

中图法分类号: TP391.9 文献标识码: A 文章编号: 1006-8961(2024)06-1575-32

论文引用格式: Du T, Hu R Z, Liu L B, Yi L and Zhao H. 2024. Research progress in human-like indoor scene interaction. Journal of Image and Graphics, 29(06):1575-1606(杜韬, 胡瑞珍, 刘利斌, 弋力, 赵昊. 2024. 室内场景拟人交互研究进展. 中国图象图形学报, 29(06):1575-1606)[DOI:10.11834/jig.240004]

室内场景拟人交互研究进展

杜韬^{1,2,3}, 胡瑞珍^{4*}, 刘利斌⁵, 弋力^{1,2,3}, 赵昊⁶

1. 清华大学交叉信息研究院, 北京 100084; 2. 上海人工智能实验室, 上海 200232; 3. 上海期智研究院, 上海 200232;
4. 深圳大学计算机与软件学院, 深圳 518061; 5. 北京大学智能学院, 北京 100871;
6. 清华大学智能产业研究院, 北京 100084

摘要: 人类智能是在与环境交互中进化的, 因而如何实现智能体与环境的自主交互是推进智能演化的关键。环境自主交互是一项涉及计算机图形学、计算机视觉和机器人等多个学科领域的研究课题, 引起广泛的关注和探究, 学术界已围绕这一热点研究问题从不同视角和技术维度开展了一系列研究工作。本文着眼于室内场景拟人交互, 全面梳理数字人与机器人在室内环境下学习完成特定交互任务过程中需要涉及的仿真交互平台、场景交互数据和交互生成算法3方面基本要素的研究进展。在仿真交互环境搭建方面, 本文梳理了仿真环境涉及的仿真技术和研究进展, 并对代表性的拟人交互仿真平台进行了介绍; 在场景交互数据构建方面, 本文从场景交互感知数据集、场景交互运动数据集以及交互数据规模的高效扩充3方面对国内外研究现状进行了详细介绍; 在拟人交互感知与生成方面, 本文介绍了以交互为导向的场景可供性分析的相关工作, 并以交互生成为线索, 分别梳理了数字人一场景交互生成、机器人一场景交互生成的相关工作。基于对国内外相关工作的梳理和讨论, 最后从交互仿真、交互数据、交互感知和交互生成4个方面, 总结了该领域目前仍面临的挑战, 并对未来的发展趋势进行了展望。

关键词: 环境交互; 交互仿真; 交互数据; 交互感知; 交互生成

Research progress in human-like indoor scene interaction

Du Tao^{1,2,3}, Hu Ruizhen^{4*}, Liu Libin⁵, Yi Li^{1,2,3}, Zhao Hao⁶

1. Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China;
2. Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China; 3. Shanghai Qi Zhi Institute, Shanghai 200232, China;
4. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518061, China;
5. School of Intelligence Science and Technology, Peking University, Beijing 100871, China;
6. Institute for AI Industry Research, Tsinghua University, Beijing 100084, China

Abstract: Human intelligence evolves through interactions with the environment, which makes autonomous interaction between intelligent agents and the environment a key factor in advancing intelligence. Autonomous interaction with the environment is a research topic that involves multiple disciplines, such as computer graphics, computer vision, and robotics, and has attracted significant attention and exploration in recent years. In this study, we focus on human-like interaction in indoor environment and comprehensively review the research progress in the fundamental components including

收稿日期: 2024-01-08; 修回日期: 2024-04-01; 预印本日期: 2024-04-08

所有作者为共同一作, * 通信作者: 胡瑞珍 ruizhen.hu@szu.edu.cn

基金项目: 国家自然科学基金项目(62322207)

Supported by: National Natural Science Foundation of China (62322207)

simulation interaction platforms, scene interaction data, and interaction generation algorithms for digital humans and robots. Regarding simulation interaction platforms, we comprehensively review representative simulation methods for virtual humans, objects, and human-object interaction. Specifically, we cover critical algorithms for articulated rigid-body simulation, deformable-body and cloth simulation, fluid simulation, contact and collision, and multi-body multi-physics coupling. In addition, we introduce several popular simulation platforms that are readily available for practitioners in the graphics, robotics, and machine learning communities. We classify these popular simulation platforms into two main categories: simulators focusing on single-physics systems and those supporting multi-physics systems. We review typical simulation platforms in both categories and discuss their advantages in human-like indoor-scene interaction. Finally, we briefly discuss several emerging trends in the physical simulation community that inspire promising future directions: developing a full-featured simulator for multi-physics multi-body physical systems, equipping modern simulation platforms with differentiability, and combining physics principles with insights from learning techniques. Regarding scene interaction data, we provide an in-depth review of the latest developments and trends in datasets that support the understanding and generation of human-scene interactions. We focus on the need for agents to perceive scenes with a focus on interaction, assimilate interactive information, and recognize human interaction patterns to improve simulation and movement generation. Our review spans three areas: perception datasets for human-scene interaction, datasets for interaction motion, and methods for scaling data efficiently. Perception datasets facilitate a deeper understanding of 3D scenes, which highlights geometry, structure, functionality, and motion. They offer resources for interaction affordances, grasping poses, interactive components, and object positioning. Motion datasets, which are essential for crafting interactions, delve into interaction movement analysis, including motion segmentation, tracking, dynamic reconstruction, action recognition, and prediction. The fidelity and breadth of these datasets are vital for creating lifelike interactions. We also discuss scaling challenges, with the limitations of manual annotation and specialized hardware, and explore current solutions like cost-effective capture systems, dataset integration, and data augmentation to enable the generation of extensive interactive models for advancing human-scene interaction research. For robot-scene interaction, this study emphasizes the importance of affordance, that is, the potential action possibilities that objects or environments can provide to users. It discusses approaches for detecting and analyzing affordance at different granularities, as well as affordance modeling techniques that combine multi-source and multimodal data. In the aspect of digital human-scene interaction, this study provides a detailed introduction to the simulation and generation methods of human motion, especially focusing on technologies based on deep learning and generative models in recent years. Building on this foundation, the study reviews ways to represent a scene and recent successful approaches that achieve high-quality human-scene interaction simulation. Finally, we discuss the challenges and future development trends in this field.

Key words: environment interaction; interaction simulation; interaction data; interaction perception; interaction generation

0 引言

近年人工智能研究取得了巨大进展,尤其在计算智能和认知智能领域。例如,机器学习使得计算机能够从大量数据中学习和预测;自然语言处理使得计算机能够理解和生成人类语言;计算机视觉使得计算机能够识别和理解图像和视频等。然而,这些都只是人类智能的一部分。人的认知和智力活动并不仅仅是大脑孤立的计算,它还涉及大脑、身体和环境的相互作用。例如,人类可以通过眼睛看到环境,通过耳朵听到声音,通过身体感受到温度和触

感,然后大脑会根据这些信息做出决策,引导与环境的交互,并在与环境的交互中进化反馈学习,因此,如何实现智能体与环境的自主交互是推进智能演化的关键。如何让智能体理解并适应环境、在复杂环境中做出正确的决策,是一项涉及计算机图形学、计算机视觉和机器人等多个学科领域的研究课题,近些年引起了广泛的关注和探索,学术界已围绕这一热点研究问题从不同视角和技术维度开展了一系列研究工作。

智能体通过与其所处环境的互动来学习和进化,以便更好地完成其任务,这种互动可以发生在各种环境中,无人驾驶汽车和家庭服务机器人分别是

在室外和室内场景中的典型应用,无人驾驶汽车需要与其所处的环境进行持续的互动,以便学习如何在各种交通情况下安全地行驶;而家庭服务机器人则需要与家庭环境进行交互,以便学习如何完成各种复杂家务任务。由于室外场景与室内场景在诸多方面存在显著差异,包括环境中主要组成要素、引起复杂环境动态变化的因素来源、所需执行的任务类别、对人机交互方式的倾向性、操作执行的安全性需求等,而这些差异对智能体与环境交互带来了不同的挑战和要求,也使得相应的技术研究有不同的侧重点。本文着眼于室内场景拟人交互,旨在全面梳理数字人与机器人在室内环境下学习完成特定交互任务过程中需要涉及的平台、数据和算法3方面基本要素的研究进展。

仿真环境在拟人交互研究中扮演着重要角色。一方面,利用仿真环境可以生成大规模训练数据,避免了真实环境中采集数据的人力成本和安全隐患;另一方面,高质量仿真环境可以作为拟人交互算法的评测平台,降低算法开发和部署的测试开销和安全风险。此外,无论是对于场景的交互感知还是对于场景交互运动的理解都离不开数据集支持,为了应对开放环境、多样任务的需求,如何在保证数据质量的前提下高效采集并扩充相关数据的规模也成为了重要的研究问题。

1 仿真交互环境搭建

本节介绍拟人交互仿真环境的相关内容。首先,对拟人交互仿真环境涉及的仿真技术和研究进展进行梳理,包括如何仿真交互主体(数字人或机器人)、如何仿真交互物体对象以及如何仿真二者的交互。然而,从零开始实现上述算法需要对计算物理、数值分析和并行计算等众多方向有深入了解,门槛较高。幸运的是,经过多年的发展,一些科研机构和企业为整个社区集成并维护了众多高质量的拟人交互仿真平台,提供了上述仿真算法的高质量工程代码实现,为拟人交互的研究提供了极大便利。因此,简要介绍近期涌现的几种有代表性的拟人交互仿真平台,包括几种面向特定物理场景(刚体、软体和流体)的仿真平台和一些面向多种物理系统的通用仿真平台。

1.1 拟人交互仿真技术概述

1.1.1 数字人和机器人仿真

传统上,数字人和机器人在仿真环境中被视为铰接刚体(articulated rigid body)进行建模和解算。铰接刚体由一系列刚体组成,刚体和刚体之间通过关节来连接以建立运动约束。例如,早期物理引擎中使用的“布娃娃物理系统(Gästrin, 2004)”将数字人视为多个刚体以及刚体间约束的集合。由于早期物理引擎求解器能力有限,这种数字人的关节往往会像布娃娃一样随意弯折,“布娃娃物理系统”因此得名。

当前铰接刚体的主流仿真方法分为两类:基于最大坐标(maximal coordinates)和基于简化坐标(reduced coordinates,也称为广义坐标(generalized coordinates))的仿真方法。在最大坐标中,铰接刚体系统的每一个刚体享有完整的平移和旋转自由度,关于每个刚体的平移和旋转自由度列出动力学方程,并将关节以约束方程的形式加入到动力学模型中,在求解时利用Lagrange乘子法或其变种进行处理(Nocedal和Wright, 1999)。最大坐标法可以直观地操作每个刚体上的自由度,且可以通过约束方程灵活处理刚体之间不同的拓扑连接关系。然而,由于数值方法带来的离散化误差,最大坐标法往往需要在解算过程中不断将全体自由度投影回约束方程的可行域内(Bargteil等, 2020),需要额外的数值方法去修正(Bender等, 2014; Weinstein等, 2006)。

与最大坐标法赋予每个刚体自由度的做法相比,简化坐标法采取了不同的策略。简化坐标法考虑系统中的独立自由度,这些自由度通常以关节自由度的形式存在。基于分析力学可以列出针对这些独立自由度的动力学方程(Lanczos, 2012),并利用经典的Featherstone算法快速求解(Featherstone, 1984)。与最大坐标法相比,简化坐标法的推导和实现一般更加复杂,在处理铰接刚体环等特定拓扑结构时也更加困难。但是简化坐标法从源头上自动满足了关节带来的运动学约束,避免了最大坐标法为了满足约束所要求的数值修正,因而具备更高的精确度。

上述两种方法在目前的主流物理引擎中均有广泛应用。例如,NVIDIA公司推出的PhysX物理引擎同时支持最大坐标和简化坐标,并指出最大坐标法牺牲了一定的准确度,适合面向游戏应用;简化坐标

具备更高的精确度,适用于机器人领域的应用。

尽管铰接刚体模型应用广泛,相关仿真技术经过多年的发展也趋于成熟,数字人和机器人仿真仍然有部分技术需求是铰接刚体模型难以胜任的,例如对肌肉等柔性材料的建模和解算。一些仿真研究提出了多种将铰接刚体与软体模型耦合的仿真方法,为仿真更加生物真实的数字人或仿生机器人提供了可行的技术方案。这一方向的早期工作(Kim和Pollard, 2011; Liu等, 2013; Shinar等, 2008; Sifakis等, 2007)针对前述的最大坐标或简化坐标表示下的铰接刚体各自提出了经典的耦合问题解法。在这些方法的基础上, RedMax(Wang等, 2019b)提供了一种更加通用的耦合方法,可以灵活适配在最大坐标或简化坐标表示下的铰接刚体与软体耦合问题,是当前本方向的最好解法之一。

上述方法面向一般铰接刚体与软体耦合场景。在仿真数字人这一特定问题上,一些相关研究在处理耦合问题的同时利用高质量的医疗成像数据对人体肌肉和骨骼进行精细建模,以实现更为逼真的数字人仿真效果(Lee等, 2019; Teran等, 2005)。由于数字人与物的交互往往通过手来进行,近期工作(Zheng等, 2022)关注了对手部肌肉和骨骼进行精细建模,构建了复杂而灵活的人手仿真模型。这些工作所使用的人体医学数据集(Ackerman, 1998; Wang等, 2020)在众多研究领域中都发挥着重要作用,体现了构建高质量数据集在学术研究中的辐射效应和长期价值。

1.1.2 交互对象仿真

交互对象可以按照其物理类型分为刚体、软体和流体。刚体是目前交互任务中最常见的交互物体类型。这里面既包含同单个刚体的交互(如夹爪抓取茶杯),也包含同铰接刚体的交互(如机械手开门或拉抽屉)。无论是单个刚体还是铰接刚体,其仿真技术和工具都比较成熟,具体可以参见1.1.1小节。

除了刚体之外,一些工作也开始考虑数字人或机器人与软体的交互。这里用软体统称在交互过程中可以发生形变的固体。按照软体的空间维度可以将其分为体(如橡皮泥)、面(如布料)和线(如绳子)3类,分别对应三维空间中余维度(codimension)为0、1、2的情形。软体的仿真技术以连续介质力学中的固体力学为理论基础,一般选择在Lagrange视角的空间离散化方法上进行数值解算。常见的空间离

散化方法包括粒子法(Becker等, 2009; Müller等, 2004)、有限元法(Sifakis和Barbic, 2012)以及粒子—网格混合法(Jiang等, 2016)等。由于软体系统自由度多、计算量大,一些后续工作从状态表示、数值求解和并行计算等多个角度提出了软体仿真的加速方法(Bouaziz等, 2014; Lan等, 2022b; Müller等, 2007; Shen等, 2021; Sifakis和Barbic, 2012; Luo等, 2019; 梅春晖和石教英, 2003)。此外,一些近期工作针对布料等余维度不为零的软体提出了快速解法(Wang, 2021; Wu等, 2022; 任靖雯等, 2024),同样值得关注。

一些拟人交互工作也开始考虑数字人或机器人与流体进行交互的任务(Lin等, 2021; Xian等, 2022)。流体仿真是图形学和计算物理中的经典课题,国内外对此均多有研究(Lyu等, 2023; Li等, 2022c; Yang等, 2015; Batty等, 2007)。流体的仿真技术以连续介质力学中的流体力学为理论基础,根据其离散化方法可以大致分类为基于Lagrangian粒子的模拟方法(Bender和Koschier, 2015; Ihmsen等, 2014)和基于Eulerian网格的模拟方法(Foster和Fedkiw, 2001; Stam, 2023)。由于拟人交互一般不直接与流体发生作用,而是通过使用固体器具与流体交互(如倒水),这类任务的流体仿真通常需要解算流体与固体的耦合(Batty等, 2007; Peskin, 2002; Robinson-Mosher等, 2008; Takahashi和Batty, 2022; Ruan等, 2021)。

1.1.3 人一物交互仿真

在拟人交互仿真问题中,数字人或机器人与物体的交互一般是通过仿真固体间的接触模型(contact model)来实现的。接触模型的仿真是图形学和机械工程等领域的经典研究问题,此前的大量相关研究给出过众多解法。这些解法大致可以按照将接触条件视为动力学方程的硬约束或是软约束分为约束方法(constraint-based method)和惩罚方法(penalty-based method)两类。SIGGRAPH(special interest group for computer graphics)课程(Andrews和Erleben, 2021)对上述两类方法的代表性工作以及优劣做了极为详细的介绍。

在图形学领域,一般认为目前仿真接触的最好方法是由Li等人(2020)给出的增量势能接触(incremental-potential contact, IPC)。IPC在内点法(interior point method)(Nocedal和Wright, 1999)的启发下基于接触面的距离构造了对数形式的障碍函数

(barrier function)确保碰撞不穿模。通过一系列关键算法步骤的设计,IPC有效避免了以往惩罚方法的缺陷,提供了稳定而精确的软体接触仿真解法。IPC在众多复杂的软体仿真算例上表现出色,并在图形学和机械工程等学科的研究前沿受到了广泛关注。

在IPC的基础上,一系列后续工作从多方面对IPC进行了改进,在继承IPC不穿模和求解稳定等优点的同时将其适用场景从软体拓展到了更复杂的物理环境。Ferguson等人(2021)将IPC的思想拓展到刚体间的接触与碰撞,能够稳定求解复杂的刚体接触与碰撞场景。另一项相关工作(Lan等,2022a)在基于仿射变换的近似刚体上应用了IPC求解接触与碰撞。类似地,Li等人(2021)将IPC的思路应用到布料系统中以稳定求解其常见的自碰撞。在上述工作的基础上,Chen等人(2022d)给出了多固体系统仿真的统一框架,能够稳定处理包含刚体、软体和布料的复杂多固体系统中的接触,是目前该方向的标志性解法之一。

1.2 拟人交互仿真平台简介

1.2.1 面向特定物理场景的仿真平台

首先介绍3种只面向刚体物理场景的仿真平台。Box2D(Catto,2023)是早期比较具有代表性的刚体物理仿真引擎,由暴雪娱乐的Erin Catto开发,主要面向二维场景下的游戏应用。比较知名的刚体仿真平台是由Google研究院开发的Brax(Freeman等,2021)和由美国佐治亚理工学院主导开发的DART(dynamic animation and robotics toolkit)(Lee等,2018)。Brax的定位是面向大规模刚体场景的快速可微分物理仿真引擎,其中微分功能是借助Google提供的JAX(just after execution)框架的自动求导功能实现的。DART的定位是一款精确、稳定的刚体物理仿真,主要面向机器人和动画应用。

另一些物理仿真平台专注处理软体物理系统,其中影响力比较大的是美国南加州大学开发的VEGA(non-linear FEM deformable object simulator)(Sin等,2013),一款稳定、高性能的开源三维软体仿真引擎,主要面向熟悉软体仿真的学术界和工业界研发人员。近期比较典型的工作是PlasticineLab(Huang等,2021),它本身并不是一款物理引擎,而是基于材质点法(material point method,MPM)搭建的简单夹爪与三维弹塑性体(橡皮泥)交互的仿真

