doi:10.6041/j.issn.1000-1298.2025.09.018

具身智能农业机器人关键技术与发展趋势

苗中华[1]　朱子煜[1]　张　伟[1]　薛振锋[1]　孙　腾[1]　张异凡[1]　谢　涛[1]
何创新[1]　李　楠[1]　苑　进[2]　赵春江[3]　刘成良[4]
(1.上海大学机电工程与自动化学院,上海 200444;2.山东农业大学机电工程学院,泰安 271018;
3.北京市农林科学院,北京 100097;4.上海交通大学机械与动力工程学院,上海 200072)

摘要:全球农业智能化转型背景下,传统农业机器人面临非结构化环境、作业对象不确定性及控制参数时变等技术瓶颈,导致农业机器人实用效果差,产业化落地困难。具身智能作为实现通用人工智能的关键路径,为突破上述瓶颈提供了新范式。本文系统阐述具身智能理论驱动的农业机器人技术体系与发展路径,创新性构建“感知-决策-模拟-进化-诊断”五位一体研究框架。具体而言:具身感知聚焦开放环境场景理解、作业目标主动感知及多模态融合感知;具身决策涵盖端到端导航、大模型驱动的机械臂操作及多机协同算法;具身模拟着力于高保真场景重建、生成及虚实迁移技术;具身进化整合无监督学习、强化学习及模仿学习机制;具身诊断构建作业状态监测与质量评估体系。最后,展望未来发展方向,为农业 4.0 时代的农业机器人智能化升级提供理论支撑与技术路线。

关键词:农业机器人;具身智能;基础模型;大语言模型;多模态感知;强化学习
中图分类号:S2　　**文献标识码**:A　　**文章编号**:1000-1298(2025)09-0212-28　　**OSID**:

Key Technologies and Development Trends of Embodied Intelligence Agricultural Robots

MIAO Zhonghua[1]　ZHU Ziyu[1]　ZHANG Wei[1]　XUE Zhenfeng[1]　SUN Teng[1]　ZHANG Yifan[1]　XIE Tao[1]
HE Chuangxin[1]　LI Nan[1]　YUAN Jin[2]　ZHAO Chunjiang[3]　LIU Chengliang[4]
(1. *School of Mechatronic and Automation*, *Shanghai University*, *Shanghai* 200444, *China*
2. *College of Mechatronic Engineering*, *Shandong Agricultural University*, *Tai'an* 271018, *China*
3. *Beijing Academy of Agriculture and Forestry Sciences*, *Beijing* 100097, *China*
4. *School of Mechanical Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200072, *China*)

Abstract: In the context of the global transformation towards agricultural intelligence, traditional agricultural robots face technological bottlenecks such as unstructured environments, uncertainty of operational objects, and time-varying control parameters, resulting in poor practical performance and difficulties in industrial implementation. Embodied intelligence, as a key pathway to achieving general artificial intelligence, offers a paradigm to overcome these challenges. The agricultural robot technology system and development path driven by embodied intelligence theory were systematically discussed, and a “perception - decision - simulation - evolution - diagnosis” integrated research framework was innovatively constructed. Specifically, embodied perception focused on large model-driven visual foundation models, multimodal fusion perception, and open environment scene understanding; embodied decision-making encompasses end-to-end navigation, large model-driven robotic arm operations, and multi-robot collaboration algorithms; embodied simulation emphasized high-fidelity scene reconstruction, generation, and virtual-to-real transfer technologies; embodied evolution integrated continuous learning, unsupervised learning, and reinforcement learning mechanisms; and embodied diagnosis built a work status monitoring and quality evaluation system. Finally, the future development directions were

收稿日期:2025-07-12　修回日期:2025-07-20
基金项目:国家自然科学基金项目(52375107)和上海市科委创新行动计划项目(24N32800100)
作者简介:苗中华(1977—),男,教授,博士生导师,主要从事 AI+农业机器人研究,E-mail:zhhmiao@shu.edu.cn
通信作者:赵春江(1964—),男,教授,博士生导师,主要从事智慧农业技术装备研究,E-mail:zhaocj@nercita.org.cn
刘成良(1964—),男,教授,博士生导师,主要从事农业机器人研究,E-mail:chlliu@sjtu.edu.cn

explored, providing theoretical support and technical routes for the intelligent upgrade in the era of agriculture 4.0.

Key words: agricultural robot; embodied intelligence; foundation model; large language model; multi modal perception; reinforcement learning

0　引言

农业机器人是实现农业现代化、乡村振兴和农业强国战略的关键生产工具，也是有效应对农业劳动力短缺挑战的重要途径，其智能化水平已成为衡量国家农业科技发展水平的核心标志。然而，历经数十年的发展，尽管涌现出多种农业机器人原型样机，但其实际应用效果仍普遍欠佳，产业化进程面临显著瓶颈。究其根源，一方面由于农业生产环境复杂，例如农业作业环境的非结构化、作业对象的不确定性以及控制模型的时变性等；另一方面则源于农业机器人本体作业技能的不足与技术泛化能力的薄弱。因此，亟需引入具身智能理念，探索人工智能驱动的农业机器人技术研究新范式，以推动农业机器人技术创新和产业化变革。

机器人技术研究历经半个世纪的发展，始终在“专用性”与“通用性”的悖论中艰难求索[1-2]。传统技术范式建立在功能模块解耦的基础之上：感知系统负责环境理解，决策系统生成行为策略，执行系统完成物理操作[3-4]。这种分层架构在结构化场景中展现出工程实用性，却难以应对开放环境中的不确定性挑战[5]。究其本质，功能解耦设计割裂了智能体与环境的交互闭环，导致系统陷入“见木不见林”的认知局限 ——视觉传感器聚焦物体几何特征却忽视力学约束，运动控制器追求轨迹跟踪精度而漠视任务意图，决策系统规划理想路径但无视本体能力边界[1]。

当前技术研究范式的不足在具身智能领域尤为凸显。波士顿动力 Atlas 机器人虽能完成高难度体操动作，但其预设动作库难以适应非标准场景；特斯拉 Optimus 在工厂演示中展现灵巧操作能力，却仍需人工定义每个抓取位姿。这些案例折射出深层矛盾：基于规则与模板的控制系统，无法实现人类般随境而变的适应能力[6]。破局之道在于重构技术范式，从“感知-决策-执行”的线性流程转向“多模态认知-动态决策-本体适配”的闭环体系[7]。这种转变不仅需要算法创新，更要求对智能系统架构进行根本性变革。

传统机器人研究的“大脑-身体”割裂，本质上是控制理论与人工智能学科分野的产物[8-9]。最优控制理论追求给定约束下的轨迹精度，强化学习专注任务目标的策略优化，二者在技术实现层面长期处于平行发展状态。这种割裂导致智能系统陷入“理想规划”与“现实执行”的永恒矛盾：规划系统基于理想化本体模型生成天马行空的行动方案，执行系统在物理定律约束下艰难挣扎。

作为人工智能向物理世界延伸的关键桥梁，具身智能正成为重塑智能系统设计理念的核心驱动力，通过将感知、认知、决策和执行统一于智能体与环境的交互过程中，实现“身体即大脑、行动即思考”的知行合一技术研究新范式。具身智能强调智能体通过物理身体与环境的持续交互来构建对世界的理解，将抽象符号推理与具体物理操作有机融合，形成“感知-决策-行动”的统一闭环。

近期，已有一些具身智能研究相关的优秀论著[10-16]，但均未涉及到农业机器人领域。然而，农业场景的开放性、复杂性以及作业对象的形态位姿等不确定性因素，迫切需要赋予农业机器人多模态感知认知、动态推理决策、本体精准执行以及技能持续进化的具身智能能力。因此，本文旨在探讨具身智能在农业机器人中的研究范式、关键技术及发展趋势，填补具身智能与农业机器人之间的研究空白，以推动农业机器人技术创新和产业化变革。

1　农业机器人挑战与具身智能机遇

1.1　全球农业智能化转型的迫切需求

作为高度自主化的作业装备，农业机器人能够替代人工自主执行高重复性劳动任务，并实现农业生产精准化作业。基于对全球农业机器人技术现状与应用水平的分析，尽管近些年农业机器人研究已取得显著进展，但当前技术整体仍停留于基础自动化阶段，存在明显局限：部分作业环节仍需依赖人工干预；场景适应性不足、功能同质化问题突出；实用效果不理想，产业化落地困难。现有系统多面向特定、固定的作业类型（如大田喷洒、温室巡检等）或单一工况设计[17]，普遍缺乏应对农业环境复杂多变特性的灵活性及跨场景作业的迁移能力。就市售产品而言，同质化现象突出，核心功能高度趋同，功能布局多集中于简单的基础性作业（如喷洒、巡检、监测等），难以有效响应农业生产场景日益增长的多元化与精细化技能性操作需求。

针对当前农业机器人存在的技能泛化能力不

足、场景适应性差等关键问题,实现从自动化向智能化的范式转变已成为农业机器人发展的必然选择。《全国智慧农业行动计划(2024—2028 年)》[18]明确将推进农业智能化转型列为智慧农业建设的核心目标。这一转型将促使农业机器人突破传统自动化作业的局限,构建“感知-决策-执行”的智能闭环系统,使其具备基于环境交互的自主学习和持续进化能力[19]。在复杂多变的非结构化农业场景中,智能化的农业机器人能够通过持续积累多模态数据和环境交互经验,实现算法模型的自主迭代优化,从而不断提升作业精度和作业效率;同时,这种基于实践的学习机制使其能够有效增强对未知场景的适应能力,并将学习到的知识迁移应用于相似任务,最终实现跨作物、跨场景、跨任务的泛化性能显著提升。

1.2 农业机器人产业落地的技术瓶颈

农业生产迫切需要用自动化乃至智能化的装备来代替人工。然而,我国果蔬生产机器人的研发进度相对缓慢,绝大部分仅仅停留在样机阶段,无法实现复杂农业场景下的连续可靠生产作业。其根本原因在于:①农业生产场景非结构化:农业环境通常是非结构化和动态变化的,充满了各种复杂和不确定的因素。其生产环境中光照、温湿度等环境参数的动态性与不确定性并存,以及空间布局的结构化与非结构化交织。②作业对象的不确定性:农业机器人面临的作业对象通常具有不确定性和多样性。例如,在任务级别果蔬形状、姿态、大小、成熟度等高度不均,经常伴随藤蔓遮挡、果实娇嫩等挑战,都可能影响农业机器人的工作表现。③生产设备通用性有限:设施农业生产可能涉及植保、打叶、收获和巡检等多环节繁杂任务的多流程长周期系统性工程,现有大多数农业机器人主要集中在单一功能的开发上,不具备多样任务的执行能力,通用性差造成了资源浪费和成本剧增。④智能泛化性程度低:现有自动化的农业机器人的技能固化,无法实现经验积累和技能增长,人机交互和机器-环境交互范式仍处于传统定式化模式,即工程师将专业的知识提炼并“翻译”给机器人,机器人只依赖于预设的规则和模型机械地执行人类设定好的程序,并不具备环境认知、任务学习和技能迁移能力,一旦系统化部署便难以通过与环境及任务交互持续提升自身能力,导致多任务执行泛化性能差,智能化程度不足。上述瓶颈严重影响了农业机器人的产业落地。

1.3 具身智能对农业机器人的变革性价值

作为人工智能的重要分支,具身智能(Embodied intelligence)强调通过与物理环境的实时交互实现智能化作业。在农业机器人领域,该技术展现出解决现有技术瓶颈的潜力,其优势主要体现于 3 方面:显著提升机器人在非结构化、动态农业环境中的适应性,使机器人能够根据实时交互灵活调整行为。通过紧密结合感知与动作闭环,增强机器人在复杂、不确定环境中的感知与决策能力,促使其做出更适切的响应。有助于提高机器人与传统农艺的融合性,通过对农业作业流程进行智能化建模,机器人得以理解并模拟传统农艺的操作方式,同时提供更高效的自动化解决方案。综上所述,具身智能的引入为农业机器人带来了更强的环境适应性、更高的作业精度以及更优的农艺兼容性。随着技术的持续演进,具身智能有望成为驱动未来农业机器人发展的核心力量。

2 具身智能概念与核心内涵

2.1 具身智能基本概念与理论基础

作为实现通用人工智能(Artificial general intelligence, AGI)的重要路径,具身智能(Embodied AI)已成为近年来学术研究和产业发展的热点领域[16]。实际上具身智能并非全新的概念,1950 年,英国计算机科学家艾伦·图灵(Alan Turing)首次提出具身智能的概念[20],探索如何使机器感知和理解世界,并作出相应的决策和行动。然而,受限于当时的技术,低精度传感器、有限计算能力及欠缺的算法理论无法支持这一构想的发展。

20 世纪 80 年代中期,随着符号主义局限性的日益显现,以 MIT 人工智能实验室的罗德尼·布鲁克斯(Rodney Brooks)为代表的研究者逐渐认识到,智能行为可以直接从自主机器与其环境的简单物理交互中产生,而这种交互不依赖于预先设定的复杂算法[21]。他主张智能不应仅仅被视为对抽象知识的被动处理过程,而应当通过智能体与真实物理环境的持续、动态交互过程中自然涌现[22],这一观点为具身智能奠定了理论基础。

2.2 具身智能发展历程与技术演进

具身智能的探索与发展已走过了 70 多年的历程,然而在过去的几十年里,由于人工智能技术的限制,具身智能一直处于概念化阶段。近年来在高性能算力和大规模标注数据的支持下,深度学习、强化学习、无监督学习等人工智能技术在图像识别[23]、智能决策[24]、自然语言处理[25]领域取得了突破性进展。然而这些“离身”的智能体被限制在赛博空间,不拥有物理身体,缺乏与真实世界的交互,无法驱动机器人实体完成特定任务[22]。

近年来,以 ChatGPT 为代表的大语言模型(Large language model, LLM)[26]通过对大规模互联

网语料数据的预训练，展现了卓越的泛化与常识理解能力，引领了大模型在计算机视觉、自然语言处理和多模态领域的创新浪潮。CLIP[27]、DINOv2[28]、SAM[29]、SAM2[30]等视觉基础模型在图像分割、目标检测、特征提取等任务中展现出强大的零样本和少样本推理能力。LLaMA 2[31]、Gemini 1.5[32]、DeepSeek－V3[33]等大语言模型在语言理解、逻辑推理、代码编写、知识问答等任务上持续刷新性能基准，结合思维链技术可实现复杂问题的推理。BLIP[34]、BLIPv2[35]、GPT4－V[36]等多模态模型通过预训练视觉编码器和跨模态对齐技术，将视觉输入映射到语言模型的嵌入空间，使大语言模型能够处理多模态数据，实现了更深层次的语义理解。

