

具身智能自主无人系统技术

孙长银^{1,2,3,4} 袁心³ 王远大³ 柳文章^{2,4}

摘要 自主无人系统是一类具有自主感知和决策能力的智能系统,在国防安全、航空航天、高性能机器人等方面有着广泛的应用.近年来,基于 Transformer 架构的各类大模型快速革新,极大地推动了自主无人系统的发展.目前,自主无人系统正迎来一场以“具身智能”为核心的新一代技术革命.大模型需要借助无人系统的物理实体来实现“具身化”,无人系统可以利用大模型技术来实现“智能化”.本文阐述具身智能自主无人系统的发展现状,详细探讨包含大模型驱动的多模态感知、面向具身任务的推理与决策、基于动态交互的机器人学习与控制、三维场景具身模拟器等具身智能领域的关键技术.最后,指出目前具身智能无人系统所面临的挑战,并展望未来的研究方向.

关键词 自主无人系统,具身智能,大语言模型,人工智能

引用格式 孙长银,袁心,王远大,柳文章.具身智能自主无人系统技术.自动化学报,2025,51(4):762-777

DOI 10.16383/j.aas.c240456 **CSTR** 32138.14.j.aas.c240456

Embodied Intelligence Autonomous Unmanned Systems Technology

SUN Chang-Yin^{1,2,3,4} YUAN Xin³ WANG Yuan-Da³ LIU Wen-Zhang^{2,4}

Abstract Autonomous unmanned systems are intelligent systems with autonomous perception and decision-making capabilities, widely applied in areas such as defense security, aerospace, and high-performance robotics. In recent years, the rapid advancements of various large models based on the Transformer architecture have significantly accelerated the development of autonomous unmanned systems. Currently, these systems are undergoing a new technological revolution centered on “embodied intelligence”. Large models require the physical embodiment of unmanned systems to achieve “embodiment”, while unmanned systems can leverage large model technologies to achieve “intelligence”. This paper outlines the current state of development in embodied intelligent autonomous unmanned systems and provides a detailed discussion of key technologies in the field of embodied intelligence, including large-model-driven multimodal perception, reasoning and decision-making for embodied tasks, robot learning and control based on dynamic interaction, and 3D embodied simulators. Finally, the paper identifies existing challenges in embodied intelligence unmanned systems and explores future research directions.

Key words Autonomous unmanned systems, embodied intelligence, large language models, artificial intelligence

Citation Sun Chang-Yin, Yuan Xin, Wang Yuan-Da, Liu Wen-Zhang. Embodied intelligence autonomous unmanned systems technology. *Acta Automatica Sinica*, 2025, 51(4): 762-777

具身智能的核心内涵是要求系统具备完整的自

主环境感知与认知能力、流畅的人机交互能力、可靠的智能决策和运动操纵规划能力^[1],能够通过与环境交互实现能力的泛化和对新场景的适应^[2].具身智能的发展可以追溯到 20 世纪中期, Wiener^[3]提出的系统自我调节理念,以及 Turing^[4]提出的智能需要通过环境交互才能涌现的观点,均强调了智能系统与物理世界互动的重要性,为具身智能的发展提供了关键指导.

随着新一轮科技革命和产业变革的到来,自主无人系统逐渐成为具身智能技术的主要载体和应用平台,在国防军事、城市治理、精准医疗等多个领域发挥着不可替代的重要作用^[5-7].在俄乌军事冲突中,俄罗斯将“柳叶刀”无人机投入战场,作为打击装甲车辆和火炮系统的重要武器.截至 2024 年 2 月 28 日,俄军使用搭载高爆炸弹头的“柳叶刀”无人

收稿日期 2024-06-30 录用日期 2024-09-27

Manuscript received June 30, 2024; accepted September 27, 2024

国家自然科学基金创新研究群体(61921004),国家自然科学基金重点项目(62236002),国家自然科学基金(62203113)资助

Supported by National Natural Science Foundation of China for Creative Research Groups (61921004), Key Projects of National Natural Science Foundation of China (62236002), and National Natural Science Foundation of China (62203113)

本文责任编辑 温广辉

Recommended by Associate Editor WEN Guang-Hui

1. 安徽大学自主无人系统技术教育部工程研究中心 合肥 230601
2. 安徽大学安徽省无人系统与智能技术工程研究中心 合肥 230601
3. 东南大学自动化学院 南京 210096 4. 安徽大学人工智能学院 合肥 230601

1. Engineering Research Center of Autonomous Unmanned System Technology, Ministry of Education, Anhui University, Hefei 230601 2. Anhui Provincial Engineering Research Center for Unmanned System and Intelligent Technology, Anhui University, Hefei 230601 3. School of Automation, Southeast University, Nanjing 210096 4. School of Artificial Intelligence, Anhui University, Hefei 230601

机进行了 1163 次攻击, 共摧毁了 363 个目标, 并严重破坏了 615 个目标, 这使得“柳叶刀”无人机成为俄军中最有效率的精确制导武器之一。与此同时, 反无人机系统的发展也在加速, 成为保护军事和民用设施免受无人机威胁的关键手段^[8]。2023 年美国研制生产了“模块化监视侦察反无人机系统”(亦称“吸血鬼”系统)。该系统可安装在地面机动平台或固定地点, 主要探测装置为球状光电模块化传感器和激光指示器, 能够快速识别并拦截多种类型的无人机, 确保在复杂环境中的有效防御^[9]。自主无人系统已经成为提升综合国力的重要技术支撑^[10-12]。

目前, 自主无人系统正迈入以“具身智能”为核心的新一代技术革命阶段^[13-14]。传统的面向特定任务、封闭场景的无人系统设计思路已不能满足社会生产与军事应用的需求。相比之下, 面向开放交互环境的具身智能无人系统成为了未来的发展趋势^[15]。非结构化、未知、动态、开放的任务环境要求自主无人系统具有自主学习能力, 可以在与环境交互中提取有效信息, 实时调整和优化自身行为策略。近年来, 基于 Transformer 架构的各类大模型快速革新^[16-19], 使得无人系统不仅可以准确理解自然语言指令、视觉图像以及连续传感器状态等感知信息, 还能驱动系统完成与开放环境的交互, 这极大地推动了无人系统具身智能的发展。在大语言模型的驱动下, 新一代具身智能无人系统的发展具有如下特征: 1) 在应用场景上, 从封闭单一任务场景向开放任务场景发展; 2) 在适用范围上, 从特定单一任务向通用任务发展; 3) 在系统设计理念上, 从孤立的感知、控制、决策模块向大模型驱动下各模块深度融合发展(见图 1)。

本文内容安排如下: 第 1 节介绍具身智能无人系统的发展现状, 概述近年来的重要成果; 第 2 节针对具身智能无人系统的关键技术展开论述; 第 3 节综述具身智能无人系统两项典型的研究任务; 第 4 节展望具身智能无人系统的未来研究方向; 第 5 节总结全文。

1 具身智能无人系统发展现状

无人系统具身智能的实现依赖于感知层、决策层、控制层三个层面的发展与相互融合。以往制约具身智能发展的主要原因包括: 1) 多模态信息的处理存在天然鸿沟; 2) 知识难表征与零样本推理难实现。处理多模态信息的一大阻碍是无人系统难以将开放词域下的目标对象准确地锚定到视觉图像中的对应物体, 反之亦然。这一问题严重限制了系统在开放、非结构化任务场景的部署。而知识难表征与零样本推理难实现则导致了无人系统通常只能解决特定的问题, 无法将有限的经验泛化到其他任务场景, 因此无人系统难以具备通用任务的处理能力。

随着 Transformer 网络架构的问世^[20], 使得解决上述两大难题成为可能。基于 Transformer 的大模型在计算机视觉^[21-23]、自然语言处理^[24]、多模态学习^[25-28]等领域已取得了革命性的突破。在多模态信息处理方面, 视觉语言模型 CLIP^[29] (Contrastive language-image pre-training) 实现了开放词域下的语言文本模态和视觉图像模态的对齐, 使得机器人系统不仅可以“看到”开放场景中的物体, 还能够“看懂”物体。该工作所使用的双流 (Two stream) 结构与对比学习方法在后续的研究中得到进一步发展和推广, 扩展到了语音、视频、点云等模态的对齐

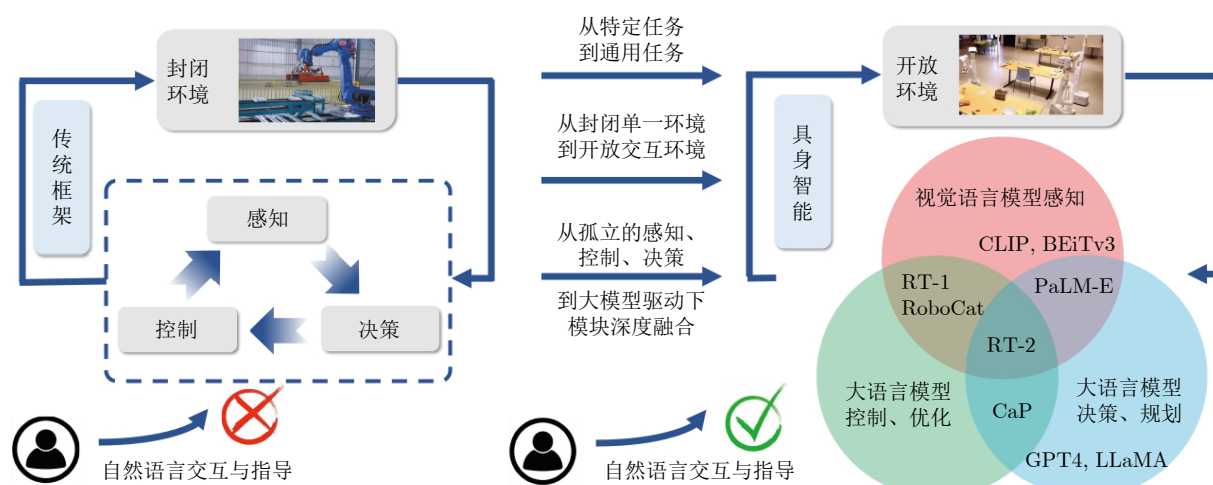


图 1 自主无人系统体系架构发展趋势

Fig.1 Architecture development trend of autonomous unmanned systems

任务. 在知识表征和零样本推理方面, 大语言模型 InstructGPT^[30] 具有强大的自然语言理解、语义表征与知识推理能力, 使得机器人系统不仅可以“听懂”自然语言指令, 还能够完成任务的决策与规划. 此外, 大语言模型的出现使得传统系统架构中的感知模块、控制模块、决策模块的边界逐渐模糊, 各模块深度融合以及感知-控制-决策一体化的趋势逐渐凸显. 表 1 所示为具有代表性的具身智能模型架构.