平台。

专注处理流体物理系统的仿真平台也涌现出来,其中代表性工作是PhiFlow(Holl等,2020)。PhiFlow与此前的流体仿真相比,主要的特点是提供了可微分的流体仿真并接入了主流机器学习框架,因此它是一款主要面向机器学习和优化应用的流体仿真引擎。在拟人交互方面,FluidLab(Xian等,2022)主要基于MPM搭建了固体与流体交互的仿真平台,在仿真环境中复现了倒水和咖啡拉花等日常生活涉及的流体交互任务。

1.2.2 面向通用物理场景的仿真平台

上述仿真平台都面向单一的物理系统。仿真单一物理系统足以处理相当多的拟人交互任务,例如刚体机械手抓取刚体物件。然而,更加复杂的拟人交互任务往往涉及多个物理系统,例如一个厨房机器人可能从事的切菜、和面、洗碗等任务涉及刚体、软体和流体等多物理系统的耦合。针对通用多物理场景涌现了若干有影响力的仿真平台,本节将对它们当中的代表性工作进行简要介绍。

目前,在学术界和工业界影响力比较大的通用物理场景仿真平台首推NVIDIA开发的Isaac Sim(NVIDIA Corporation,2023a)。Isaac Sim主要面向刚体物理系统,也对软体仿真做了一定支持。Isaac Sim面向机器人社区的研究和应用进行了深度的GPU(graphics processing unit)并行加速,其仿真速度十分出色。与Isaac Sim定位类似的是目前由DeepMind维护并开源的MuJoCo仿真平台。MuJoCo(Todorov等,2012)同样是一款面向机器人社区的物理仿真平台,在主要面向刚体物理系统的同时也在一定程度上支持软体仿真。MuJoCo也强调快速的物理仿真,并在性能优先的基础上定义了自己的接触模型。近期发布的MuJoCo 3版本基于JAX后端在GPU和TPU(tensor processing unit)上进行了大规模的硬件加速,性能十分瞩目(Google DeepMind,2023)。

另一款在机器人社区中比较经典的开源仿真平台是Bullet(Coumans和Bai,2021),同样主要支持刚体和软体物理系统的仿真,在Google Brain,Stanford AI实验室,OpenAI等研究机构的众多工作中被使用过(Sadeghi等,2018;Tan等,2018;Xu等,2018;Yan等,2018)。

关于上述3种常用仿真平台早期版本的比较和

讨论可以参见 Erez 等人(2015)的成果。

除了上述主要面向刚体机器人应用的通用仿真平台,一些以软体或流体为主要场景的仿真平台也值得关注,这里简要介绍两款在学术界有代表性的相关平台: PolyFEM (poly-spline finite-element method)(Schneider等,2019)是由美国纽约大学主导开发的开源有限元仿真引擎,主要面向学术研究中的软体和流体场景的高精度解算,目前仍在积极开发和维护中;SOFA(simulation open framework architecture)(Faure等,2012)是由法国国立计算机及自动化研究院(INRIA)支持开发的开源物理引擎,主要面向医学仿真和机器人应用,以软体和流体仿真为主。

最后简要介绍两种编写高性能图形学程序的编程语言 Taichi(Hu等,2020)和 Warp(NVIDIA Corporation,2023c)。Taichi 起源于美国麻省理工学院(Massachusetts Institute of Technology, MIT)的计算机与人工智能实验室(Computer Science and Artificial Intelligence Laboratory, CSAIL),目前由太极图形维护;Warp 由 NVIDIA 开发并维护。尽管它们本身不是物理仿真平台,但是它们为仿真引擎的开发者提供了良好的高性能编程框架,将开发者从对 GPU 等计算资源的底层管理中解放出来。此外,两种编程框架都支持自动微分功能,编写的物理仿真引擎自动可微,能够无缝接入当前主流的机器学习框架和应用中。

2 场景交互数据构建

室内场景拟人交互的生成离不开对场景以及对交互运动本身的理解。一方面智能体需要从交互的角度出发理解场景,对于场景中可支持交互的信息进行提取;另一方面智能体需要理解类人交互的运动流程,从而更好地模拟、规划和生成类似的运动。本节将从场景交互感知数据集、场景交互运动数据集以及数据规模高效扩充3方面综述国内外研究进展与趋势。

2.1 场景交互感知数据集

为了更好地支持交互生成任务,智能体需要从交互角度出发理解三维场景。这种理解超越了传统计算机视觉关注的语义识别、检测与分割,而更需要几何化、结构化、功能性与运动性的理解。近年来涌

现了一系列专注于感知可支持交互信息的数据集。根据下游交互策略的差异,这些数据集的聚焦点也不尽相同,包括交互可供性、多指抓取位姿、交互部件结构与运动性、物体位姿等信息。

交互可供性是决定物体如何被交互和使用的基本属性。现有的数据集通常将该类信息记录为物体或场景局部区域与交互动作之间的映射。例如, ContactDB (analyzing and predicting grasp contact via thermal imaging)(Brahmbhatt等,2019)利用热力图记录人在抓取不同物体时所覆盖的接触区域,并将这些区域以可抓取区域的形式记录,描述了人如何在以特定功能为交互目标的前提下与物体进行接触,为拟人功能性抓取提供了模仿信息。3D AffordanceNet(Deng等,2021)考虑更丰富的可供性类别,描述了人与物体的交互以及物体与物体之间的交互可供性。EGO-TOPO (environment affordances from egocentric video)(Nagarajan等,2020)从物体扩展至场景,标注了第一人称视角视频中不同场景区域的交互动作与对象,为理解场景交互可供性和第一视角人物交互提供了有效的数据支持。

交互可供性通常涵盖了较为丰富的交互类型,但是标注往往以区域热力图、评分图为主,信息粒度较为粗略;多指抓取作为拟人交互的基石,更直接地刻画了交互过程中多指手的可能位姿,并在一系列数据集工作中得到广泛关注。例如, ContactPose (Brahmbhatt等,2020)利用动捕系统采集了大量真人手物抓取示例,并提供了配套的 RGB-D 视频、物体多边形网格、物体表面接触图和抓取位姿等信息,为拟人交互的生成提供了有效的学习素材。然而,受限于数据采集与标注的困难,真实数据规模往往非常有限。AffordPose(Jian等,2023)基于部件级可供性,利用算法生成并人工调整了大量多指抓取位姿,以扩大数据规模并更好地反映多指抓取的功能性。

除了几何化的交互标注信息,对物体本身结构的理解对交互至关重要。随之而来的是大量以三维物体结构与运动性信息为中心的数据集。早期描述物体部件运动性的数据集(Hu等,2017)将物体的三维多边形网格划分为可运动部件,并根据运动类型对数据进行分类,以支持三维几何的运动性分析。尽管数据集规模有限,但仍然启发了后续工作的发展。SAPIEN(a simulated part-based interactive envi-

ronment)(Xiang等,2020)在大规模三维数据集ShapeNet(Chang等,2015)和PartNet(Mo等,2019)的基础上,按照物体的语义类别挑选出可交互的铰接物体,并对物体的部件分割(孙刘杰等,2024;Shi等,2023;李新叶和陈丁,2023)、铰接关节类别与属性、部件运动范围等信息进行了标注,推出的PartNet-Mobility数据集迄今为止仍是领域中最大规模的铰接物体数据集之一。GAPartNet(Geng等,2023b)在PartNet-Mobility数据集基础之上,按照交互可供性对部件进行了归类,并按照交互方式定义了通用的部件位姿,以便更好地连通几何数据与拟人交互。

事实上,位姿在几何与交互的连通上所发挥的作用早已得到了机器人领域的充分认可,智能体通常会借助对物体位姿的估计以便更好地完成抓取的规划,所以场景交互感知一直以来也将位姿估计放在非常核心的位置。PoseCNN(a convolutional neural network for 6D object pose estimation in cluttered scenes)(Xiang等,2018)作为早期利用深度学习进行位姿估计的工作之一,推出了YCB-Video数据集,该数据集以YCB(Yale-CMU-Berkeley)物体为基础,精确标注了视频帧中不同YCB物体的6自由度位姿,以评估位姿估计方法在复杂环境中的准确性。YCB-Video聚焦于实例级别的物体位姿估计,难以迁移至多样全新的几何体上,NOCS(normalized object coordinate space)(Wang等,2019a)为了解决这一位姿估计的泛化性问题,提出了类别级位姿的概念,并推出了包含类内几何多样性的6自由度位姿数据集,用于测试位姿估计算法在已知类别的全新几何体上的位姿估计性能。

总的来说,为了从交互角度理解三维场景并支持交互生成任务,往往需要综合利用多个数据集,涵盖交互可供性、多指抓取位姿、交互部件结构与运动性、物体位姿等信息。这些数据集为研究人员提供了多样化的数据资源,支持更准确、丰富的交互生成算法开发。

2.2 场景交互运动数据集

拟人交互的生成一方面需要从交互的角度出发更充分地对环境进行理解;另一方面则需要模仿对象即人与场景的交互运动样例支持。拟人交互运动的生成离不开对人与场景交互运动的分析,这种分析通常覆盖了运动分割、交互运动追踪、动态交互重

建、系统识别、行为分割与理解以及行为预测等。为了支持以上分析,大量聚焦于三维场景中交互运动的数据集应运而生。按照数据采集和获取方式,可以分为基于动作捕捉与基于遥操作的交互运动数据集。

2.2.1 基于动作捕捉系统

获取人与场景的交互运动数据最直接的策略是利用动作捕捉系统与算法,相比于非交互场景,交互过程中交互主客体之间的强遮挡、运动多样性也给动作捕捉带来了相应的挑战。早期工作如FPHA(first-person hand action)(Garcia-Hernando等,2018)通过在人手与物体上粘贴磁传感器结合逆运动学算法来追踪人手与物体的位姿的变化,然而,此类型的数据集受到标识传感器的影响,数据的视觉真实性较差,难以用于分析真实自然场景中手物交互运动的观测。为了去除标识点对视觉数据真实性的影响,HO3D(honnotate 3D)(Hampali等,2020)提出了一种利用多视图RGBD同步视频对手物交互过程中手与物体位姿进行估计追踪的方法,并基于此方法收集了包含27段视频78 K视频帧的手物交互运动数据集。DexYCB(Chao等,2021)进一步通过人工修正提升交互数据质量并将视频段数提升至1 000段。H2O(two hands manipulation objects)(Kwon等,2021)突破单手交互的限制,聚焦于双手交互过程中的左右手协同,而ARCTIC(a dataset of two hands that dexterously manipulate objects)(Fan等,2023)则是收集了双手操作铰接物体的几何数据,展现了更加丰富的交互行为,用以支撑智能体灵巧操控运动的生成。不仅仅是手物交互,智能体拟人交互也需要全身的协调。GRAB(a dataset of whole-body human grasping of objects)(Taheri等,2020)利用基于标识点的动捕系统捕获了1 334段日常生活中人完成全身物体抓取的运动序列,用以支撑数据驱动的全身交互生成。同样为了去除标识点对视觉数据的影响,并覆盖更加丰富的交互模式(如除手外身体其他部位与环境的交互),BEHAVE(dataset and method for tracking human object interactions)(Bhatnagar等,2022)利用多视图RGBD同步视频捕获了321段全身人物交互序列。PiGraph(proto-typical interaction graphs)数据集(Savva等,2016)使用定机位的Kinect.v2 RGB-D相机来捕捉人们的日常行为,包括30个场景,63段视频,5名不同的角色,

总时长约 2 h。PROX (proximal relationships with object exclusion) 数据集 (Hassan 等, 2019) 使用 Kinect-One RGB-D 相机拍摄 20 名不同演员在 12 个不同场景中不断与场景进行各种交互的 RGB-D 数据, 并基于该数据对场景进行了重建, 对场景中角色进行了 SMPLify-X 模型 (Pavlakos 等, 2019) 估计。HUMANISE (a large-scale and semantic-rich synthetic HSI dataset) 数据集 (Wang 等, 2022c) 同样聚焦人与场景的交互, 大大扩充了拍摄的场景和数据量, 并且进行了丰富的语义标注和文本描述。以上交互数据通常聚焦于运动本身的多样性, 而忽略了交互对象多样性对交互带来的影响, 交互几何对象数量通常在几十个以下。与之相对的, 大量 RGB 视频数据集如 Something-Something (Goyal 等, 2017) 以及 EpicKitchen (Damen 等, 2021) 记录了日常生活中人与各种复杂物体交互的过程, 这些数据体量丰富, 对于行为的识别理解大有裨益, 但是由于缺少三维几何标注, 对于面向生成的应用场景使用较为困难。为了突破交互过程中高质量几何数据缺少物体多样性的局限性, HOI4D (human-object interaction 4D) (Liu 等, 2022a) 采集了大量类别级手物交互视觉数据, 描述了人在与多类物体交互过程中, 交互随类内物体几何差异而产生的变化, 并进行了精细的几何标注, 为增强交互生成算法的几何泛化性提供了基础。

2.2.2 基于遥操作系统

基于动捕系统获取到的交互运动数据往往缺少对交互过程的动力学刻画, 智能体对于交互的模仿只能停留在运动学的层面, 再考虑到智能体与人体的形态差异、运动重定向所内蕴的歧义性, 这种运动的模仿往往伴随着很多技术上的挑战。为了更好地破除形态差异所带来的模仿困难, 也为了从动力学层面获取对交互的描述, 遥操作系统近年来得到了广泛的关注。其基本原理在于利用视觉算法或硬件设计, 将人的操作行为直接映射到智能体执行器上, 由人来操作智能体完成各式交互任务。Wong 等人 (2022) 利用手机遥控智能体在虚拟环境中进行导航、抓取等交互任务, 收集人类遥操作数据以支持智能体模仿学习。RH20T (Fang 等, 2023b) 通过硬件平台设计使得操作员可以通过触觉设备与脚踏板完成对机械臂的遥操作, 用以完成大量富接触操作任务, 并基于此平台采集了一个支持多样技能学习的大规模遥操作数据集。ContactArt (Zhu 等, 2023) 则是利

用基于手型识别与运动重定向的视觉遥操作系统采集了铰接物体操控的交互数据, 用以支持复杂的交互生成任务。

总的来说, 这些不同类型的交互运动数据集在拟人交互生成中具有重要作用。基于动作捕捉的数据集关注真实场景中的交互运动, 通过动作捕捉系统获取数据。而基于遥操作的数据集则通过遥操作机器人手臂或虚拟手模拟交互行为。这些数据集质量的提高、规模的扩大对于研究人员理解人与环境之间的交互运动、生成逼真的拟人交互至关重要。

2.3 交互数据规模高效扩充

随着大模型对人工智能领域的推进, 研究人员越来越重视高质量大规模数据对拟人交互生成的价值, 很多研究者认为大规模场景交互数据是推动通用拟人交互生成的关键因素。然而传统交互数据的获取往往需要大量人工参与和投入, 也需要高成本专用化硬件平台的支持, 为数据的高效扩充带来了很大阻力。现阶段交互数据远达不到互联网文本与图像的量级, 也阻碍了交互大模型的产生。研究者面临的核心科学问题之一便是如何高效扩大现有交互数据的规模, 而目前的主流研究思路包含低成本交互捕捉系统设计、数据集组合以及交互数据增强等方面。

2.3.1 低成本交互捕捉系统设计

为了扩充基于动作捕捉获取的交互数据, 核心问题在于需要降低动捕算法对硬件系统的需求, 提升动捕算法的性能与鲁棒性。在过去几年中, 动捕领域的研究重心从基于标识点的动捕过渡到弱标识甚至不需要标识点的动捕, 从基于多相机系统的动捕演变为基于单相机系统的动捕, 从实验室环境下的动捕迈向自然环境下的动捕, 从非交互场景中的动捕扩展到复杂交互行为下的动捕。FrankMocap (Rong 等, 2021) 利用参数化模型表示人体与人手, 利用神经网络预测模型参数并结合优化算法完成从单 RGB 图像中对人体位姿的鲁棒估计。InterCap (Huang 等, 2022) 利用多视角 RGBD 视频捕获人与物体的交互运动, 在不借助标识点的情况下, 通过利用接触信息来同时提升人与物体位姿的估计准确度。Xie 等人 (2023) 仅利用单 RGB 相机来同时跟踪交互过程中人与物体的位姿, 基于深度学习通过较为鲁棒的人体运动估计, 结合物体可见视频帧对物体位姿的估计, 来恢复物体遮挡帧物体的位姿。

扩充基于遥操作系统的交互数据,核心问题在于设计更为高效、低成本的遥操作系统,以增加数据采集效率,降低数据采集成本。Zhao等人(2023b)推出了一套价格低于2万美元的开源遥操作系统,支持用户通过逆驱动机械臂精准高效完成诸如穿束线带、装电池等复杂、富接触的双手操控任务。Fang等人(2023a)推出了一款价格低于600美元的可穿戴机械外骨骼,可以支持全臂操控数据的低成本高效采集,用于学习复杂的机械臂操控。AnyTeleop(Schneider等,2019)推出了一套基于视觉的遥操作系统,只需要借助低成本相机,便可以支持不同遥操作主体与对象、不同相机配置的需求,借助视觉算法完成对人行为的追踪与重定向。曾芬芳等人(2000)借助数据手套实现了在虚拟环境中对物体的抓取、释放等操作。Oprea等人(2019)借助虚拟现实系统,在头戴设备与运动控制器的支持下,完成虚拟环境中用户对虚拟物体的真实操控。

无论是动捕系统还是遥操作系统,软硬件协同设计以降低系统成本、提升系统效率是高效扩充交互数据采集的关键。

2.3.2 数据集组合

扩充交互数据的方式是对已有数据集进行集成整合,看似简单,然而充分集成类型格式不同、分布差异较大、标注信息不同、体量不同的异质化数据集,从而保证数据的有效性仍然充满了挑战。AMASS(archive of motion capture as surface shapes)(Mahmood等,2019)数据集对已有的小体量动捕数据集进行了整合,统一了对人物运动的参数化表达,使得该数据集可以更好地支持数据驱动的学习方法。Ego4D(a massive-scale egocentric video dataset and benchmark suite)(Grauman等,2022)聚焦第一人称场景交互视频数据,通过调动来自全球9个国家74个位置931名数据采集者,将不同区域所采集得到的数据汇集在一起,保留不同数据的不同标注信息,分数据设定基准任务,形成了大规模交互视频数据。Padalkar等人(2023)通过收集来自21所研究机构采集自22种不同机器人的操作样例数据,对数据的格式进行统一,对示例数据的状态空间与行为空间分别进行粗对齐,以形成涵盖500多种机器人操作技能的大规模机器人交互数据集。在可预见的未来,世界范围研究力量的整合对大规模交互数据集的发展仍至关重要。