上述进展不仅展示了大模型的潜力，也为其与具身智能的实现提供了新的技术路线。谷歌团队在具身智能领域的系列突破性工作展现了大模型与机器人融合的巨大潜力。SayCan 模型首次将大语言模型与机器人技能紧密结合[37]，利用大语言模型解析复杂的自然语言指令，生成可执行的操作计划，实现了从抽象语言指令到具体物理操作的无缝转换。RT－1 模型基于 Transformer 构建了端到端机器人模型[38]，通过大规模数据预训练，实现了根据用户的语言指令和图像输入控制机器人完成任务，验证了大模型在机器人上的可行性。RT－2 模型[39]进一步将预训练视觉－语言模型（Vision language model，VLM）应用于机器人控制系统，借助视觉-语言模型 PaLM－E[40]在海量跨模态数据上预训练获得的强大表征能力，展现出显著的零样本迁移性能，验证了视觉-语言模型在具身智能领域的应用潜力。清华大学团队提出的双臂操作模型 RDT－1B 在双臂任务上取得了突破[41]，基于扩散模型和可扩展的 Transformer 架构，实现了双臂机器人复杂任务执行、零样本泛化和少样本学习。Physical Intelligence 公司研发的模型在预训练的视觉-语言模型的基础上加入了独立的动作专家模块，通过流匹配生成连续动作，提升了模型的多任务泛化和跨本体适应能力[42]。智元机器人提出的通用具身基座大模型 GO－1，采用 ViLLA 架构，结合多模态大模型与混合专家系统，实现了小样本快速泛化、跨本体迁移和持续进化，显著提升机器人在复杂任务中的执行效率。

研究者通过结合大模型的知识与机器人的物理交互能力，在具体场景中通过“本体-模型-环境”的三元交互，与模仿学习、强化学习结合，提升策略的泛化能力和对环境的适应能力，使机器人从单纯的指令执行者逐步迈向具备自主性和适应性的具身智能体。图 1 列举了近年来具身智能领域的代表性成果，这些里程碑式成果共同构建了一个从感知理解、语言交互到动作执行的完整技术栈。

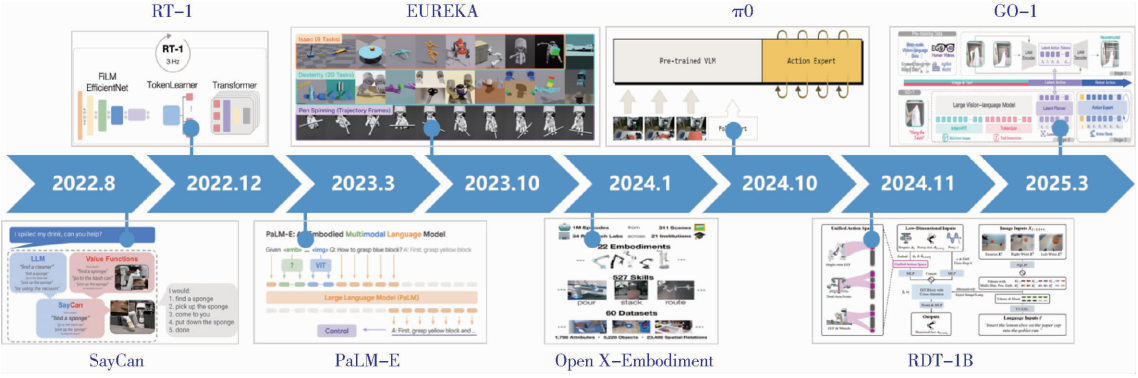


图 1　近年来具身智能领域的代表性成果

Fig. 1　Representative achievements in field of embodied intelligence in recent years

2.3　具身智能与离身智能的本质差异

具身智能与离身智能代表了人工智能发展的两条截然不同的技术路径。离身智能依托大规模数据训练获得智能能力，其活动范围局限于赛博空间；而具身智能则强调智能体通过与在物理环境的交互中获得智能能力。这种技术范式决定了两者在学习范式和认知深度上的本质区别。

在学习范式上，离身智能体通过大量数据的预训练和针对下游任务的微调，虽然在语言理解、图像识别等任务上表现卓越，但这种离线学习范式缺乏与环境的实时交互反馈，难以处理复杂多变的现实场景需求。而具身智能体除了理解语言和图像信息外，还需要将环境感知、任务理解转化为具体的物理动作输出，实现从感知到行动的完整闭环。通过在线交互学习，具身智能体能够在实际操作中不断积累技能、实时修正行为策略并适应环境变化。

在认知深度上，离身智能虽然能够在语义层面获得丰富的知识表征，但缺乏对真实世界物理规律的直观体验和深层理解。相比之下，具身智能将感知、认知、行动统一在闭环系统中，通过身体与环境的直接物理接触，能够深入理解真实世界的物理定

律,这种基于身体经验的学习使具身智能实现对物理世界运行机制的本质性理解。

2.4 具身智能的核心技术要素

具身智能的核心技术要素体现在集成化的具身本体、智能化的具身模型、高质量的具身数据和自适应的具身学习 4 个关键维度。

集成化的具身本体:本体是与物理世界交互的基础;通过模块化设计实现功能组件的灵活配置,支持多样化的任务需求;集成视觉、触觉等多模态传感器,实现对环境的全方位感知;高精度执行器与智能控制算法协同,确保动作的准确性和流畅性。本体的机械结构与机电性能直接影响了具身系统的运动灵活性、环境适应性及任务执行效率。

智能化的具身模型:具身模型是具身系统的“大脑”,将感知、认知、决策和控制深度融合,整合视觉、语言、触觉等多模态信息,实现任务理解并生成决策指令,控制具身本体完成“感知-行动”闭环。

高质量的具身数据:作为模型学习进化的原材料,训练数据有多种来源:基于大规模物理仿真平台的高效合成数据生成,解决数据规模瓶颈;结合人类专家遥操作提供的精准行为示范,建立可靠技能先验;真实环境中的自主探索数据,补偿仿真与现实间的物理特性差异。

自适应的具身学习:具身智能体依托多元化学习方法,实现对环境的自适应泛化和本体技能的持续进化。通过强化学习、模仿学习、自监督学习等学习机制的有机融合,赋予具身智能系统持续学习和自我进化的核心能力。

这 4 个核心要素构成了一个有机统一的技术体系,它们相互依存、协同工作,共同支撑着具身智能系统的整体功能。集成化的具身本体为系统提供可靠的硬件基础,智能化的具身模型构建认知决策核心,高质量的具身数据驱动模型训练优化,自适应的具身学习实现智能体持续升级。通过这种深度集成的技术架构,具身智能系统能够在复杂真实环境中完成从“感知-行动”闭环,为实现真正的通用智能机器人奠定了基础。

3 具身智能农业机器人研究范式

具身智能农业机器人研究范式构建了一个系统性的技术体系。如图 2 所示,该体系以基础模型和学习机制为核心驱动引擎,构建了多层次、多模态的智能农业解决方案。整个系统围绕“感知认知-规划决策-模拟生成-学习进化-诊断运维”5 个核心环节,形成了完整的技术闭环。感知认知环节负责处理多传感器数据输入,利用视觉基础模型和视觉-语言模型实现开放农业场景的语义理解和作业对象的主动感知识别。规划决策环节负责分析语义结构化数据和任务需求,利用大语言模型的推理能力将复杂任务分解为多个可执行的子任务,并基于端到端架构和视觉-语言-动作模型实现农业机器人的精准导航与动作执行。模拟生成环节通过高保真场景渲染技术,构建可复用的虚拟农业环境,并利用扩散模型对场景进行动态建模和数据增强。学习进化环节基于虚拟农业仿真环境的机器人技能进化框架,通过大规模并行仿真训练构建农业操作技能库,结合跨本体迁移学习方法,经过真实农业场景下的闭环验证实现农业机器人技能的学习进化。诊断运维环节深度分析作业过程中产生的多维数据,实现机器人系统的实时状态诊断、故障预警和作业质量评估。

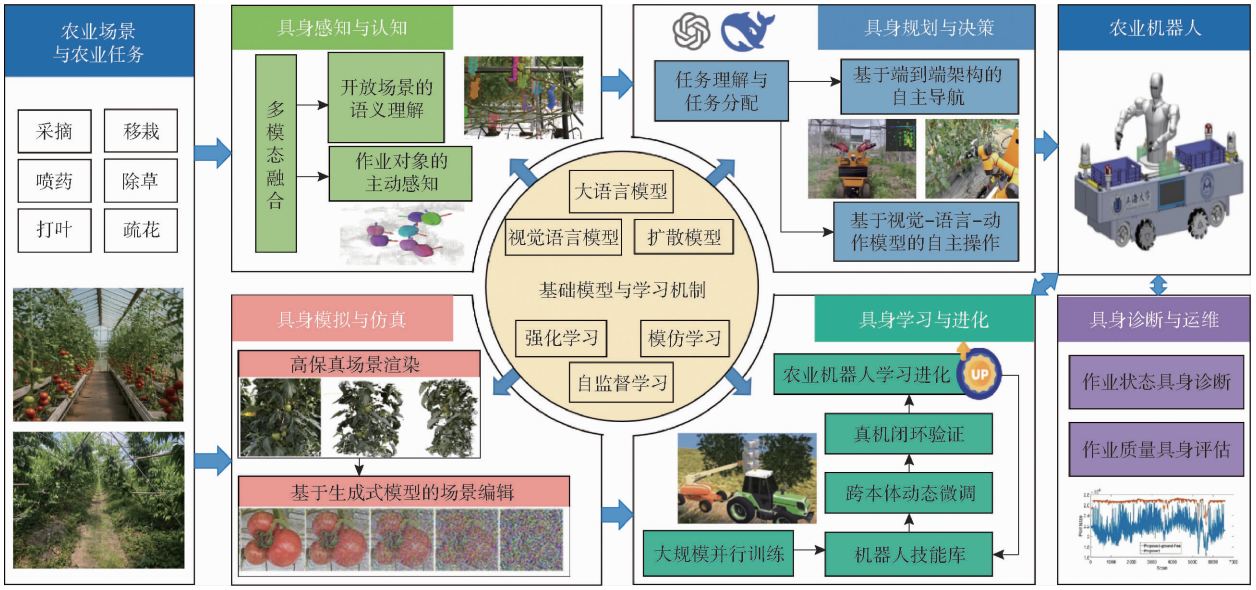


图 2 具身智能农业机器人研究范式

Fig. 2 Research paradigm for embodied intelligent agricultural robots

4　具身感知与认知关键技术

农业环境具有高度的动态性和不确定性，包括多样化的作物品种、复杂的植物生长状态、不规则的地形地貌以及多变的光照条件等，这些特征对农业机器人的感知能力提出了极高的要求。传统的静态感知方法难以应对如此复杂多变的环境，而具身智能的引入为解决这一难题提供了新的思路。本节将从感知层次的不同维度展开深入分析，分别从对开放环境的场景理解、对作业目标的主动识别、对多源异构信息的有效整合 3 方面，阐述具身智能如何赋能农业机器人构建更加鲁棒、精准和自适应的感知认知能力。

4.1　开放环境场景理解

农业机器人作为精准农业和智能农机的核心载体，正经历从结构化环境向开放环境的范式转变。传统农业机器人多依赖于预设规则和封闭类别的视觉算法，其应用场景局限于温室、标准化果园等受控环境。然而，真实的农田环境具有动态性（昼夜光照变化、作物生长周期）、强干扰性（杂草、枝叶遮挡）和长周期性（跨季节表型演变）等特征，这对机器人在开放环境下的场景理解能力提出了严峻挑战。

传统深度学习模型在开放农业环境的场景理解中存在显著局限性。首先，传统模型严重依赖大量人工标注的训练数据，在面对开放环境的多样性和复杂性时，获取全面的标注数据往往成本高昂且不切实际。其次，每个模型通常只能理解特定类型的场景，具有强烈的任务特异性，难以应对开放环境中的多元化场景需求。此外，传统模型的泛化能力有限，只能识别训练过程中见过的有限场景类别，无法适应开放环境中不断出现的新场景和未知目标。最后，不同场景理解任务的模型之间无法有效共享知识，形成了知识孤岛，制约了机器人在农业环境下跨场景、跨类别认知能力的构建。

视觉基础模型（Visual foundation model，VFM）为开放农业环境的场景理解提供了新思路。这类模型通过在大规模多样化数据上的自监督预训练，学习到了具有强泛化性的通用视觉表征，具备优异的零样本推理能力[43]。在农业应用场景中，模型能够有效识别和理解训练阶段未曾接触的新场景与新目标类别，实现对开放环境下陌生场景和物体的准确理解。这一特性不仅显著减少了人工标注的成本负担，更使模型能够基于全局上下文信息，动态调整其理解策略，展现出良好的环境适应性。

Segment anything model（SAM）[29] 作为视觉基础模型的重要突破，通过 1 100 万图像和 11 亿掩码的大规模预训练，实现了通用的零样本分割能力。不同于传统视觉 Transformer 模型，SAM 创新性地引入了提示驱动机制，能够根据点、框或掩码等空间提示自适应地完成各类分割任务，这种特性使农业机器人可动态调整分割策略以适应不同的作业需求。其 3 层架构设计实现了特征提取与任务解码的有效解耦，图像编码器输出的通用视觉特征可同时农业场景认知、作业目标定位等多种农业感知任务。SAM 展现出卓越的零样本跨域迁移能力，这得益于其在大规模多样化数据上预训练构建的通用视觉表征空间，使农业机器人无需针对特定场景进行额外训练即可实现准确分割，显著降低了在新环境中的部署门槛。

4.2　作业目标主动感知

在作物生长环境中，叶片、枝干和果实之间普遍存在交叉遮挡现象，如图 3 所示。这些遮挡物不仅影响视觉感知的准确性，还构成了机器人到达目标路径上的物理障碍，直接干扰其作业路径的合理规划。因此，准确检测当前环境的遮挡状态，并制定有效的遮挡规避与障碍物绕行策略，是农业机器人需要解决的关键问题之一。



图 3　农业作物受不同程度遮挡示例

Fig. 3　Examples of agricultural crops with varying degrees of occlusion

在现代农业实践中，尽管可以通过整枝打叶等田间管理措施显著缓解目标遮挡问题，但此类方法往往依赖大量人力物力，难以实现自动化作业流程。同时，在当前多数农业机器人采用固定视角相机的前提下，目标识别过程中遮挡现象仍不可避免。遮挡不仅会降低目标检测的置信度，严重时甚至导致目标无法被识别。现有研究多采用两类方法缓解此问题：通过引入包含遮挡状态的数据扩充训练集以增强模型鲁棒性[44]；利用先验知识对被遮挡区域进行重建与补全[45]。然而，对于如何通过主动策略规避遮挡干扰，相关研究仍相对有限。在机器人感知

领域,主动感知被认为是解决此类问题的潜在方向。通过不断探索、获取新视角信息,机器人可逐步感知受遮挡或不确定的目标[46]。在采摘机器人目标识别研究中,视觉伺服控制与多视角重复采样等主动感知策略被广泛采用。这类方法受到人类主动感知机制的启发,通过动态调整摄像机姿态、反复观测目标区域,实现信息获取的最大化,从而显著提升目标检测的置信度与准确率。

(1)遮挡状态检测与分类

在进行主动感知减轻遮挡问题之前,检测当前目标所受遮挡状态(程度和种类等)至关重要,只有这样才可以做出对应的最佳动作。为此,GAO 等[47]根据遮挡条件将苹果分为 4 类,即无遮挡、叶片遮挡、树枝遮挡和果实遮挡,便于后续采摘。WU 等[48]使用 YOLO v4 识别模型,用背景透明的真实叶片图像随机覆盖在苹果图像上,构成叶片遮挡果实,然后利用插图数据扩充方法对原始数据集进行批量扩充。ZHENG 等[49]基于 YOLO v8n - CK 网络,提出了不同遮挡度下白菜头关键点检测与直径估计方法。ÍÑIGUEZ 等[50]提出不同遮挡条件下无创葡萄束检测的深度学习建模方法。这些研究成果的不断涌现,充分体现了农业环境遮挡问题解决方案的迫切需求以及该领域技术攻关的挑战性。