2022 年底, 谷歌机器人团队首先在无人系统的多模态感知和控制方面取得了突破, 发布了机器人 Transformer 模型^[31] (Robotics Transformer, RT-1). RT-1 以图像帧序列和自然语言指令作为系统输入, 输出机械臂运动的目标位姿与移动基座的控制指令. RT-1 通过模仿学习的方法具备了抓取与放置物品、打开与关闭柜门等基本技能. 2023 年 3 月, 谷歌团队发布了面向开放场景通用任务的具身多模态大模型^[32] (Embodied multimodal language model, PaLM-E). PaLM-E 以视觉图像信息、连续传感器数据和自然语言文本作为系统输入, 输出文本形式的回答或者下层技能选择的分词 (Token), 并驱动执行层完成机器人操作. PaLM-E 将真实世界的连续传感器模态信息直接融入语言模型中, 从而建立多模态感知信息的统一表征框架, 并通过端到端的训练方式, 使得机器人能够利用语言模型内化的世界知识完成复杂动态场景下的移动与操纵、视觉问答和视觉语言导航等具身任务. 图 2 所示为 PaLM-E 指导机器人完成长程任务. 该项研究成果开启了具身智能研究的新阶段, 是该领域里程碑式的工作. 2023 年 7 月, 谷歌团队将感知层、决策层与控制层进行深度结合, 提出具有控制闭环特性的视觉语言动作模型 (Vision-language-action model, VLA), 即 RT-2-PaLM-E^[33]. 该模型框架将

视觉图像、语言指令以及动作序列嵌入到高维语义空间, 使得机器人控制底层同样能够利用大语言模型抽象的语义概念以及丰富的世界知识来改善机器人控制性能. 图 3 列举了具身智能无人系统关键技术.

机器人是实现具身智能的重要载体, 其中以全尺寸的人形机器人最具代表性. 人形机器人具备人类的外观形态以及行动能力, 可以使用双腿直立行走或奔跑, 通过四肢与身体的协调完成复杂场景下的通用任务, 如搬运、扫地和整理等. 2023 年 5 月, 特斯拉发布了电驱人形机器人 Optimus, 实现了工业生产复杂场景中的自主移动、物料搬运、组装操作等生产任务. 2023 年 8 月, 国内智元团队发布第一代通用型具身智能机器人远征 A1, 采用了自研软件框架 AgiROS, 搭载了语言任务模型 WorkGPT, 可以理解自然语言指令, 能够实现复杂任务的多级推理. 2024 年 3 月, Figure 发布全球首个搭载 ChatGPT 的人形机器人 Figure 01. Figure 01 具有强大的自然语言理解能力, 能够进行任务指令的逻辑推理, 并完成物体抓握与重排任务. 同样在 2024 年 3 月, 宇树科技发布人形机器人 H1. H1 采用深度强化学习的方法掌握了后空翻技巧, 成为全球首款实现后空翻的全尺寸电驱人形机器人. 图 4 所示为具有代表性的全尺寸人形机器人. 机器人硬件实体的快速发展为具身智能的落地与普及奠定了坚实基础.

2 具身智能无人系统关键技术

近年来, 具身智能技术的快速发展促使自主无人系统在多模态感知、自主决策、交互式学习等方面取得了突破, 提高了系统解决开放环境下复杂任

表 1 具身智能模型架构
Table 1 Embodied intelligence model architecture

名称	模型参数	响应频率 (Hz)	模型架构说明
SayCan ^[34]	—	—	SayCan 利用价值函数表示各个技能的可行性, 并由语言模型进行技能评分, 能够兼顾任务需求和机器人技能的可行性
RT-1 ^[31]	350 万	3	RT-1 采用 13 万条机器人演示数据的数据集完成模仿学习训练, 能以 97% 的成功率执行超过 700 个语音指令任务
RoboCat ^[35]	12 亿	10 ~ 20	RoboCat 构建了基于目标图像的可迁移机器人操纵框架, 能够实现多个操纵任务的零样本迁移
PaLM-E ^[32]	5620 亿	5 ~ 6	PaLM-E 构建了当时最大的具身多模态大模型, 将机器人传感器模态融入语言模型, 建立了端到端的训练框架
RT-2 ^[33]	550 亿	1 ~ 3	RT-2 首次构建了视觉-语言-动作的模型, 在多个具身任务上实现了多阶段的语义推理
VoxPoser ^[36]	—	—	VoxPoser 利用语言模型生成关于当前环境的价值地图, 并基于价值地图进行动作轨迹规划, 实现了高自由度的环境交互
RT-2-X ^[37]	550 亿	1 ~ 3	RT-2-X 构建了提供标准化数据格式、交互环境和模型的数据集, 包含 527 种技能和 16 万个任务



图 2 PaLM-E 完成长程任务

Fig.2 The PaLM-E completes long range tasks

务的能力. 图 5 所示为具身智能自主无人系统框架示意图. 本节将以具身智能与大语言模型技术的最新进展及其在无人系统中的应用为主要脉络, 简述该领域的关键技术, 主要包括大语言模型驱动的多模态感知、基于场景重建的场景表征与开放环境的场景理解、面向具身任务的决策与规划、基于动态交互的机器人学习与控制以及面向具身智能的三维场景仿真模拟器.

2.1 大语言模型驱动的多模态感知

人类通过眼睛、耳朵、皮肤等感官从周围环境中收集图像、声音、压力等模态的信息, 上述感知信息在经过神经传递与大脑处理后, 成为人类感知与

认知环境的重要依据. 因此, 如何实现与人类相类似的多模态感知与理解能力是实现无人系统智能化和自主化的基础前提^[38]. 预训练的基础模型可以对文本、图像、传感器数据等多模态输入信息进行统一编码, 能够表征跨模态信息之间的高级语义关联, 在实现无人系统的多模态感知与信息融合方面具有巨大潜力^[39-41]. 本节将介绍两类预训练的基础模型, 分别是视觉语言大模型和多模态大模型.

1) 视觉语言大模型

视觉语言大模型 (Vision-language model, VLM) 是一类有着广泛应用的基础模型, 主要用于理解自然语言指令和视觉图像信息, 可以解决开放词域下的目标分类、目标检测、全景分割等问题. 视觉语言大模型通过在大规模数据集上预训练来学习不同模态之间的语义对应关系, 能够将不同模态的信息统一编码到高维语义空间进行处理^[42]. 相比于单纯的视觉模型或语言模型, 视觉语言大模型的特点是可以将高度抽象的自然语言与视觉图像信息进行对齐, 以便理解相同对象不同模态表征下的关联概念. 在图像分类任务中, 视觉语言大模型不再是利用固定数目的分类头来判断图像所属类别, 而是通过理解自然语言的抽象概念, 判断任一给定词语与图像之间的关联性 (如嵌入向量的余弦相似度) 来完成分类任务. 因此, 视觉语言大模型不再受限于分类头数目, 可以实现开放词域下的图像分类. 目前, 视觉语言大模型在解决开放词域下的图像分类^[29, 43]、目标检测^[44-46]、语义分割^[47-51]等感知任务方面已经取得了重要进展.

在开放词域下的图像分类方面, 视觉语言模型

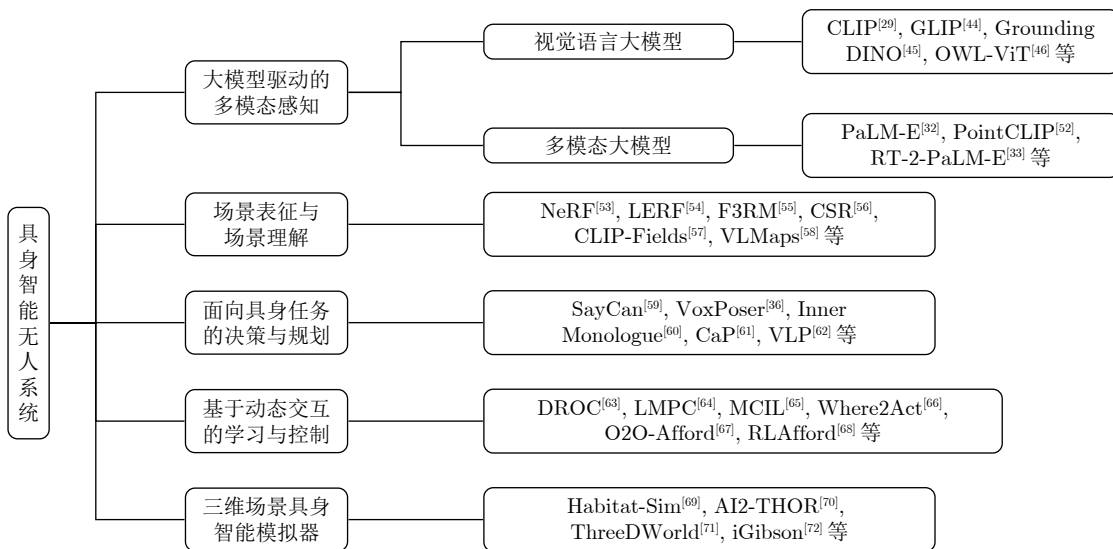


图 3 具身智能无人系统关键技术示意图

Fig.3 Schematic diagram of key technologies in embodied intelligence unmanned systems