2.3.3 交互数据增强

除了采集获取真实场景中的交互数据,扩充交互数据规模的另一行之有效的策略便是利用计算合成、生成学习等技术对真实数据进行增强。Obman(Hasson等,2019)通过传统的多指抓取算法在给定几何体上生成了大量的合成抓取位姿,并基于启发式排序挑选出其中一部分与人真实抓取更加匹配的位姿形成了一个大规模的多指抓取数据集,并证明了该合成数据集在抓取位姿估计等真实场景视觉应用中的价值。DexGraspNet(Wang等,2023)同样聚焦多指抓取问题,利用优化算法,高效生成了大量多指抓取样例,与Obman相比规模提升了近50倍,也在机器人抓取规划等问题中印证了数据的价值。OakInk(Yang等,2022a)利用抓取重定向,通过对物体几何特征的分析,将真实数据中采集到的多指抓取信息,成功迁移到同类别不同物体上,大大扩充了真实抓取数据的规模,并在手物交互重建、手物交互位姿估计等任务上证明了合成数据所带来的裨益。在交互运动数据方面,诸如像CAMS(canonicalized manipulation spaces)(Zheng等,2023)这种交互生成模型,可以通过刻画训练数据分布,在全新几何物体上生成更多多样逼真拟人的交互数据,用以满足交互理解、交互重定向等下游应用的需求。Zhang等人(2023a)则是通过构建几何课程,将收集自简单几何体的交互运动,以物理真实的方式逐步重定向到多样复杂的几何体之上,大大增加了交互运动的拓展性。

基于计算合成或是生成学习所得的交互数据往往能够以极低的成本增加交互数据的多样性,给深度学习带来支持,然而这些数据与真实交互数据的差异仍是不可忽视的问题,增强数据所带来的知识必须是可迁移至真实场景的。为了弥合增强数据与真实数据的差异,一种策略是通过对真实交互数据的分析,如DiffCloud(Sundaresan等,2022),完成系统识别,以更加贴近真实的方式生成增强数据;另一种常见的策略则是利用域随机化,代表工作包括Akkaya等人(2019)训练灵巧手完成魔方操作的方法,通过对仿真环境进行随机扰动生成分布广泛的增强数据,一定程度上覆盖真实数据的分布,并使得学习算法对弱相关环境变量鲁棒,从而实现到真实场景的迁移。迄今为止,如何利用好合成数据仍是解决交互数据体量不足的重要研究问题。

3 拟人交互感知与生成

在上述仿真交互环境搭建和场景交互数据构建的基础上,本节主要介绍室内场景拟人交互感知和生成方面的工作。首先,对交互导向的场景可供性分析方面的研究工作从不同角度进行分类介绍。然后,对数字人与场景以及机器人与场景交互生成的相关技术和方法进行梳理。在数字人与场景交互生成方面,详细介绍了数字人动作的表达与生成、交互的表达与生成以及常用的交互数据集。在机器人与场景交互生成方面,以交互任务作为切入点,详细介绍了不同类型的任务以及对应的生成方法。

3.1 交互导向的场景感知

可供性是一种描述物体或环境特性对个体行动可能性的概念,指的是物体或环境能提供给使用者的行动潜能。这个概念在人机交互、机器人工程和认知科学等多个领域中极为重要,因为它涉及如何识别和利用周围环境中的物体来完成特定任务。例如,一个椅子提供了坐的可供性,而门把手提供了转动或拉动的可供性。

可供性研究可以根据粒度来区分,主要分为细粒度可供性和粗粒度可供性。细粒度可供性关注于特定物体的详细和具体的行动潜能,例如机械臂在操作物体时,需要识别物体的哪一部分可以被抓取或操作。粗粒度可供性则关注于更广泛和一般的行动潜能,例如识别一个场景中所有物体的一般用途。

另一种区分方法是根据可供性信息的训练来源,可以分为多模态可供性、单样本可供性、开集可供性和弱监督可供性。多模态可供性研究如何结合不同类型的输入(如图像、文本、视频)来识别和理解可供性;单样本可供性关注于如何从单一样本中学习并推广到其他情境;开集可供性则探索如何处理未在训练数据集中出现的新物体或场景;弱监督可供性关注于如何在标注数据有限的情况下进行有效学习。

3.1.1 多粒度可供性分析

细粒度可供性任务是指在给定操作物体和机械臂状态的情况下,输出物体可供性信息。这些信息包括物体每个部分的可供性概率、执行任务所需的操作方向以及任务成功的概率。过去的研究主要关注单个机械臂的操纵任务,而实际场景中许多任务

需要两只机械臂协作完成,例如稳定地提起一个没有提手的水桶(图1(a))。为了解决这个问题,Zhao等人(2023c)提出了一个新的框架。该框架允许两只机械臂协作完成更具挑战性的任务,不仅在模拟环境中,还在真实环境中。针对可变形物体操纵问题(如毛巾,图1(b)),Wu等人(2023)提出了一种具有前瞻性密集视觉可供性的方法。这种物体的操作空间非常庞大,存在许多局部最优值,因此算法容易陷入局部最优值,如图1(b)右下角所示,可供性得分只有0.792,但是很难在有限步骤内跳出这个局部最优值。为了避免这种情况,Wu等人(2023)在计算当前帧的可供性分数反馈时,考虑了操作后下一帧的可供性分数,确保每次反馈都考虑了未来的步骤。此外,Wu等人(2023)采用了自主生成训练数据的方法,使得算法具有较高的通用性。然而,Wu等人(2023)的方法需要经历多个训练阶段,训练效率很低。为了解决这个问题,Liang等人(2023)引入了自编码器的方法,将训练过程缩减为一阶段(如图1(c)最右侧所示),并且只需要正样本而无需负样本,大大降低了训练过程和时间的复杂度。此外,Liang等人(2023)还提出了一种中间特征融合策略,通过在中间阶段融合多模态数据,提升了模型对机械臂和操作物体之间关系的感知能力。最后,Zhong等人(2023)提出了首个3D版本的transporter。该方法只需两帧点云即可自监督提取时序上稳定的三维特征点集合,并能用于任意未知铰链物体的闭环操作。

粗粒度可供性是指在给定场景下,识别场景中所有物体的一般可供性,如椅子可供坐、水龙头可供扭等,不关注具体操作细节。Chen等人(2022c)提出了一个多任务联合训练框架,将语义分割、物体属性检测(如材质)和粗粒度可供性检测整合在一起(图1(e))。实验证明,这3个任务之间的信息具有相互促进的作用。Li等人(2022a)提出了一种名为TOIST(task oriented instance segmentation Transformer)的新方法,用于解决任务导向检测中的挑战性问题,即语言引导的可供性检测。如图1(f)所示,给定图像以及文本“smear butter”,算法将涂抹黄油的工具框出来。为了解决这个问题,Li等人(2022a)引入了名词—代词蒸馏框架,充分利用预训练的名词理解模型。通过在训练过程中访问名词的真值,以无监督方式生成名词,并训练代词特征来选

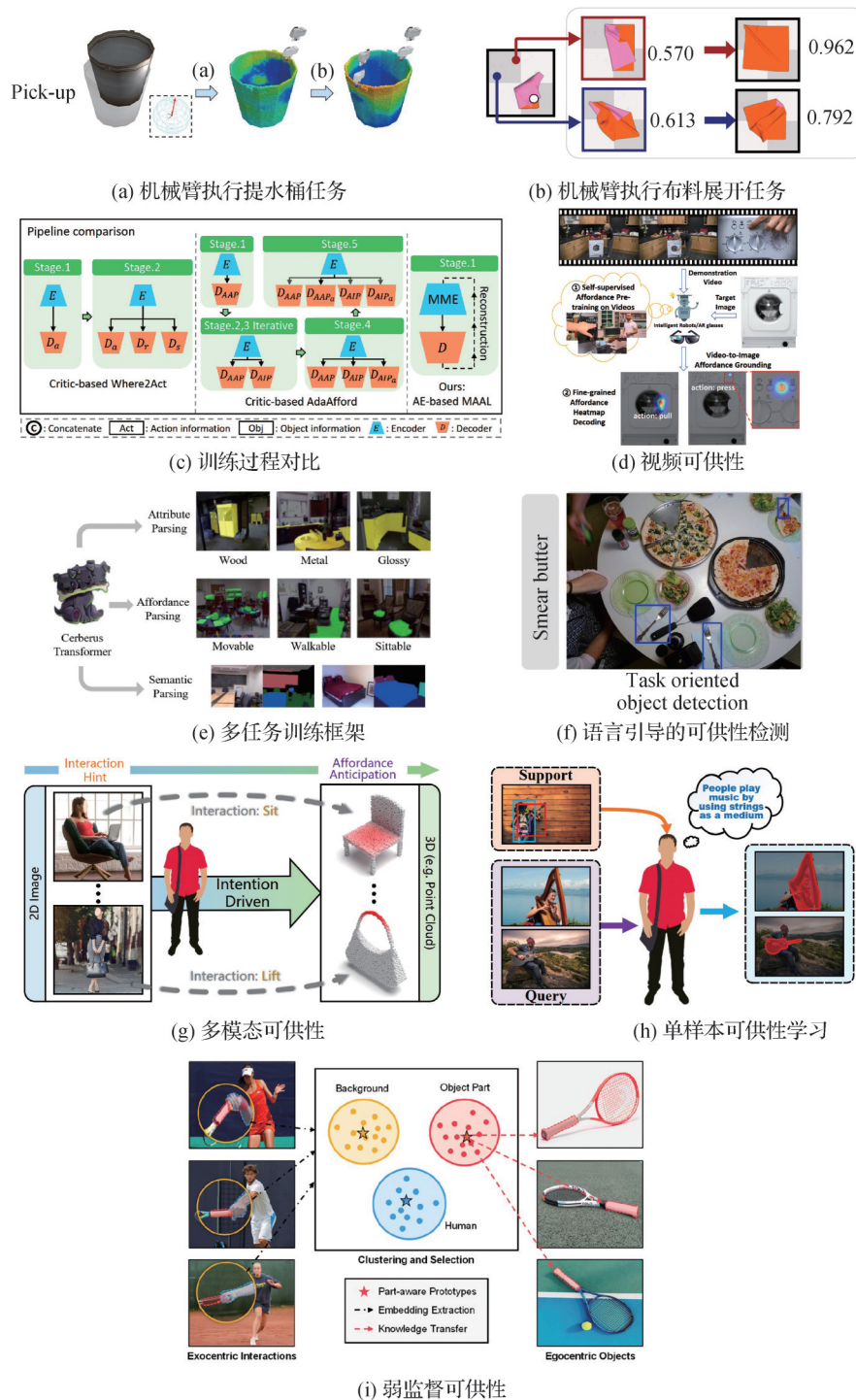


图1 代表性可供性分析方法对比示意

Fig. 1 Representative affordance analysis methods ((a) robot arm performs bucket lifting task; (b) robotic arm performs fabric unfolding tasks; (c) comparison of training processes; (d) video affordance; (e) multi task training framework; (f) language guided affordance detection; (g) multimodal affordance; (h) single affordance availability learning; (i) weakly supervised affordance)

择这些名词。因此,在推理过程中,网络对名词保持不可知状态,从而增强了其灵活性和适应性。此外, Tang 等人(2023a)提出了一种多层次的链式思考提示方法,用于从大型语言模型中提取可供性知识。

该方法涉及多个推理步骤,从任务级别到对象级别的示例,并进一步深入研究伴随理由的基本视觉属性。同时,还引入了一种名为 CoTDet (knowledge-conditional detection framework) 的基于知识的检测

框架,该框架通过将检测器与已获取的知识进行条件化,生成对象查询并执行边界框回归。

3.1.2 多源可供性学习

基于认知科学的启示, Yang 等人(2023b)指出人类对所观察到的物体具有感知经验,从而促进他们组织感知和操作新物体的能力(可供性)。因此,该研究提出了一个全新的任务设置。通过从二维人物和物体交互的图像中学习可供性,实现对三维数据的可供性进行推断(如图1(g)所示)。为了实现这一目标,研究团队设计了一种多模态可供性算法,并构建了一个包含图像一点云对的数据集。另外, Zhai 等人(2021)定义了一个具有挑战性的单样本可供性检测问题。如图1(h)所示,该问题要求算法从支撑集中学习的知识能够泛化到查询集,其中支撑集包括1幅支撑图像和相应的可供性标注信息,而查询集包括多幅查询图像。为了解决这个问题, Zhai 等人(2021)提出了一种新颖的 OSAD-Net(one-shot affordance detection network)模型。该模型能够有效地学习可供性知识,并利用合作学习方法来捕捉具有相同感知能力的物体之间的关系。此外, Zhai 等人(2021)还提供了一个大规模的基准数据集 PADv2,以促进该领域的研究。该数据集包括多样化的图像、可供性遮罩标注、深度信息、人/物体边界框以及支撑图像标注。 Nguyen 等人(2023)通过结合自然语言作为输入,成功实现了开集点云可供性检测。此外, Li 等人(2023a)提出了一种新的弱监督可供性学习方法,该方法只需要提供图像级别的标注(例如握取动作),即可识别全局视角图像集的交集部分。然后,通过聚类筛选出与可供性相关的物体预测,并将其用于监督第一视角图像集,从而实现弱监督训练,如图1(i)所示。

除了从图像中学习可供性,视频—图像可供性任务是指从视频数据中学习,并将所学技能应用于新的目标图像。如图1(d)所示,观看洗衣机操作视频,然后利用所学知识将新场景的可供性信息呈现在AR(augmented reality)眼镜中。然而,这个问题的复杂性在于演示视频和图像通常是在不同环境中捕获的,这使得在图像中检测可供性变得具有挑战性。此外,对视频—图像数据对进行标注是一项耗时的任务,需要标注者查看每个视频的全部内容,导致该任务数据的稀缺。为了解决这些挑战, Chen 等人(2023a)提出了一种解决方案,利用多尺度特征金字

塔之间的交叉注意力,以便逐步解码细粒度的可供性热图。此外, Chen 等人(2023)通过设计自训练框架来解决有限的训练数据问题,该框架利用大量的在线视频来提升视频—图像可供性的性能。 Mur-Labadia 等人(2023)提出一个自动标注粗粒度可供性的框架,利用环境的三维地图从第一人称交互视频中自动提取基于交互的可供性,确保像素级精确定位可供性。此外, Mur-Labadia 等人(2023)利用这种框架构建了基于现有数据集(Damen 等, 2018)的一个规模最大、最完整的可供性数据集 EPIC-Aff。 EPIC-Aff 提供基于交互、多标签的可供性注释。为了促进多个可供性在同一空间中的共存,特别是当它们与同一对象关联时, Mur-Labadia 等人(2023)提出了基于多标签检测的可供性分割方法。

具有代表性的可供性分析方法的类别与特点分析如表1所示。

3.2 数字人—场景交互生成

人是室内交互场景的核心参与者。在拟人交互生成的领域研究中,人的动作和行为作为研究对象具有极其重要的地位。然而,生成真实的人与场景的交互是一个非常困难的问题。一方面,人的动作具有极高的多样性,且不同场景环境下的交互行为进一步增加了这种多样性;另一方面,人们对日常动作的真实性非常敏感,因此对生成动作的质量提出了更高的要求。本节将数字人与数字场景的交互作为目标,探讨有效表达和生成人的动作的方法,以及生成人与场景对象交互动作的技术。

3.2.1 数字人动作的表达与生成

1)动作的表达与生成。人的动作受到多种主客观因素共同影响,因此对动作的真实性、自然程度等属性进行量化描述往往非常困难。为了解决这一问题,大部分的动作生成方法往往采用基于数据的思路,即从大量人的动作数据中建立动作生成模型,并利用模型在特定条件下进行动作生成。

实时的动作生成要求生成的动作可以实时响应不同的输入,因此利用自回归(autoregressive)网络逐帧/逐块的生成是一类常用的方法。一些早期的工作使用轻量的多层感知机(multilayer perceptron, MLP)作为自回归模型,在特定数据集上可以产生高质量的动作(Holden 等, 2017; Starke 等, 2019, 2022)。不过,为了减少动作歧义性问题,这类方法通常会在自回归模型里加入周期性“相位”信息来帮

表 1 代表性可供性分析方法类别与特点分析

Table 1 Categories and characteristics of representative affordance analysis methods

方法	类别	特点
Wu 等人(2023)	细粒度可供性	通过一种具有前瞻性的可行性评估方法来处理可变形物体操纵任务;利用自主生成的方法生成训练数据对。
Zhao 等人(2023c)	细粒度可供性	两个机械臂协作完成任务;第 2 个机械臂的推理过程依赖于第 1 个机械臂的状态,而第 1 个机械臂的监督也依赖于第 2 个机械臂的状态;尽管多了一个机械臂,但操作空间复杂度仍然维持在线性复杂度。
Zhong 等人(2023)	细粒度可供性	只需两帧点云即可自监督提取时序上稳定的三维特征点集合;能用于任意未知铰链物体的闭环操作。
Liang 等人(2023)	细粒度可供性	训练过程只有一个阶段,并且只需要正样本,训练效率大幅提升;引入自编码器。
Chen 等人(2022c)	粗粒度可供性	多任务联合优化框架。
Li 等人(2022a)	粗粒度可供性	语言引导的可供性检测;提出名词一代词蒸馏框架实现名词教代词理解动词。
Tang 等人(2023a)	粗粒度可供性	从大型语言模型中提取可供性知识。
Yang 等人(2023b)	多模态可供性	基于二维图像中人物与物体的交互,学习推断实现三维数据的可供性。
Zhai 等人(2021)	单样本可供性	定义单样本可供性问题;合作学习框架;单样本可供性数据集。
Nguyen 等人(2023)	开集可供性	CLIP 文本特征提取器。
Li 等人(2023a)	弱监督可供性	全局视角的图像集教第一视角图像的弱监督框架。
Chen 等人(2023a)	视频可供性	采用自训练框架解决视频可供性任务训练数据稀缺问题。
Mur-Labadia 等人(2023)	视频可供性	一个用于自动标注可供性标签的框架;构建了一个规模最大、最完整的可供性数据集。

助网络对位移等动作进行有效建模(Holden 等, 2017)。后续工作尝试将生成模型引入自回归框架,例如 Normalizing Flow(Henter 等, 2020)以及变分自编码器(variational autoencoder, VAE)(Ling 等, 2020),从而允许生成的动作具有一定的多样性。此外,基于“最近邻查找”思想而诞生的 Motion Matching 技术(Büttner, 2015)因其能确保生成动作质量较高的同时实现优秀的运行效率,从而近些年获得了工业界的青睐。

非实时的动作序列合成允许一次性生成整段动作,因此动作模型可以更加准确地建模动作的时序关系。例如可以通过基于 Transformer 的 VAE 模型将一个动作片段编码成一个单独的向量表示,然后再进行生成(Petrovich 等, 2021)。生成式对抗网络(generative adversarial network, GAN)及其变种(Li 等, 2022b)是另一种可行的方法。基于给定的一小段示例动作,使用层次化的 GAN 网络可以在保留示例动作基本“语义”的情况下实现多样化生成。与 Li 等人(2022b)类似, Motion Matching 也可以与 GAN 结合实现层次化的动作序列生成,并相较于基

于神经网络的方法展现出了更高的可控性(Li 等, 2023c)。受到扩散模型在图像生成任务上强大能力的启发,近年来很多工作利用扩散模型实现了高质量的动作生成,如 Azadi 等人(2023); Chen 等人(2023b); Dabral 等人(2023); Tevet 等人(2023)等。