(2)基于视觉伺服的主动感知策略

在成功检测到遮挡状态后,如何通过主动调整视角来获取更佳的观测条件成为关键问题。视觉伺服是指使用计算机视觉数据来控制机器人的运动。LEHNERT 等[51]开发了 3D Move to See(3DMTS),能够使用自制的相机矩阵同时获取场景的多个视角的图像,并评估当前视角下相机向何方向移动可以更快增加目标在图像中的面积。在 3DMTS 的基础上,文献[52 - 53]使用中央相机图像作为输入,将相机矩阵获得的目标区域梯度用作标签,建立数据集,并训练一种新的 CNN 模型 Deep - 3DMTS,它用单幅图像就可以指导机械臂的移动。值得注意的是,上述研究主要是在仿真环境下进行,而且构建摄像机矩阵成本高,要求位置精度极高,因此在作物生长环境下实现还存在一定困难。

(3)基于多视角采样的主动感知策略

针对非严重遮挡,SUN 等[54]提出了基于启发式的主动感知方法,首先检测到遮挡物相对于果实本体的位置,然后利用在机械臂上的相机向遮挡物相反的方向移动,避开遮挡。此方法在一定程度内减轻了遮挡,提高了识别成功率,但是由于该方法缺乏边界条件,容易造成错过最佳位置,而且多步重采样也带来了效率降低的问题。针对启发式方法的缺陷,SUN 等[55]提出了深度主动感知方法,通过构建深度学习网络,学习当前目标遮挡情况与绕开遮挡最佳观测位置的相对关系,最终实现在检测出当前遮挡状态后,快速预测最佳观测位置,实现了相对高效的主动感知(图 4)。与此类似,LUO 等[56]基于深度确定性策略梯度算法(Deep deterministic policy gradient, DDPG),通过机器人自主运动寻找最佳观测位置,初步验证了深度强化学习和主动感知结合的有效性(图 5)。虽然其工作只适用于串采果实,而且无法针对遮挡较为严重的情况,但为遮挡处理和主动感知提供了新的理论思路和方法基础。因此,为了进一步提高机器人主动感知的有效性和实

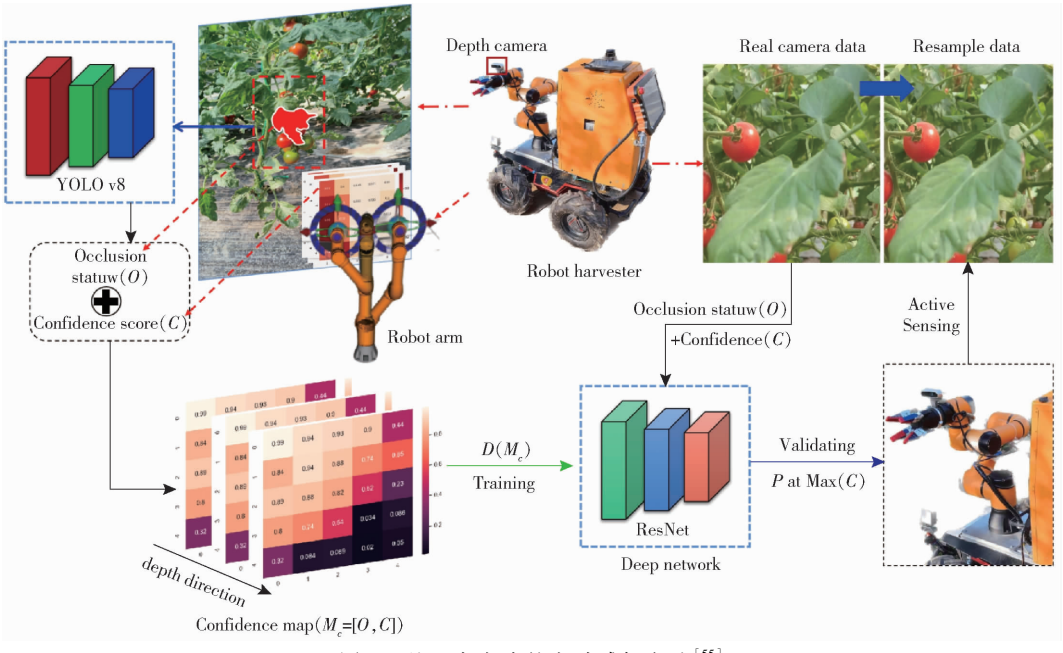


图 4 基于启发式的主动感知方法[55]

Fig. 4 Heuristic-based active perception method

用性，可以考虑将深度强化学习与主动感知模式相结合，让机器人通过自主环境交互与探索，提炼出针对遮挡、障碍物存在情况下的主动感知运动策略，提高其感知精度和最终的作业成功率。

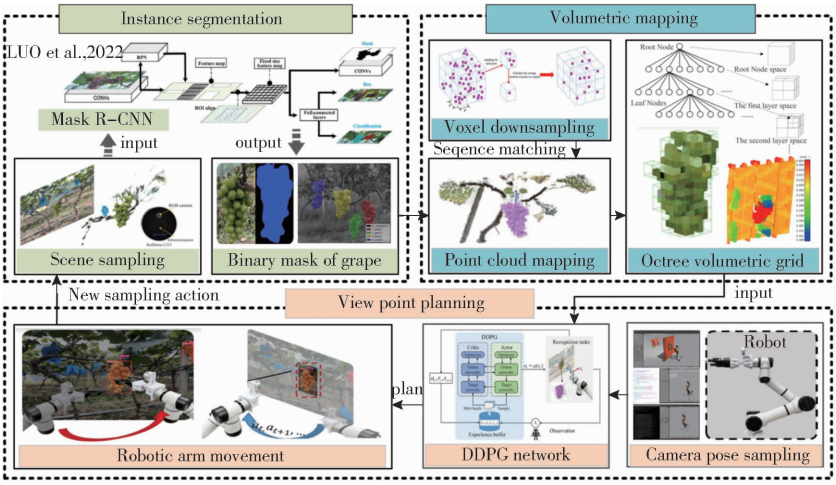


图 5　动态调整相机视角实现主动感知[56]

Fig. 5　Active perception through dynamic camera viewpoint adjustment

4.3　多模态融合感知

多模态融合感知是实现农业场景理解与精准作业的关键技术。农业环境的复杂性和多样性决定了单一感知模态的局限性，不同农业场景（大田、温室、果园）具有差异化的感知需求，动态变化的光照条件、植株遮挡、环境干扰等因素会影响传感器性能，而多样化的作业任务更需要不同感知模态的协同配合。RGB－D 相机可同时获取色彩信息与深度数据，是实现目标识别与避障的基础传感器[57]；高光谱成像设备通过窄波段光谱分析能力，能够检测作物早期病虫害特征[58]；热成像相机则通过监测冠层温度分布来评估植株水分胁迫状态。激光雷达通过三维点云重建技术为果园树冠建模提供空间信息；超声波传感器以其成本优势实现近距离避障功能；力/触觉传感器可精确量化机械手的操作力度，避免采摘过程中对脆弱农产品的损伤；气体传感器则通过检测土壤中的特定气体成分来优化施肥决策。这些异构传感器的协同工作构成了具身智能农业机器人多模态感知的硬件基础。

在多模态信息融合过程中，不同传感器和数据源之间存在显著的异构性差异，这给数据的有效整合带来了重大挑战。如何克服这些异构特性，实现多源信息的协同融合，是当前多模态感知技术面临的核心问题。如图 6 所示，根据信息融合层次的不同，现有的多模态信息融合方法主要分为数据级融合、特征级融合和决策级融合 3 种技术方案[59]。

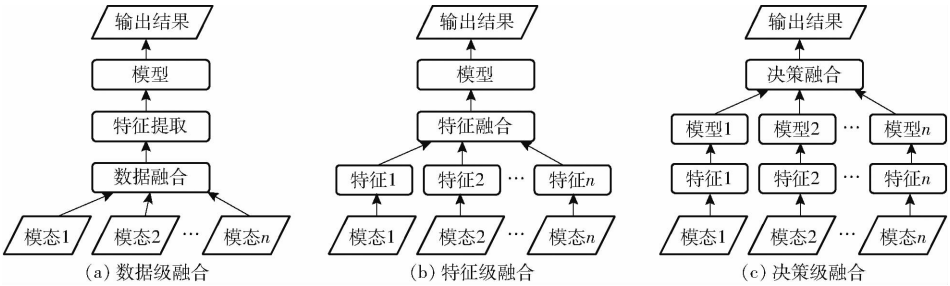


图 6　3 种多模态信息融合架构

Fig. 6　Three multimodal information fusion architectures

（1）数据级融合

数据级融合是最底层的融合方式，可以最大程度地保留每种模态的原始信息，提供其他融合方法无法获得的细节，且实现相对简单。VORA 等[60]提出用图像语义分割的结果给原始点云“染色”的方法，将这种经过图像语义信息增强后的点云输入到常规结构的 3D 目标检测器以达到更高的性能。然而，数据级融合在处理过程中缺乏模态间的深层互动和协调机制，导致无法充分挖掘不同模态之间的互补特性和潜在关联性，因此难以实现最优的信息整合。尽管该方法提供了丰富的信息，但在动态环境中，尤其是在光照、植物遮挡等变化较大的环境中可能具有局限性。

（2）特征级融合

特征级融合是指在多个模态数据的基础上，提取各自的关键特征，再将这些特征进行整合，最终形

成一个统一的表示。以基于 RGB 相机和激光雷达的跨模态 3D 目标检测任务为例,特征级融合方法通常采用两阶段框架,将图像特征和与之对应的点云特征进行融合。在点级融合方面,HUANG 等[61]通过将点云投影到图像平面,检索相应的语义得分或特征。LIANG 等[62]提出了连续融合层,能够为每个鸟瞰图特征图中的网格检索与最近 3D 点对应的图像特征。MMF 进一步通过将深度图生成的伪点与点云进行结合,实现了密集的点对点特征融合。LIU 等[63]提出了 EPNet + +,使用 LI - Fusion 层增强两种模态之间的交互,从而获取更全面的特征。如图 7 所示,WANG 等[64]则通过从 2D 检测模型中提取的中间 CNN 特征对点云进行装饰。为了应对点云数据的稀疏性问题,从而实现高效准确的 3D 目标检测(图 7),QI 等[65]通过结合 RGB 图像中的 2D 边界框和深度信息生成 3D 截锥体,并利用 PointNet + +网络处理截锥体内的点云数据,从而实现高效准确的 3D 目标检测。特征级融合能够帮助农业机器人更精确地进行目标识别,如病虫害的早期检测、作物生长状态评估等。

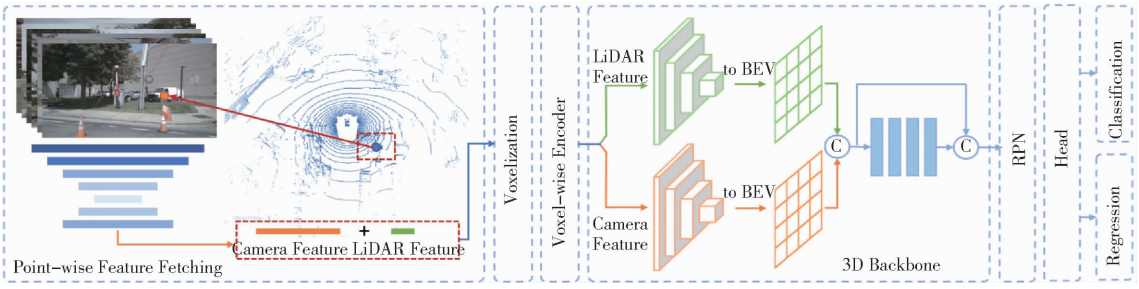


图 7 特征级融合示意图[64]

Fig. 7 Feature level fusion schematic

(3)决策级融合

决策级融合是一种高层次的融合方法,也称之为后融合,其输出是一个统一的决策结果。在这一过程中,首先对每个传感器获取的信息进行处理,包括预处理、特征提取、识别或判断,以形成对目标的初步结论。最终,通过综合各个判定结果及其可信度,依据一定的准则做出最优决策。决策级融合方法从具体任务需求出发,充分利用特征层融合中提取的各类信息,理论上能够比单一传感器方案提供更为精确或明确的结果。该方法具有较高的灵活性,且对信息传输的带宽要求较低,能够有效融合反映目标或环境各方面信息的不同类型数据,且具备处理非同步信息的能力。然而,由于农业背景下环境和目标的动态变化特性,以及先验知识难以获取等因素,决策级融合技术的发展仍面临一定的挑战。

在跨模态 3D 目标检测中,决策级融合通过将不同模态感知结果进行综合,以得出统一的决策输出。例如,PANG 等[66]利用 2D 检测结果与 3D 检测结果的几何和语义一致性来提升单模态检测器的性能。具体来说,点云数据和图像数据分别进行处理,得到两组不同的检测结果,再进行融合,以实现更准确的目标识别。为了提升可见光与红外图像在决策级融合目标检测中的性能,宁大海等[67]提出了一种基于模型可靠性的决策级融合策略。该方法首先通过图像预处理技术改善红外图像的整体质量,然后分别对可见光图像和热红外图像的目标检测模型进行独立训练和测试。

5 具身规划与决策关键技术

农业生产作业具有高度的复杂性和系统性特征,涉及多样化的田间任务、复杂的路径规划、精细化的操作控制以及多机器人的协同配合等多个层面,这些特征对农业机器人的规划决策能力提出了极高的要求。下面将从规划决策的不同技术维度展开深入分析,分别从导航、操作、多机协同 3 个方面,阐述具身智能如何赋能农业机器人构建更加智能、高效和协同的规划决策能力。

5.1 具身导航关键技术

5.1.1 端到端导航

传统的农业机器人导航框架普遍采用模块化架构。该架构将导航任务解耦为感知、规划和控制等多个离散子模块。其中,感知模块负责环境信息的采集与处理;规划模块基于感知结果进行路径规划与任务决策;控制模块则负责将规划指令转化为具体的执行动作。

各子模块依赖特定的传感器和算法独立处理数据,并通过串行流程逐级传递信息。然而,这种串行处理模式存在一个固有缺陷:模块间的误差会沿处理链向下级联传递并累积,导致系统整体性能的下降。这一问题已成为制约农业机器人技术发展的关键瓶颈。

相比之下,端到端导航模型采用了一体化的学习框架,能够同时处理感知、路径规划与运动控制等分离的子任务(图 8)。这种端到端的架构设计展现

出多方面的显著优势：首先，它有效规避了模块化架构中误差累积与传播的问题；其次，模型能够直接从原始输入数据中学习任务相关的、最优化的特征表示，从而避免了人工设计特征的主观性与局限性；此外，统一的网络结构通常能带来更高的计算效率。当前，端到端导航的主流方法根据其学习范式可分为两大类：基于模仿学习的方法和基于强化学习的方法。

