。然而，实现完美的人机协作存在沟通和协调障碍、过度依赖和安全问题等挑战。

研究界已经在解决人机协作挑战方面取得了一些进展。例如，Jin等人提出了一种基于深度强化学习的两级分层控制框架，以建立最优的人机合作策略。Wang等人介绍了一种名为Co-GAIL的策略训练方法，该方法基于人机交互演示和交互式学习过程中的协同优化。然而，这些方法通常在模拟环境中实施或只能执行有限数量的任务，因此不适合实际应用。未来，人机协作将继续成为重要的研究领域，需要不断探索以提高效率和安全性。

D. 模型压缩和机器人加速

在嵌入式系统、移动设备和边缘计算等应用中，具有具身智能系统的机器人通常具有有限的计算资源。这使得优化和压缩深度模型以满足存储空间、实时性和准确性的要求变得至关重要。虽然基于LLMs的方法在具身AI方面取得了显著进展，但也导致了计算资源需求的增加，这对在计算能力有限的设备上实施构成了挑战。

因此，未来的模型压缩研究有望促进智能机器人的实际应用。在现实应用中，长时间的等待往往导致用户体验不佳。因此，期望机器人能够快速完成任务。然而，许多当前的主流模型操作频率较低。例如，Google的RT-2模型根据使用的VLMs的参数规模，其决策频率在1-5 Hz之间，表明在实际应用之前仍存在巨大差距。最近，人形机器人Figure 014能够以200 Hz的频率生成动作指令，这得益于OpenAI的LLMs和高效的端到端网络架构。这一成就为未来关于机器人加速的研究带来了更大的乐观情绪。

E. 模型可解释性和应用安全性

基于深度学习的方法通常被称为“黑箱”，因为难以直观地理解其决策过程。对于基于深度学习的智能机器人来说，这种黑箱特性可能导致用户的怀疑和不信任。特别是在机器人与人类紧密互动的环境中，缺乏透明度还可能引发对个人安全的担忧。因此，研究具身学习方法的可解释性至关重要，这有助于人们理解模型的决策过程并增加对机器人的信任。


除了模型可解释性之外，还需要从其他角度保证智能机器人的安全性，包括实施更可靠的在线学习和控制技术以防止机器人运动可能造成的潜在伤害。此外，还需要采用对抗性训练来保护机器人免受攻击，并设计稳健的安全监控方法来检测可能的安全风险。这些领域的进一步研究有望提高机器人在实际应用中的安全性和可靠性。

4. 总结

综上所述，本文全面综述了以对象为中心的机器人操作中具身学习的现有方法。我们首先介绍了该任务的概念及其基本组成部分，然后将其与相关综述文章进行了比较。接下来，我们系统地介绍了三个类别的主要工作。然后，我们探讨了常用的数据集和评估指标，并突出了一些代表性应用。

对更多实验结果和文章细节感兴趣的读者，可以阅读一下论文原文~

这里给大家推荐一门我们最新的课程《具身智能，从入门到实战系统教程！》：

3D视觉工坊
www.3dcver.com

国内首个面向具身智能方向的 理论与实战课程

公众号 · 3D视觉工坊

• 课程特色 •

本课程从学术研究和实际应用两方面，带你从零入门具身智能的原理学习、论文阅读、代码梳理等内容。

课程由具身智能领域的资深专家主讲，他们先后担任研究所、国企、大厂具身智能负责人，拥有丰富的理论知识和实践经验。

公众号 · 3D视觉工坊

• 学后收获 •

入门具身智能领域：掌握具身智能的基础知识和核心技术

科研能力提升：从论文阅读、代码复现到发现问题、提出改进的全面科研入门

代码精通：逐行注释具身智能相关代码，掌握每个实现细节，动手复现并作改进

顶会研究：通过系列专题分析，发掘顶会idea并作出SOTA工作

收获高薪offer：进入具身智能风口领域，收获高薪offer

公众号 · 3D视觉工坊

• 主讲介绍 •

木木老师

主要负责智能方向，拥有985本硕学历，并具备11年的机器人行业经验。她先后在研究所、国企及大型企业担任具身智能负责人，具有丰富的ACT、RT2、3D_diffuser_act等前沿算法的复现和改进经验。此外，木木老师在面试与被面试方面也有着丰富经验，能够提供实用的职业建议。

宇哥

主要负责本体方向，同样拥有985本硕学历，具有10年机器人相关项目的从业经验。他长期从事机器人与人工智能相关研究，精通机器人一体化关节控制、运动控制及具身智能相关领域。宇哥的丰富实践经验和深入的技术理解，将为学员们提供全面的知识和指导。

• 课程大纲 •



3D视觉工坊
www.3dcver.com

第一章：预备知识与概述

第1节 具身智能简介

- 具身智能定义
- 行业发展趋势

第2节 预备知识与技能

- Docker基础
- Python编程基础
- PyTorch深度学习框架基础

第3节 课程规划

- 具身智能课程规划

第二章：机器人技术基础

第1节 机器人基本硬件组成

- 常见传感器介绍
- 常见执行器介绍
- 控制系统

第2节 坐标与位姿变换

- 坐标系介绍
- 齐次变换矩阵
- 正逆运动学

第3节 轨迹规划入门

- 轨迹生成
- 路径规划
- 实时轨迹控制

第三章：RT详解

第1节 Transformer原理详解

- Transformer算法架构
- 自注意力机制与位置编码
- 代码实战理解

第2节 RT1原理详解

- RT1模型概览
- 机器人如何使用Transformer
- RT1的优缺点

第3节 RT2原理详解

- RT2基于RT1的改进之处
- RT2的优缺点

第四章：RT1实战

第1节 数据采集方案

- 数据格式分析
- 数据预处理
- 数据质量分析

第2节 实战

- RT1项目框架讲解
- 代码实战理解

第5章 Aloha ACT 算法详解

第1节 CVAE算法详解

- 变分自编码器 (VAE)
- 条件变分自编码器 (CVAE)
- 代码实战理解

第2节 ACT算法详解

- ACT模型概览
- 机器人如何使用CVAE
- ACT的优缺点

第6章 Aloha ACT实战

第1节 数据采集方案

- 数据格式分析
- 数据预处理
- 数据质量分析

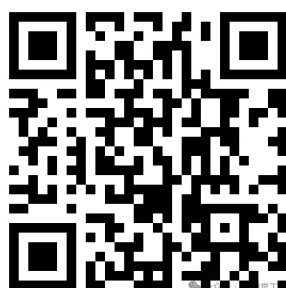
第2节 实战

- ACT项目框架讲解
- 代码实战理解

公众号 · 3D视觉工坊

课程答疑

本课程答疑主要在本课程对应的鹅圈子中答疑，学员学习过程中，有任何问题，可以随时在鹅圈子中提问。



▲长按购买课程



▲长按添加小助理微信
cv3d007，咨询更多