2)动作的条件合成。在动作生成的过程中,条件控制同样非常重要。关键帧是一种常见的动作生成条件。通过给出合适的关键帧,并让动作模型生成补间动画,可以有效地控制动作的关键细节,并生成高质量的整体动作。其中关键帧间运动的补全,可以通过在自回归模型中加入位置编码和关键帧信息的方式(Harvey 等, 2020),或者引入“相位”信息引导补全(Tang 等, 2023b),或者在扩散模型中使用“修复”的方式(Chen 等, 2023b; Tevet 等, 2023),或者利用 Classifier-Free Guidance 的方式(Karunratanakul 等, 2023)进行生成。而对于在线的动作合成,往往由用户提供控制信号,生成未来期望的运动轨迹,再基于运动学的原理,根据当前的状态生成未来可能的运动(Holden 等, 2017; Zhang 等, 2018)。在动作生成的过程中加入“相位”信息能够有效提高动作的

质量和自然性,实现复杂的篮球、足球、拳击等运动的动作生成(Starke等,2019,2020,2021,2022)。层级化的运动控制先使用变分自编码器VAE训练一个动作表征空间,再根据不同的任务通过强化学习训练控制器(Ling等,2020),这样在满足控制信号的同时让生成的动作更加多样。

类似于“文生图”的任务,基于文字的动作生成可以允许人类通过自然语言的方式与数字人角色互动,甚至可以通过语言接口利用语言大模型的能力完成更复杂的任务。基于扩散模型的文本动作生成取得很大的进展。文本输入可以经过预训练语言模型,例如T5(Raffel等,2020),提取特征,然后以该特征作为条件,训练条件扩散模型,或采用Classifier-Free Guidance(Ho和Salimans,2022)的方式进行采样生成,可以得到语义与语言描述相近的动作。一些工作受到Stable Diffusion启发,先训练一个动作的自编码器,然后在隐空间上进行扩散采样,由于特征空间被显著压缩,训练成本降低,更加容易收敛,采样的速度也大幅度提高(Barquero等,2023;Chen等,2023d;Kong等,2023)。此外,连续动作序列可以被VQ-VAE(vector quantised VAE)(van den Oord等,2017)等网络离散化为一个个词元(token),并将该词元(token)序列视做一门新的“语言”,之后将该“语言”对应的词元(token)序列空间和自然语言空间进行对齐,从而实现灵活地模态切换和跨模态生成(Jiang等,2023)。

除了文本生成动作,动作的扩散模型也可以用来实现音乐生成舞蹈的任务(Li等,2023b;Tseng等,2023),与文本生成动作时直接将整段描述进行输入有所区别,音乐生成往往先划分成固定长度的片段,并提取节拍等信息作为条件,以提升生成舞蹈动作与音乐之间的关联性。此外,如何分别显式地让生成模型感知音乐和舞蹈之间的“节奏关联”和“编舞结构关联”,从而提升生成效果,是一个值得深入研究的方向。针对节奏,可以设计专门衡量音乐和舞蹈间节奏匹配程度的奖励函数,并通过强化学习框架来微调生成模型,使得其生成的舞蹈动作节奏与音乐更加适配(Siyao等,2022)。与音乐生成舞蹈任务相似的是语音/文本生成手势动作。同样地,扩散模型也成功地用来生成高质量的手势动作(Alexander等,2023;Ao等,2023)。与其他任务有所区别,手势生成需要克服“语音和手势间相关性较弱”

的混淆问题。一种直观的想法是显式地对语音和手势之间的“韵律”和“语义”进行显式建模,从而提升生成效果(Ao等,2022;Liu等,2022b)。此外,利用CLIP等大规模多模态预训练模型来实现风格化的手势生成也是一个值得探索的研究方向(Ao等,2023;Zhi等,2023)。

3)基于仿真的动作控制。在与场景的交互过程中,一个很重要的方面就是交互的物理真实性。没有物理约束的环境交互动作生成方法通过数据进行学习,往往对环境的物理属性进行了简化,如箱子、篮球等都只具有形状属性,或者抽象成了一个点,无法实现根据箱子重量变化生成不同动作的效果(Starke等,2019)。同时,学习到的动作生成模型很难对动态的环境变化进行反馈,学习到的交互鲁棒性也高度依赖数据集的多样性。一种真实但也更有挑战性的方式,是通过物理仿真真实世界环境,指定训练任务目标,控制智能体学习主动与环境交互,在仿真环境下学习到的控制策略更加真实,面对环境变化的反应也更加鲁棒。在仿真环境里,智能体同样能够学习到与真实人物相似的行为策略(Peng等,2018,2021,2022;Yao等,2022),并且能够实现打篮球(Liu和Hodgins,2018)、打网球(Zhang等,2023a)、使用筷子(Yang等,2022b)等复杂且精细的交互运动。

在仿真环境中需要解决一个非常复杂的控制问题。人形机器人领域通常使用较为简化的模型来实现在运动中保持自身的平衡。然而这种方法往往会智能生成机器风格的动作,缺少真实的人的动作的细节。为了解决这一问题,利用基于强化学习的模仿运动数据是一类比较常用的方法(Liu等,2016;Peng等,2018)。早期的工作通常需要为每一段动作训练单独的控制策略(Peng等,2018),效率和泛化性较差。越来越多的工作开始关注将生成模型与运动控制策略相结合的方法。通过对大量动作样本或者控制策略样本的学习,可以实现直接对控制策略进行编码,进而支持各种下游任务(Peng等,2022;Won等,2022;Yao等,2022)。

3.2.2 数字人—场景交互的表达与生成

数字人与场景的交互可以看做一类特殊的条件动作生成问题。在室内场景中,数字人通常需要完成避障漫游并且基于场景物体(例如床、凳等)的功能语义与其互动(Hassan等,2021a;Huang等,

2023a; Starke 等, 2019; Zhao 等, 2023a)。除此之外, 搬运、移动物体(Liu 和 Hodgins, 2018; Merel 等, 2020), 以及对小物件进行灵巧操作(Chen 等, 2023c; Zhang 等, 2021)等任务也是很有挑战性的场景交互问题。

面对不同的交互任务时, 场景中的对象往往采用不同的几何表示方式。由于交互和生成都在数字场景中进行, 许多数字人一场景交互生成工作假设场景状态信息可以从仿真器中准确获取, 从而极大地降低问题难度。例如当操作对象为刚体时, 可以很方便地使用物体的位置、朝向等信息作为物体的表达方式(Hassan 等, 2023; Liu 和 Hodgins, 2018; Zhang 等, 2023a)。更加通用的交互任务则需要使用可泛化的物体形状表达, 例如进行运动避障时, 可以用较大的体素来描述场景中存在的障碍物(Hassan 等, 2021a; Starke 等, 2019; Zhao 等, 2023a), 而在与物体进行交互时, 则采用精细的面片(Hassan 等, 2021b; Wang 等, 2021)、较小的体素(Zhang 等, 2021)、点云(Wang 等, 2022a; Wang 等, 2022c; Zhao 等, 2022)等更加细致的局部表达。也有部分工作采用深度图(Zhang 等, 2020b)、基点集(Zhang 等, 2020a), 甚至直接通过渲染模拟视觉输入(Merel 等, 2020)等方式来表达场景。而数字人本身在交互过程中则通常会采用简化的刚性碰撞体(Starke 等, 2021)、骨骼姿态(Starke 等, 2020; Zhang 等, 2021)、人体表面采样点(Zhao 等, 2023a), 或是精细的表面皮肤面片表达(Hassan 等, 2019, 2021b; Wang 等, 2021; Wang, 2022b)。

基于场景交互的运动生成的一大难点在于交互方式的多样性。对于相同的物体, 角色的交互方式有很多种可能, 由此产生的数据需求和歧义性问题给生成模型的训练带来较大难度。条件变分自编码器(conditional variational autoencoder, cVAE)(Zhang 等, 2020a; Zhang, 2020b)等生成模型可以有效解决这一问题。以场景作为输入条件, 对符合要求的角色姿态和全局位置信息进行压缩编码, 在应用时, 只需要给定场景的信息, 就能自动生成对应可能的角色信息。但是, 以场景为条件来预测角色交互的方式往往具有泛化性的问题。因为场景的表达非常复杂, 现有的数据规模有限, 训练容易过拟合。另一个研究思路则是由角色的姿态来预测与场景发生交互的肢体部位分布以及交互的物体类别, 再根据交互

信息在场景中寻找合适的放置位置。例如以角色的姿态和体型参数作为条件, 训练cVAE来预测身体表面顶点与环境的交互信息(Hassan 等, 2021b), 编码器对身体表面顶点坐标以及接触物体标签进行压缩编码, 预测时以当前姿态作为条件, 采样得到各种可能的交互情况。再到场景中寻找对应交互的物体, 通过优化的方式寻找最合适的放置位置。

相比于生成瞬时的交互姿态, 基于环境约束生成一整段交互运动则更加具有挑战。一种直观的方式是将问题拆分成关键帧生成和帧间补全两项任务(Wang 等, 2021), 先使用上述瞬时交互姿态生成的方法, 根据场景的点云预测关键帧处角色的姿态, 然后利用条件动作生成技术生成中间的运动。对于复杂场景, 可以先使用A*算法进行路径规划, 然后根据路径目标生成中间动作(Hassan 等, 2021a)。而另一种方式则是通过强化学习来主动学习如何在场景约束下进行运动(Zhao 等, 2023a), 先基于大规模数据训练运动表征空间, 然后以环境信息作为状态, 运动表征空间作为行为空间, 与环境的交互、避障等作为目标来定义奖励函数, 学习与环境交互的运动策略。除此之外, 基于语言的动作生成任务也可以扩展到场景交互任务上。通过自然语言描述角色与场景的交互方式, 例如:“坐在椅子上”、“触摸桌子”, 并通过交互方式和对象的语言描述来预测角色的姿态以及交互的部位(Zhao 等, 2022), 也可以根据场景和给定的运动描述生成全程的运动(Wang 等, 2022a; Wang, 2022c)。

3.2.3 挑战与机遇

数字人一场景交互生成领域发展迅速, 但仍然存在很多问题和挑战。由于在仿真空间中获取准确环境信息较为容易, 许多现有系统感知环境信息的方法是利用抽象过的简单特征(如, 人与物体表面的距离等)。如何结合大规模多模态模型直接通过图像甚至视频来感知室内环境上下文, 有希望挖掘出更深的语义, 便于处理更复杂的行为。此外, 人与场景的交互难以仅靠外部信息(如, 角色物理状态、空间环境信息等)进行完整的精确建模, 因为完成交互的过程还涉及人的内在心理意图, 如何挖掘并结合这部分信息是一项挑战。在数据驱动的大背景下, 足够量级且多元的数据是确保构建一个足够鲁棒的拟人交互系统的必要条件。如何提高数据获取效率, 以及如何使用其他模态(如视频、文字等)信息进

行学习是一个值得探索的问题。在算法方面,目前该领域的算法尚处于“百家争鸣”阶段,许多工作在网络结构和训练方法上不断创新。不过,受语言、视觉大模型在“规模化(*scaling*)”上成果的启发,扩大模型的规模,构建动作、交互的大模型也是一个潜在的研究方向。但如何对整个环境、角色的信息进行有效表征,对于构建大模型来说仍是一个困难的问

题。最后,目前该领域缺少统一的、受广泛认可的测试基准(*benchmark*)。参考其他领域的发展规律,良好的测试基准有助于整个领域知识的快速迭代和改进。在数字人—场景交互生成领域,任务的多样性给建立通用的测试基准带来很大的挑战,如何建立公认有效的测试和评估方法则是一个非常有意义的问题。数字人—场景交互生成方法如表2所示。

表2 数字人—场景交互生成方法
Table 2 Digital human-scene interaction generation methods

方法	交互对象	交互任务	生成方法	场景表达	物理仿真
Holden 等人(2017),Zhang 等人(2018)	无	指令移动	自回归	轨迹高度值	否
Starke 等人(2019)	箱子,座椅	指令移动	自回归	体素	否
Starke 等人(2020)	篮球	指令移动	自回归	交互点	否
Starke 等人(2021)	拳击	指令移动	自回归	接触点	否
Zhang 等人(2021)	手持物体	物体交互	自回归	体素,SDF	否
Ling 等人(2020)	无	指令移动	层次化控制	轨迹,目标	否
Zhang 等人(2020b)	场景物体	场景内放置	单帧生成+cVAE	深度图	否
Zhang 等人(2020a)	场景物体	场景内放置	单帧生成+cVAE	Basis Point Sets (BPS)	否
Hassan 等人(2021b)	场景物体	场景内放置	单帧生成+cVAE	无	否
Zhao 等人(2022)	场景物体	场景内放置	语言描述+self attention	点云	否
Wang 等人(2021)	场景物体	场景内移动	单帧生成+帧间补全	面片	否
Hassan 等人(2021a)	场景物体	场景内移动	轨迹规划+自回归	体素	否
Wang 等人(2022a)	场景物体	场景内移动	语言描述+单帧生成+帧间补全	点云	否
Wang 等人(2022c)	场景物体	场景内移动	语言描述+自回归	点云	否
Zhao 等人(2023a)	场景物体	场景内移动	强化学习	体素	否
Hasson 等人(2019)	场景物体	场景内交互	强化学习	包围盒	是
Peng 等人(2018,2021,2022), Yao 等人(2022)	无	指令移动	强化学习	无	是
Yang 等人(2022b)	筷子	指令移动	强化学习	无	是
Liu 和 Hodgins (2018)	篮球	指令移动	强化学习	位置,速度	是
Zhang 等人(2023a)	网球	指令移动	强化学习	位置,速度	是

3.3 机器人—场景交互生成

本节讨论机器人与场景交互生成方面的工作,并将重点放在机器人对各种不同类型物体的近距离操控任务上。从该领域目前主要的研究任务和解决任务使用的通用方法对现有研究工作进行简要介绍。

3.3.1 机器人—场景交互生成任务

根据交互生成层级的差异性,机器人和场景的交互任务一般可以分为两个层面,一个是原子技能层面,一个是组合技能层面。原子技能层面的交互致力于改变单个操作对象的位置、状态等自身属性;

而组合技能层面的交互则更加侧重于完成一个整体的目标,这个过程一般通过组合若干个原子技能级任务来完成。

在原子技能交互任务中,由于交互物体自身性质不同,机器人与其交互方式也会有所不同。对于最常见的刚性物体,机器人往往考虑的是如何抓取该物体。过去的很多工作主要研究如何使用夹爪(Kalashnikov 等,2018;Mahler 等,2017)或者灵巧手(Liu 等,2020;Mandikal 和 Grauman,2021;She 等,2022;Wan 等,2023;Xu 等,2023)来完成物体抓取,

以及基于抓取的操作(Liu等, 2022c; Liu等, 2023; Qin等, 2022a, b)。灵巧手因其灵活的手指结构, 可以完成一些复杂的手内物体操作, 这一方面被广泛研究的问题是如何使用手指旋转物体(Andrychowicz等, 2020; Chen等, 2022a; Yin等, 2023), 如转笔、盘核桃等任务(Nagabandi等, 2020; Rajeswaran等, 2018)。此外, 最近一个双手操作任务的模拟平台被提出并得到了广泛关注(Chen等, 2022e)。除了单个刚性的物体, 机器人和一些具有关节物体的交互也得到了广泛的关注(Geng等, 2023a; Geng等, 2023c; Wang等, 2022b; Xu等, 2022)。这些关节物体包含了很多日常生活中的物体, 如抽屉、门、笔记本等。机器人需要学会这些物体的使用方式, 让其按照关节设计的方式运动。机器人和可形变物体的交互也是近年比较活跃的研究领域。其中, 与衣物等布料类物体的交互比较有代表性, 任务包括如何使用单臂(Wu等, 2023)或者双臂(Ha和Song, 2022)将布料摊平、折叠(Maitin-Shepard等, 2010)等。此外, 另一类研究较多的可形变物体是弹塑性(elasto-plastic)物体, 如橡皮泥、面团等, 这里研究的问题一般是如何让机器人将通过挤压(Qi等, 2022; Shi等, 2022)以及组合方法(Lin等, 2022; Shi等, 2023)将其变为指定形状。更多关于可形变物体的交互可以参考Yin等人(2021)的综述。

组合技能交互任务的主要侧重点在于如何规划和组合不同技能级交互方式, 进而完成复杂的任务, 如家具装配等。随着大模型的兴起, 如何借助大模型强大的常识推理能力来解决长序列任务的规划成为一个十分热门的话题(Brohan等, 2023; Driess等, 2023)。在这一类任务中, 机器人往往可能需要在室内的不同地点完成不同的任务。为了扩大机器人交互的空间, 一些工作会将机械臂安装到一个可移动的基座上, 使得机器人能够在更大的室内空间移动完成复杂的连续交互任务。这些工作已经展示机器人能够在一个地点拿起物体, 在另外一个地点放下物体(Yang等, 2023a; Yenamandra等, 2023; Yokoyama等, 2023)。此外, 一些工作发现在执行交互任务时考虑移动模块本身的运动范围(Jauhri等, 2022; Zhang等, 2023b), 甚至利用移动模块额外提供的灵巧性(Fu等, 2023), 可以帮助操作模块更好地完成交互任务。

3.3.2 机器人—场景交互生成方法

在机器人与场景交互的具体方法层面, 一般可以

根据其是否会根据感知反馈即时调整目标, 分为开环(open-loop)方法和闭环(closed-loop)方法。开环方法一般需要先设定机器人和场景的交互位置, 再控制自身姿态运动到指定的交互位置, 在执行过程中不会调整目标, 因此交互对象状态的变化常常会导致执行失败, 比较适用于静态的、低频次的交互任务; 闭环方法中, 机器人一般不会提前规划交互位置, 而是会根据当前的感知调整自身行为和目标, 对各种误差和干扰更加鲁棒, 因此适用于动态、高频次的交互任务。

在基于开环方法的原子技能交互任务中, 交互位置一般根据用户要求和感知信息计算得到。一些抓取工作使用无模型的搜索算法(Miller和Allen, 2004)求解交互位置。对于更广泛层面交互任务来说, 常见的做法是通过数据训练模型去预测场景的交互位置(Brohan等, 2023; Liu等, 2020; Liu, 2022c)。由于完成相同任务的交互方式可能有多种, 为了保证生成结果的多样性, 一些方法选择预测给定位置的成功率, 然后配合搜索算法得到交互位置(Mahler等, 2017); 另外一些方法则使用生成式模型来保证预测位置的多样性(Liu等, 2023)。控制机器人到达指定位置的方法可以使用现有运动规划器来完成。

在基于闭环方法的原子技能交互任务中, 经常会使用强化学习方法来学习一个感知到动作映射的策略函数来控制机器人(Chen等, 2022a; Geng等, 2023c; Mandikal和Grauman, 2021; Qi等, 2022; Wan等, 2023)。模仿学习是另一类学习策略的方法, 往往会和强化学习一起使用以提升其学习效率和结果质量。一些灵巧手操作工作通过将人类手部轨迹迁移到特定的灵巧手上, 使用示例增强的强化学习方法(Rajeswaran等, 2018)让机器人能够更快更好地学会抓取和开门等任务(Qin等, 2022a, b)。另外一些工作尝试从人类轨迹中构建出强化学习所需的奖励函数来帮助机器人学习推动物体等任务(Kumar等, 2023; Zakka等, 2022)。还有一些工作免除了强化学习的训练, 它们的策略是直接通过最近邻搜索当前感知和示例感知的最相似的帧对应的动作并执行(Arunachalam等, 2023; Pari等, 2021)。不同类型物体的原子技能级交互如图2所示。