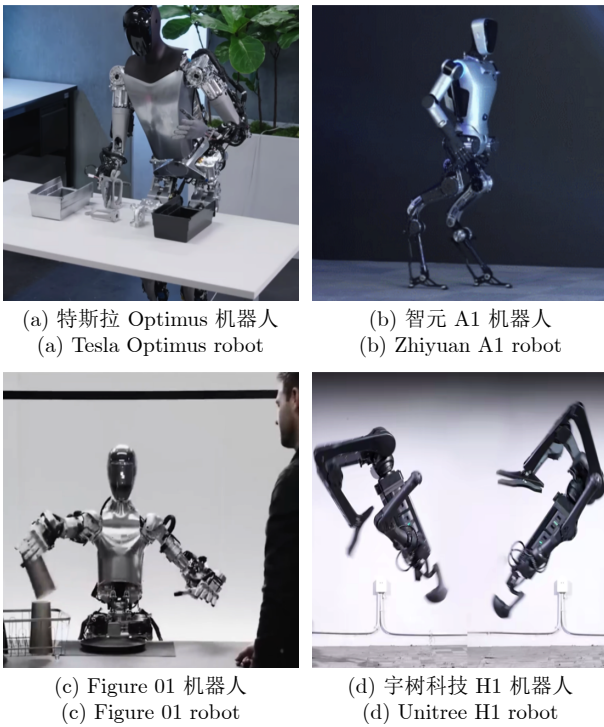


图 4 各类人形机器人

Fig. 4 Various humanoid robots

CLIP^[29] 突破了封闭式监督学习范式的限制, 实现了自然语言与视觉图像的细粒度对齐, 是该领域里程碑式的工作. CLIP 采用对比表征学习的方法, 设计了文本-图像匹配的辅助任务 (Proxy task), 通过联合训练图像编码器和文本编码器, 最大化相匹配图像和文本嵌入的余弦相似度, 同时最小化不相匹

配图像和文本嵌入的相似度. 预训练模型所输出的视觉图像编码可以表征图像的高级语义信息, 这使得模型能够零样本转移到不同类型的下游任务. 模型训练采用了 4 亿多条互联网图像-文本关联数据对, 需要在 256 块 V100 GPU 上训练 12 天. 该研究成果在超过 30 个机器视觉任务上进行了测试, 涵盖了图像分类、目标检测、动作识别等. 测试结果表明, CLIP 的性能达到甚至超越了通过监督训练所得模型的性能. 在 ImageNet 零样本图像分类任务中, CLIP 可以在没有使用任何一条 ImageNet 的图像数据进行训练的前提下, 具有与 ResNet-50 相同的分类准确性. CLIP 能够实现开放环境下的视觉图像与自然语言的细粒度对齐, 因此在机器人导航任务中, CLIP 能够有效地将视觉感知与自然语言指令结合, 提高系统的语义理解与决策能力, 尤其在非结构化、动态环境下表现更突出. LM-Nav^[73] 采用 CLIP 计算当前的视觉图像信息与大语言模型输出的地标文本匹配相似度, 得出各个地标的概率分布, 进而指导系统下一步的导航决策. CoWs^[74] 采用 CLIP 图像编码蕴含的语义信息, 实现了开放词域下的目标定位. CoWs 首先将图像离散化成若干个较小的图像块, 并为每个图像块标识一个空间位置. 然后, 将目标物体的语言描述嵌入与每一个图像块的嵌入进行相似度匹配, 获取文本-图像的相似度评分. 最后, 选取得分最高的区域作为导航的目标位置.

在开放词域下的目标检测方面, 语言图像预训

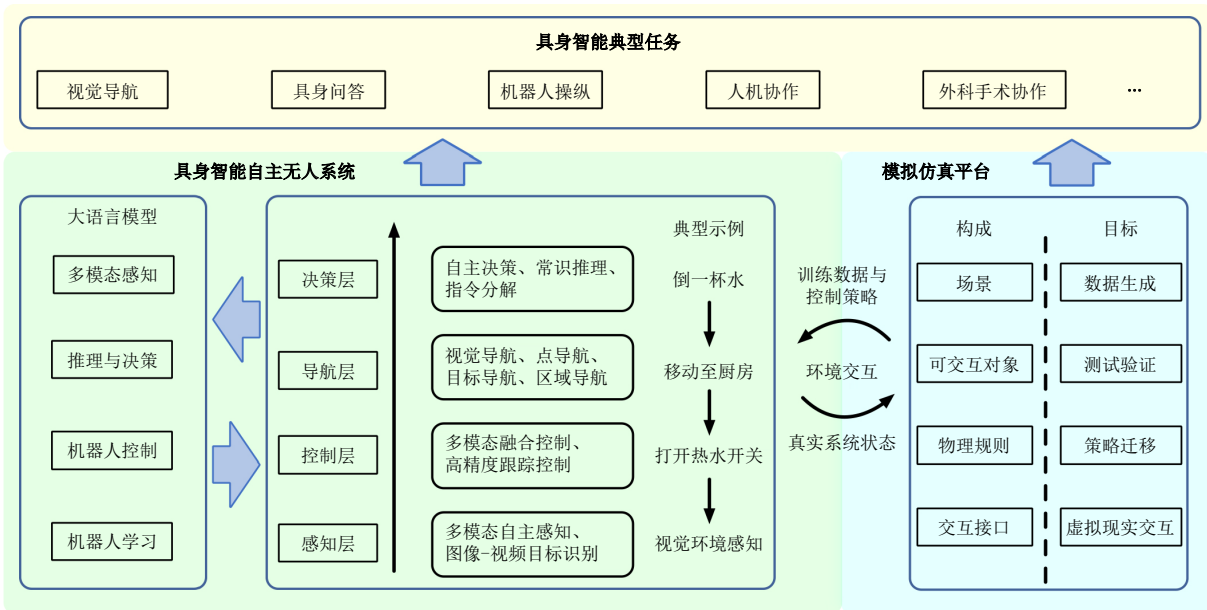


图 5 具身智能自主无人系统框架示意图及典型应用

Fig. 5 Framework diagram and typical application of embodied intelligence autonomous unmanned systems

练基础模型 GLIP^[44] (Grounded language-image pre-training)、Grounding DINO^[45] (Self-distillation with no labels, DINO) 与 OWL-ViT^[46] (Vision Transformer for open-world localization) 均采用了与 CLIP 相似的双流结构, 利用抽象的自然语言作为目标特征的语义指导. 基础模型通过学习图像和语言的特征嵌入, 不仅能够完成已知类别的目标检测, 还能够处理开放词域下的新类别, 从而提高系统在复杂和动态环境下的泛化能力. GLIP 采用了自训练 (Self-training) 的方式来提升样本利用率, 模型根据网络爬取的图像生成与之对应的文本信息, 构成新的可用于训练的图像-文本对. GLIP 在 2 700 万条文本-图像数据对上进行预训练, 其中 300 万条来自人工注释, 2 400 万条来自网络爬取的图像-文本对. 在 COCO 数据集的零样本目标检测任务上, GLIP 达到了 61.5 平均精度 (Average precision, AP), 超过了之前的 SOTA 模型. 相比于 GLIP, Grounding DINO 与 OWL-ViT 借鉴了 DETR^[75] 的框架设计思路, 将目标检测问题等价地转换为集合预测问题, 构建了端到端的训练框架, 避免了候选区域生成 (Region proposal generation) 和非极大值抑制 (Non-maximum suppression) 等传统流程^[76-77]. OWL-ViT 采用 ViT 编码器将图片编码为固定数目的分词 (Token), 并利用每一个分词预测其对应目标类别以及边界框位置. Grounding DINO 则从文本图像的特征融合与文本表征形式两个角度改进了双流框架, 分别设计了多阶段的视觉语言模态融合机制与细粒度的文本表征方式. 目前, 目标检测领域的基础模型已经应用于真实场景下的机器人导航与操纵任务. VoxPoser^[36] 构建了基于开放词域目标检测器 OWL-ViT^[46]、图像分割模型 SAM^[51] 与长视频分割模型 XMem^[78] 的具身感知框架. 该框架首先调用 OWL-ViT 输出目标物体的边界框, 然后将其输入到 SAM^[51] 中获取目标掩码, 最后利用 XMem^[78] 实现对目标物体的准确跟踪.

2) 多模态大模型

对于具身智能无人系统而言, 单纯依赖对图像和文本的感知与理解能力, 难以满足系统在开放环境中执行任务的需求. 系统需要融合传感器数据以及其他状态信息 (如三维点云、压力、温度等), 以实现对环境更加全面和精确的感知. 如果采用类似 CLIP 的双流编码框架为每个模态单独设计编码器, 将面临两大关键难题: a) 针对每一个模态单独编码的框架将会导致不同模态在浅层的信息融合不足^[41]; b) 基于双流编码的框架需要针对不同模态间

的对齐和融合问题分别设计相应的损失函数^[40]. 然而, 随着感知模态数目的增加, 设计合理的目标函数将会变得十分困难.

针对上述多模态问题, 谷歌团队提出了具身多模态大模型 PaLM-E^[32]. 在骨干网络构建和预训练目标函数设计方面, PaLM-E 提出了基于语言嵌入空间的多模态统一表征方式, 设计了基于自回归形式的目标函数. PaLM-E 架构的核心思想是将连续的具身感知信息 (如图像、状态估计及其他传感器模态) 进行分词化 (Tokenization) 处理, 将其映射到与语言模态共享的高维空间中. 通过这种方式, 所有模态的信息都被统一转换为分词向量, 使得视觉、语言和传感器数据能够在同一表征空间内进行联合处理. 这种多模态信息的表示方法能够充分利用预训练模型的语言理解与逻辑推理能力, 提升了系统在复杂环境下的性能表现. 此外, PaLM-E 将所有训练任务的目标函数统一设计为类似 GPT 模型^[16, 18-19] 的自回归生成概率函数, 输出结果根据具体任务类型解码为文本序列或机器人技能序列. PaLM-E 通过构建统一的多任务训练框架, 实现了视觉、语言和传感器数据的高效融合, 确保不同任务之间的共享知识能够通过跨模态的表示方式得到有效传递. 测试结果表明, PaLM-E 在机器人操纵、视觉问答、视觉语言导航等多种具身任务中具有持平甚至超越之前 SOTA 框架的性能, 是当前最先进的具身多模态大模型之一.