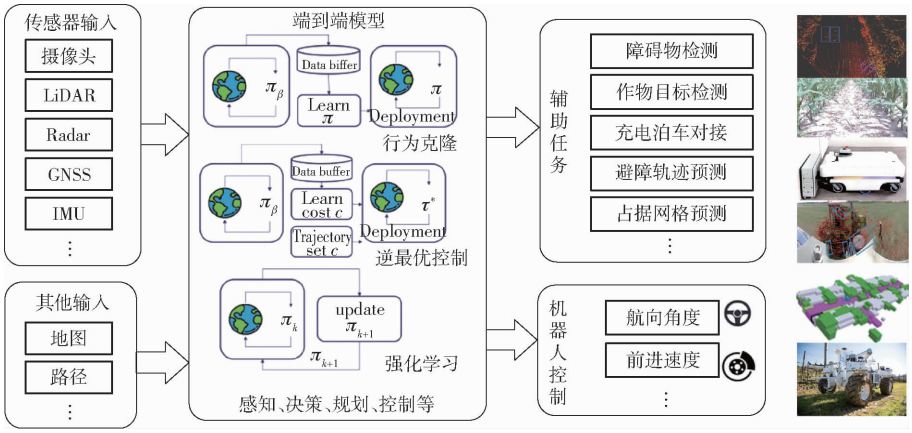


图 8　端到端导航架构图

Fig. 8　End to end navigation architecture diagram

（1）基于强化学习的端到端导航

强化学习（Reinforcement learning，RL）是一种通过智能体与环境的交互来学习最优决策策略的机器学习范式。与依赖标注数据的监督学习不同，强化学习让智能体在环境中通过试错进行自主探索学习。其核心框架是：智能体感知当前状态，据此选择动作；环境响应该动作，返回奖励信号并转移到下一个状态。智能体的终极目标是学习一个策略，以最大化长期获得的累积奖励。正是由于这种与环境的直接、持续交互机制，强化学习在处理分布偏移问题时展现出独特优势。在训练和测试环境存在差异的情况下，强化学习智能体并非仅仅模仿固定的演示数据，而是能够通过持续的探索和学习，主动适应新的环境条件并调整其行为策略，从而更好地应对现实世界中的动态变化。

在精准农业作业导航任务中，智能体需要在复杂农田环境中与土壤、作物、地形等进行交互，收集反馈信息，如作物生长状态、地形起伏等，然后基于这些反馈调整自身行为策略，使特定农业任务中的奖励函数达到最大化，进而实现精准农业作业导航目标。强化学习技术根据学习机制的不同，主要包括基于值函数的方法（深度 Q 网络（Deep Q - Network，DQN））以及基于 Actor - Critic 模型的方法（深度确定性策略梯度（Deep deterministic policy gradient，DDPG）、异步优势 Actor - Critic 模型（Asynchronous advantage actor-critic，A3C））。由于强化学习在农业等复杂环境下的应用需求日益增长，众多研究致力于提升其稳定性和效率。

DQN 通过深度神经网络来近似 Q 值函数，为农业无人系统的自主决策和导航提供了新的解决方案。RONECKER 等[68]首次将 DQN 应用在无人驾驶的决策过程中，结合 DQN 和控制理论来安全地导航自动驾驶车辆的方法。而在复杂农田环境中，农业机器人需要具备自主学习和适应能力，以应对多变的土壤条件、作物布局和障碍物等挑战。深度强化学习作为一种强大的机器学习技术，能够使机器人通过与环境的交互，基于奖励信号自主学习最优策略。PENG 等[69]提出了一种基于深度强化学习算法 Dueling Double DQN 的端到端方法，创新性地将摄像机图像和车辆运动信息共同构成状态空间，使车辆能够自主学习端到端驾驶。

在农业自动化领域，农业机器人面临在复杂多变的农田环境中执行精准作业的挑战。基于 Actor - Critic 模型的深度强化学习方法，为提升农业机器人的自主导航和决策能力提供了重要支持。其中 DDPG 结合策略梯度和深度学习，适用于连续动作空间，CUI 等[70]提出了一种基于 DDPG 模型的无人机自主运动规划端到端方法，构建了一个以 DDPG 网络作为智能体核心，利用卷积神经网络处理深度图像并提取关键特征以支持决策的方法。LI 等[71]提出了一种基于 A3C 模型的面对未知环境端到端自主导航算法，将激光雷达数据映射到控制命令。采用 LM - TD3 网络，结合 TD3 策略梯度网络和 LSTM 神经网络，提升导航经验和障碍物处理能力。通过结合各种强化学习模型，农业机器人能够更有效地应对多变的农田条件，实现精准的路径规划和障碍物避让，从而提高作业效率和安全性，不仅提升了农业机器人的智能化水平，还为农业自动化和智

能化的发展提供了强有力的支持。

(2)基于模仿学习的端到端导航

模仿学习(Imitation learning, IL)是一种通过模仿专家示范来训练智能体的机器学习方法(图 9),广泛应用于自动驾驶、机器人导航等领域。其核心思想是让智能体学习专家的决策策略,从而在复杂环境中实现高效的自主决策。其中模仿学习根据学习机制的不同分为行为克隆(Behavior cloning,BC)和逆强化学习(Inverse reinforcement learning,IRL)。

行为克隆:模仿学习中的行为克隆作为一种直

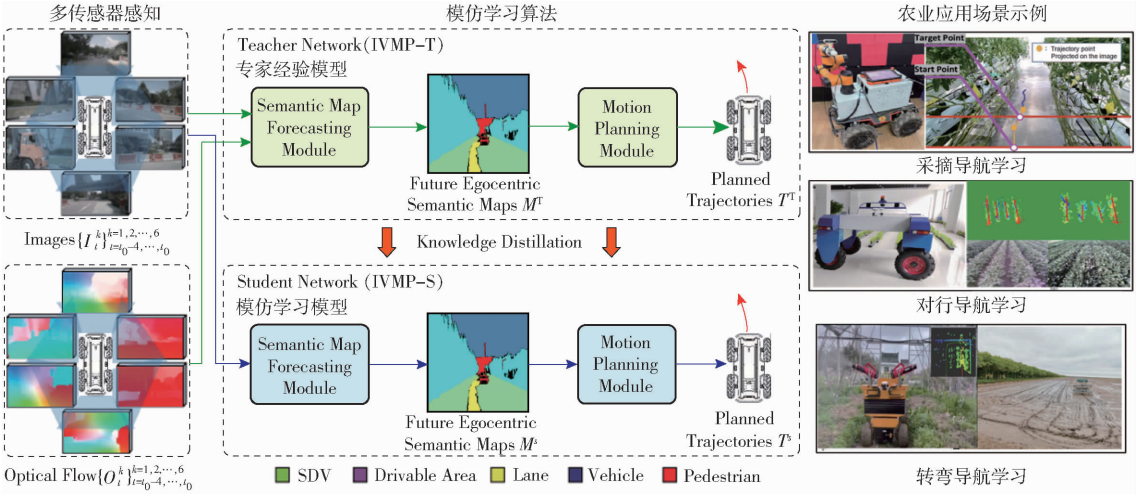


图 9 基于模仿学习的农业机器人应用示例

Fig. 9 Application examples of agricultural robots based on imitation learning

接的方法,本质为有监督学习问题,通过收集专家在各状态下采取的动作所形成的状态-动作对,训练策略网络来拟合专家行为,具有实现简单、训练高效的优势,但存在“分布偏移”问题,即测试时遭遇未见过状态会导致性能大幅下滑。在农业领域,农业机器人执行田间路径跟踪、作物采摘等任务时,可借助行为克隆。以人类驾驶员或遥控操作员为专家,收集其示范数据,训练机器人模仿操作,实现自动导航和作业,降低操作难度与学习成本,提升农业自动化与智能化水平。

条件行为克隆(Conditional behavior cloning)算法以当前状态及额外条件变量为输入。在农业场景下,条件变量可表示为开始喷洒、绕过水洼、加速通过等指令。这使得单一模型能够依据不同的高级指令执行相应的子任务,从而显著提升其在农业自动化应用中的实用性和灵活性。

数据集聚合(DAgger)是一种迭代式在线算法。先用专家数据训练策略,再部署策略收集新状态,请专家标注动作,将新数据加入训练集重新训练。如此循环可缓解分布偏移。在农业中,若机器人因光线变化偏离作物行,DAgger 记录此刻画面,由专家标定转向角度,使机器人学会修正路线。农业机器人通过模仿人类驾驶员在田间行走的轨迹,实现对作物行间的自动导航。HUANG 等[72]利用摄像头或激光雷达感知环境,结合行为克隆方法训练神经网络,使机器人能够在不同作物行间稳定行驶,显著提升了导航的鲁棒性和适应性。

逆最优控制(Inverse optimal control, IOC):逆最优控制,也称逆强化学习,假设专家的行为是某种最优策略的结果,目标是通过专家演示反推出其隐含的奖励函数,再用强化学习方法学习最优策略。相比行为克隆,逆最优控制能更好地泛化到未见过的场景,但实现复杂、对专家演示的质量和多样性要求更高。

为了解决奖励函数不唯一的问题,最大熵逆最优控制策略寻找一个既能解释专家行为,又能让其他非专家行为尽可能随机(熵最大)的奖励函数。这使得学习到的奖励函数更具鲁棒性,不会过拟合到专家演示的特定细节。

生成对抗模仿学习(Generative adversarial imitation learning, GAIL)是一种类逆强化学习方法,规避了显式重建奖励函数的繁琐步骤。在 GAIL 框架下,生成器由机器人的策略网络构成,负责输出动作序列以模仿专家行为;判别器则担当“鉴赏家”角色,专注于区分状态-动作对来源于真实专家还是生成器。训练阶段,二者开启对抗模式:判别器不断学习鉴别,生成器持续优化改进。此时,判别器的反馈转化为一种实时动态调整的奖励信号,生成器借助 PPO 等强化学习算法最大化该信号,促使自身行为逐步逼近专家策略,直至达到以假乱真的效果。

在复杂农田环境下,机器人需理解专家的意图或目标,而不仅仅是模仿其动作。逆最优控制能助其推断隐含目标,以在新环境做出合理决策。农业任务多为多目标的,例如高效率、低能耗、零作物损

伤等。专家操作时会在这些目标间权衡，多目标 逆最优控制可从专家数据中反推出各目标权重，使机器人学会在不同情况下做出合理权衡决策，如兼顾作物保护、土壤保护和作业效率等。相比单纯行为克隆，逆最优控制能让机器人在未见过的田块、不同作物类型、不同气候条件下，保持较好的作业表现。

综上，模仿学习作为智能体学习专家行为的重要方法，在农业机器人领域展现出巨大潜力。行为克隆通过专家示范直接训练策略网络，适用于规则性强的农业作业场景，如行间导航和自动驾驶。逆最优控制则通过推断专家的隐含目标和奖励函数，使机器人能够在复杂和动态的农业环境中实现更高层次的自主决策。随着深度学习和传感技术的发展，模仿学习方法逐步推动农业机器人向更高效、更智能的方向发展。

5.1.2　基于视觉-语言模型的机器人导航

近年来，基于视觉-语言模型（Visual language model，VLM）的农业机器人导航研究取得显著进展，其核心在于利用大语言模型的多模态理解、语义推理和泛化能力，解决传统方法在复杂环境中的局限性。

相较于传统导航方法，VLM 的优势在于其出色的泛化能力和语义理解能力，凭借其在大规模数据上的预训练，能够更好地适应不同的农业场景和任务需求，为智能农业机器人的实际应用提供了新的技术路径。

图 10 展示了视觉-语言模型在农业机器人上的应用案例，其中 VLM 在农业机器人导航中的应用主要体现在 3 个关键环节：场景描述、场景分析和层次规划。在场景描述阶段，VLM 能够综合处理视觉信息（如作物状态、天气条件、地形特征等）和语言指令，生成对农田环境的全面理解。相比传统的单模态感知方法，VLM 能够更好地处理复杂的语义信息和上下文关系。

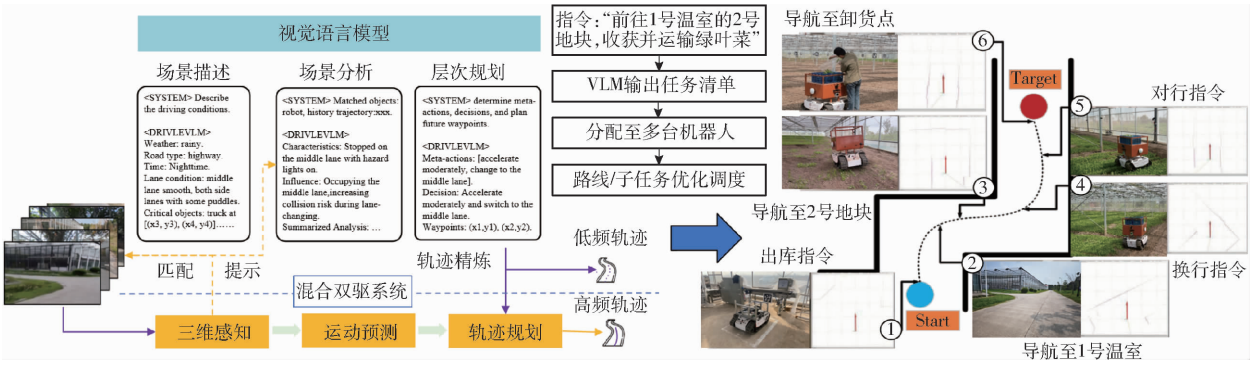


图 10　基于视觉-语言模型的导航应用示例

Fig. 10　Example of vision language model based navigation application

在场景分析环节，VLM 展现出强大的推理能力，能够识别关键对象、分析空间关系，并评估环境中的风险因素。例如，系统能够自动识别作物行、障碍物位置，并理解"前往 1 号温室的 2 号地块获取运输绿叶菜"这样的复杂指令。这种语义理解能力使得机器人能够处理更加自然和灵活的人机交互场景。

在层次规划阶段，VLM 能够将高层语义目标分解为可执行的具体动作序列。图 10 展示了基于视觉-语言模型的导航应用示例，模型能够生成从起始位置到目标位置的完整导航路径，包括路径点选择、避障策略和任务执行指令。这种分层决策架构有效提升了系统的智能化水平。

5.2　具身操作关键技术

5.2.1　大模型驱动的任务分解

近年来，大语言模型（Large language model，LLM）在自然语言处理领域实现突破性进展，其强大的语义理解与生成能力，为机器人任务规划开辟了全新范式。相比之下，传统机器人任务规划方法多依赖预定义规则和人工编写的知识库，难以应对复杂多变的现实环境。而大语言模型凭借大规模预训练积累的世界知识与推理能力，不仅为实现更具通用性和灵活性的机器人任务规划提供了可能，更为农业机器人适配复杂作业环境开辟了新的规划路径。后面将综述大语言模型在机器人任务规划领域的最新研究进展，剖析现有方法的优势与局限，并探究其在具身智能农业机器人中的发展潜力。

回顾近年来的相关研究，众多学者基于大语言模型提出了新型机器人任务规划框架。其中，LIU 等[73]提出的 LLM + P 框架，创新性地将经典规划器的优势与大语言模型深度融合：通过将自然语言描述的规划问题转换为规划域定义语言（PDDL）文件，借助经典规划器快速求解后，再将结果转回自然语言形式。实验结果显示，LLM + P 框架能够为多数问题提供最优解决方案，而单纯依赖大语言模型时，在多数问题中甚至无法生成可行规划。CHEN