对于组合技能交互任务中的规划问题, 经典方法通过符号规划(Nau等, 1999)和优化方法来解决(Toussaint, 2015), 但是这种方法往往需要显式地定义下层任务以及约束条件。基于学习的方法可以尝



图2 不同类型物体的原子技能级交互

Fig. 2 Skill-level interaction with different kinds of objects ((a) in-hand manipulation with the rigid object (Andrychowicz et al., 2020); (b) manipulation with the articulated object (Geng et al., 2023a); (c) manipulation with the deformable object (Shi et al., 2023))

试用于解决这个问题。最近,大模型被运用到这个问题中并取得良好的效果(Huang等,2019;Xu等,2018)。大模型的优势在于具有强大的推理能力和常识储备,对各种长序列任务都能够做到很好的拆分和规划,但是由于缺乏对场景以及机器人本体的认知能力,生成的规划步骤很难成功在场景中执行。为了解决这个问题,一些工作侧重增强大模型对场

景的感知能力(Brohan等,2023;Driess等,2023),而另外一些工作则将大模型与任务层面(Ahn等,2022)和场景层面(Huang等,2023b)的可行性预测相结合来保证生成任务的可执行性。图3和图4分别展示了具有移动能力的机器人完成组合技能任务和提升大模型规划任务在特定场景中可执行性的方法。机器人一场景交互生成方法如表3所示。



图3 具有移动能力的机器人完成组合技能任务

Fig. 3 The mobile robot achieves complex scene interactions by combining various manipulation skills ((a) find the object within the initial container; (b) pick up the object from the initial container; (c) find the target container; (d) place the object in the target container)

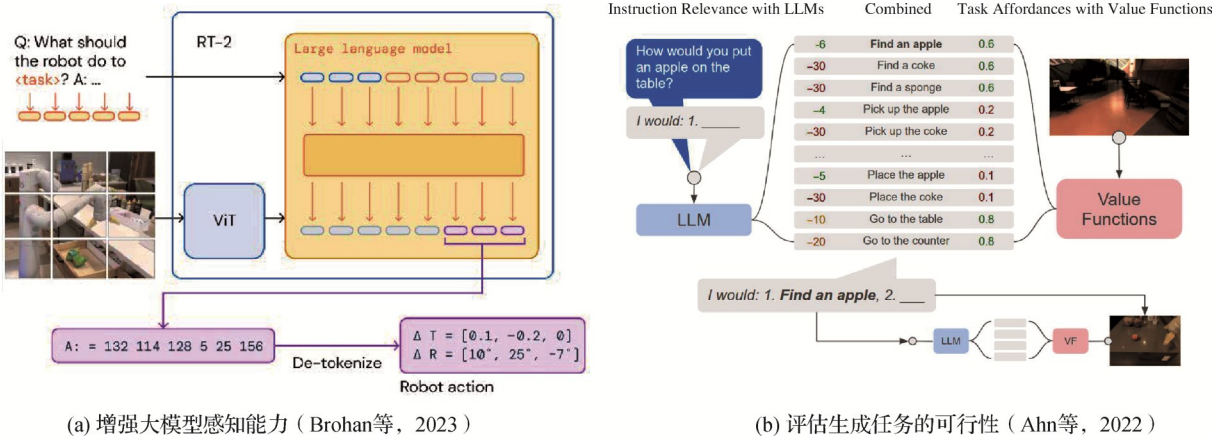


图4 提升大模型规划任务在特定场景中可执行性的方法

Fig. 4 Methods to improve the feasibility of large-scale model planning tasks in specific scenarios ((a) enhance the perception ability of large models (Brohan et al., 2023); (b) evaluate the feasibility of generating tasks (Ahn et al., 2022))

表 3 机器人—场景交互生成方法

Table 3 Robot-scene interaction generation methods

方法	交互主体	交互对象	任务类型	生成方法	数据类型	是否依赖仿真
Kalashnikov 等人(2018)	平行夹具	刚性物体	抓取	闭环方法	数据集 + 自监督	否
Mahler 等人(2017)	平行夹具	刚性物体	抓取	开环方法	数据集	否
Liu 等人(2022c)	平行夹具	刚性物体	调整位置	开环方法	数据集	是
Liu 等人(2023)	平行夹具	刚性物体	调整位置	开环方法	数据集	是
Zakka 等人(2022)	平行夹具	刚性物体	接近/推动	闭环方法	数据集 + 自监督	是
Kumar 等人(2023)	平行夹具	刚性物体	接近/推动	闭环方法	数据集 + 自监督	是
Liu 等人(2020)	灵巧手	刚性物体	抓取	开环方法	数据集 + 自监督	否
Mandikal 和 Grauman(2021)	灵巧手	刚性物体	抓取	闭环方法	数据集 + 自监督	是
Qi 等人(2022)	灵巧手	刚性物体	抓取	闭环方法	数据集 + 自监督	是
Wan 等人(2023)	灵巧手	刚性物体	抓取	闭环方法	自监督	是
Qin 等人(2022b)	灵巧手	刚性物体	操作	闭环方法	数据集 + 自监督	是
Qin 等人(2022a)	灵巧手	刚性物体	操作	闭环方法	数据集 + 自监督	是
Andrychowicz 等人(2020)	灵巧手	刚性物体	手内操作	闭环方法	自监督	是
Chen 等人(2022a)	灵巧手	刚性物体	手内操作	闭环方法	自监督	是
Yin 等人(2023)	灵巧手	刚性物体	手内操作	闭环方法	自监督	是
Arunachalam 等人(2023)	灵巧手	刚性物体	手内操作	闭环方法	数据集 + 自监督	是
Rajeswaran 等人(2018)	灵巧手	刚性物体	手内操作	闭环方法	数据集 + 自监督	是
Nagabandi 等人(2020)	灵巧手	刚性物体	手内操作	闭环方法	自监督	是
Geng 等人(2023a)	平行夹具	关节物体	关节运动	闭环方法	自监督	是
Geng 等人(2023c)	平行夹具	关节物体	关节运动	闭环方法	自监督	是
Wang 等人(2022b)	平行夹具	关节物体	关节运动	闭环方法	自监督	是
Xu 等人(2022)	平行夹具	关节物体	关节运动	闭环方法	自监督	是
Pari 等人(2021)	平行夹具	关节物体	关节运动	闭环方法	数据集	是
Maitin-Shepard 等人(2010)	双臂系统	布料类物体	折叠	开环方法	启发式方法	否
Ha 和 Song(2022)	双臂系统	布料类物体	展开	闭环方法	自监督	是
Wu 等人(2023)	平行夹具	布料类物体	展开	闭环方法	自监督	是
Qi 等人(2022)	定制工具	弹塑类物体	改变成指定形状	闭环方法	自监督	是
Shi 等人(2022)	平行夹具	弹塑类物体	改变成指定形状	闭环方法	自监督	是
Lin 等人(2022)	定制工具	弹塑类物体	改变成指定形状	闭环方法	自监督	是
Shi 等人(2023)	定制工具	弹塑类物体	改变成指定形状	闭环方法	自监督	是
Ahn 等人(2022)	移动机器人	各种物体	完成人类指令	开环方法	大模型	是
Driess 等人(2023)	移动机器人	各种物体	完成人类指令	开环方法	大模型	是
Brohan 等人(2023)	移动机器人	各种物体	完成人类指令	闭环方法	大模型	是
Yenamandra 等人(2023)	移动机器人	各种物体	拿起—放下	开环方法	大模型(开放词)	是
Huang 等人(2023b)	平行夹具	各种物体	完成人类指令	开环方法	大模型	是

4 发展趋势与展望

1)交互仿真。随着计算机算力的提升以及拟人交互不断产生的新兴需求,仿真领域也在不断更新和发展,主要包括以下发展趋势:可扩展、可微分、可学习。首先,近期的拟人交互工作逐渐开始探索在复杂物理系统上的学习任务(Heiden等,2021;Huang等,2021;Lin等,2021;Xian等,2022)。不断涌现的新任务对仿真平台的可扩展性提出了更高的要求。从某种意义上讲,拟人交互社区期待一个统一的仿真框架可以处理刚体、软体、流体及其耦合的物理系统。其次,仿真引擎逐渐开始关注提供或利用仿真的微分信息(Freeman等,2021;Werling等,2021;Chen等,2022b;Li等,2023d,Shen等,2021)。这些微分输出提供了关于动力系统的高阶信息,为由动力学方程约束的优化问题提供了新的工具。如何利用好可微仿真提供的这些额外信息依然是一个开放的研究问题,相关工作仍在探索中。最后,随着仿真任务越来越复杂,计算难度越来越高,不少工作开始探索将经典的物理仿真和深度学习方法的结合(Chu和Thuerey,2017;Karniadakis等,2021;Pfaff等,2021;Raissi等,2019,2020),一方面加速乃至替代昂贵的数值计算,另一方面让仿真工具和感知、控制等模块的深度学习方法无缝衔接,其未来值得关注。

2)交互数据。目前国内外关于如何获取大规模高质量的场景交互相关数据主要聚焦在可交互数字场景的自动生成以及交互运动的高效采集与自动扩充上。在交互感知数据的构建方面,大家越来越多地意识到传统三维数字资产的局限性,无论是从物体还是场景的数量及多样性上,都距离服务真实场景需求很远。可交互数字场景的自动生成也随之成为重要的研究方向,从物体层面,借助基础生成模型,越来越多的工作可以仅依靠文本或图像输入生成细节丰富几何逼真的三维模型,保证这些三维数字资产的可交互性,使其具有可控的结构、运动、可供性特征也随之成为重要的研究课题;从场景层面,基础生成模型一样带来了许多新的机遇与挑战,如何突破现有场景数据集的局限,生成在几何、外观、布局、物理属性等层面逼真的三维场景,更好地服务从虚拟到真实的交互感知迁移得到了广泛关注。在

交互运动数据的构建方面,一方面降低现有交互数据采集系统的成本、增强其易用性以方便大规模部署采集数据仍是亟待解决的问题;另一方面,利用仿真、重定向和生成等手段,更加自动化大规模的生成交互运动示例,也为突破现有交互数据局限性带来了新的可能。

3)交互感知。未来的可供性研究后续几个重要趋势发展包括:开集学习、小样本学习和自监督学习。首先,随着技术的发展,开集学习在可供性研究中的重要性将日益增强。这涉及到让机器学习模型能够有效处理和识别在训练数据中未出现的新物体或场景,增强模型在现实世界多变环境中的适用性和鲁棒性。其次,在可供性研究中,小样本学习的重要性将继续增长,特别是在资源有限或数据稀缺的环境中,通过提高模型从少量数据中学习和推广知识的能力,可以大幅提升算法的效率和实用性。最后,自监督学习方法在可供性研究中的应用将继续扩大,这种方法可以减少对大量标注数据的依赖,使模型能够从未标注的数据以及仿真环境中自主学习和提取有用信息,这对于处理复杂和多样化的实际场景尤其重要。伴随这些发展趋势,可供性研究将进一步推动人工智能和机器人技术的发展,尤其是在提高自动化系统对复杂环境的适应性和灵活性方面。这将有助于创造出更加智能、直观且能够与人类高效协作的机器和系统。

4)交互生成。拟人交互领域后续发展方向主要是往更多的自主性、更强的适应能力、更深的人机交互、更高的交互安全和更广的应用领域。随着机器学习和人工智能的发展,智能体正在变得更加自主,它们能够在更大程度上理解和适应环境、自我学习和改进其行为。智能体需要在各种不确定和动态的环境中工作,包括对未知环境的快速建模和理解、对环境变化的实时响应、对异常情况的处理等。此外,随着社区和家庭机器人的发展,机器人需要更深入地理解和与人互动,包括理解人的语言、情绪和需求,用自然和友好的方式与人交互,为人提供个性化的服务和帮助。与此同时,随着智能体与环境交互的复杂性增加,如何确保交互的安全性,仿真智能体对人或环境造成伤害也是需要考虑的关键问题。最后,随着技术的进步,能够进行自主交互的智能体的应用领域将会继续扩大,如何挖掘其在工业生产、医疗健康、家庭服务、娱乐休闲和教育培训等各大应用

领域的潜力并产生深远影响是需要长期思考的问题。

5 结 语

本文聚焦室内场景拟人交互,从平台、数据和算法3方面对近几年国内外发展现状进行系统梳理和概述。尽管室内场景拟人交互的研究近年来取得长足发展,但离真正的自主交互和智能演化还有非常大的距离,仍存在诸多亟待解决的问题和极大的提升空间。希望本文能帮助读者对室内场景拟人交互的研究发展过程中涉及到的多方面要素有较为全面的了解和认识,并在未来研究思路方面有所启发。

致 谢 本文由中国图象图形学学会智能图形专委会组织撰写,该专委会链接为<https://www.csig.org.cn/16/201801/49334.html>。太极图形的刘天添对仿真部分提出了很多宝贵的修改意见;蒋鸿达、敖腾隆、许少聪、余奇晋、苏向宇和乐齐钦参与了资料整理工作,一并致谢。

参考文献(References)

- Ackerman M J. 1998. The visible human project. *Proceedings of the IEEE*, 86(3): 504-511 [DOI: 10.1109/5.662875]
- Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, Finn C, Fu C Y, Gopalakrishnan K, Hausman K, Herzog A, Ho D, Hsu J, Ibarz J, Ichter B, Irpan A, Jang E, Ruano R J, Jeffrey K, Jesmonth S, Joshi N J, Julian R, Kalashnikov D, Kuang Y H, Lee K H, Levine S, Lu Y, Luu L, Parada C, Pastor P, Quiambao J, Rao K, Rettinghouse J, Reyes D, Sermanet P, Sievers N, Tan C, Toshev A, Vanhoucke V, Xia F, Xiao T, Xu P, Xu S C, Yan M Y and Zeng A. 2022. Do as I can, not as I say: grounding language in robotic affordances [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2204.01691.pdf>
- Akkaya I, Andrychowicz M, Chociej M, Litwin M, McGrew B, Petron A, Paino A, Plappert M, Powell G, Ribas R, Schneider J, Tezak N, Tworek J, Welinder P, Weng L L, Yuan Q M, Zaremba W and Zhang L. 2019. Solving rubik's cube with a robot hand [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/1910.07113.pdf>
- Alexanderson S, Nagy R, Beskow J and Henter G E. 2023. Listen, denoise, action! Audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics*, 42(4): #44 [DOI: 10.1145/3592458]
- Andrews S and Erleben K. 2021. Contact and friction simulation for computer graphics//ACM SIGGRAPH 2021 Courses. [s.l.]: ACM: #2 [DOI: 10.1145/3450508.3464571]
- Andrychowicz O M, Baker B, Chociej M, Józefowicz R, McGrew B, Pachocki J, Petron A, Plappert M, Powell G, Ray A, Schneider J, Sidor S, Tobin J, Welinder P, Weng L L and Zaremba W. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1): 3-20 [DOI: 10.1177/0278364919887447]
- Ao T L, Gao Q Z, Lou Y K, Chen B Q and Liu L B. 2022. Rhythmic gesticulator: rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics*, 41(6): 1-19 [DOI: 10.1145/3550454.3555435]
- Ao T L, Zhang Z Y and Liu L B. 2023. GestureDiffuCLIP: gesture diffusion model with CLIP latents. *ACM Transactions on Graphics*, 42(4): #42 [DOI: 10.1145/3592097]
- Arunachalam S P, Silwal S, Evans B and Pinto L. 2023. Dexterous imitation made easy: a learning-based framework for efficient dexterous manipulation//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, England: IEEE: 5954-5961 [DOI: 10.1109/icra48891.2023.10160275]
- Azadi S, Shah A, Hayes T, Parikh D and Gupta S. 2023. Make-animation: large-scale text-conditional 3D human motion generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 14993-15002 [DOI: 10.1109/ICCV51070.2023.01381]
- Bargteil A W, Shinar T and Kry P G. 2020. An introduction to physics-based animation//SIGGRAPH Asia 2020 Courses. [s.l.]: ACM: #5 [DOI: 10.1145/3415263.3419147]
- Barquero G, Escalera S and Palmero C. 2023. BelFusion: latent diffusion for behavior-driven human motion prediction//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2317-2327 [DOI: 10.1109/ICCV51070.2023.00220]
- Batty C, Bertails F and Bridson R. 2007. A fast variational framework for accurate solid-fluid coupling//ACM SIGGRAPH 2007 Papers. San Diego, USA: ACM: #100 [DOI: 10.1145/1275808.1276502]
- Becker M, Ihmsen M and Teschner M. 2009. Corotated SPH for deformable solids//Proceedings of the 5th Eurographics conference on Natural Phenomena. Munich, Germany: Eurographics Association: 27-34
- Bender J, Erleben K and Trinkle J. 2014. Interactive simulation of rigid body dynamics in computer graphics. *Computer Graphics Forum*, 33(1): 246-270 [DOI: 10.1111/cgf.12272]
- Bender J and Koschier D. 2015. Divergence-free smoothed particle hydrodynamics//The 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Los Angeles, USA: ACM: 147-155 [DOI: 10.1145/2786784.2786796]
- Bhatnagar B L, Xie X H, Petrov I A, Sminchisescu C, Theobalt C and Pons-Moll G. 2022. BEHAVE: dataset and method for tracking human object interactions//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans,