除了图像和视频等二维视觉信息, 三维深度图数据同样能够有效提升无人系统对场景的全面感知与理解能力. 深度图通过提供环境的精确三维几何信息, 使得系统能够更加准确地进行自身定位, 并识别目标物体的空间位置. 针对三维物体识别任务, PointCLIP^[79] 采用了与 CLIP 多模态对齐类似的处理方法, 构建了面向三维点云信息与类别文本的双流编码结构, 再利用对比学习的方式完成不同模态嵌入向量的对齐. 点云信息的处理利用 CLIP 视觉图像编码器对多视角的点云投影图进行编码, 同时设计了可调节参数的残差适配模块 (Adapter) 来确保每个视图都能够整合全局信息. 在训练过程中, PointCLIP 冻结了 CLIP 图像和文本编码器的网络参数, 仅针对适配模块进行训练, 以此避免由于模型规模过大所导致的过拟合问题. 在三维物体分类数据集 ModelNet10、ModelNet40^[52] 测试中, PointCLIP 能够在少样本学习条件下, 表现出与在完整数据集上训练的 CurveNet^[80] 相当的性能. 由于 PointCLIP 需要将三维数据投射到二维平面才能进行编码, 因此难以避免原始三维点云特征的损失. 针对

上述问题, ULIP^[81] 构建了三维点云、文本和图像的统一表示框架, 对每个模态单独编码, 其中图像和文本使用 CLIP 编码器, 三维点云采用 PointNet++^[82] 和 PointMLP^[83]. 在 ModelNet40 数据集的零样本三维分类任务中, ULIP 达到了 49.9% 的准确率, 显著优于 PointCLIP.

2.2 基于场景重建的表征与开放环境的场景理解

场景表示的主要目的是将任务场景表示为计算机可以存储和处理的结构性形式. 场景表示不仅应包括场景中物体的形状、大小、位置和可交互性等特征信息, 还应涵盖物体之间的空间关系, 如距离、遮挡等. 场景表示有助于无人系统准确理解周围环境的空间结构, 掌握场景的动态变化, 从而实现高效的推理与决策. 三维场景重建则指根据任务场景已知的数据 (如图像、视频或三维点云) 构建三维数字模型的过程. 场景重建为场景表示提供了多视角的描述方式, 超越了传统二维图像或视频的限制性.

神经辐射场^[84] (Neural radiance field, NeRF) 是一种利用少量二维图像重建高质量三维场景的技术, 能够生成复杂场景下的新视角图像. 该方法通过深层全连接神经网络隐式存储静态场景的三维信息, 以场景中的五维坐标 (空间位置和视角方向) 作为输入, 输出对应坐标的体素密度与 RGB 值, 并借助体渲染技术还原视角下物体的几何形状与颜色. 然而, NeRF 生成的仅是 RGB 密度场, 并未包含任何高级语义信息, 这限制了其在机器人导航、操纵等需要语义信息任务中的应用. Kerr 等^[53] 提出了语言嵌入辐射场 (Language embedded radiance field, LERF), 利用特征金字塔和 CLIP 图像编码器提取多个视角下二维图像的多尺度特征, 并采用 NeRF 监督训练的方式重建三维场景. 基于 CLIP 特征重建的三维场景包含了丰富的语义信息, 实现了场景物体与自然语言的对齐, 能够支持开放场景下的物体重排、视觉问答等任务. Shen 等^[54] 提出了基于三维特征场的机器人操纵方法 (Feature fields for robotic manipulation, F3RM), 设计了融合二维图像语义特征和三维几何特征的场景表示方法, 即蒸馏特征场 (Distilled feature fields, DFF). DFF 首先利用 CLIP 提取二维图像特征, 再通过 Mask-CLIP^[50] 的重参数化技巧获得场景密集特征, 最终将其嵌入三维体积 (3D volume) 中. 基于 DFF 的场景表示融合了三维几何特征与语义特征, 为机器人进行复杂环境下的精确操纵提供了丰富的感知信息和可靠的语义理解支持.

基于 NeRF 的场景表示虽然能够较好地表征

物体的形状、姿态等几何特征, 但也不可避免地继承了 NeRF 的局限性. 首先, NeRF 是一种静态的三维场景重建方法, 因此当场景发生动态变化时, 必须重新进行场景扫描和建模, 而重建过程通常十分耗时. 其次, NeRF 依赖于多视角图像来完成场景重建, 然而在机器人导航任务中, 往往要求机器人仅利用自我中心视角下的数据实现场景表征与理解. 为了解决动态场景表征问题, Gadre 等^[55] 提出了基于自我中心 RGB 图像连续场景表示方法 (Continuous scene representations, CSR). CSR 是一种基于图结构的场景表示方法, 其中节点表示场景中的物体, 边表示物体之间的关系, 节点和边均以连续值的形式表征. CSR 在导航过程中能够将当前观测的局部子图嵌入到全局图中, 并在线更新节点和边的状态, 以反映物体及其关系的变化. 这种灵活的图结构及其更新方式使得 CSR 适用于存在动态变化的任务场景.

2.3 面向具身任务的决策与规划

在机器人学领域, 具身任务指的是机器人在现实世界中执行的涉及物理交互的任务, 例如视觉语言导航、物体重排、多机协作等. 此类任务要求机器人不仅能够与物理环境进行交互, 还需具备适应环境动态变化的能力. 目前, 视觉语言模型和多模态大模型^[12, 58] 能够理解自然语言指令, 完成对多模态信息的对齐与融合, 甚至可以通过重建场景实现更为全面的三维场景表征. 然而, 要实现人类水平的推理与决策, 系统还需要具有自然语言指令锚定、零样本逻辑推理、实时决策与规划等能力^[85].

大语言模型经过大规模互联网数据集的预训练, 包含了丰富的世界知识, 能够完成文本生成、代码生成和常识推理等任务^[86]. 然而, 目前的大多数大语言模型主要应用于语言类任务, 无法直接用于驱动现实环境中的机器人行为与动作^[39]. 如果大语言模型未能将机器人行为指令准确锚定到合理的可执行技能上, 可能会导致机器人执行与现实情境不符的操作. 针对上述问题, Huang 等^[87] 提出了一种基于提示词的指令锚定方法, 将自然语言指令精确规划为一系列可执行的机器人技能序列. 该方法首先利用大语言模型自回归地生成自然语言短语序列, 然后使用 RoBERTa 模型^[88] 计算自然语言短语与可执行技能的语义相似度, 通过选取相似度最高的技能, 完成机器人指令锚定. 为提高规划的准确性, 作者采用了上下文学习 (In-context learning), 通过示例提示提高指令与技能的匹配度. 不同于基于语义相似度的锚定方法, Ichter 等^[34] 设计了一种

基于可供性函数 (Affordance functions) 的指令锚定方法. 该方法通过综合计算上层指令的重要性得分和下层技能的可供性, 来选择最终执行的技能. 此外, VoxPoser^[36] 通过构建密集型的价值地图的方式来实现机器人技能的规划与底层控制序列的映射. 该方法首先利用大语言模型生成用于计算三维体素价值地图的 Python 程序, 该程序能够调用视觉语言模型来获取目标对象的空间几何信息, 并根据任务需求和约束要求输出三维空间任一点的价值分数. 最终, VoxPoser 采用模型预测控制实时求解最优控制序列. 相比于将语言指令锚定到有限数目技能集的传统方法, VoxPoser 能够将指令映射为更高细粒度的控制序列, 显著提升了系统的灵活性.

在处理自然语言任务时, 大语言模型展现了零样本推理与决策能力. 为了将大语言模型的推理与决策能力应用于机器人具身任务中, Huang 等^[59] 提出了一种基于大语言模型的内部对话机制 (Inner Monologue). Inner Monologue 是一种利用大语言模型对多个反馈源信息进行闭环处理的实时规划机制, 其中反馈信息包括自然语言反馈、场景描述以及技能执行的成功检测等. 大语言模型的闭环处理机制允许系统实时将各类反馈信息纳入规划过程中, 形成动态调整的“内部对话”, 从而实现有效的决策和任务重规划. 然而受限于给定的技能集, Inner Monologue 的规划通常缺乏技能的灵活性与控制的精确性. 针对上述问题, Liang 等^[89] 提出了一种基于语言模型编程的机器人规划与控制框架 (Code as policies, CaP). CaP 利用大语言模型生成的程序逻辑结构以及参数化的底层控制基元, 实现了精确的控制输出. CaP 使得规划过程不再受限于特定的技能集, 而是能够根据任务需求规划机器人的行为动作, 生成细粒度的控制序列 (如机械臂向前移动 1 cm). 在阻抗控制、基于视觉的拾取与放置以及轨迹控制等任务中, CaP 均展示了出色的操纵性能. 除了利用大模型直接完成规划与决策外, Du 等^[90] 还探索了如何利用符合客观世界物理规律的视频生成模型辅助机器人完成规划. 该研究提出了视频语言规划框架 (Video language planning, VLP), 其核心思想是将视频生成模型视为一种简化的世界模型^[61], 利用该模型预测机器人执行给定动作的视频帧, 并根据任务的最终目标计算视频帧的价值分数, 最后采用树搜索的方法选取当前状态下的最优动作.

2.4 基于动态交互的机器人学习与控制

在复杂、未知且开放的环境中, 通常无法预先

为机器人设计涵盖所有可能情况的控制策略. 因此, 执行此类任务要求机器人具备自适应的学习能力^[60]. 机器人不应仅依赖于封闭的数据集进行学习, 而应通过与外界的交互, 理解和适应动态环境, 并据此优化原有的控制策略. 以下将具体探讨基于自然语言交互的机器人学习以及基于视觉可供性学习的控制方法.