等[74]开发了 RoboGPT 系统(图 11),是专为执行指令跟随任务设计的具身长期决策系统,整合了 3 个核心组件:RoboPlanner、RoboSkill 和 Re - Plan。其中 RoboPlanner 通过 67 000 条具身规划数据对 Llama 模型进行微调,可将任务分解为逻辑子目标;RoboSkill 针对各子目标规划具体任务,有效提升导航与操作能力;Re - Plan 模块则依据实时环境反馈动态调整子目标。在 ALFRED 基准测试[75]及其他领域任务中,RoboGPT 的成功率分别达到 60.7% 和 78.0%,性能优于其他先进方法。LIU 等[76]提出了 DELTA 方法,这是一种基于大语言模型的高效分解式长期机器人任务规划方案。该方法以场景图作为大语言模型内部的环境表示,可快速生成精确的规划问题描述,并借助大语言模型将长期任务目标分解为自回归子目标序列。在 5 个不同领域的评估中,DELTA 的成功率达 98%,规划时间较现有方法缩短数千倍,充分展现了其在复杂环境下处理长期任务的高效性。

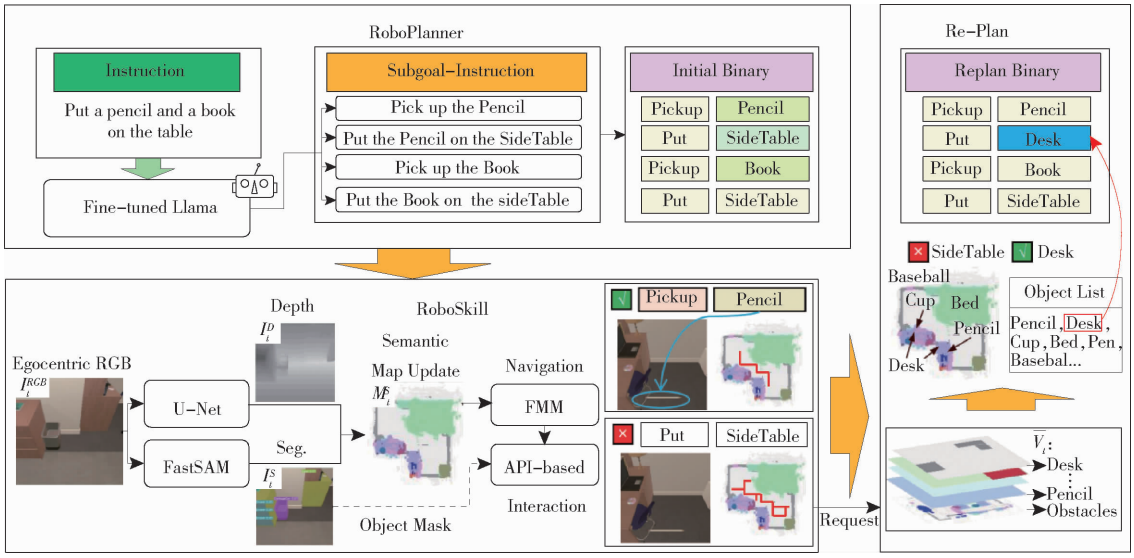


图 11 LLM 与经典任务规划相结合示例[74]

Fig. 11 Example of combining LLM with classic task planning

同时,还有学者通过融合大语言模型与经典任务规划方法,提出了改进的机器人任务分解与规划方案,具体如下:AHN 等[37]提出的 SayCan 方法,通过大语言模型生成动作序列,并结合 Affordance 函数评估动作可行性。其操作流程为:首先将自然语言形式的服务请求输入系统,由大语言模型完成任务理解并生成潜在动作序列,再通过 Affordance 函数筛选出可行动作。该方法虽在特定任务中取得一定成效,但在复杂长期任务处理中仍存在局限。SONG 等[77]开发的 LLM - Planner 框架,依托大语言模型的少样本学习能力,将自然语言指令转换为可执行的规划步骤,并结合实时环境反馈动态调整。实验表明,该框架在视觉-语言导航任务中展现出较好的泛化能力,但其在复杂操作任务中的成功率仍有提升空间。LIN 等[78]提出的 Text2Motion 方法,可将自然语言指令转换为可执行的运动计划,其核心机制为:借助大语言模型解析指令语义,结合机器人运动学与动力学模型生成合理的运动轨迹。该方法在简单物体重排任务中能生成有效的运动规划,但在复杂环境中的适应性仍需进一步增强。RANA 等[79]提出的 SayPlan 方法,通过 3D 场景图对大语言模型进行扩充,实现了可扩展的机器人任务规划。其流程为:先借助大语言模型开展语义搜索,确定与任务相关的子场景图;再利用大语言模型生成任务计划,并依据环境反馈进行迭代重规划。该方法在处理含依赖关系的复杂子任务时表现出较好适应性。CHEN 等[80]提出的 AutoTAMP 方法,将大语言模型同时作为翻译器与检查器,实现自回归任务与运动规划。该方法先将自然语言指令转换为规划问题,再经大语言模型校验可行性并动态调整。实验表明,其在复杂操作任务中可生成有效规划,但处理长期任务的效率仍待提升。

当前,通过融合大语言模型与经典规划方法、环境表示及具身交互技术,机器人任务规划领域的相关研究已取得显著进展。但现有方法仍存在多重挑战与局限:大语言模型固有的幻觉现象可能导致生成的任务规划方案缺乏可行性;对大规模人工标注数据的强依赖性,严重制约了方法的跨领域泛化能力;在动态环境中,系统对不确定性的处理及适应性调整能力仍需大幅提升。具体来看,LLM + P 方法在新领域应用中往往依赖大量人工干预才能有效运行;RoboGPT 在极端复杂任务场景下的成功率仍有

提升空间；DELTA 系统在动态环境中的自适应能力则亟待增强。

针对农业机器人领域，要充分释放大语言模型（LLM）在提升任务分解精度与智能规划效能上的潜力，需在现有技术框架基础上，从以下几个关键方向深化研究：①构建适配农业场景的鲁棒性规划范式。通过将作物生长模型、土壤墒情感知等农业专属约束嵌入大语言模型推理过程，结合多模态验证机制（如视觉传感器实时反馈校验），从源头抑制幻觉问题对规划可行性的干扰。②探索农业数据高效利用的学习框架。鉴于农田数据标注成本高、场景差异大的特点，可利用大语言模型的少样本学习能力，通过构建跨作物、跨地域的农业任务元知识库，降低对大规模人工标注数据的依赖。同时，创新性融合深度学习的端到端决策优势与大语言模型的符号推理能力，形成“数据驱动 + 知识引导”的混合规划模式，大幅提升复杂农艺流程（如精准灌溉、病虫害防治）的规划效率。③研发动态农业环境的自适应规划机制。针对农田环境中光照变化、作物生长动态等不确定性因素，可将时空连续的农田场景图与大语言模型的时序推理能力结合，构建“环境状态实时解析-规划方案在线演化”的闭环系统。例如，通过大语言模型对倒伏作物的视觉信号进行语义理解，动态调整收割机路径规划，这种具身智能与符号推理的深度耦合，显著增强机器人在非结构化农业环境中的应变能力。

5.2.2　视觉触觉融合抓取

在机器人抓取与操作任务中，视觉与触觉模态的融合对提升柔性物体操作的稳定性和适应性至关重要[81]。其中，视觉传感器（如 RGB - D 相机）可提供物体的全局几何信息、空间位置及表面状态；触觉传感器（如 GelSight[82]）则能捕捉局部接触力、纹理形变等物理特性，从而有效弥补视觉在微观交互感知层面的局限。

研究者通过 Transformer、CNN 等深度学习模型融合视觉与触觉模态数据，以更全面地推断物体物理属性（如摩擦因数、质心）并优化抓取策略[83]。例如，吴培良等[84]提出的多感知局部增强 Transformer，通过 Feature - to - Patch 模块提取边缘信息，结合 Local - Enhanced 模块增强局部特征相关性，将视觉与触觉特征深度融合，最终使抓取准确率提升至 85%。类似地，LI 等[85]开发的 TGCNN 模型，利用图卷积网络实现视觉-触觉数据的融合，在复杂场景中显著提升了抓取鲁棒性。

此外，LI 等[86]提出多感官融合感知的机器人学习框架，整合视觉、音频及触觉传感数据，用于解决密集包装与倾倒这两类挑战性操作任务。实验结果显示：视觉主要提供物体位置信息，但易受遮挡影响；音频可提供即时反馈并表征物体材质等特性，但其信息在非接触状态下较为有限；触觉则能捕捉物体动力学信息，尤其适用于表征局部几何形状及接触力的方向与大小。受人类触觉-视觉联觉机制启发，YUAN 等[87]提出一种基于点云的新型触觉表示方法（图 12）。该方法将触觉传感器信号与深度相机获取的点云数据融合至统一 3D 空间，使机器人具备对自身触觉交互的“视觉化感知”能力。这种融合方式从根本上实现了视觉与触觉信息的有机整合，不仅提供更丰富的空间表征，更有助于机器人对动作进行精准推理。

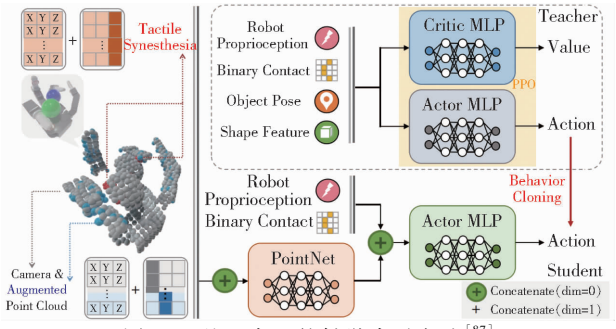


图 12　基于点云的触觉表示方法[87]

Fig. 12　Tactile representation method based on point cloud

5.3　农业机器人多机协同关键技术

农业作业场景通常具有任务环境复杂（如不规则地形、多变光照）、区域广阔（大田、果园）、对象多样（作物、土壤、障碍物）等特点。受限于自身能力，单一农业机器人难以高效应对这些复杂农业环境带来的挑战。因此，多农业机器人协同作业成为突破单台设备能力瓶颈、提升农业智能化水平和作业效率的关键发展方向（图 13）。通过针对农业任务特点（如播种、植保、收获）的高效任务分配与调度，多农业机器人系统可实现田间协同作业的高效性与精准性[88]。随着具身智能理念的引入，农业机器人逐步具备感知-决策-执行的闭环适应能力，系统协作方式也由静态任务分工转向动态自组织协同，进一步增强了其在复杂农业场景中的自主性与灵活性。



图 13　智慧农业多机器人任务分配与调度示意图

Fig. 13　Schematic diagram of task allocation and scheduling for multiple robots in smart agriculture

在多农业机器人协同系统中，农业作业任务的

分配与调度构成了决策环节的核心支柱,其制定需全面权衡作业效率的最大化、各农业机器人能耗的均衡化,以及对动态农业环境的适应性等关键要素。

在农业机器人的协同作业体系中,集中式分配与调度技术特指由控制中心对全局作业任务进行系统性分解,再将拆分后的子任务精准指派给各台农业机器人执行。GUO 等[89]针对多农业机器人除草任务分配问题提出了协同离散人工蜂群算法,该方法引入了协同跟随策略避免局部最优,并针对瓶颈机器人设计了局部搜索算子以最小化所有除草机器人的作业时间。试验结果表明了该方法在农业除草机器人任务分配与调度环境中的优越性。MIAO 等[90]采用强化学习选择最优分配策略,并利用进化算法优化作业时间和总成本,适用于农业除草机器人作业任务分配与调度。但这种基于强化学习的方法时间成本高。DAI 等[91]研究了多采摘机器人任务分配问题,旨在实现最大完成时间和所有机器人总行驶距离的双目标优化,提出了一种有效的多目标离散人工蜂群算法。但这种群智能方法计算量较大,需要较长时间才能收敛。尽管集中式方法能获得较优解,但对通信与计算资源依赖较高,难以适应农业环境中的动态变化和大规模部署。

在农业机器人协同作业中,分布式任务分配与调度技术是指各农业机器人依托自身传感器获取的局部环境与作业信息,独立完成任务计算与规划。王猛等[92]聚焦同构农机机群的动态作业场景,提出了一种基于改进合同网算法的任务分配方法。该方法在构建包含作业时间、油耗及行驶路程的综合代价函数基础上,引入农机间的分布式招投标机制,通过设定科学的招标阈值、中标者任务再分配及任务交换等策略,显著增强了农业机器人任务分配的鲁棒性与灵活性。

综上所述,农业机器人多机协同作业的任务分配与调度技术已在各类农业任务中得到广泛应用。其中,集中式任务分配虽能实现全局最优解,但受限于通信瓶颈与环境静态假设的约束;而分布式方法凭借更强的环境适应性与系统扩展性,更适用于复杂动态的农业作业场景。随着智能农业向高度自主化方向演进,未来的研究可进一步融合具身智能的环境感知能力与自主决策机制,探索出动态农业环境下更高效、更具弹性的任务调度策略,为农业机器人协同作业提供更有力的技术支撑。

6 具身模拟与生成关键技术

农业机器人的智能化发展,需在复杂多变的真实农业环境中完成大量训练与验证,但实际田间试验成本高昂、周期漫长,且存在作物损伤、设备故障等风险,这使得高质量的具身模拟环境成为农业机器人研发的关键基础设施。同时,农业场景的季节性、作物生长的动态演化、多样的天气条件以及不同地域的环境差异,对模拟系统的真实性、灵活性和泛化能力提出了极高技术要求。图 14 展示了具身模拟与生成的技术路线。

下面将从具身模拟与生成的核心技术维度展开深入分析,围绕高保真场景重建、场景编辑与生成、Sim - to - real 迁移、具身仿真模拟器这 4 个关键方向,阐述如何构建高质量农业模拟环境,实现从虚拟训练到真实应用的有效迁移。

6.1 高保真场景重建

在面向农业场景的具身模拟与生成研究中,高保真场景重建是支撑技术落地的核心研究方向[93],图 15 展示了典型的农业场景高保真场景流程。这是一项融合多学科技术与农业生态复杂性的系统性工程。其核心是构建物理真实、生物动态与功能适