- USA: IEEE: 15914-15925 [DOI: 10.1109/cvpr52688.2022.01547]
- Bouaziz S, Martin S, Liu T T, Kavan L and Pauly M. 2014. Projective dynamics: fusing constraint projections for fast simulation. *ACM Transactions on Graphics*, 33(4): #154 [DOI: 10.1145/2601097.2601116]
- Brahmbhatt S, Ham C, Kemp C C and Hays J. 2019. ContactDB: analyzing and predicting grasp contact via thermal imaging//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA: IEEE: 8701-8711 [DOI: 10.1109/cvpr.2019.00891]
- Brahmbhatt S, Tang C C, Twigg C D, Kemp C C and Hays J. 2020. ContactPose: a dataset of grasps with object contact and hand pose//*Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer: 361-378 [DOI: 10.1007/978-3-030-58601-0_22]
- Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Chormanski K, Ding T L, Driess D, Dubey A, Finn C, Florence P, Fu C Y, Arenas M G, Gopalakrishnan K, Han K H, Hausman K, Herzog A, Hsu J, Ichter B, Irpan A, Joshi N, Julian R, Kalashnikov D, Kuang Y H, Leal I, Lee L, Lee T W E, Levine S, Lu Y, Michalewski H, Mordatch I, Pertsch K, Rao K, Reymann K, Ryoo M, Salazar G, Sanketi P, Sermanet P, Singh J, Singh A, Soricut R, Tran H, Vanhoucke V, Vuong Q, Wahid A, Welker S, Wohlhart P, Wu J L, Xia F, Xiao T, Xu P, Xu S C, Yu T H and Zitkovich B. 2023. RT-2: vision-language-action models transfer web knowledge to robotic control [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2307.15818.pdf>
- Büttner M. 2015. Motion matching-the road to next gen animation [EB/OL]. [2023-12-20]. https://www.youtube.com/watch?v=z_wpgHFSWss&t=658s
- Catto E. 2023. Box2D [EB/OL]. [2023-12-20]. <https://github.com/erincatto/box2d>
- Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, Savarese S, Savva M, Song S R, Su H, Xiao J X, Yi L and Yu F. 2015. Shapenet: an information-rich 3d model repository [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/1512.03012.pdf>
- Chao Y W, Yang W, Xiang Y, Molchanov P, Handa A, Tremblay J, Narang Y S, Van Wyk K, Iqbal U, Birchfield S, Kautz J and Fox D. 2021. DexYCB: a benchmark for capturing hand grasping of objects//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA: IEEE: 9040-9049 [DOI: 10.1109/cvpr46437.2021.00893]
- Chen J, Gao D F, Lin K Q and Shou M Z. 2023a. Affordance grounding from demonstration video to target image//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, Canada: IEEE: 6799-6808 [DOI: 10.1109/CVPR52729.2023.00657]
- Chen L H, Zhang J W, Li Y W, Pang Y R, Xia X B and Liu T L. 2023b. HumanMAC: masked motion completion for human motion prediction//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE: 9510-9521 [DOI: 10.1109/ICCV51070.2023.00875]
- Chen S R, Wu A and Liu C K. 2023c. Synthesizing dexterous nonprehensile pregrasp for ungraspable objects//*Proceedings of 2023 ACM SIGGRAPH Conference*. Los Angeles, USA: Association for Computing Machinery: #10 [DOI: 10.1145/3588432.3591528]
- Chen T, Xu J and Agrawal P. 2022a. A system for general in-hand object Re-orientation//*Proceedings of 2022 Conference on Robot Learning*. London, UK: PMLR: 297-307
- Chen X, Jiang B, Liu W, Huang Z L, Fu B, Chen T and Yu G. 2023d. Executing your commands via motion diffusion in latent space//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada: IEEE: 18000-18010 [DOI: 10.1109/cvpr52729.2023.01726]
- Chen X W, Ni X Y, Zhu B, Wang B and Chen B Q. 2022b. Simulation and optimization of magnetoelastic thin shells. *ACM Transactions on Graphics*, 41(4): #61 [DOI: 10.1145/3528223.3530142]
- Chen X X, Liu T Y, Zhao H, Zhou G Y and Zhang Y Q. 2022c. Cerberus Transformer: joint semantic, affordance and attribute parsing//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA: IEEE: 19617-19626 [DOI: 10.1109/cvpr52688.2022.01903]
- Chen Y N, Li M C, Lan L, Su H, Yang Y and Jiang C F F. 2022d. A unified newton barrier method for multibody dynamics. *ACM Transactions on Graphics*, 41(4): #66 [DOI: 10.1145/3528223.3530076]
- Chen Y P, Wu T H, Wang S J, Feng X D, Jiang J C, Lu Z Q, McAleer S, Dong H, Zhu S C and Yang Y D. 2022e. Towards human-level bimanual dexterous manipulation with reinforcement learning//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans, USA: NeurIPS: 5150-5163
- Chu M Y and Thuerey N. 2017. Data-driven synthesis of smoke flows with CNN-based feature descriptors. *ACM Transactions on Graphics*, 36(4): #69 [DOI: 10.1145/3072959.3073643]
- Coumans E and Bai Y. 2021. PyBullet, a python module for physic simulation for games, robotics and machine learning [EB/OL]. [2023-12-20]. <http://pybullet.org>
- Dabral R, Mughal M H, Golyanik V and Theobalt C. 2023. MoFusion: a framework for denoising-diffusion-based motion synthesis//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada: IEEE: 9760-9770 [DOI: 10.1109/cvpr52729.2023.00941]
- Damen D, Doughty H, Farinella G M, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W and Wray M. 2018. Scaling egocentric vision: the EPIC-KITCHENS dataset//*Proceedings of the 15th European Conference on Computer Vision*. Munich, Germany: Springer: 753-771 [DOI: 10.1007/978-3-030-01225-0_44]

- Damen D, Doughty H, Farinella G M, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W and Wray M. 2021. The EPIC-KITCHENS dataset: collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 4125-4141 [DOI: 10.1109/tpami.2020.2991965]
- Deng S H, Xu X, Wu C Z, Chen K and Jia K. 2021. 3D AffordanceNet: a benchmark for visual object affordance understanding//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA: IEEE: 1778-1787 [DOI: 10.1109/cvpr46437.2021.00182]
- Drissi D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T H, Huang W L, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I and Florence P. 2023. PaLM-E: an embodied multimodal language model//*Proceedings of the 40th International Conference on Machine Learning*. Honolulu, USA: PMLR: 8469-8488
- Erez T, Tassa Y and Todorov E. 2015. Simulation tools for model-based robotics: comparison of bullet, Havok, MuJoCo, ODE and PhysX//*Proceedings of 2015 IEEE International Conference on Robotics and Automation (ICRA)*. Seattle, USA: IEEE: 4397-4404 [DOI: 10.1109/icra.2015.7139807]
- Fan Z C, Taheri O, Tzionas D, Kocabas M, Kaufmann M, Black M J and Hilliges O. 2023. ARCTIC: a dataset for dexterous bimanual hand-object manipulation//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada: IEEE: 12943-12954 [DOI: 10.1109/cvpr52729.2023.01244]
- Fang H J, Fang H S, Wang Y M, Ren J J, Chen J J, Zhang R, Wang W M and Lu C W. 2023a. Low-cost exoskeletons for learning whole-arm manipulation in the wild [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2309.14975.pdf>
- Fang H S, Fang H J, Tang Z Y, Liu J R, Wang C X, Wang J B, Zhu H Y and Lu C W. 2023b. RH20T: a comprehensive robotic dataset for learning diverse skills in one-shot//*Proceedings of the 7th Conference on Robot Learning (CoRL 2023)*. Atlanta, USA: CoRL: #9
- Faure F, Duriez C, Delingette H, Allard J, Gilles B, Marchesseau S, Talbot H, Courtécuisse H, Bousquet G, Peterlik I and Cotin S. 2012. SOFA: a multi-model framework for interactive physical simulation//Payan Y, ed. *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*. Berlin Heidelberg, Germany: Springer: 283-321 [DOI: 10.1007/8415_2012_125]
- Featherstone R. 1984. *Robot dynamics algorithms*. Edinburgh, UK: The University of Edinburgh
- Ferguson Z, Li M C, Schneider T, Gil-Ureta F, Langlois T, Jiang C F, Zorin D, Kaufman D M and Panozzo D. 2021. Intersection-free rigid body dynamics. *ACM Transactions on Graphics*, 40(4): #183 [DOI: 10.1145/3450626.3459802]
- Foster N and Fedkiw R. 2001. Practical animation of liquids//*Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. Los Angeles, USA: ACM: 23-30 [DOI: 10.1145/383259.383261]
- Freeman D, Frey E, Raichuk A, Girgin S, Mordatch I and Bachem O. 2021. Brax—a differentiable physics engine for large scale rigid body simulation//*Proceedings of the 1st Neural Information Processing Systems Track on Datasets and Benchmarks 1*. [s.l.]: NeurIPS: #404
- Fu Z P, Cheng X X and Pathak D. 2023. Deep whole-body control: learning a unified policy for manipulation and locomotion//*Proceedings of the 6th Conference on Robot Learning*. Auckland, New Zealand: PMLR: 138-149
- Garcia-Hernando G, Yuan S X, Baek S and Kim T K. 2018. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE: 409-419 [DOI: 10.1109/cvpr.2018.00050]
- Gästrin J. 2004. *Physically based character simulation—rag doll behaviour in computer games*. Stockholm, Sweden: Royal Institute of Technology
- Geng H R, Li Z M, Geng Y R, Chen J Y, Dong H and Wang H. 2023a. PartManip: learning cross-category generalizable part manipulation policy from point cloud observations//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada: IEEE: 2978-2988 [DOI: 10.1109/cvpr52729.2023.00291]
- Geng H R, Xu H L, Zhao C Y, Xu C, Yi L, Huang S Y and Wang H. 2023b. GAPartNet: cross-category domain generalizable object perception and manipulation via generalizable and actionable parts//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada: IEEE: 7081-7091 [DOI: 10.1109/cvpr52729.2023.00684]
- Geng Y R, An B S, Geng H R, Chen Y P, Yang Y D and Dong H. 2023c. RLAfford: end-to-end affordance learning for robotic manipulation//*Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, England: IEEE: 5880-5886 [DOI: 10.1109/icra48891.2023.10161571]
- Google DeepMind. 2023. MuJoCo 3 [EB/OL]. [2023-12-20]. <https://github.com/google-deeppmind/mujoco/discussions/1101>
- Goyal R, Ebrahimi Kahou S, Michalski V, Materzynska J, Westphal S, Kim H, Haenel V, Fruend I, Yianilos P, Mueller-Freitag M, Hoppe F, Thureau C, Bax I and Memisevic R. 2017. The “something something” video database for learning and evaluating visual common sense//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE: 5843-5851 [DOI: 10.1109/iccv.2017.622]
- Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X Y, Martin M, Nagarajan T, Radosavovic I, Ramakrishnan S K, Ryan F, Sharma J, Wray M,

- Xu M M, Xu E Z, Zhao C, Bansal S, Batra D, Cartillier V, Crane S, Do T, Doulaty M, Erapalli A, Feichtenhofer C, Fu Q C, Gebreselasie A, González C, Hillis J, Huang X H, Huang Y F, Jia W Q, Khoo W, Kolányi J, Kottur S, Kumar A, Landini F, Li C, Li Y H, Li Z Q, Mangalam K, Modhugu R, Munro J, Murrell T, Nishiyasu T, Price W, Puentes P R, Ramazanova M, Sari L, Somasundaram K, Southerland A, Sugano Y, Tao R J, Vo M, Wang Y C, Wu X D, Yagi T, Zhao Z W, Zhu Y Y, Arbeláez P, Crandall D, Damen D, Farinella G M, Fuegen C, Ghanem B, Ithapu V K, Jawahar C V, Joo H, Kitani K, Li H Z, Newcombe R, Oliva A, Park H S, Rehg J M, Sato Y, Shi J B, Shou M Z, Torralba A, Torresani L, Yan M F and Malik J. 2022. Ego4D: around the world in 3,000 hours of egocentric video//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18973-18990 [DOI: 10.1109/CVPR52688.2022.01842]
- Ha H and Song S. 2022. FlingBot: the unreasonable effectiveness of dynamic manipulation for cloth unfolding//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 24-33
- Hampali S, Rad M, Oberweger M and Lepetit V. 2020. HONotate: a method for 3d annotation of hand and object poses//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3193-3203 [DOI: 10.1109/cvpr42600.2020.00326]
- Harvey F G, Yurick M, Nowrouzezahrai D and Pal C. 2020. Robust motion in-betweening. ACM Transactions on Graphics, 39(4): #60 [DOI: 10.1145/3386569.3392480]
- Hassan M, Ceylan D, Villegas R, Saito J, Yang J M, Zhou Y, Black M J. 2021a. Stochastic scene-aware motion prediction//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 11354-11364 [DOI: 10.1109/iccv48922.2021.01118]
- Hassan M, Choutas V, Tzionas D and Black M. 2019. Resolving 3D human pose ambiguities with 3D scene constraints//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2282-2292 [DOI: 10.1109/iccv.2019.00237]
- Hassan M, Ghosh P, Tesch J, Tzionas D and Black M J. 2021b. Populating 3D scenes by learning human-scene Interaction//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: 14703-14713 [DOI: 10.1109/cvpr46437.2021.01447]
- Hassan M, Guo Y R, Wang T W, Black M, Fidler S and Peng X B. 2023. Synthesizing physical character-scene interactions//Proceedings of 2023 ACM SIGGRAPH Conference. Los Angeles, USA: Association for Computing Machinery: #63 [DOI: 10.1145/3588432.3591525]
- Hasson Y, Varol G, Tzionas D, Kalevtykh I, Black M J, Laptev I and Schmid C. 2019. Learning joint reconstruction of hands and manipulated objects//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11799-11808 [DOI: 10.1109/cvpr.2019.01208]
- Heiden E, Macklin M, Narang Y, Fox D, Garg A and Ramos F. 2021. DiSECT: a differentiable simulation engine for autonomous robotic cutting//Proceedings of the 17th Robotics: Science and Systems. [s.l.]: Robotics: Science and Systems: #67 [DOI: 10.15607/RSS.2021.XVII.067]
- Henter G E, Alexanderson S and Beskow J. 2020. MoGlow: probabilistic and controllable motion synthesis using normalising flows. ACM Transactions on Graphics, 39(6): #236 [DOI: 10.1145/3414685.3417836]
- Ho J and Salimans T. 2022. Classifier-free diffusion guidance [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2207.12598.pdf>
- Holden D, Komura T and Saito J. 2017. Phase-functioned neural networks for character control. ACM Transactions on Graphics, 36(4): #42 [DOI: 10.1145/3072959.3073663]
- Holl P, Koltun V, Um K and Thuerey N. 2020. phiflow: a differentiable PDE solving framework for deep learning via physical simulations//Workshop on Differentiable Vision, Graphics, and Physics in Machine Learning at NeurIPS 2020. [s.l.]: [s.n.]
- Hu R Z, Li W C, Van Kaick O, Shamir A, Zhang H and Huang H. 2017. Learning to predict part mobility from a single static snapshot. ACM Transactions on Graphics, 36(6): #227 [DOI: 10.1145/3130800.3130811]
- Hu Y M, Anderson L, Li T M, Sun Q, Carr N, Ragan-Kelley J and Durand F. 2020. DiffTaichi: differentiable programming for physical simulation//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR: 1-18
- Huang D A, Nair S, Xu D F, Zhu Y K, Garg A, Li F F, Savarese S and Niebles J C. 2019. Neural task graphs: generalizing to unseen tasks from a single video demonstration//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8557-8566 [DOI: 10.1109/cvpr.2019.00876]
- Huang S Y, Wang Z, Li P H, Jia B X, Liu T Y, Zhu Y X, Liang W and Zhu S C. 2023a. Diffusion-based generation, optimization, and planning in 3D scenes//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 16750-16761 [DOI: 10.1109/CVPR52729.2023.01607]
- Huang W L, Wang C, Zhang R H, Li Y Z, Wu J J and Li F F. 2023b. VoxPoser: composable 3D value maps for robotic manipulation with language models//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 540-562
- Huang Y H, Taheri O, Black M J and Tzionas D. 2022. InterCap: joint markerless 3D tracking of humans and objects in interaction//Proceedings of the 44th DAGM German Conference on Pattern Recognition. Konstanz, Germany: Springer: 281-299 [DOI: 10.1007/

- 978-3-031-16788-1_18]
- Huang Z A, Hu Y M, Du T, Zhou S Y, Su H, Tenenbaum J B and Gan C. 2021. PlasticineLab: a soft-body manipulation benchmark with differentiable physics//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR; 1-18
- Ihmsen M, Orthmann J, Solenthaler B, Kolb A and Teschner M. 2014. SPH fluids in computer graphics//Proceedings of the 35th Annual Conference of the European Association for Computer Graphics. Strasbourg, France: Eurographics; 21-42 [DOI: 10.2312/egst.20141034]
- Jauhari S, Peters J and Chalvatzaki G. 2022. Robot learning of mobile manipulation with reachability behavior priors. IEEE Robotics and Automation Letters, 7(3): 8399-8406 [DOI: 10.1109/lra.2022.3188109]
- Jian J T, Liu X P, Li M Y, Hu R Z and Liu J. 2023. AffordPose: a large-scale dataset of hand-object interactions with affordance-driven hand pose//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE; 14667-14678 [DOI: 10.1109/ICCV51070.2023.01352]
- Jiang B, Chen X, Liu W, Yu J Y, Yu G and Chen T. 2023. Motion-GPT: human motion as a foreign language//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA; NeurIPS; #14795
- Jiang C F F, Schroeder C, Teran J, Stomakhin A and Selle A. 2016. The material point method for simulating continuum materials//ACM SIGGRAPH 2016 Courses. Anaheim, USA: ACM; #24 [DOI: 10.1145/2897826.2927348]
- Kalashnikov D, Irpan A, Pastor P, Ibarz J, Herzog A, Jang E, Quillen D, Holly E, Kalakrishnan M, Vanhoucke V and Levine S. 2018. QT-opt: scalable deep reinforcement learning for vision-based robotic manipulation [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2018.10293.pdf>
- Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S F and Yang L. 2021. Physics-informed machine learning. Nature Reviews Physics, 3(6): 422-440 [DOI: 10.1038/s42254-021-00314-5]
- Karunratnakul K, Preechakul K, Suwajanakorn S and Tang S Y. 2023. Guided motion diffusion for controllable human motion synthesis//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE; 2151-2162 [DOI: 10.1109/ICCV51070.2023.00205]
- Kim J and Pollard N S. 2011. Fast simulation of skeleton-driven deformable body characters. ACM Transactions on Graphics, 30(5): #121 [DOI: 10.1145/2019627.2019640]
- Kong H Y, Gong K H, Lian D Z, Mi M B and Wang X C. 2023. Priority-centric human motion generation in discrete latent space//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE; 14760-14770 [DOI: 10.1109/ICCV51070.2023.01360]
- Kumar S, Zamora J, Hansen N, Jangir R and Wang X L. 2023. Graph inverse reinforcement learning from diverse videos//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR; 55-66
- Kwon T, Tekin B, Stühmer J, Bogo F and Pollefeys M. 2021. H2O: two hands manipulating objects for first person interaction recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE; 10118-10128 [DOI: 10.1109/iccv48922.2021.00998]
- Lan L, Kaufman D M, Li M C, Jiang C F F and Yang Y. 2022a. Affine body dynamics: fast, stable and intersection-free simulation of stiff materials. ACM Transactions on Graphics, 41(4): #67 [DOI: 10.1145/3528223.3530064]
- Lan L, Ma G Q, Yang Y, Zheng C X, Li M C and Jiang C F F. 2022b. Penetration-free projective dynamics on the GPU. ACM Transactions on Graphics, 41(4): #29 [DOI: 10.1145/3528223.3530069]
- Lanczos C. 2012. The Variational Principles of Mechanics. North Chelmsford: Courier Corporation
- Lee J, Grey M X, Ha S, Kunz T, Jain S, Ye Y, Srinivasa S S, Stilman M and Liu C K. 2018. Dart: dynamic animation and robotics toolkit. The Journal of Open Source Software, 3(22): #500 [DOI: 10.21105/joss.00500]
- Lee S, Park M, Lee K and Lee J. 2019. Scalable muscle-actuated human simulation and control. ACM Transactions on Graphics, 38(4): #73 [DOI: 10.1145/3306346.3322972]
- Li G, Jampani V, Sun D Q and Sevilla-Lara L. 2023a. LOCATE: localize and transfer object parts for weakly supervised affordance grounding//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE; 10922-10931 [DOI: 10.1109/CVPR52729.2023.01051]
- Li M C, Ferguson Z, Schneider T, Langlois T, Zorin D, Panozzo D, Jiang C F F and Kaufman D M. 2020. Incremental potential contact: intersection-and inversion-free, large-deformation dynamics. ACM Transactions on Graphics, 39(4): #49 [DOI: 10.1145/3386569.3392425]
- Li M C, Kaufman D M and Jiang C F F. 2021. Codimensional incremental potential contact. ACM Transactions on Graphics, 40(4): #170 [DOI: 10.1145/3450626.3459767]
- Li P F, Tian B W, Shi Y L, Chen X X, Zhao H, Zhou G Y and Zhang Y Q. 2022a. TOIST: task oriented instance segmentation transformer with noun-pronoun distillation//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA; NeurIPS; 17597-17611
- Li P Z, Aberman K, Zhang Z H, Hanocka R and Sorkine-Hornung O. 2022b. GANimator: neural motion synthesis from a single sequence. ACM Transactions on Graphics, 41(4): #138 [DOI: 10.1145/3528223.3530157]
- Li R H, Zhao J F, Zhang Y C, Su M Y, Ren Z P, Zhang H, Tang Y S and Li X. 2023b. FineDance: a fine-grained choreography dataset