基于自然语言交互的机器人学习是通过人类的自然语言指导, 优化机器人的行为策略, 帮助其快速适应当前环境. Zha 等^[62] 提出了一种基于大语言模型的机器人在线纠错框架 (Distillation and retrieval of online corrections, DROC), 该框架能够接收自然语言形式的反馈信息, 在纠错过程中提炼出通用的知识, 以提升机器人在新环境中的表现. 由于大语言模型的上下文长度有限, 原有的知识可能被后续交互信息覆盖, 为解决这一问题, DROC 设计了基于文本和视觉相似性的知识检索机制. 提炼后的知识被存入知识库, 通过检索与当前任务语义特征和视觉特征最为相似的历史知识, 指导机器人的当前行为. 随着 DROC 知识库的扩展, 人为干预的次数逐渐下降, 机器人在未知环境的任务成功率不断提升. 凭借有效的知识存储和检索机制, DROC 在多个长程操作任务中的表现性能超越了之前的基线算法. 然而, DROC 本质上仍依赖于上下文学习, 随着任务复杂性的增加, 所需检索的知识提示也会相应增加, 这限制了其应用场景的广泛性. 为解决上述问题, Liang 等^[91] 提出了结合上下文学习与模型微调的语言模型预测控制框架 (Language model predictive control, LMPC). LMPC 的核心思想是将人机交互建模为部分可观测的马尔科夫决策过程, 通过监督微调 (Supervised fine-tuning, SFT) 构建人机交互的“动力学模型”, 然后使用启发式搜索和回退时域控制求解最优动作序列. LMPC 利用了上下文学习的在线指导与语言模型预测控制滚动优化的优势, 提高了机器人在未知环境中的任务成功率. LMPC 在 5 种机器人和 78 个具身任务上进行了测试, 相较于传统算法, 将任务的成功率提升了 26.9%, 同时将人类干预次数从 2.4 次降低到 1.9 次. 由于获取自然语言标注的机器人训练数据的成本高昂, 因此如何高效利用已有的数据完成训练, 最大化策略的泛化能力和有效性, 是该研究领域亟需解决的问题之一. Lynch 和 Sermanet^[92] 提出了基于语言条件的多上下文模仿学习 (Multicontext imitation learning, MCIL) 方法, 构建了基于共享隐目标空间 (Latent goal space) 的端到端训练框架. MCIL 通过设计多模态任务描述 (包括图像

描述、自然语言描述等)的关联编码器,将具有相同语义的任务描述映射到共享隐目标空间中的相邻位置。由于图像目标描述可以通过抽取视频帧的方式大量获取,因此 MCIL 极大地减少了对语言标注训练数据的依赖。

基于视觉可供性学习 (Visual affordance learning) 的控制是该领域的重要研究方向之一^[63-64]。视觉可供性是指依据视觉感知的输入推理得到可执行的目标物体交互方式或行为,为实现机器人与环境的高效交互提供支持。Mo 等^[65]提出了基于可供性预测网络的机器人控制框架 Where2Act。该框架首先将机器人的操纵任务分解为若干操纵基元(如推、拉等),然后针对目标物体的每一个像素点以及点云数据,预测机器人操纵行为的交互成功率。Where2Act 设计了基于 U-net^[93]以及 PointNet++^[82]的编码网络和解码网络,输出空间中每个像素点的高维特征,以实现视觉可供性的密集型预测。该框架成功实现了对 15 种物品类别(共计 972 种形状)的 6 类常见交互操作,并具有对未知物品类别的泛化能力。为了推理物体间的交互可能性,Mo 等^[94]进一步提出了物体-物体的可供性学习框架 O2O-Afford。物体-物体可供性学习的主要目标是判断物体间交互的合理性,例如,在机器人使用工具完成目标的任务场景中,除了需要实现机器人自身与工具的交互外,还需确定工具与目标物体之间的交互合理性。O2O-Afford 框架采用了目标核(Object-kernel)点卷积网络,将目标物体的点云特征作为卷积核,在采样的场景中进行点卷积操作,聚合目标物体与场景之间的像素点特征,以实现视觉可供性的预测。

2.5 具身智能模拟器

实现无人系统的具身智能不仅依赖于互联网获取的静态数据(如图像、视频或文本),还需要通过系统与环境的实时交互来获取类似人类自我中心感知(Ego-centric perception)的多模态数据。通常情况下,在真实场景中直接获取此类数据的成本较高。此外,真实场景中的机器人数据采集受限于训练场景的单一性,可能导致控制策略的泛化能力有限,难以适应新的任务场景。因此,如何获取和生成长序列多模态交互数据是该领域亟待解决的关键问题之一。相较于直接在现实场景中获取数据的方法,构建高保真的具身智能模拟器已成为高效获取训练数据的主要手段之一。具身智能模拟器利用计算机图形学技术与物理引擎重建真实的物理环境,模拟客观的物理规律,从而真实反映物体(包括流体与固体)之间的相互作用。由于具身智能模拟器不受

物理时空限制,可以采用多种方式构建虚拟场景,且能够并行执行多个环境的仿真与数据采集,因此具备可扩展性强、成本低以及效率高等优势。目前,常见的具身智能模拟器包括 Habitat-Sim^[95]、AI2-THOR^[67]、ThreeDWorld^[68]和 iGibson^[96]等。以下将从场景与可交互对象的建模、交互过程的建模以及人机交互接口的设计等方面介绍模拟器的主要特点。

场景与可交互对象建模:场景与可交互对象建模是智能体获取自我中心感知数据的基础。模拟器需要提供结构合理且逼真的虚拟场景,以反映真实环境的特征。虚拟场景的构建可以通过程序生成^[69],也可以通过对真实世界进行扫描来实现。通过程序生成的方式可以高效地生成大批量训练场景,从而丰富采集数据的多样性。ProcTHOR-10K 模拟器通过程序生成 10 000 个语义上合理的室内场景,用于训练机器人导航和操纵策略,实验结果表明,采用 ProcTHOR-10K 进行训练的机器人策略具有较强的泛化性能。与基于程序生成的方法相比,通过扫描构建的场景具有更高的保真度,能够更准确地反映真实世界的图像信息,从而更有利于实现智能体的虚实迁移。Matterport3D 模拟器基于 10 800 张密集采样的全景 RGB-D 图像构建了 90 个真实建筑的室内场景模拟环境^[70],确保场景中每个三维对象(如咖啡杯、盆栽和壁纸纹理)的独特性。可交互对象模型的主要来源包括 Gibson 数据集、Matterport3D 数据集、SUNCG 数据集以及互联网的三维对象模型。目前,各类模拟器的对象数据库正在迅速增加,其中 Gibson 与 AI2-THOR 的数据库分别包含超过 500 个和 3 000 个可交互对象。

交互过程建模:模拟器不仅需要构建逼真的三维场景,还需建模物体间的交互过程,以反映真实世界客观的物理规律。交互过程可根据交互细节的建模程度分为抽象交互和基于物理特性的交互。抽象交互不关注底层控制的具体实现,而是为智能体预先设定了可执行的运动与操纵技能集,包括移动、旋转、打开、拾取、放置等一系列连贯的控制序列^[67]。智能体直接通过选择技能与环境进行交互,只要智能体与交互对象之间的距离满足要求,所选择的技能便会成功执行。相比之下,基于物理特性的交互要求对客观的物理规律进行建模,例如刚体动力学模型和流体动力学模型^[68]。此类建模旨在客观反映交互过程中目标物体的物理状态变化,以满足不同任务场景对细粒度模拟的需求。

人机交互接口设计:模拟器可通过虚拟-现实控制器提供人机交互接口,使用户能够直接控制智能体,以进行模仿学习的数据采集或实时交互。文

献 [71] 在 AI2-THOR 中引入了用于人机交互的手势指令. 用户可以利用虚拟现实设备操控模拟器中的手势, 从而向智能体下达任务指令, 例如, 用户可以通过指向某个可交互对象来指示智能体移动该物体.

3 具身智能无人系统典型任务

随着研究重点从“互联网人工智能”逐渐扩展到“具身人工智能”, 智能体的学习范式不再局限于静态数据集, 还需要能够利用实时的环境交互数据在线优化自身策略. 本节结合具身智能交互式学习的特点, 重点探讨具身智能在机器人领域的两项典型任务——视觉导航和基于语言模型驱动的机器人操纵, 并对相关研究进展进行概述.

3.1 视觉导航

视觉导航 (Visual navigation) 是具身智能领域中的典型任务之一, 其要求智能体根据任务指令和图像信息导航至指定目标, 如位置点、物体或区域 [72]. 视觉导航任务的核心在于感知、决策与执行的紧密结合, 智能体需要从视觉输入中理解环境、识别目标, 并完成路径规划. 在该过程中, 多模态感知信息的处理能力尤为重要, 包括 RGB 图像、深度图、连续传感器信息等. 这些感知信息可以帮助智能体实现对环境的语义理解和空间几何理解, 从而有效应对复杂、动态的非结构化环境. 近年来, 随着大语言模型技术的快速发展, 视觉-语言导航 (Vision-and-language navigation, VLN) 逐渐成为视觉导航领域的重要研究方向之一 [97]. VLN 任务要求智能体能够理解语言指令, 并将其与视觉感知中的目标物体进行有效关联, 完成目标位置或物体的定位和导航. 此外, 视觉导航的任务形式还在不断扩展. 音频-视觉-语言导航 (Audio-visual-language navigation, AVLN) 引入了音频感知模态, 使得智能体能够利用声音源定位进行导航 [98]. 这为智能体在视觉受限条件下的导航提供了新路径, 进一步增强了智能体在真实世界环境中的适应能力. 具身问答 (Embodied question answering, EQA) 要求智能体在导航过程中通过自然语言问答的形式与环境进行交互, 并根据反馈信息完成目标定位与推理任务 [99].

在复杂、开放任务场景中, 首先要求智能体具备零样本目标导航能力, 即在缺乏特定目标物体的训练数据时, 仍能够通过语言描述和视觉感知准确识别、定位新目标, 并完成路径规划, 这依赖于智能体的自然语言理解和视觉-语言信息的对齐能力. Gadre 等 [74] 研究了面向开放场景的零样本目标导航算法, 提出了 CLIP on wheels (CoW) 框架. 该框架

通过引入视觉语言模型 CLIP, 以实现开放词域下的目标识别与定位. 然而, 在未知的任务场景中, 通常无法获取足够的标注的训练样本. 针对上述问题, Majumdar 等 [100] 提出了一种基于学习的零样本目标导航框架. 该框架首先利用预训练的 CLIP 将目标图像嵌入高维语义空间中, 并训练智能体执行语义目标导航 (Semantic-goal navigation). 经过训练的导航策略只需将自然语言目标同样嵌入语义空间, 即可完成自然语言指令的导航任务. 与传统的训练方法相比, 该方法的一大优势在于通过随机采样的方式便能够快速获取大量基于目标图像的训练样本, 从而显著降低数据标注的负担. 然而, 直接采用预训练视觉语言模型进行视觉导航往往无法理解和定位语言描述中空间几何信息. 例如, 当任务目标为“在沙发和电视机之间的物品”时, CLIP 模型很难有效定位目标物体. 因此, 如何实现视觉语言模型与空间几何信息的有效关联是视觉语言导航研究的一个重要方向. Huang 等 [57] 提出了一种开放词域下的三维空间视觉-语言地图表示 (VLMaps), 通过构建场景语义地图的方式来解决基于空间描述的目标定位问题. VLMaps 利用 LSeg 获取的密集像素特征, 再结合深度图将高维特征投射到对应坐标, 从而生成场景语义地图. 此外, 该研究还利用大语言模型生成可执行的 Python 机器人代码, 代码可调用参数化的导航基元, 进而实现基于空间描述的精确导航.