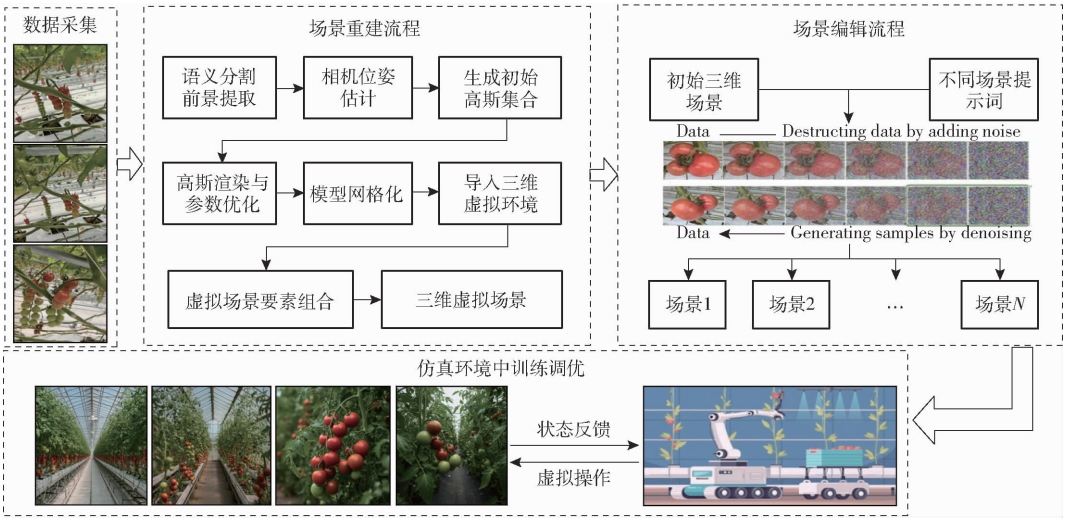


图 14 具身模拟与生成的技术路线

Fig. 14 Technical roadmap for embodied simulation and generation

配三位一体的数字孪生环境，为农业机器人的感知训练、决策优化、任务验证等提供全要素仿真平台。

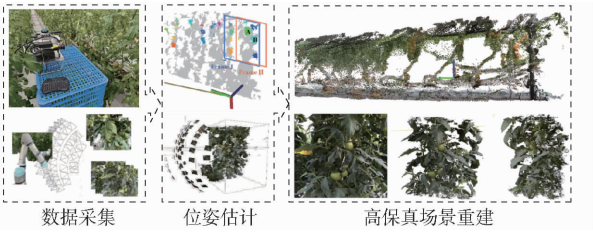


图 15　农业场景高保真场景重建流程图

Fig. 15　High fidelity scene reconstruction process for agricultural scenes

从物理真实性角度，高保真场景重建需精准复刻农业环境的空间结构与物理属性。通过多源异构数据融合技术，将激光雷达的厘米级地形高程数据[94]、多光谱相机的作物冠层光谱信息[95]、超声波传感器的土壤墒情数据[96]等进行时空对齐，构建包含地形起伏、植株形态、土壤质地等要素的三维模型。例如，在果园场景中，模型需精确呈现果树的枝桠分布、果实挂果密度等，这些微观特征直接影响机器人视觉识别准确率与机械臂采摘路径规划[97]。

生物动态性的复现是农业场景重构的独特挑战。农作物的生长周期[98]、病虫害导致的形态变异[99]、环境因素引发的生理响应[100]等动态过程，需通过时间序列建模技术实现数字化表达。依托植物生长模型与传感器数据的耦合，将作物从播种到收获的全过程转化为可交互虚拟场景，使机器人能在模拟环境中学习不同生长阶段的作业策略。例如，番茄植株在不同生长周期的茎秆硬度差异会直接影响机械臂夹持力度的控制参数[101]，高保真场景需实时反馈此类生物力学特性变化。

功能适配性着重强调场景模型对机器人任务的支撑能力，重建过程需将农业作业的功能需求融入模型设计。例如，在杂草识别场景中[102]，不仅要呈现杂草与作物的形态差异，还需通过语义分割技术标注不同植物的类别信息，为机器人视觉算法提供训练标签；在农机导航场景中[103]，则需对田间道路、灌溉沟渠等功能性要素进行拓扑建模，生成可支持路径规划算法验证的数字地图。

此外，高保真场景重建还需具备环境扰动模拟能力。借助气候模拟引擎生成不同光照、风速风向、降水条件下的场景变体，结合物理引擎模拟雨滴击叶的力学效果、风载下植株的摆动轨迹，使机器人能在极端环境中适应性得到充分训练。这种动态环境建模能力突破传统静态场景局限，为农业机器人应对多变自然条件提供了仿真验证平台。

6.2　场景编辑与场景生成

场景编辑与生成是具身模拟研究的关键技术，其核心在于给定数字场景及对应的目标场景描述，通过规则驱动、数据驱动、生成式模型及多模态融合等方法，快速创建逻辑自洽、逼真多样的场景，为具身智能体提供丰富的模拟训练环境。具体方法包括：

（1）基于规则的场景生成[104]：依托预设规则与知识图谱构建场景。例如，构建果园时，依据果树行间距适配农机通行、果实挂果高度集中于 1.5 ~ 3.0 m 等规则，从作物模型库选取品种，按梯田或平地布局规则确定植株位置与朝向，生成符合采摘需求的果园拓扑；设施农业场景中，根据温室栽培规则，快速生成适配蔬菜定植、巡检机器人训练的温室布局。

（2）数据驱动的场景生成[105]：利用农田实测数据或历史作业数据学习生成新场景。基于海量农田三维点云数据（如无人机激光雷达采集的作物冠层形态、田间障碍物分布），学习玉米、水稻等作物不同生长周期的空间结构特征，生成多样化田间场景；杂草识别场景中，通过分析农田图像中杂草与作物的像素分布模式，生成含不同种类作物与杂草的虚拟农田，为除草机器人视觉算法提供训练样本。

（3）生成式模型驱动的场景生成[106]：运用生成对抗网络（GAN）、变分自编码器（VAE）、扩散模型等生成场景。利用 GAN 生成不同光照、天气下的农田场景，生成器学习多云、晴天、黄昏等实景特征，生成对应虚拟麦田，助力机器人视觉系统适应复杂光照下的作物识别；借助扩散模型从噪声中迭代生成高细节作物生长动态场景，如模拟番茄从幼苗到结果期的茎秆弯曲、果实膨大过程，用于机械臂抓取力度控制算法训练。

（4）多模态融合的场景生成[107]：融合多模态数据生成场景，提升真实感与丰富度。结合文本描述与多光谱图像数据，生成含冠层光谱特征（如叶绿素含量分布）和杂草语义标签的三维大豆田场景；农机导航场景中，融合 GPS 轨迹、土壤墒情传感器数据与田间实景图像，生成带土壤阻力分布、作物倒伏风险标注的虚拟农田，为无人农机路径规划提供多维度训练环境。

6.3　Sim－to－Real 迁移

具身智能农业机器人的 Sim－to－Real 迁移，核心在于将机器人在虚拟农业环境中通过模拟训练习得的感知、决策与操作能力，精准高效地迁移至真实农业场景，以此提升机器人在真实场景的泛化能力。

在虚拟环境中，依托高精度农业场景模拟系统，可构建涵盖农田地形、作物生长状态、农机具作业环境等要素的逼真虚拟场景。机器人能围绕播种、除草、灌溉、采摘等核心农事任务开展大规模模拟训

练。例如在虚拟果园中反复优化机械臂的果实采摘路径规划,或在虚拟农田里仿真无人驾驶农机的自主导航与播种作业流程,从而快速积累任务处理经验。

然而,虚拟农业环境与真实农业场景之间存在显著的 Sim - to - Real 迁移“域间隙”。真实场景中,土壤质地的复杂分层、作物生长的个体差异、光照强度波动、降雨等天气变化,以及田间杂物的随机干扰等因素,均难以在虚拟环境中完全复刻。典型如虚拟模拟中设定的作物成熟度视觉识别标准,可能因真实作物表面色泽、纹理的细微差异而失效;模拟的土壤阻力参数,也会与实际土壤在湿度、颗粒度动态变化下的阻力产生偏差,进而导致农机具行进控制策略出现执行误差。

为应对虚实之间的域间隙,研究人员采用了多重针对性技术手段:

(1)域随机化[108]:在虚拟训练阶段,通过随机改变作物的颜色、形态、种植密度,模拟不同土壤湿度、光照角度与强度,以及添加随机分布的田间杂物等方式,迫使机器人适应农业环境的多样性,显著增强其对复杂真实场景的鲁棒性。例如训练除草机器人时,通过随机化杂草的种类、生长形态与分布密度,使其学习识别不同状态的杂草,有效避免因虚拟与现实中杂草特征差异导致的误判。

(2)多阶段训练与微调策略[109]:先在虚拟环境中完成大规模基础训练,使机器人掌握农业任务的基础操作逻辑与决策模型;随后利用少量真实农业场景数据对模型进行针对性微调。以果蔬采摘机器人为例,在虚拟环境中完成基础训练后,只需在真实果园采集少量果实样本数据,即可调整视觉识别参数与机械臂抓取力度,快速优化模型以适应真实环境。

(3)农业数字孪生技术[110]:通过创建与真实农田、果园高度同步的虚拟孪生模型,在虚拟空间中开展算法优化与策略训练,再将成熟方案迁移至真实农业机器人,既能保证训练效果,又能大幅降低实际测试的成本与风险。

这种 Sim - to - Real 迁移技术在农业生产中具有重要应用价值:在果园采摘场景,经虚拟训练与迁移的机器人可精准识别不同成熟度、不同生长位置的果实,并以适配力度实现无损采摘,显著提升采摘效率与果实品质;在农田管理领域,经虚拟训练的无人驾驶农机迁移至真实农田后,能根据实际地形、土壤墒情自动调整作业参数,实现精准播种、变量施肥,大幅提升资源利用效率,为农业生产的智能化、精准化升级提供有力支撑。

6.4 具身仿真模拟器

近年来,针对具身智能研究的新型模拟器不断涌现,为智能体的训练与开发提供了多重核心价值:①构建安全高效的学习环境,使智能体可在无物理风险的虚拟空间中开展高强度训练。②能低成本生成海量数据 Sim - to - Real 迁移。这类数据若在现实场景中采集,往往需投入极高的时间与经济成本。③可作为标准化基准测试工具,为不同算法与方法的效果评估提供统一对比维度。

表 1 为常用于具身智能体研究的模拟器,模拟器中使用的数据既可以是合成的,也可以是基于现实世界扫描得到的。例如 Matterport3D[111] 和 Habitat - Sim[112] 模拟器是基于现实世界环境扫描构建的,这类现实世界模拟器极大地增强了智能体从模型到现实应用的能力。NVIDIA 开发的 Isaac - Sim[113] 仿真器是基于 PhysX 物理引擎构建,支持对农业机器人机械臂抓取、作物茎秆弯曲等场景进行高精度动力学仿真,甚至能模拟不同土壤湿度下农机行进的阻力差异。该模拟器可无缝对接迁移 ROS(机器人操作系统),允许研究人员在虚拟环境中测试多传感器融合算法,例如将双目视觉与触觉传感器数据结合,优化果实采摘时的力度控制策略。而 Genesis 作为轻量级通用仿真平台,采用模块化设计,用户可通过拖放式操作快速构建包含农田、温室、果园等场景的虚拟环境,其内置的植物生长引擎能根据温度、光照等参数动态生成不同生长阶段的作物模型,如玉米拔节期的冠层结构或葡萄藤的攀爬形态,特别适合农业机器人的播种、除草等任务的算法原型验证。

尽管上述模拟器已在农业机器人的具身智能研究中展现出重要价值,但当前领域内仍缺乏专为农业场景设计的具身智能专用仿真器。现有工具多为通用型模拟器或针对特定工业场景开发,难以充分适配农业环境的独特属性:既无法精准复现作物生长的动态演化过程,也难以全面模拟土壤特性的复杂交互,更难以覆盖农业生产中多样化的作业任务需求。

因此,研发具备农业场景深度适配能力的专用具身智能仿真器,已成为推动农业机器人技术向高精度、强鲁棒性发展的关键方向,其核心价值在于为农业具身智能的算法训练、系统验证提供更贴近真实生产场景的虚拟试验平台。

7 具身学习与进化关键技术

农业机器人在实际作业过程中面临着高度动态和不确定的环境挑战,包括不断变化的作物生长状

表 1　具身智能体的模拟器与数据集

Tab. 1　Simulators and datasets for embodied intelligent agents

模拟器	数据集	任务	环境
Matterport3D	基于现实室内环境扫描数据	室内场景导航、物体交互、空间理解等，如家庭、办公室场景中的路径规划	室内
Habitat－Sim	融合三维重建与传感器数据	支持室内外场景的导航与操作模拟，如机器人在虚拟房屋中避障或抓取物体	室内/室外
Isaac－Sim	基于物理引擎的合成数据集	专注于机器人操作与复杂环境模拟，可构建工业机械臂、农业机器人等多类型智能体，支持高精度动力学仿真，并能生成包含视觉、触觉等多模态传感器数据的训练场景	室内/室外
Genesis	基于生成式模型的通用数据引擎	轻量化、快速、易用的仿真平台，可以仿真各类物体与材质的通用物理引擎	室内/室外

态、多样化的气候条件、复杂的地形地貌以及突发的异常情况等，这些因素使得传统的预编程和静态学习方法难以满足农业机器人长期稳定运行的需求。同时，农业作业的长周期性特征要求机器人具备持续学习、自主适应和知识积累的能力，能够在缺乏大量标注数据的情况下不断优化自身性能，并在面对新环境和新任务时实现快速适应与智能决策。下面从具身学习与进化的核心技术维度展开深入分析，分别从自监督学习、强化学习、模仿学习 3 个关键方面，阐述如何构建农业机器人的自主学习与持续进化能力，使其能够在复杂农业环境中实现真正的智能化与环境自适应。

7.1　自监督学习

监督预训练技术已经成为现代人工智能领域的一项关键突破，它使得模型能够在没有人工标注数据的情况下，通过自动生成标签来学习有用的特征。与传统的监督学习方法不同，后者需要大量的人工标注数据来进行训练，自监督学习则依赖大量未标注的数据，利用数据本身的结构性信息进行学习，进而提取出通用的、具有迁移能力的表示，图 16 展示了自监督学习的原理。在语言处理领域，自监督学习方法已广泛应用，其中包括自回归模型和遮蔽自编码模型。自回归方法的目标是根据先前的文本序列预测下一个单词，而遮蔽自编码方法则是通过遮蔽掉输入序列中的一部分元素，迫使模型学习推断这些缺失部分的内容。这些方法能够在大规模无标注数据上进行训练，从而获得丰富的语言文本特征，这些特征能够很好地迁移到各种下游任务，如文本分类、情感分析等。

自监督学习的最大优势在于它能够在缺乏标注数据的环境中，利用海量的原始数据进行预训练，进而为下游任务提供强大的泛化能力。这些通用表示可以通过微调（fine-tuning）在特定任务中进一步优化，也可以通过零样本学习（zero-shot learning）直接应用到新的任务中，显著提高了模型在多种任务上

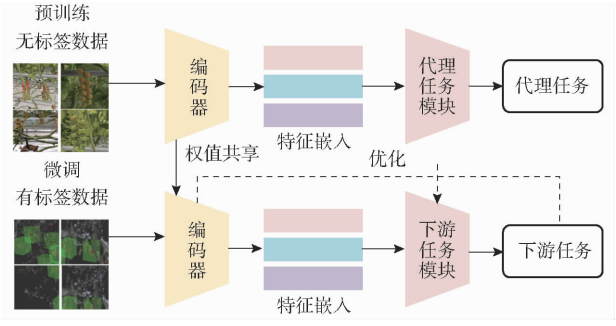


图 16　自监督学习原理图

Fig. 16　Principle diagram of self supervised learning

的表现。

在具身智能的领域，自监督学习的应用也展现了广泛的潜力。具身智能体需要在复杂的环境中进行感知、推理和决策，自监督学习作为一种学习方法，可以帮助这些智能体通过与环境的互动，自行学习并优化其行为策略。例如，通过与物理世界的互动，智能体可以通过自监督的方式从未标注的数据中学到物体的运动规律、触觉反馈和视觉信息，从而实现自主的技能学习与适应能力。这种方法不仅减轻了人工标注数据的需求，也使得具身智能体能够在更加多样化和动态的环境中获得更强的学习能力。