- for 3D full body dance generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10200-10209 [DOI: 10.1109/ICCV51070.2023.00939]
- Li W, Ma Y H, Liu X P and Desbrun M. 2022c. Efficient kinetic simulation of two-phase flows. *ACM Transactions on Graphics*, 41(4): #114 [DOI: 10.1145/3528223.3530132]
- Li W Y, Chen X L, Li P Z, Sorkine-Hornung O and Chen B Q. 2023c. Example-based motion synthesis via generative motion matching. *ACM Transactions on Graphics*, 42(4): #94 [DOI: 10.1145/3592395]
- Li X Y and Chen D. 2023. The improved atrous spatial pyramid pooling and polarized self-attention based bottom-up panoptic segmentation. *Journal of Image and Graphics*, 28(8): 2410-2419 (李新叶, 陈丁. 2023. 融合改进ASPP和极化自注意力的自底向上全景分割. *中国图象图形学报*, 28(8): 2410-2419) [DOI: 10.11834/jig.220279]
- Li Z H, Xu Q Y, Ye X H, Ren B and Liu L G. 2023d. DiffFR: differentiable SPH-based fluid-rigid coupling for rigid body control. *ACM Transactions on Graphics*, 42(6): #179 [DOI: 10.1145/3618318]
- Liang Y Z, Wang X H, Zhu L C and Yang Y. 2023. MAAL: multimodality-aware autoencoder-based affordance learning for 3D articulated objects//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 217-227 [DOI: 10.1109/ICCV51070.2023.00027]
- Lin X Y, Qi C, Zhang Y C, Huang Z A, Fragkiadaki K, Li Y Z, Gan C and Held D. 2022. Planning with spatialtemporal abstraction from point clouds for deformable object manipulation//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 1640-1651
- Lin X Y, Wang Y F, Olkin J and Held D. 2021. SoftGym: benchmarking deep reinforcement learning for deformable object manipulation//Proceedings of 2020 Conference on Robot Learning. Cambridge, USA: PMLR: 432-448
- Ling H Y, Zinno F, Cheng G and Van De Panne M. 2020. Character controllers using motion VAEs. *ACM Transactions on Graphics*, 39(4): #40 [DOI: 10.1145/3386569.3392422]
- Liu H Y, Iwamoto N, Zhu Z H, Li Z Q, Zhou Y, Bozkurt E and Zheng B. 2022a. DisCo: disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM: 3764-3773 [DOI: 10.1145/3503161.3548400]
- Liu L B and Hodgins J. 2018. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics*, 37(4): #142 [DOI: 10.1145/3197517.3201315]
- Liu L B, van de Panne M and Yin K K. 2016. Guided learning of control graphs for physics-based characters. *ACM Transactions on Graphics*, 35(3): #29 [DOI: 10.1145/2893476]
- Liu L B, Yin K K, Wang B and Guo B N. 2013. Simulation and control of skeleton-driven soft body characters. *ACM Transactions on Graphics*, 32(6): #215 [DOI: 10.1145/2508363.2508427]
- Liu M, Pan Z R, Xu K, Ganguly K and Manocha D. 2020. Deep differentiable grasp planner for high-DOF grippers//Proceedings of the 16th Robotics: Science and Systems. Corvallis, USA: Robotics: Science and Systems: #66 [DOI: 10.15607/rss.2020.xvi.066]
- Liu W Y, Du Y L, Hermans T, Chernova S and Paxton C. 2023. StructDiffusion: language-guided creation of physically-valid structures using unseen objects//Proceedings of the 19th Robotics: Science and Systems. Daegu, Korea (South): Robotics: Science and Systems: #3 [DOI: 10.15607/rss.2023.xix.031]
- Liu W Y, Paxton C, Hermans T and Fox D. 2022b. StructFormer: learning spatial structure for language-guided semantic rearrangement of novel objects//Proceedings of 2022 International Conference on Robotics and Automation. Philadelphia, USA: IEEE: 6322-6329 [DOI: 10.1109/icra46639.2022.9811931]
- Liu Y Z, Liu Y, Jiang C, Lyu K, Wan W K, Shen H, Liang B Q, Fu Z J, Wang H and Yi L. 2022c. HOI4D: a 4D egocentric dataset for category-level human-object interaction//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20981-20990 [DOI: 10.1109/cvpr52688.2022.02034]
- Luo R, Xu W W, Shao T J, Xu H Y and Yang Y. 2019. Accelerated complex-step finite difference for expedient deformable simulation. *ACM Transactions on Graphics*, 38(6): #160 [DOI: 10.1145/3355089.3356493]
- Lyu C Y, Bai K, Wu Y H, Desbrun M, Zheng C X and Liu X P. 2023. Building a virtual weakly-compressible wind tunnel testing facility. *ACM Transactions on Graphics*, 42(4): #125 [DOI: 10.1145/3592394]
- Mahler J, Liang J, Niyaz S, Laskey M, Doan R, Liu X Y, Ojea J A and Goldberg K. 2017. Dex-Net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics//Proceedings of Robotics: Science and Systems XIII. Cambridge, USA: Robotics: Science and Systems: #58 [DOI: 10.15607/rss.2017.xiii.058]
- Mahmood N, Ghorbani N, Troje N F, Pons-Moll G and Black M. 2019. AMASS: archive of motion capture as surface shapes//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5441-5450 [DOI: 10.1109/iccv.2019.00554]
- Maitin-Shepard J, Cusumano-Towner M, Lei J N and Abbeel P. 2010. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding//Proceedings of 2010 IEEE International Conference on Robotics and Automation. Anchorage, USA: IEEE: 2308-2315 [DOI: 10.1109/robot.2010.5509439]
- Mandikal P and Grauman K. 2021. Learning dexterous grasping with object-centric visual affordances//Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an,

- China; IEEE: 6169-6176 [DOI: 10.1109/icra48506.2021.9561802]
- Mei C H and Shi J Y. 2003. Boundary element method based soft object simulation. *Chinese Journal of Computers*, 26(12): 1709-1716 (梅春晖, 石教英. 2003. 基于边界元素法的柔软物体变形模拟. *计算机学报*, 26(12): 1709-1716) [DOI: 10.3321/j.issn:0254-4164.2003.12.014]
- Merel J, Tunyasuvunakool S, Ahuja A, Tassa Y, Hasenclever L, Pham V, Erez T, Wayne G and Heess N. 2020. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics*, 39(4): #39 [DOI: 10.1145/3386569.3392474]
- Miller A T and Allen P K. 2004. Graspit! A versatile simulator for robotic grasping. *IEEE Robotics and Automation Magazine*, 11(4): 110-122 [DOI: 10.1109/MRA.2004.1371616]
- Mo K C, Zhu S L, Chang A X, Yi L, Tripathi S, Guibas L J and Su H. 2019. PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA: IEEE: 909-918 [DOI: 10.1109/cvpr.2019.00100]
- Müller M, Heidelberger B, Hennix M and Ratcliff J. 2007. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2): 109-118 [DOI: 10.1016/j.jvcir.2007.01.005]
- Müller M, Keiser R, Nealen A, Pauly M, Gross M and Alexa M. 2004. Point based animation of elastic, plastic and melting objects//2004 ACM SIGGRAPH/Eurographics symposium on Computer animation. Grenoble, France: Eurographics Association: 141-151 [DOI: 10.1145/1028523.1028542]
- Mur-Labadia L, Guerrero J J and Martinez-Cantin R. 2023. Multi-label affordance mapping from egocentric vision//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*. Paris, France: IEEE: 5215-5226 [DOI: 10.1109/ICCV51070.2023.00483]
- Nagabandi A, Konolige K, Levine S and Kumar V. 2020. Deep dynamics models for learning dexterous manipulation//*Proceedings of the 3rd Annual Conference on Robot Learning*. Osaka, Japan: PMLR: 1101-1112
- Nagarajan T, Li Y H, Feichtenhofer C and Grauman K. 2020. Ego-topo: environment affordances from egocentric video//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: IEEE: 160-169 [DOI: 10.1109/cvpr42600.2020.00024]
- Nau D, Cao Y, Lotem A and Munoz-Avila H. 1999. SHOP: simple hierarchical ordered planner//*Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*. Stockholm, Sweden: Morgan Kaufmann Publishers Inc: 968-973
- Nguyen T, Vu M N, Vuong A, Nguyen D, Vo T, Le N and Nguyen A. 2023. Open-vocabulary affordance detection in 3d point clouds//*Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Detroit, USA: IEEE: 5692-5698 [DOI: 10.1109/IROS55552.2023.10341553]
- Nocedal J and Wright S J. 1999. *Numerical Optimization*. New York, USA: Springer
- NVIDIA Corporation. 2023a. NVIDIA Isaac Sim [EB/OL]. [2023-12-20]. <https://developer.nvidia.com/isaac-sim>
- NVIDIA Corporation. 2023c. NVIDIA Warp [EB/OL]. [2023-12-20]. <https://developer.nvidia.com/warp-python>
- Oprea S, Martinez-Gonzalez P, Garcia-Garcia A, Castro-Vargas J A, Orts-Escolano S and Garcia-Rodriguez J. 2019. A visually realistic grasping system for object manipulation and interaction in virtual reality environments. *Computers and Graphics*, 83: 77-86 [DOI: 10.1016/j.cag.2019.07.003]
- Padalkar A, Pooley A, Jain A, Bewley A, Herzog A, Irpan A, Khazatsky A, Rai A, Singh A, Brohan A, Raffin A, Wahid A, Burgess-Limerick B, Kim B, Schölkopf B, Ichter B, Lu C W, Xu C, Finn C, Xu C F, Chi C, Huang C G, Chan C, Pan C, Fu C Y, Devin C, Driess D, Pathak D, Shah D, Büchler D, Kalashnikov D, Sadigh D, Johns E, Ceola F, Xia F, Stulp F, Zhou G Y, Sukhatme G S, Salhotra G, Yan G, Schiavi G, Kahn G, Su H, Fang H S, Shi H C, Amor H B, Christensen H I, Furuta H, Walke H, Fang H J, Mordatch I, Radosavovic I, Leal I, Liang J, Abou-Chakra J, Kim J, Peters J, Schneider J, Hsu J, Bohg J, Bingham J, Wu J J, Wu J L, Luo J L, Gu J Y, Tan J, Oh J, Malik J, Tompson J, Yang J, Lim J J, Silvério J, Han J, Rao K, Pertsch K, Hausman K, Go K, Gopalakrishnan K, Goldberg K, Byrne K, Oslund K, Kawaharazuka K, Zhang K, Rana K, Srinivasan K, Chen L Y, Pinto L, Tan L, Ott L, Lee L, Tomizuka M, Du M, Ahn M, Zhang M T, Ding M Y, Srirama M K, Sharma M, Kim M J, Kanazawa N, Hansen N, Heess N, Joshi N J, Suenderhauf N, Di Palo N, Shafiullah N M N, Mees O, Kroemer O, Sankeeti P R, Wohlhart P, Xu P, Sermanet P, Sundaresan P, Vuong Q, Rafailov R, Tian R, Doshi R, Martín-Martín R, Mendonca R, Shah R, Hoque R, Julian R, Bustamante S, Kirmani S, Levine S, Moore S, Bahl S, Dass S, Sonawani S, Song S R, Xu S C, Haldar S, Adebola S, Guist S, Nasiriany S, Schaal S, Welker S, Tian S, Dasari S, Belkhale S, Osa T, Harada T, Matsushima T, Xiao T, Yu T H, Ding T L, Davchev T, Zhao T Z, Armstrong T, Darrell T, Jain V, Vanhoucke V, Zhan W, Zhou W X, Burgard W, Chen X, Wang X L, Zhu X H, Li X L, Lu Y, Chebotar Y, Zhou Y F, Zhu Y F, Xu Y, Wang Y X, Bisk Y, Cho Y, Lee Y, Cui Y C, Wu Y H, Tang Y J, Zhu Y K, Li Y Z, Iwasawa Y, Matsuo Y, Xu Z and Cui Z F. 2023. Open X-embodiment: robotic learning datasets and RT-X models [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2310.08864.pdf>
- Pari J, Shafiullah N M, Arunachalam S P and Pinto L. 2021. The surprising effectiveness of representation learning for visual imitation//*18th Robotics: Science and Systems*. New York City, USA: Robotics: Science and Systems: #10 [DOI: 10.15607/rss.2022.xviii.010]
- Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman A A, Tzionas

- D and Black M J. 2019. Expressive body capture: 3d hands, face, and body from a single image//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10967-10977 [DOI: 10.1109/cvpr.2019.01123]
- Peng X B, Abbeel P, Levine S and Van de Panne M. 2018. Deep-Mimic: example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4): #143 [DOI: 10.1145/3197517.3201311]
- Peng X B, Guo Y R, Halper L, Levine S and Fidler S. 2022. ASE: large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics*, 41(4): #94 [DOI: 10.1145/3528223.3530110]
- Peng X B, Ma Z, Abbeel P, Levine S and Kanazawa A. 2021. AMP: adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics*, 40(4): #144 [DOI: 10.1145/3450626.3459670]
- Peskin C S. 2002. The immersed boundary method. *Acta Numerica*, 11: 479-517 [DOI: 10.1017/S0962492902000077]
- Petrovich M, Black M J and Varol G. 2021. Action-conditioned 3D human motion synthesis with Transformer VAE//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10965-10975 [DOI: 10.1109/ICCV48922.2021.01080]
- Pfaff T, Fortunato M, Sanchez-Gonzalez A and Battaglia P. 2021. Learning mesh-based simulation with graph networks//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR: 1-18
- Qi C, Lin X Y and Held D. 2022. Learning closed-loop dough manipulation using a differentiable reset module. *IEEE Robotics and Automation Letters*, 7(4): 9857-9864 [DOI: 10.1109/lra.2022.3191239]
- Qin Y Z, Su H and Wang X L. 2022a. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4): 10873-10881 [DOI: 10.1109/lra.2022.3196104]
- Qin Y Z, Wu Y H, Liu S W, Jiang H W, Yang R H, Fu Y and Wang X L. 2022b. DexMV: imitation learning for dexterous manipulation from human videos//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 570-587 [DOI: 10.1007/978-3-031-19842-7_33]
- Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y Q, Li W and Liu P J. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *The Journal of Machine Learning Research*, 21(1): #140 [DOI: 10.1109/cvpr52729.2023.00941]
- Raissi M, Perdikaris P and Karniadakis G E. 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378: 686-707 [DOI: 10.1016/j.jcp.2018.10.045]
- Raissi M, Yazdani A and Karniadakis G E. 2020. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481): 1026-1030 [DOI: 10.1126/science.aaw4741]
- Rajeswaran A, Kumar V, Gupta A, Vezzani G, Schulman J, Todorov E and Levine S. 2018. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations//Proceedings of the 14th Robotics: Science and Systems. Pittsburgh USA: Robotics: Science and Systems: #49 [DOI: 10.15607/rss.2018.xiv.049]
- Ren J W, Dai J F and Lin H W. 2024. Simulation of cloth with thickness based on isogeometric continuum elastic model. *Journal of Image and Graphics*, 29(1): 243-255 (任靖雯, 戴俊飞, 蔺宏伟. 2024. 等几何连续介质弹性模型的带厚度布料仿真方法. *中国图象图形学报*, 29(1): 243-255) [DOI: 10.11834/jig.221199]
- Robinson-Mosher A, Shinar T, Gretarsson J, Su J and Fedkiw R. 2008. Two-way coupling of fluids to rigid and deformable solids and shells. *ACM Transactions on Graphics*, 27(3): 1-9 [DOI: 10.1145/1360612.1360645]
- Rong Y, Shiratori T and Joo H. 2021. FrankMocap: a monocular 3D whole-body pose estimation system via regression and integration//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada, : IEEE: 1749-1759 [DOI: 10.1109/iccvw54120.2021.00201]
- Ruan L W, Liu J Y, Zhu B, Sueda S, Wang B and Chen B Q. 2021. Solid-fluid interaction with surface-tension-dominant contact. *ACM Transactions on Graphics*, 40(4): #120 [DOI: 10.1145/3450626.3459862]
- Sadeghi F, Toshev A, Jang E and Levine S. 2018. Sim2Real viewpoint invariant visual servoing by recurrent control//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4691-4699 [DOI: 10.1109/cvpr.2018.00493]
- Savva M, Chang A X, Hanrahan P, Fisher M and Nießner M. 2016. PiGraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics*, 35(4): #139 [DOI: 10.1145/2897824.2925867]
- Schneider T, Dumas J, Gao X F, Botsch M, Panozzo D and Zorin D. 2019. Poly-spline finite-element method. *ACM Transactions on Graphics*, 38(3): #19 [DOI: 10.1145/3313797]
- She Q J, Hu R Z, Xu J Z, Liu M, Xu K and Huang H. 2022. Learning high-DOF reaching-and-grasping via dynamic representation of gripper-object interaction. *ACM Transactions on Graphics*, 41(4): #97 [DOI: 10.1145/3528223.3530091]
- Shen S Y, Yang Y, Shao T J, Wang H, Jiang C F F, Lan L and Zhou K. 2021. High-order differentiable autoencoder for nonlinear model reduction. *ACM Transactions on Graphics*, 40(4): #68 [DOI: 10.1145/3450626.3459754]
- Shi H C, Xu H Z, Clarke S, Li Y Z and Wu J J. 2023. RoboCook: long-horizon elasto-plastic object manipulation with diverse tools//Pro-

- ceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 642-660
- Shi H C, Xu H Z, Huang Z A, Li Y Z and Wu J J. 2022. RoboCraft: learning to see, simulate, and shape elasto-plastic objects with graph networks//18th Robotics: Science and Systems. New York City, USA: Robotics: Science and Systems: #8 [DOI: 10.15607/rss.2022.xviii.008]
- Shinar T, Schroeder C and Fedkiw R. 2008. Two-way coupling of rigid and deformable bodies//Proceedings of 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Dublin, Ireland: Eurographics Association: 95-103
- Sifakis E and Barbic J. 2012. FEM simulation of 3D deformable solids: a practitioner's guide to theory, discretization and model reduction//ACM SIGGRAPH 2012 Courses. Los Angeles, USA: ACM: #20 [DOI: 10.1145/2343483.2343501]
- Sifakis E, Shinar T, Irving G and Fedkiw R. 2007. Hybrid simulation of deformable solids//Proceedings of 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. San Diego, USA: Eurographics Association: 81-90
- Sin F S, Schroeder D and Barbič J. 2013. Vega: non-linear FEM deformable object simulator. Computer Graphics Forum, 32(1): 36-48 [DOI: 10.1111/j.1467-8659.2012.03230.x]
- Siyao L, Yu W J, Gu T P, Lin C Z, Wang Q, Qian C, Loy C C and Liu Z W. 2022. Bailando: 3D dance generation by actor-critic GPT with choreographic memory//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11040-11049 [DOI: 10.1109/cvpr52688.2022.01077]
- Stam J. 2023. Stable fluids. Seminal Graphics Papers: Pushing the Boundaries, 2: #81
- Starke S, Mason I and Komura T. 2022. DeepPhase: periodic autoencoders for learning motion phase manifolds. ACM Transactions on Graphics, 41(4): #136 [DOI: 10.1145/3528223.3530178]
- Starke S, Zhang H, Komura T and Saito J. 2019. Neural state machine for character-scene interactions. ACM Transactions on Graphics, 38(6): #209 [DOI: 10.1145/3355089.3356505]
- Starke S, Zhao Y W, Komura T and Zaman K. 2020. Local motion phases for learning multi-contact character movements. ACM Transactions on Graphics, 39(4): #54 [DOI: 10.1145/3386569.3392450]
- Starke S, Zhao Y W, Zinno F and Komura T. 2021. Neural animation layering for synthesizing martial arts movements. ACM Transactions on Graphics, 40(4): #92 [DOI: 10.1145/3450626.3459881]
- Sun L J, Zeng T F, Fan J X and Wang W J. 2024. Double-view feature fusion network for LiDAR semantic segmentation. Journal of Image and Graphics, 29(1): 205-217 (孙刘杰, 曾腾飞, 樊景星, 王文举. 2024. 大场景双视点云特征融合语义分割方法. 中国图象图形学报, 29(1): 205-217) [DOI: 10.11834/jig.220943]
- Sundaresan P, Antonova R and Bohgl J. 2022. DiffCloud: real-to-sim from point clouds with differentiable simulation and rendering of deformable objects//Proceedings of 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Kyoto, Japan: IEEE: 10828-10835 [DOI: 10.1109/iros47612.2022.9981101]
- Taheri O, Ghorbani N, Black M J and Tzionas D. 2020. GRAB: a dataset of whole-body human grasping of objects//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 581-600 [DOI: 10.1007/978-3-030-58548-8_34]
- Takahashi T and Batty C. 2022. ElastoMonolith: a monolithic optimization-based liquid solver for contact-aware elastic-solid coupling. ACM Transactions on Graphics, 41(6): 1-19 [DOI: 10.1145/3550454.3555474]
- Tan J, Zhang T N, Coumans E, Iscen A, Bai Y F, Hafner D, Bohez S and Vanhoucke V. 2018. Sim-to-real: learning agile locomotion for quadruped robots//Proceedings of the 14th Robotics: Science and Systems. Pittsburgh, USA: Robotics: Science and Systems: #10 [DOI: 10.15607/RSS.2018.XIV.010]
- Tang J J, Zheng G, Yu J Y and Yang S B. 2023a. CoTDet: affordance knowledge prompting for task driven object detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3045-3055 [DOI: 10.1109/ICCV51070.2023.00285]
- Tang X J, Wu L J, Wang H, Hu B, Gong X, Liao Y C, Li S N, Kou Q L and Jin X G. 2023b. RSMT: real-time stylized motion transition for characters//Proceedings of 2023 ACM SIGGRAPH 2023 Conference Proceedings. Los Angeles, USA: ACM: #38 [DOI: 10.1145/3588432.3591514]
- Teran J, Sifakis E, Blemker S S, Ng-Thow-Hing V, Lau C and Fedkiw R. 2005. Creating and simulating skeletal muscle from the visible human data set. IEEE Transactions on Visualization and Computer Graphics, 11(3): 317-328 [DOI: 10.1109/tvcg.2005.42]
- Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D and Bermano A H. 2023. Human motion diffusion model//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR: #11970
- Todorov E, Erez T and Tassa Y. 2012. MuJoCo: a physics engine for model-based control//Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura-Algarve, Portugal: IEEE: 5026-5033 [DOI: 10.1109/iros.2012.6386109]
- Toussaint M. 2015. Logic-geometric programming: an optimization-based approach to combined task and motion planning//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina: IJCAI: 1930-1936
- Tseng J, Castellon R and Liu C K. 2023. EDGE: editable dance generation from music//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 448-458 [DOI: 10.1109/cvpr52729.2023.00051]
- van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Con-

- ference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6309-6318
- Wan W K, Geng H R, Liu Y, Shan Z K, Yang Y D, Yi L and Wang H. 2023. UniDexGrasp++: improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3868-3879 [DOI: 10.1109/ICCV51070.2023.00360]
- Wang B H, Mateuk G and Barbič J. 2020. Hand MRI dataset [EB/OL]. [2023-12-20]. <http://www.jernejbarbic.com/hand-mri-dataset>
- Wang H, Sridhar S, Huang J W, Valentin J, Song S R and Guibas L J. 2019a. Normalized object coordinate space for category-level 6D object pose and size estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2637-2646 [DOI: 10.1109/cvpr.2019.00275]
- Wang H M. 2021. GPU-based simulation of cloth wrinkles at submillimeter levels. *ACM Transactions on Graphics*, 40(4): #169 [DOI: 10.1145/3450626.3459787]
- Wang J B, Rong Y, Liu J Y, Yan S J, Lin D H and Dai B. 2022a. Towards diverse and natural scene-aware 3d human motion synthesis//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20428-20437 [DOI: 10.1109/cvpr52688.2022.01981]
- Wang J S, Xu H Z, Xu J W, Liu S F and Wang X L. 2021. Synthesizing long-term 3D human motion and interaction in 3d scenes//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9396-9406 [DOI: 10.1109/cvpr46437.2021.00928]
- Wang R C, Zhang J L, Chen J Y, Xu Y Z, Li P H, Liu T Y and Wang H. 2023. DexGraspNet: a large-scale robotic dexterous grasp dataset for general objects based on simulation//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, England: IEEE: 11359-11366 [DOI: 10.1109/icra48891.2023.10160982]
- Wang Y, Weidner N J, Baxter M A, Hwang Y, Kaufman D M and Sueda S. 2019b. REDMAX: efficient & flexible approach for articulated dynamics. *ACM Transactions on Graphics*, 38(4): #104 [DOI: 10.1145/3306346.3322952]
- Wang Y A, Wu R H, Mo K C, Ke J Q, Fan Q N, Guibas L J and Dong H. 2022b. AdaAfford: learning to adapt manipulation affordance for 3D articulated objects via few-shot interactions//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 90-107 [DOI: 10.1007/978-3-031-19818-2_6]
- Wang Z, Chen Y X, Liu T Y, Zhu Y X, Liang W and Huang S Y. 2022c. HUMANISE: language-conditioned human motion generation in 3D scenes//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS: 14959-14971
- Weinstein R, Teran J and Fedkiw R. 2006. Dynamic simulation of articulated rigid bodies with contact and collision. *IEEE Transactions on Visualization and Computer Graphics*, 12(3): 365-374 [DOI: 10.1109/tvcg.2006.48]
- Werling K, Omens D, Lee J, Exarchos I and Liu C K. 2021. Fast and feature-complete differentiable physics for articulated rigid bodies with contact//17th Robotics: Science and Systems. [s.l.]: Robotics: Science and Systems
- Won J, Gopinath D and Hodgins J. 2022. Physics-based character controllers using conditional VAEs. *ACM Transactions on Graphics*, 41(4): #96 [DOI: 10.1145/3528223.3530067]
- Wong J, Tung A, Kurenkov A, Mandlekar A, Li F F, Savarese S and Martín-Martín R. 2022. Error-aware imitation learning from teleoperation data for mobile manipulation//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 1367-1378
- Wu B T, Wang Z D and Wang H M. 2022. A GPU-based multilevel additive schwarz preconditioner for cloth and deformable body simulation. *ACM Transactions on Graphics*, 41(4): #63 [DOI: 10.1145/3528223.3530085]
- Wu R H, Ning C R and Dong H. 2023. Learning foresightful dense visual affordance for deformable object manipulation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10913-10922 [DOI: 10.1109/ICCV51070.2023.01005]
- Xian Z, Zhu B, Xu Z J, Tung H Y, Torralba A, Fragkiadaki K and Gan C. 2022. FluidLab: a differentiable environment for benchmarking complex fluid manipulation//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR: 1-19
- Xiang F B, Qin Y Z, Mo K C, Xia Y K, Zhu H, Liu F C, Liu M H, Jiang H X, Yuan Y, Wang H, Yi L, Chang A X, Guibas L J and Su H. 2020. SAPIEN: a simulated part-based interactive environment//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11094-11104 [DOI: 10.1109/cvpr42600.2020.01111]
- Xiang Y, Schmidt T, Narayanan V and Fox D. 2018. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes//14th Robotics: Science and Systems. Pittsburgh, USA: Robotics: Science and Systems: #19
- Xie X H, Bhatnagar B L and Pons-Moll G. 2023. Visibility aware human-object interaction tracking from single RGB camera//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4757-4768 [DOI: 10.1109/cvpr52729.2023.00461]
- Xu D F, Nair S, Zhu Y K, Gao J L, Garg A, Li F F and Savarese S. 2018. Neural task programming: learning to generalize across hierarchical tasks//Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE: 3795-3802 [DOI: 10.1109/ICRA.2018.8460689]
- Xu Y Z, Wan W K, Zhang J L, Liu H R, Shan Z K, Shen H, Wang R

- C, Geng H R, Weng Y J, Chen J Y, Liu T Y, Yi L and Wang H. 2023. UniDexGrasp: universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4737-4746 [DOI: 10.1109/cvpr52729.2023.00459]
- Xu Z J, He Z P and Song S R. 2022. Universal manipulation policy network for articulated objects. IEEE Robotics and Automation Letters, 7(2): 2447-2454 [DOI: 10.1109/lra.2022.3142397]
- Yan X C, Hsu J, Khansari M, Bai Y F, Pathak A, Gupta A, Davidson J and Lee H. 2018. Learning 6-DOF grasping interaction via deep geometry-aware 3D representations//Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE: 3766-3773 [DOI: 10.1109/ICRA.2018.8460609]
- Yang L X, Li K L, Zhan X Y, Wu F, Xu A R, Liu L and Lu C W. 2022a. OakInk: a large-scale knowledge repository for understanding hand-object interaction//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20921-20930 [DOI: 10.1109/cvpr52688.2022.02028]
- Yang T, Chang J, Ren B, Lin M C, Zhang J J and Hu S M. 2015. Fast multiple-fluid simulation using Helmholtz free energy. ACM Transactions on Graphics, 34(6): #201 [DOI: 10.1145/2816795.2818117]
- Yang T Y, Arnaud S, Shah K, Yokoyama N, Clegg A W, Truong J, Undersander E, Maksymets O, Ha S, Kalakrishnan M, Mottaghi R, Batra D and Rai A. 2023a. LSC: language-guided skill coordination for open-vocabulary mobile pick-and-place [EB/OL]. [2023-12-20]. <https://languageguidedskillcoordination.github.io/>.
- Yang Y H, Zhai W, Luo H C, Cao Y, Luo J B and Zha Z J. 2023b. Grounding 3D object affordance from 2D interactions in images//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10871-10881 [DOI: 10.1109/ICCV51070.2023.01001]
- Yang Z S, Yin K K and Liu L B. 2022b. Learning to use chopsticks in diverse gripping styles. ACM Transactions on Graphics, 41(4): #95 [DOI: 10.1145/3528223.3530057]
- Yao H Y, Song Z H, Chen B Q and Liu L B. 2022. ControlVAE: model-based learning of generative controllers for physics-based characters. ACM Transactions on Graphics, 41(6): #183 [DOI: 10.1145/3550454.3555434]
- Yenamandra S, Ramachandran A, Yadav K, Wang A S, Khanna M, Gervet T, Yang T Y, Jain V, Clegg A, Turner J M, Kira Z, Savva M, Chang A X, Chaplot S D, Batra D, Mottaghi R, Bisk Y and Paxton C. 2023. HomeRobot: open-vocabulary mobile manipulation//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 1975-2011
- Yin H, Varava A and Kragic D. 2021. Modeling, learning, perception, and control methods for deformable object manipulation. Science Robotics, 6(54): #8803 [DOI: 10.1126/scirobotics.abd8803]
- Yin Z H, Huang B H, Qin Y Z, Chen Q F and Wang X L. 2023. Rotating without seeing: towards in-hand dexterity through touch//19th Robotics: Science and Systems. Daegu, Korea (South): Robotics: Science and Systems: #36 [DOI: 10.15607/rss.2023.xix.036]
- Yokoyama N, Clegg A W, Undersander E, Ha S, Batra D and Rai A. 2023. Adaptive skill coordination for robotic mobile manipulation [EB/OL]. [2023-12-20]. <https://arxiv.org/abs/2304.00410v1>
- Zakka K, Zeng A, Florence P, Tompson J, Bohg J and Dwibedi D. 2022. XIRL: cross-embodiment inverse reinforcement learning//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 537-546
- Zeng F F, Liang B L, Liu Z and Wang J H. 2000. An interactive environment design based on digital glove. Journal of Image and Graphics, 5(2): 153-157 (曾芬芳, 梁柏林, 刘镇, 王建华. 2000. 基于数据手套的人机交互环境设计. 中国图象图形学报, 5(2): 153-157) [DOI: 10.11834/jig.20000214]
- Zhai W, Luo H C, Zhang J, Cao Y and Tao D C. 2021. One-shot object affordance detection in the wild. International Journal of Computer Vision, 130(1): 2472-2500 [DOI: 10.1007/s11263-022-01642-4]
- Zhang H, Starke S, Komura T and Saito J. 2018. Mode-adaptive neural networks for quadruped motion control. ACM Transactions on Graphics, 37(4): #145 [DOI: 10.1145/3197517.3201366]
- Zhang H, Ye Y T, Shiratori T and Komura T. 2021. ManipNet: neural manipulation synthesis with a hand-object spatial representation. ACM Transactions on Graphics, 40(4): #121 [DOI: 10.1145/3450626.3459830]
- Zhang H T, Yuan Y, Makovychuk V, Guo Y R, Fidler S, Peng X B and Fatahalian K. 2023a. Learning physically simulated tennis skills from broadcast videos. ACM Transactions on Graphics, 42(4): #95 [DOI: 10.1145/3592408]
- Zhang J Z, Gireesh N, Wang J L, Fang X M, Xu C Y, Chen W G, Dai L and Wang H. 2023b. GAMMA: graspability-aware mobile manipulation policy learning based on online grasping pose fusion [EB/OL]. [2023-12-20]. <https://arxiv.org/pef/2309.15459.pdf>
- Zhang S W, Zhang Y, Ma Q L, Black M J and Tang S Y. 2020a. PLACE: proximity learning of articulation and contact in 3D environments//Proceedings of 2020 International Conference on 3D Vision (3DV). Fukuoka, Japan: IEEE: 642-651 [DOI: 10.1109/3dv50981.2020.00074]
- Zhang Y, Hassan M, Neumann H, Black M J and Tang S Y. 2020b. Generating 3D people in scenes without people//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6193-6203 [DOI: 10.1109/cvpr42600.2020.00623]
- Zhao K F, Wang S F, Zhang Y, Beeler T and Tang S Y. 2022. Compositional human-scene interaction synthesis with semantic control//Proceedings of the 17th European Conference on Computer Vision. Tel

- Aviv, Israel: Springer: 311-327 [DOI: 10.1007/978-3-031-20068-7_18]
- Zhao K F, Zhang Y, Wang S F, Beeler T and Tang S Y. 2023a. Synthesizing diverse human motions in 3D indoor scenes//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 14692-14703 [DOI: 10.1109/ICCV51070.2023.01354]
- Zhao T Z, Kumar V, Levine S and Finn C. 2023b. Learning fine-grained bimanual manipulation with low-cost hardware//19th Robotics: Science and Systems. Daegu, Korea(South): Robotics: Science and Systems: #16
- Zhao Y, Wu R H, Chen Z H, Zhang Y R, Fan Q N, Mo K C and Dong H. 2023c. DualAfford: learning collaborative visual affordance for dual-gripper manipulation//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR: #1971
- Zheng J T, Zheng Q Y, Fang L X, Liu Y and Yi L. 2023. CAMS: CAnonicalized manipulation spaces for category-level functional hand-object manipulation synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 585-594 [DOI: 10.1109/cvpr52729.2023.00064]
- Zheng M L, Wang B H, Huang J T and Barbič J. 2022. Simulation of hand anatomy using medical imaging. ACM Transactions on Graphics, 41(6): #273 [DOI: 10.1145/3550454.3555486]
- Zhi Y H, Cun X D, Chen X L, Shen X, Guo W, Huang S L and Gao S H. 2023. LivelySpeaker: towards semantic-aware co-speech gesture generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20750-20760 [DOI: 10.1109/ICCV51070.2023.01902]
- Zhong C L, Zheng Y H, Zheng Y P, Zhao H, Yi L, Mu X D, Wang L, Li P F, Zhou G Y, Yang C, Zhang X L and Zhao J. 2023. 3D implicit transporter for temporally consistent keypoint discovery//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3846-3857 [DOI: 10.1109/ICCV51070.2023.00358]
- Zhu Z H, Wang J S, Qin Y Z, Sun D Q, Jampani V and Wang X L. 2023. ContactArt: learning 3D interaction priors for category-level articulated object and hand poses estimation [EB/OL]. [2023-12-20]. <https://arxiv.org/pdf/2305.01618.pdf>

作者简介

- 杜韬,男,助理教授,主要研究方向为计算机图形学中的计算设计与物理仿真。E-mail:taodu@tsinghua.edu.cn
- 胡瑞珍,通信作者,女,教授,主要研究方向为计算机图形学和具身智能。E-mail:ruizhen.hu@szu.edu.cn
- 刘利斌,男,助理教授,主要研究方向为计算机图形学、数字人仿真与具身智能。E-mail: libin.liu@pku.edu.cn
- 弋力,男,助理教授,主要研究方向为三维视觉、交互规划与生成和人机协作。E-mail:eric yi@mail.tsinghua.edu.cn
- 赵昊,男,助理教授,主要研究方向为几何与认知层面的场景理解及其在机器人中的应用。E-mail:zhaohao@air.tsinghua.edu.cn