如何利用预训练大模型完成导航任务中的常识推理与决策规划, 同样是视觉导航领域的研究热点之一. Zhou 等 [101] 提出了基于大语言模型的导航框架 NavGPT, 利用大模型解决导航任务中的指令分解、常识推理、导航进程跟踪以及异常情况下的策略调整等决策与规划问题. 该框架将视觉观察的文本描述、导航历史信息和未来可探索方向作为输入, 推理智能体的当前位置, 并输出下一步的决策规划. Shah 等 [73] 进一步研究了户外开放场景下的长距离机器人导航问题, 提出了自然语言驱动的机器人导航框架 LM-Nav. 该系统结合了三个预训练模型: 大语言模型 GPT-3 [19], 用于将自然语言指令解析为一系列地标目标; 视觉语言模型 CLIP [29], 通过联合概率评估地标与拓扑节点之间的关系, 并在拓扑地图中定位地标; 视觉导航模型 ViNG [102], 负责执行导航指令并输出底层控制动作. LM-Nav 实现了真实复杂场景中的长距离 (可达 800 m) 视觉语言导航.

3.2 基于语言模型驱动的操纵

机器人操纵是一类极具挑战性的具身任务 [36],

主要原因在于交互的复杂性以及精确控制的高要求。在操纵任务中，机器人不仅需要在复杂环境中识别、抓取、移动各类物体，还必须处理目标物体在操纵过程中可能发生的形变。为此，机器人不仅需要具备多模态信息的处理能力，包括视觉、触觉和连续传感器等感知信息，还需在实时环境中作出高效且合理的动作决策。

近年来，各类大模型在多模态信息理解和常识推理等方面取得了显著进展，展现出其在机器人操控领域的广阔应用前景。这主要得益于大模型通过大规模互联网数据进行训练，具备强大的多模态信息处理能力以及对物理世界的常识推理能力。然而，尽管大语言模型在多个任务中表现出色，直接将其应用于机器人操控任务仍存在一定局限性。这是因为大语言模型缺乏对机器人技能的理解，其推理与决策能力尚未与实际机器人的可执行操作技能进行有效对齐。Iceter 等^[34]提出了 SayCan 框架，通过将技能的选择与技能的执行划分为两个独立层次结构进行处理。上层利用大语言模型筛选一定数量的机器人候选技能，并根据任务的最终目标对每个技能进行评分；下层则使用可供性函数来评估每个技能在当前环境下的执行可行性。技能的最终选择将通过结合大语言模型的评分与可行性函数的概率推断来确定。然而在 SayCan 框架中，上层的技能选择与下层的技能可行性评估是分离的。这种分离结构限制了 SayCan 在动态环境中的适应性，上层的技能选择难以针对环境的实时变化做出改变，可能导致决策滞后或失误。为了解决上述问题，谷歌团队构建了多模态具身大模型 PaLM-E^[32]，实现了视觉、

语言和机器人传感器数据的高效融合。与 SayCan 框架不同，PaLM-E 不再依赖于可行性函数的辅助模型，而是构建了一个端到端的系统框架，将多模态感知信息直接映射为可执行技能。该框架使得大语言模型中丰富的世界知识和强大的推理能力能够直接应用于机器人技能的规划，从而显著提升机器人在复杂场景中的任务执行能力。尽管 PaLM-E 已经取得显著成功，但其输出仍然受到预定义技能集的限制，无法根据实际场景动态调整底层技能，这一问题成为制约机器人操纵的主要瓶颈之一。针对上述问题，谷歌团队提出了视觉-语言-动作模型 RT-2^[33]，将机器人底层细粒度的动作进行分词化处理，并直接整合进模型的训练过程。

4 具身智能的未来研究方向

尽管具身智能已经取得了飞速发展，其仍面临诸多亟待解决的问题与挑战，这些问题不仅限制了具身智能技术的广泛应用，也深刻影响着未来的研究方向。图 6 列出了具身智能目前存在的 key 问题以及未来潜在的研究方向。

4.1 虚实迁移方法

实现有效的虚实迁移是具身智能研究中的一大挑战。虚拟仿真环境虽然能够具有逼真的视觉画面，但在物理引擎方面仍与现实世界存在显著差异，尤其是在模拟柔性和流体物体方面的表现与实际情况相差甚远。虚拟到现实的差异严重限制了机器人学习精细操纵任务的能力，阻碍了从仿真到现实世界的迁移。因此，提升仿真环境中的物理真实性，缩小

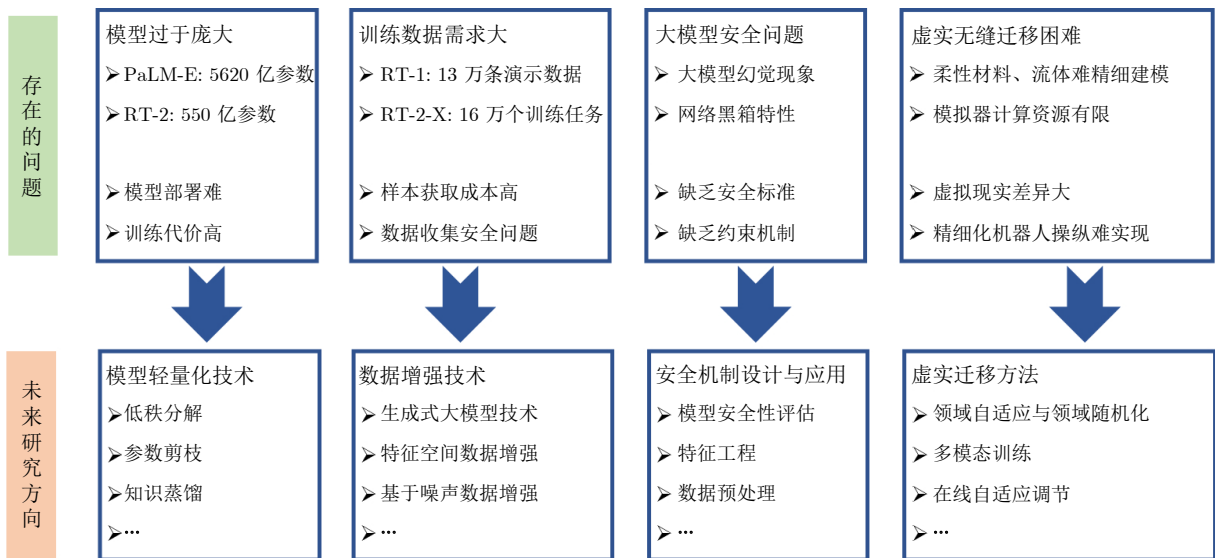


图 6 具身智能未来研究方向
 Fig.6 Future research direction of embodied intelligence

虚拟与现实之间的差距, 对于提高机器人在现实世界中的操作性能和适应性至关重要. 未来的研究应考虑开发更精确的物理模拟技术, 并探索更加鲁棒的虚实迁移策略, 逐步降低虚实迁移的难度.

4.2 模型轻量化技术

尽管大语言模型、多模态大模型已经在各类机器人任务中取得了重要突破, 但是此类模型通常基于 Transformer 框架设计, 这导致了此类模型通常参数量庞大, 对计算资源的需求较高, 因此难以部署在边缘设备和移动设备上. 模型压缩技术是一个具有显著潜力的解决方案, 其以牺牲部分性能为代价, 换取模型在计算资源受限的条件下依然能够保持较高的运行效率. 未来的研究应考虑探索不同的模型压缩方法, 如权重共享、低秩分解、参数剪枝、知识蒸馏和轻量级网络架构设计, 以及相应的模型性能评估方法, 以精确量化模型的综合能力.

4.3 大语言模型安全机制

研究人员发现大语言模型存在生成与实际来源不符的、不真实的甚至潜在危险内容的倾向. 此类幻觉现象的存在以及大语言模型固有的黑箱特性, 使得其存在无法预知的安全隐患, 这阻碍了大模型在具有高度安全性要求领域的应用. 因此, 如何完善大语言模型的安全机制仍是一个重要的研究方向. 目前, 导致大模型幻觉的可能原因包括训练样本偏差、模型过拟合、缺乏安全约束机制等. 因此, 进行适当的数据预处理、特征工程和模型安全性评估, 是确保模型输出安全可信内容的重要手段. 此外, 设计针对性防护机制和引入更加透明的解释性模型框架, 也将有助于提升大语言模型在实际应用中的安全性与可靠性.

4.4 多智能体系统

随着具身智能在单智能体应用方向上的突破, 其在多智能体系统上的应用也逐渐受到关注. 在多智能体领域, 研究主要聚焦于如何高效协同多个自主体分工合作, 共同完成单智能体无法胜任的复杂任务. 然而, 具身智能在单智能体的研究成果通常无法简单地直接应用在多智能体系统领域, 这是因为多智能体系统面临着一系列独特的挑战^[103].

首先, 高效协同和智能体间的合作策略需要精心设计. 每个智能体需要在复杂的环境中自适应地调整自身行为, 以实现整体系统的最佳性能. 这涉及到动态环境下的实时决策和在线自适应, 这种复杂性远超单智能体的情况. 其次, 不确定性是多智

能体系统中的核心问题. 环境的不确定性、其他智能体行为的不可预测性以及自身状态的不确定性, 都需要通过相应的算法进行建模与处理, 如贝叶斯方法、强化学习和模糊逻辑等, 以提高系统在不稳定条件下的决策能力和稳定性. 此外, 随着多智能体系统规模的不断扩大, 可扩展性问题变得愈发重要. 解决多智能体间通信、协调和资源分配的高效性, 避免计算复杂度的指数级增长, 成为当前亟待解决的挑战之一.