7.1.1　基于自回归预训练的自监督学习

自回归预训练方法正被应用于具身智能任务，核心思想是将智能体与环境交互的轨迹（状态、动作、观测）建模为序列。REID 等[114]提出使用 Transformer 模型自回归预测轨迹序列，决策 Transformer 借此在离线强化学习中实现了性能提升与快速收敛。Gato 模型[115]将多模态、多任务数据统一为 token 序列，通过共享 Transformer 架构在数百项任务上预训练，展示了强大通用性。PaLM－E 模型[39]融合大型语言模型与具身智能体，通过自回归训练实现现实世界决策与执行，其机制是将观察转化为语言单元输入，再将输出语言序列解码为动作。RoboCat 模型[116]基于自回归多模态序列建模，

结合视觉编码与自我迭代数据增强,提升了持续学习能力和真实场景零样本泛化性能。

7.1.2 基于掩码自编码预训练的自监督学习

自监督学习另一个典型算法是基于掩码自编码预训练的具身智能模型,掩码自编码预训练通过重构被遮蔽的多模态输入学习跨模态表征,显著提升下游任务的样本效率与泛化能力。MVP 模型[117]利用掩码图像自编码预训练通用视觉表征,并冻结编码器以强化学习微调控制器。SMART 模型[118]融合多任务自监督目标预训练控制变换器,提升模仿与强化学习效率。MaskDP 模型[119]通过预测缺失状态动作序列实现高效强化学习。Voltron 模型[120]结合掩码重建与语言生成进行语言驱动的视觉表征学习。RPT 模型[121]通过预测被遮蔽的感知运动内容学习跨任务与环境的鲁棒表征。

7.1.3 基于对比学习的自监督学习

基于对比学习的方法通过对比不同数据样本之间的相似性与差异性,来学习有效的特征表征。例如,NUNES 等[122]提出了一种基于对比学习的 3D 自监督学习方法,专注于理解场景结构。他们通过从点云中提取与类别无关的片段,并使用对比损失来区分结构上相似和不相似的元素,从而帮助模型更好地理解复杂场景中的物体关系和空间结构。ViLBERT[123]、UNITER[124]等视觉-语言模型采用对比学习方法将视觉和语言模态有效地连接起来,在视觉问答等下游任务上展现了显著的性能提升。CLIP 模型[27]采用从网页抓取的大量视觉-语言数据进行预训练,通过对比学习方法在视觉和语言之间建立联系。这些模型的成功在于其能够处理丰富但嘈杂的数据,表现出强大的任务泛化能力和跨任务学习能力。对于具身智能体而言,这类模型能够有效提升智能体在视觉和语言理解方面的能力,使其在与人类进行自然交互或执行多任务时表现出更强的适应性。

进一步地, VLN - BERT 模型[125]通过视觉-语言预训练,证明了无身智能体上学习到的视觉表征在具身智能体上的有效性。这种方法可以增强农业机器人在复杂环境中的视觉能力,使其能够理解不同的场景、标识特定的物体,并与周围环境进行交互。此外,R3M 模型[126]结合时间对比学习、视频语言对齐和正则化惩罚训练了一种通用视觉表示模型,通过预训练从多样的人类视频数据中学习,提升下游机器人操作任务的数据效率和学习效果。

7.2 强化学习

近年来,强化学习(Reinforcement learning, RL)作为机器学习的重要分支,凭借其具备让机器人通过与环境交互自主学习最优决策的能力,在机器人领域展现出较大潜力。强调如何基于环境而行动,以取得最大化的预期利益,其核心思想是智能体(Agent)通过与环境交互,基于奖励信号(Reward)学习最优策略(Policy),以最大化长期累积回报[127]。强化学习主要分为 3 类:

(1)基于值函数的方法:这类方法通过估计状态或动作的价值函数间接优化策略。经典算法包括 Q - Learning,其无模型特性使其在未知环境中具有强大的学习能力。深度 Q 网络(DQN)[128]的提出标志着深度强化学习时代的开启,通过经验回放和目标网络解决了深度神经网络在强化学习中的不稳定问题。

(2)基于策略的方法:这类方法直接优化策略函数,特别适用于连续动作空间。策略梯度定理为这类方法提供了理论基础。REINFORCE 算法作为最基础的策略梯度方法,虽然方差较大但为后续发展奠定了基础。Actor - Critic 算法通过结合值函数估计减少了方差。近端策略优化(PPO)[129]通过限制策略更新幅度,在保持样本效率的同时提高了训练稳定性,成为当前最广泛使用的策略优化算法之一。

(3)混合方法:结合了值函数和策略梯度的优势。深度确定性策略梯度(DDPG)[130]将 DQN 的思想扩展到连续控制领域。SAC 算法[131]通过最大熵强化学习框架,在性能和探索之间达到了良好平衡。TD3 算法[132]通过目标策略平滑和延迟更新进一步提升了连续控制的稳定性。

强化学习作为一种通过与环境交互、在试错中学习最优策略的机器学习范式,凭借其卓越的序列决策优化能力,在从感知到行动的端到端控制和跨模态智能决策中实现了突破。

苏黎世联邦理工学院机器人系统实验室[133]借助深度强化学习技术,成功攻克了四足机器人仿真到现实转移的核心难题,实现了敏捷运动控制的重大突破。通过在仿真环境中训练神经网络控制策略,并成功部署至 ANYmal 四足机器人平台,使机器人掌握了远超传统方法的动态运动技能。该方法的技术创新体现在 3 个关键维度:策略架构上,巧妙融合机器人内感受数据与高层运动指令,实现精准的感知-决策耦合;奖励设计上,构建兼顾速度跟踪精度、能耗优化和运动稳定性的多目标优化函数;训练策略上,通过对质量分布、摩擦特性、传感器噪声等关键物理参数的系统性随机化,显著提升了策略的现实适应性。这项突破性研究不仅验证了强化学习在真实机器人系统中的实用性,更标志着从理论仿

真向实际应用的历史性跨越。

7.3　模仿学习

模仿学习作为强化学习的重要分支，通过学习专家演示数据获得最优策略，而非依赖环境的奖励信号，为复杂任务的快速学习提供了有效解决方案。

通过学习专家演示而非环境奖励信号来获得最优策略，为复杂任务的快速学习提供了有效途径。行为克隆（Behavior cloning）是最直接的模仿学习方法，通过监督学习直接从专家轨迹中学习策略映射。逆强化学习（Inverse reinforcement learning，IRL）试图从专家演示中推断出潜在的奖励函数，然后使用传统强化学习方法优化策略。生成对抗模仿学习（Generative adversarial imitation learning，GAIL）借鉴生成对抗网络的思想，通过判别器区分专家和智能体行为来指导策略学习。

与传统强化学习需要大量试错探索不同，模仿学习能够直接从人类专家的行为演示中提取策略知识，显著提高了学习效率。Google DeepMind 团队[38]开发的 RT－1 模型在大规模机器人控制领域取得了开创性突破。该模型基于 Transformer 架构设计，采用行为克隆方法，在包含 130 000 个真实机器人演示样本的大规模数据集上进行训练，数据涵盖 700 多个不同任务类型。RT－1 的核心创新在于将机器人控制问题转化为序列建模任务：模型接收机器人摄像头图像和自然语言指令作为输入，输出离散化的动作 token 序列，从而开创了多模态模仿学习的全新范式。

8　具身诊断与运维研究关键技术

对农业机器人而言，故障是其在非结构化农田环境中长期承受多任务、高强度作业的必然衍生结果。农机是复杂机电一体化系统，机械电气等部件的故障可以导致其感知失灵、行动失真、决策失效和执行失败。另一方面，农机的故障影响也可反映并体现在其作业质量上。因此，在具身智能大脑的引导下，环境信息、农机故障与作业质量之间的信息可以互补互济、互通有无，形成农业机器人的具身诊断和运维方法。下面分别从农业机器人作业状态具身诊断、农业机器人作业质量具身评估两个关键方面，深入分析具身诊断与运维的核心技术，图 17 阐述了如何构建基于具身智能的自主诊断与运维体系，实现农业机器人的智能化健康管理和作业质量保障。

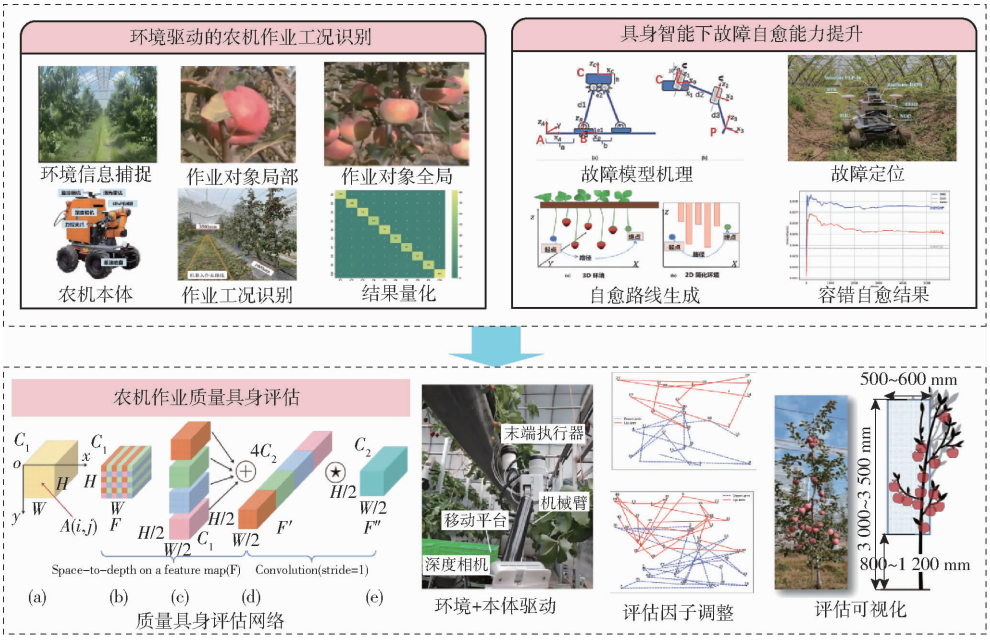


图 17　农业机器人具身诊断与运维关键技术

Fig. 17　Key technologies for body diagnosis and operation of agricultural robots

8.1　农业机器人作业状态具身诊断

确保农机内部稳定运行是应对外部复杂作业环境的基础，及时发现故障并进行自愈干涉，可以有效矫正故障农机的非预期感知、学习、规划和行动。因此，持续有效的故障监测与自愈是保障农机安全稳定作业的重要前提，也是提高作业质量“稳产保供”的重要支撑。

模型驱动、数据驱动和“数模联动”是经典机械设备故障诊断方法。对于农机故障来说，国内外学者对于这些方法进行了深入研究。在模型驱动的农机智能诊断研究方面，马楠等[134]着重阐述了农业复杂环境中多机器人运动模型常面临多传感器融合精度不足及相对位姿估计失准的问题。鉴于农机模型在未知场景下构建困难，此类方法通常仅适用于

已知农业环境和相对成熟的农机平台。在农机智能诊断的数据驱动领域,周俊等[135]开发了一种结合时频图像脊线提取与改进稀疏分量分析的 RV 减速器复合故障盲提取策略。翟长远等[136]系统评述了振动信号分析、声发射、红外热成像及油液分析等农机故障诊断技术。智能技术在农机诊断中的实践涵盖了人工智能、机器学习、大数据处理与物联网。沈伟杰等[137]构建了融合连续小波变换、卷积神经网络及支持向量机的诊断方法,此乃数据驱动农机诊断的代表性方案。邓文乾等[138]针对多机器人协同作业中的网络节点故障,提出协同检测机制,通过系统输入输出重新规划剩余机器人的任务负载,并调整作业路线或目标,实现农机故障的自愈功能,初步体现了“本体感知”特性。WANG 等[139]则考虑复杂作业环境,提出融合振动信号与共振系统数模的算法以解决故障检测难题。然而,这些诊断方法的局限性在于,其故障诊断性能取决于部署在农机上的训练网络,当诊断网络越复杂、诊断类型越丰富则农机作业才能得到较好的保证。反之,当作业环境和作业对象发生变化,部署在农机上网络将面临失效。

因此,仅仅诊断农机而不实现故障自主自愈,无法保障农机长期且高效的作业能力,无法满足高质量作业结果。而故障自愈需要对环境理解、自身认知有较为深入的探索,即达到诊断和容错的协同响应。可以说,农业机器人的作业状态具身诊断技术通过多模态感知与动态建模实现故障自愈与能源优化,构建“监测-诊断-修复”闭环系统。

8.2 农业机器人作业质量具身评估

高质量的农机作业来源于准确地感知、精确地决策和有效地执行,无论是农机传感器或是执行器发生故障都无法保证其作业质量。因此,作业质量和农机自身故障密不可分。作业质量的监测标准有作业中的含杂率、损失率、产量动态计量等,这些参数有助于对农机作业结果和收获质量形成有效评估,及时发现并解决问题。

与农机自身故障诊断不同,作业质量检测与评估的研究与应用多基于数据驱动方法,这包括基于经典评价指标的单机离身评估以及基于大数据和物联网的多机互联评估。王培等[140]选用白色光源色标传感器与高清网络摄像机作为核心检测元件,依托 Vision Assistant 视觉软件平台,开发了作业质量监测系统。赵博等[141]阐释了大数据核心内涵及其在农机作业质量管理中的独特优势。农机作业评价体系可涵盖作业效率、油耗、作业质量、重漏作业率、有效作业时长占比等指标;徐旻等[142]采用高精度 RTK GPS 接收机,实施单点静态定位测量。温昌凯等[143]融合农业大数据,结合 BP 神经网络与遗传算法对基础作业载荷数据进行分类挖掘,预测并评估了拖拉机田间旋耕作业质量。在基于物联网的作业评估研究上,一些研究[144-147]分别基于不同传感器结构提出了针对不同农业场景的质量评估框架。作业质量和农机故障、作业环境息息相关,多数研究往往关注农机数据本身的异常,而忽略引发异常的原因。具身思维能够结合环境和具身诊断结果构建更为人性化、贴近真实的作业质量评价结果。

综上所述,传统的仅通过作业数据驱动的质量监测方法无法满足智慧农业在复杂作业条件下的作业需求,知道作业质量差而不知道影响作业质量的原因,而作业质量的具身评估需要对环境理解、作业认知有较为深入的探索,需要实现内(农机)和外(环境)合一。可以说,作业质量评估技术需从单一指标检测转向“环境-操作-作物响应”协同反馈范式。

9 具身智能农业机器人未来展望

具身智能技术将驱动农业机器人向环境强适应、群体自组织、场景全贯通、能力强泛化、行为自进化等多维度突破。通过大模型专用化、知识图谱融合、行为技能泛化等关键技术突破,将实现农业机器人从单一功能向多任务适应的跨越式发展。未来的农业机器人将具备更强的环境感知能力、决策推理能力和协同作业能力,为实现精准农业和智慧农业提供强有力的技术与装备支撑。

9.1 大模型在农业场景的专用化

将通用大模型针对农业机器人领域进行深度定制化,融入作物生长规律、病虫害识别、土壤滑移特性等专业知识。构建农业领域预训练模型,融合作物生理学、土壤动力学等先验知识,通过农业数据的领域微调,构建理解农业语义、掌握农艺操作的专用智能模型。这种专用化大模型将为农业机器人提供更精准的决策支持和任务理解能力。

9.2 农业知识图谱与常识推理融合

探索多维多源知识融合架构,构建涵盖农作物、土壤、气候、农艺等多维度的大规模农业知识图谱,并与常识推理引擎深度融合。突破传统农业专家系统“重规则、轻因果”的局限,建立兼具领域知识深度与物理世界常识的农业认知智能基座,为农业机器人赋予“老农般的经验直觉与科学家级的精准决策”双重能力。基于农业认知智能基座与大模型推理能力的结合,使农业机器人能够在不同农业场景下自主地进行逻辑推断、决策优化和行为执行。