4.5 数字孪生

当前具身智能在实际落地应用中仍面临诸多问题, 其中主要包括智能体对外部环境感知精度不足、环境适应能力有限以及与复杂动态环境交互的能力差. 上述问题制约了具身智能在实际应用中的表现. 而数字孪生技术可以通过创建真实场景的虚拟副本, 来提供对环境的实时监测、模拟和预测. 其主要优势包括: 高精度环境建模和实时数据集成. 将具身智能与数字孪生相结合, 可以充分发挥两者的优势. 具身智能通过实时交互和感知来优化无人系统的决策与控制, 而数字孪生提供了高精度的虚拟环境模型来增强系统对复杂环境的理解.

5 总结

本文综述了具身智能无人系统的现状与关键技术, 深入探讨了大语言模型驱动的多模态感知、面向具身任务的推理与决策、基于交互的机器人学习与控制、以及具身智能模拟器的研究进展与发展脉络. 结合具身智能无人系统当前面临的挑战, 本文分析了未来的潜在发展方向. 目前, 具身智能的研究仍处于探索阶段, 如何确保大语言模型驱动的无人系统的安全性, 如何实现从虚拟环境到现实世界的策略迁移, 如何设计轻量化易于终端部署的模型结构, 以及如何面向多智能体系统构建高效的协同合作机制, 仍是目前亟需解决的关键问题.

References

- 1 Gupta A, Savarese S, Ganguli S, Li F F. Embodied intelligence via learning and evolution. *Nature Communications*, 2021, 12(1): Article No. 5721
- 2 Sun Chang-Yin, Mu Chao-Xu, Liu Wen-Zhang, Wang Xiao. Embodied cognitive intelligence framework of unmanned autonomous systems. *Science & Technology Review*, 2024, 42(12): 157-166
(孙长银, 穆朝絮, 柳文章, 王晓. 自主无人系统的具身认知智能框架. 科技导报, 2024, 42(12): 157-166)
- 3 Wiener N. *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge: MIT Press, 1961.
- 4 Turing A M. *Computing Machinery and Intelligence*. Oxford: Oxford University Press, 1950.
- 5 Wang Yao-Nan, An Guo-Wei, Wang Chuan-Cheng, Mo Yang,

- Miao Zhi-Qiang, Zeng Kai. Technology application and development trend of intelligent unmanned system. *Chinese Journal of Ship Research*, 2022, **17**(5): 9–26
(王耀南, 安果维, 王传成, 莫洋, 缪志强, 曾凯. 智能无人系统技术应用与发展趋势. 中国舰船研究, 2022, **17**(5): 9–26)
- 6 Kaufmann E, Bauersfeld L, Loquercio A, Müller M, Koltun V, Scaramuzza D. Champion-level drone racing using deep reinforcement learning. *Nature*, 2023, **620**(7976): 982–987
- 7 Feng S, Sun H W, Yan X T, Zhu H J, Zou Z X, Shen S Y, et al. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 2023, **615**(7953): 620–627
- 8 Zhang Peng-Fei, Cheng Wen-Zheng, Mi Jiang-Yong, He Ye-Long, Li Ya-Wen, Wang Li-Jin. Research status and prospect of key technologies for counter UAV swarm. *Journal of Gun Launch & Control*, DOI: [10.19323/j.issn.1673-6524.202311017](https://doi.org/10.19323/j.issn.1673-6524.202311017)
(张鹏飞, 程文铮, 米江勇, 和焯龙, 李亚文, 王力金. 反无人机蜂群关键技术研究现状及展望. 火炮发射与控制学报, DOI: [10.19323/j.issn.1673-6524.202311017](https://doi.org/10.19323/j.issn.1673-6524.202311017))
- 9 Zhang Lin. New insights into U.S. military anti-drone system technology. *Tank & Armoured Vehicle*, 2024(11): 22–29
(张琳. 美军反无人机系统技术新解. 坦克装甲车辆, 2024(11): 22–29)
- 10 Dong Zhao-Rong, Zhao Min, Jiang Li, Wang Zhi. Review on key technologies of autonomous collaboration in heterogeneous unmanned system cluster. *Journal of Telemetry, Tracking and Command*, 2024, **45**(4): 1–11
(董昭荣, 赵民, 姜利, 王智. 异构无人系统集群自主协同关键技术综述. 遥测遥控, 2024, **45**(4): 1–11)
- 11 Jiang Bi-Tao, Wen Guang-Hui, Zhou Jia-Ling, Zheng De-Zhi. Cross-domain cooperative technology of intelligent unmanned swarm systems: Current status and prospects. *Strategic Study of CAE*, 2024, **26**(1): 117–126
(江碧涛, 温广辉, 周佳玲, 郑德智. 智能无人集群系统跨区域协同技术研究现状及展望. 中国工程科学, 2024, **26**(1): 117–126)
- 12 Firoozi R, Tucker J, Tian S, Majumdar A, Sun J K, Liu W Y, et al. Foundation models in robotics: Applications, challenges, and the future. arXiv: 2312.07843, 2023.
- 13 Lan Feng-Bo, Zhao Wen-Bo, Zhu Kai, Zhang Tao. Development of mobile manipulator robot system with embodied intelligence. *Strategic Study of CAE*, 2024, **26**(1): 139–148
(兰沅卜, 赵文博, 朱凯, 张涛. 基于具身智能的移动操作机器人系统发展研究. 中国工程科学, 2024, **26**(1): 139–148)
- 14 Liu Hua-Ping, Guo Di, Sun Fu-Chun, Zhang Xin-Yu. Morphology-based embodied intelligence: Historical retrospect and research progress. *Acta Automatica Sinica*, 2023, **49**(6): 1131–1154
(刘华平, 郭迪, 孙富春, 张新钰. 基于形态的具身智能研究: 历史回顾与前沿进展. 自动化学报, 2023, **49**(6): 1131–1154)
- 15 Zhang Ba, Zhu Jun, Su Hang. Toward the third generation of artificial intelligence. *SCIENTIA SINICA Informationis*, 2020, **50**(9): 1281–1302
(张钹, 朱军, 苏航. 迈向第三代人工智能. 中国科学: 信息科学, 2020, **50**(9): 1281–1302)
- 16 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [Online], available: <https://gwern.net/doc/www/s3-us-west-2.amazonaws.com/d73fde5ffa8627bce44dceda2fc012da638ffb158.pdf>, January 4, 2025
- 17 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minnesota, USA: ACL, 2018. 4171–4186
- 18 Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners [Online], available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, January 4, 2025
- 19 Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv: 2005.14165, 2020.
- 20 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017. 6000–6010
- 21 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv: 2010.11929, 2021.
- 22 He K M, Chen X L, Xie S N, Li Y H, Dollár P, Girshick R, et al. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 15979–15988
- 23 Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9992–10002
- 24 Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv: 2302.13971, 2023.
- 25 Kim W, Son B, Kim I. ViLT: Vision-and-language Transformer without convolution or region supervision. arXiv: 2102.03334, 2021.
- 26 Li J N, Li D X, Xiong C M, Hoi S C H. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: ICML, 2022. 12888–12900
- 27 Yu J H, Wang Z R, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y H. CoCa: Contrastive captioners are image-text foundation models. arXiv: 2205.01917, 2022.
- 28 Bao H B, Wang W H, Dong L, Wei F R. VL-BEiT: Generative vision-language pretraining. arXiv: 2206.01127, 2022.
- 29 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv: 2103.00020, 2021.
- 30 Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv: 2203.02155, 2022.
- 31 Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. RT-1: Robotics Transformer for real-world control at scale. arXiv: 2212.06817, 2022.
- 32 Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: An embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: ICML, 2023. 8469–8488
- 33 Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choremanski K, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv: 2307.15818, 2023.
- 34 Ichter B, Brohan A, Chebotar Y, Finn C, Hausman K, Herzog A, et al. Do as I can, not as I say: Grounding language in robotic affordances. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022. 287–318
- 35 Bousmalis K, Vezzani G, Rao D, Devin C, Lee A X, Bauza M, et al. RoboCat: A self-improving foundation agent for robotic manipulation. arXiv: 2306.11706, 2023.
- 36 Huang W L, Wang C, Zhang R H, Li Y Z, Wu J J, Li F F. VoxPoser: Composable 3D value maps for robotic manipula-

- tion with language models. In: Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR, 2023. 540–562
- 37 O’Neill A, Rehman A, Gupta A, Maddukuri A, Gupta A, Padalkar A, et al. Open X-embodiment: Robotic learning datasets and RT-X models. arXiv: 2310.08864, 2024.
- 38 Zeng F L, Gan W S, Wang Y H, Liu N, Yu P S. Large language models for robotics: A survey. arXiv: 2311.07226, 2023.
- 39 Bommasani R, Hudson D A, Adeli E, Altman E, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv: 2108.07258, 2021.
- 40 Wang W H, Bao H B, Dong L, Bjorck J, Peng Z L, Liu Q, et al. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. arXiv: 2208.10442, 2022.
- 41 Bao H B, Wang W H, Dong L, Liu Q, Mohammed O K, Aggarwal K, et al. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. arXiv: 2111.02358, 2022.
- 42 Chen F L, Zhang D Z, Han M L, Chen X Y, Shi J, Xu S, et al. VLP: A survey on vision-language pre-training. *Machine Intelligence Research*, 2023, **20**(1): 38–56
- 43 Peng F, Yang X S, Xiao L H, Wang Y W, Xu C S. SgVA-CLIP: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Transactions on Multimedia*, 2024, **26**: 3469–3480
- 44 Li L H, Zhang P C, Zhang H T, Yang J W, Li C Y, Zhong Y W, et al. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 10955–10965
- 45 Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv: 2303.05499, 2023.
- 46 Minderer M, Gritsenko A A, Stone A, Neumann M, Weissenborn D, Dosovitskiy A, et al. Simple open-vocabulary object detection. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 728–755
- 47 Xu J R, de Mello S, Liu S F, Byeon W, Breuel T, Kautz J, et al. GroupViT: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 18113–18123
- 48 Li B Y, Weinberger K Q, Belongie S J, Koltun V, Ranftl R. Language-driven semantic segmentation. arXiv: 2201.03546, 2022.
- 49 Ghiasi G, Gu X Y, Cui Y, Lin T Y. Scaling open-vocabulary image segmentation with image-level labels. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 540–557
- 50 Zhou C, Loy C C, Dai B. Extract free dense labels from clip. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 696–712
- 51 Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3992–4003
- 52 Wu Z R, Song S R, Khosla A, Yu F, Zhang L G, Tang X O, et al. 3D ShapeNets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1912–1920
- 53 Kerr J, Kim C M, Goldberg K, Kanazawa A, Tancik M. LERF: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 19672–19682
- 54 Shen W, Yang G, Yu A L, Wong J, Kaelbling L P, Isola P. Distilled feature fields enable few-shot language-guided manipulation. In: Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR, 2023. 405–424
- 55 Gadre S Y, Ehsani K, Song S R, Mottaghi R. Continuous scene representations for embodied AI. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 14829–14839
- 56 Shafiqullah N M, Paxton C, Pinto L, Chintala S, Szlam A. CLIP-fields: Weakly supervised semantic fields for robotic memory. arXiv: 2210.05663, 2022.
- 57 Huang C G, Mees O, Zeng A, Burgard W. Visual language maps for robot navigation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 10608–10615
- 58 Gan Z, Li L J, Li C Y, Wang L J, Liu Z C, Gao J F. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 2022, **14**(3–4): 163–352
- 59 Huang W L, Xia F, Xiao T, Chan H, Liang J, Florence P, et al. Inner monologue: Embodied reasoning through planning with language models. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022. 1769–1782
- 60 Sun Y W, Zhang K, Sun C Y. Model-based transfer reinforcement learning based on graphical model representations. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(2): 1035–1048
- 61 Hao S, Gu Y, Ma H D, Hong J, Wang Z, Wang D, et al. Reasoning with language model is planning with world model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 8154–8173
- 62 Zha L H, Cui Y C, Lin L H, Kwon M, Arenas M G, Zeng A, et al. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 15172–15179
- 63 Hassanin M, Khan S, Tahtali M. Visual affordance and function understanding: A survey. *ACM Computing Surveys*, 2022, **54**(3): Article No. 47
- 64 Luo H C, Zhai W, Zhang J, Cao Y, Tao D C. Learning visual affordance grounding from demonstration videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(11): 16857–16871
- 65 Mo K C, Guibas L, Mukadam M, Gupta A, Tulsiani S. Where2Act: From pixels to actions for articulated 3D objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 6793–6803
- 66 Geng Y R, An B S, Geng H R, Chen Y P, Yang Y D, Dong H. RLAfford: End-to-end affordance learning for robotic manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 5880–5886
- 67 Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, et al. AI2-THOR: An interactive 3D environment for visual AI. arXiv: 1712.05474, 2017.
- 68 Gan C, Schwartz J, Alter S, Mrowca D, Schrimpf M, Traer J, et al. ThreeDWorld: A platform for interactive multi-modal physical simulation. arXiv: 2007.04954, 2020.
- 69 Deitke M, VanderBilt E, Herrasti A, Weihs L, Salvador J, Ehsani K, et al. ProcTHOR: Large-scale embodied AI using procedural generation. arXiv: 2206.06994, 2022.
- 70 Anderson P, Wu Q, Teney D, Bruce J, Johnson M, Sünderrhauf N, et al. Vision-and-language navigation: Interpreting visually-

- grounded navigation instructions in real environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3674–3683
- 71 Wu Q, Wu C J, Zhu Y X, Joo J. Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021. 4095–4102
- 72 Duan J F, Yu S, Tan H L, Zhu H Y, Tan C. A survey of embodied AI: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022, **6**(2): 230–244
- 73 Shah D, Osinski B, Levine S, Levine S. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022. 492–504
- 74 Gadre S Y, Wortsman M, Ilharco G, Schmidt L, Song S R. CoWs on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 23171–23181
- 75 Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoryuyko S. End-to-end object detection with Transformers. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 213–229
- 76 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2015. 91–99
- 77 Jiang P Y, Ergu D, Liu F Y, Cai Y, Ma B. A review of Yolo algorithm developments. *Procedia Computer Science*, 2022, **199**: 1066–1073
- 78 Cheng H K, Alexander G S. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 640–658
- 79 Zhu X Y, Zhang R R, He B W, Guo Z Y, Zeng Z Y, Qin Z P, et al. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 2639–2650
- 80 Muzahid A A M, Wan W G, Soheli F, Wu L Y, Hou L. CurveNet: Curvature-based multitask learning deep networks for 3D object recognition. *IEEE/CAA Journal of Automatica Sinica*, 2021, **8**(6): 1177–1187
- 81 Xue L, Gao M F, Xing C, Martín-Martín R, Wu J J, Xiong C M, et al. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1179–1189
- 82 Qi C R, Yi L, Su H, Guibas L J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017. 5105–5114
- 83 Ma X, Qin C, You H X, Ran H X, Fu Y. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. arXiv: 2202.07123, 2022.
- 84 Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R, Ng R. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2022, **65**(1): 99–106
- 85 Zeng A, Attarian M, Ichter B, Choromanski K M, Wong A, Welker S, et al. Socratic models: Composing zero-shot multimodal reasoning with language. arXiv: 2204.00598, 2022.
- 86 Li B Z, Nye M, Andreas J. Implicit representations of meaning in neural language models. arXiv: 2106.00737, 2021.
- 87 Huang W L, Abbeel P, Pathak D, Mordatch I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 9118–9147
- 88 Liu Y H, Ott M, Goyal N, Du J F, Joshi M, Chen D Q, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.
- 89 Liang J, Huang W L, Xia F, Xu P, Hausman K, Ichter B, et al. Code as policies: Language model programs for embodied control. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 9493–9500
- 90 Du Y L, Yang M, Florence P, Xia F, Wahid A, Ichter B, et al. Video language planning. arXiv: 2310.10625, 2023.
- 91 Liang J, Xia F, Yu W H, Zeng A, Arenas M G, Attarian M, et al. Learning to learn faster from human feedback with language model predictive control. arXiv: 2402.11450, 2024.
- 92 Lynch C, Sermanet P. Language conditioned imitation learning over unstructured data. arXiv: 2005.07648, 2020.
- 93 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention—MICCAI 2015. Munich, Germany: Springer, 2015. 234–241
- 94 Mo K C, Qin Y Z, Xiang F B, Su H, Guibas L J. O2O-Afford: Annotation-free large-scale object-object affordance learning. In: Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR, 2021. 1666–1677
- 95 Savva M, Kadian A, Maksymets O, Zhao Y L, Wijmans E, Jain B, et al. Habitat: A platform for embodied AI research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 9338–9346
- 96 Xia F, Shen W B, Li C S, Kasimbeg P, Tchapmi M E, Toshev A, et al. Interactive Gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 2020, **5**(2): 713–720
- 97 Anderson P, Chang A, Chaplot D S, Dosovitskiy A, Gupta S, Koltun V, et al. On evaluation of embodied navigation agents. arXiv: 1807.06757, 2018.
- 98 Paul S, Roy-Chowdhury A K, Cherian A. AVLEN: Audio-visual-language embodied navigation in 3D environments. arXiv: 2210.07940, 2022.
- 99 Tan S N, Xiang W L, Liu H P, Guo D, Sun F C. Multi-agent embodied question answering in interactive environments. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 663–678
- 100 Majumdar A, Aggarwal G, Devnani B, Hoffman J, Batra D. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. arXiv: 2206.12403, 2023.
- 101 Zhou G Z, Hong Y C, Wu Q. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2024. 7641–7649
- 102 Shah D, Eysenbach B, Kahn G, Rhinehart N, Levine S. ViNG: Learning open-world navigation with visual goals. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE, 2021. 13215–13222
- 103 Wen G H, Zheng W X, Wan Y. Distributed robust optimiza-

tion for networked agent systems with unknown nonlinearities. *IEEE Transactions on Automatic Control*, 2023, **68**(9): 5230–5244



孙长银 安徽大学人工智能学院教授。1996 年获得四川大学应用数学专业学士学位, 分别于 2001 年、2004 年获得东南大学电子工程专业硕士和博士学位。主要研究方向为智能控制, 飞行器控制, 模式识别和优化理论。本文通信作者。

E-mail: cysun@seu.edu.cn

(SUN Chang-Yin Professor at the School of Artificial Intelligence, Anhui University. He received his bachelor degree in applied mathematics from Sichuan University in 1996, and his master and Ph.D. degrees in electrical engineering from Southeast University in 2001 and 2004, respectively. His research interest covers intelligent control, flight control, pattern recognition, and optimal theory. Corresponding author of this paper.)



袁心 东南大学自动化学院博士后。2021 年获得东南大学控制科学与工程博士学位。主要研究方向为深度强化学习和最优控制。

E-mail: xinyuan@seu.edu.cn

(YUAN Xin Postdoctor at the School of Automation, Southeast University. He received his Ph.D. degree in control sci-

ence and engineering from Southeast University in 2021. His research interest covers deep reinforcement learning and optimal control.)



王远大 东南大学自动化学院博士后。2020 年获得东南大学控制科学与工程博士学位。主要研究方向为深度强化学习和机器人系统控制。

E-mail: wangyd@seu.edu.cn

(WANG Yuan-Da Postdoctor at the School of Automation, Southeast University. He received his Ph.D. degree in control science and engineering from Southeast University in 2020. His research interest covers deep reinforcement learning and robotic system control.)



柳文章 安徽大学人工智能学院讲师。2022 年获得东南大学控制科学与工程博士学位。主要研究方向为深度强化学习, 多智能体强化学习, 迁移强化学习, 机器人。

E-mail: wzliu@ahu.edu.cn

(LIU Wen-Zhang Lecturer at the School of Artificial Intelligence, Anhui University. He received his Ph.D. degree in control science and engineering from Southeast University in 2022. His research interest covers deep reinforcement learning, multi-agent reinforcement learning, transfer reinforcement learning, and robotics.)