9.3　农业机器人未知任务技能泛化

构建具备跨作物-跨季节-跨地域泛化能力的通用农业具身智能体，采用模型无关元强化学习（Meta-RL）框架，将耕作采收等基础技能抽象为可迁移策略元模型，在仿真环境中预训练基础操作策略集（如抓取/切割/喷施），通过少量田间示范实现新任务（如未知果蔬采摘）的少样本适应。通过强化学习、迁移学习等技术，实现从已掌握技能到新任务的快速适应与能力进化。构建农业数字孪生平台，注入各类长尾数据，通过强化学习训练农业机器人的抗扰泛化能力。实际果蔬采摘作业中，机器人遇到枝蔓障碍时可自主切换避障路径规划，其决策机制源于仿真环境中数千次类似场景的演化经验。

9.4　农场全要素仿真与机器人训练平台

构建与真实农场几何形态、物理特性及生物过程高度一致的高保真数字孪生平台，实现“作物生长-环境演化-机器作业”三大核心要素的高精度仿真：基于光合作用-养分传输耦合模型动态模拟作物全周期表型变化，集成多源数据重构温湿度梯度、风雨侵蚀及病虫害传播等环境动力学过程，并依托刚体力学引擎精准计算农机-土壤-作物交互力反馈。通过虚实闭环决策验证机制，为农业机器人提供安全高效的训练基座，利用强化学习批量训练抗干扰策略，减少物理样机技术迭代。推动研发范式从“自然条件试错”向“虚拟迭代-精准部署”跃迁，为精准农业提供持续进化的决策沙盒。

9.5　农业机器人多机协同作业机制

面向大规模农田作业的复杂需求，多机协同机制需突破传统集中式调度的刚性框架，向动态自组织、资源弹性适配、人机共生决策的三维范式演进。探索群体智能驱动的分布式架构：基于改进型拍卖算法（CBBA）与分布式模型预测控制（DMPC），构建去中心化任务-路径耦合优化模型，突破农业场景中机器人时空竞争；面对农业机器人退出或加入，采用共识制群体重组机制实现快速任务再分配。研究具身智能赋能的协同认知升维：多机雷达点云实时拼接生成厘米级障碍物共享地图（协同规避倒伏作物区），将农艺规则转化为混合整数规划约束，并借联邦学习实现跨机技能传承，促进农业机器人从单机自动化到群体智能化的质变，标志着农业真正步入自主协同的智能时代。

9.6　空地跨域农业机器人群智协同

针对开放农业场景下空地跨域机器人集群系统强干扰感知环境、弱通讯连通约束、异构动态拓扑和混杂变结构系统动力学特征，可分别从感知协同、通讯协同、规划协同和系统协同 4 个层面，来探索智慧农场跨域无人系统“感知-通信-规划-控制”闭环的协作和交互能力机制和演化规律，提出跨域开放作业环境下智慧农场无人系统群智协同的理论方法，搭建跨域多约束无人系统集群系统分布式协同控制理论分析框架，突破多模态感知、多层级通讯、多维度规划和多目标协同控制等共性关键技术，为建立非结构开放作业环境下精准、高效、灵巧集群协同无人系统的技术体系和安全理论奠定基础。

本文结合具身智能技术演进与农业机器人需求，全面探讨了具身智能在农业机器人领域的应用，从感知、决策、模拟、进化与诊断等技术维度进行了系统分析。提出了一种基于具身智能的农业机器人技术框架，涵盖基础模型驱动的感知技术、多模态信息融合的决策系统、虚拟仿真与虚实迁移技术、持续学习与进化机制以及智能诊断与评估体系。深入讨论了各类核心技术的原理、挑战与应用前景，重点分析了如何通过具身智能提升农业机器人的自主决策和学习进化能力。最后，展望了具身智能农业机器人未来技术发展方向，为未来农业机器人的智能化升级提供了理论指导和技术参考。

参　考　文　献

[1] SUN F, CHEN R, JI T, et al. A comprehensive survey on embodied intelligence: advancements, challenges, and future perspectives[J]. CAAI Artificial Intelligence Research, 2024, 3:9150042.

[2] AHMAD A, BABAR M A. Software architectures for robotic systems: a systematic mapping study[J]. Journal of Systems and Software, 2016, 122: 16-39.

[3] YU J, SADJADPOUR T, O'NEILL A, et al. MANIP: a modular architecture for integrating interactive perception for robot manipulation[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 1283-1289.

[4] PRADOS C, HERNANDO M, GAMBAO E, et al. A review and evaluation of control architectures for modular legged and climbing robots[J]. Biomimetics, 2024, 9(6): 319.

[5] CHEN W, CHI W, JI S, et al. A survey of autonomous robots and multi-robot navigation: perception, planning and collaboration[J]. Biomimetic Intelligence and Robotics, 2025, 5(2): 100203.

[6] IBARZ J, TAN J, FINN C, et al. How to train your robot with deep reinforcement learning: lessons we have learned[J]. The International Journal of Robotics Research, 2021, 40(4-5): 698-721.

[7] SONG Y, SUN P, LIU H, et al. Scene-driven multimodal knowledge graph construction for embodied AI[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6962 - 6976.
[8] CHIGNOLI M, KIM D, STANGER-JONES E, et al. The MIT humanoid robot: design, motion planning, and control for acrobatic behaviors[C]// 2020 IEEE - RAS 20th International Conference on Humanoid Robots (Humanoids). IEEE, 2021: 1 - 8.
[9] KATAYAMA S, MUROOKA M, TAZAKI Y. Model predictive control of legged and humanoid robots: models and algorithms [J]. Advanced Robotics, 2023, 37(5): 298 - 315.
[10] 白辰甲, 许华哲, 李学龙. 大模型驱动的具身智能: 发展与挑战[J]. 中国科学: 信息科学, 2024, 54(9): 2035 - 2082.
BAI Chenjia, XU Huazhe, LI Xuelong. Embodied-AI with large models: research and challenges[J]. Scientia Sinica Informationis, 2024, 54(9): 2035 - 2082. (in Chinese)
[11] 王文晟, 谭宁, 黄凯, 等. 基于大模型的具身智能系统综述[J]. 自动化学报, 2025, 51(1): 1 - 19.
WANG Wensheng, TAN Ning, HUANG Kai, et al. Embodied intelligence systems based on large models: a survey[J]. Acta Automatica Sinica, 2025, 51(1): 1 - 19. (in Chinese)
[12] 沈甜雨, 陶子锐, 王亚东, 等. 具身智能研究的关键问题: 自主感知, 行动与进化[J]. 自动化学报, 2025, 51(1): 43 - 71.
SHEN Tianyu, TAO Zirui, WANG Yadong, et al. Key problems of embodied intelligence research: autonomous perception, action, and evolution[J]. Acta Automatica Sinica, 2025, 51(1): 43 - 71. (in Chinese)
[13] 刘华平, 郭迪, 孙富春, 等. 基于形态的具身智能研究: 历史回顾与前沿进展[J]. 自动化学报, 2023, 49(6): 1131 - 1154.
LIU Huaping, GUO Di, SUN Fuchun, et al. Morphology-based embodied intelligence: historical retrospect and research progress[J]. Acta Automatica Sinica, 2023, 49(6): 1131 - 1154. (in Chinese)
[14] LIU H, GUO D, CANGELOSI A. Embodied intelligence: a synergy of morphology, action, perception and learning[J]. ACM Computing Surveys, 2025, 57(7): 1 - 36.
[15] ZHANG C, ZHANG C, XU Z, et al. Embodied intelligent industrial robotics: concepts and techniques[J]. arXiv Preprint, arXiv:2505.09305, 2025.
[16] LIU Y, CHEN W, BAI Y, et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI[J]. arXiv Preprint, arXiv:2407.06886, 2024.
[17] 刘成良,贡亮,苑进,等.农业机器人关键技术研究现状与发展趋势[J].农业机械学报,2022,53(7):1 - 22,55.
LIU Chengliang, GONG Liang, YUAN Jin, et al. Current status and development trends of agricultural robots [J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(7):1 - 22,55. (in Chinese)
[18] 农业农村部. 全国智慧农业行动计划(2024—2028 年)[EB/OL]. [2025 - 07 - 10]. https://www.gov.cn/zhengce/zhengcekv/202410/content_6983057.htm.
[19] 赵春江,范贝贝,李瑾,等.农业机器人技术进展、挑战与趋势[J].智慧农业(中英文),2023,5(4):1 - 15.
ZHAO Chunjiang, FAN Beibei, LI Jin, et al. Agricultural robots: technology progress, challenges and trends[J]. Smart Agriculture, 2023, 5(4): 1 - 15. (in Chinese)
[20] TURING A M. Computing machinery and intelligence[M]. Springer Netherlands, 2009.
[21] BROOKS R A. New approaches to robotics[J]. Science, 1991, 253(5025): 1227 - 1232.
[22] BROOKS R A. Intelligence without representation[J]. Artificial Intelligence, 1991, 47(1 - 3): 139 - 159.
[23] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097 - 1105.
[24] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484 - 489.
[25] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. (2013-02-20). http://cdn.openai.com/research-covers/Language-unsuper-vised/language_understanding_paper.pdf.
[26] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt - 4 technical report[J]. arXiv Preprint, arXiv:2303.08774, 2023.
[27] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PmLR, 2021: 8748 - 8763.
[28] OQUAB M, DARCET T, MOUTAKANNI T, et al. Dinov2: learning robust visual features without supervision[J]. arXiv Preprint, arXiv:2304.07193, 2023.
[29] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4015 - 4026.
[30] RAVI N, GABEUR V, HU Y T, et al. Sam 2: segment anything in images and videos[J]. arXiv Preprint, arXiv:2408.00714, 2024.
[31] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[J]. arXiv Preprint, arXiv:2307.09288, 2023.
[32] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context

[J]. arXiv Preprint, arXiv:2403.05530, 2024.
[33] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report[J]. arXiv Preprint, arXiv:2412.19437, 2024.
[34] LI J, LI D, XIONG C, et al. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888-12900.
[35] LI J, LI D, SAVARESE S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning. PMLR, 2023: 19730-19742.
[36] YANG Z, LI L, LIN K, et al. The dawn of lmms: preliminary explorations with GPT-4v (ision)[J]. arXiv Preprint, arXiv: 2309.17421, 2023.
[37] AHN M, BROHAN A, BROWN N, et al. Do as I can, not as I say: grounding language in robotic affordances[J]. arXiv Preprint, arXiv:2204.01691, 2022.
[38] BROHAN A, BROWN N, CARBAJAL J, et al. Rt-1: robotics transformer for real-world control at scale[J]. arXiv Preprint, arXiv:2212.06817, 2022.
[39] BROHAN A, BROWN N, CARBAJAL J, et al. Rt-2: vision-language-action models transfer web knowledge to robotic control[J]. arXiv Preprint, arXiv:2307.15818, 2023.
[40] DRIESS D, XIA F, SAJJADI M S M, et al. Palm-e: an embodied multimodal language model[C]//ICML'23: Proceedings of the 40th International Conference on Machine Learning, 2023: 8469-8488.
[41] LIU S, WU L, LI B, et al. Rdt-1b: a diffusion foundation model for bimanual manipulation[J]. arXiv Preprint, arXiv: 2410.07864, 2024.
[42] BLACK K, BROWN N, DRIESS D, et al. π0: a vision-language-action flow model for general robot control[J/OL]. Physical Intelligence, 2024. https://arxiv.org/abs/2410.24164.
[43] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[J]. arXiv Preprint, arXiv:2108.07258, 2021.
[44] LAI J, LIANG Y, KUANG Y, et al. IO-YOLOv5: improved pig detection under various illuminations and heavy occlusion [J]. Agriculture, 2023, 13(7):1349.
[45] RUSSELLO H, TOL R, KOOTSTRA G. T-leap: occlusion-robust pose estimation of walking cows using temporal information [J]. Computers and Electronics in Agriculture, 2022, 192: 106559.
[46] NELSON M E, MACIVER M A. Sensory acquisition in active sensing systems[J]. Journal of Comparative Physiology A, 2006, 192(6): 573-586.
[47] GAO F, FU L, ZHANG X, et al. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN[J]. Computers and Electronics in Agriculture, 2020, 176, 105634.
[48] WU L, MA J, ZHAO Y, et al. Apple detection in complex scene using the improved YOLOv4 model[J]. Agronomy, 2021, 11(3): 476.
[49] ZHENG J, WANG X, SHI Y, et al. Keypoint detection and diameter estimation of cabbage (*Brassica oleracea* L.) heads under varying occlusion degrees via YOLOv8n-CK network[J]. Computers and Electronics in Agriculture, 2024, 226: 109428.
[50] ÍÑIGUEZ R, GUTIERREZ S, POBLETE-ECHEVERRÍA C, et al. Deep learning modelling for non-invasive grape bunch detection under diverse occlusion conditions[J]. Computers and Electronics in Agriculture, 2024, 226: 109421.
[51] LEHNERT C, TSAI D, ERIKSSON A, et al. 3D move to see: multi-perspective visual servoing towards the next best view within unstructured and occluded environments[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 3890-3897.
[52] ZAPOTEZNY-ANDERSON P, LEHNERT C. Towards active robotic vision in agriculture: a deep learning approach to visual servoing in occluded and unstructured protected cropping environments[J]. IFAC-PapersOnLine, 2019, 52(30): 120-125.
[53] MAGISTRI F, PAN Y, BARTELS J, et al. Improving robotic fruit harvesting within cluttered environments through 3D shape completion[J]. IEEE Robotics and Automation Letters, 2024, 9(8): 7357-7364.
[54] SUN T, ZHANG W, MIAO Z, et al. Object localization methodology in occluded agricultural environments through deep learning and active sensing[J]. Computers and Electronics in Agriculture, 2023, 212: 108141.
[55] SUN T, ZHANG W, GAO X, et al. Efficient occlusion avoidance based on active deep sensing for harvesting robots[J]. Computers and Electronics in Agriculture, 2024, 225: 109360.
[56] LUO L, LIU B, CHEN M, et al. DRL-enhanced 3D detection of occluded stems for robotic grape harvesting[J]. Computers and Electronics in Agriculture, 2025, 229: 109736.
[57] 徐胜勇,李磊,童辉,等.基于 RGB-D 相机的黄瓜苗 3D 表型高通量测量系统研究[J].农业机械学报, 2023, 54(7): 204-213,281.
XU Shengyong, LI Lei, TONG Hui, et al. High-throughput measurement system for 3D phenotype of cucumber seedlings using RGB-D camera[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(7):204-213,281. (in Chinese)
[58] LAI J, LIANG Y, KUANG Y, et al. IO-YOLOv5: improved pig detection under various illuminations and heavy occlusion

[J]. Agriculture, 2023, 13(7): 1349.
[59] 贾明达,杨金明,孟维亮,等.融合点云与图像的环境目标检测研究进展[J].中国图象图形学报,2024,29(6):1765 - 1784.
JIA Mingda, YANG Jinming, MENG Weiliang, et al. Survey on the fusion of point clouds and images for environmental object detection[J]. Journal of Image and Graphics, 2024,29(6):1765 - 1784. (in Chinese)
[60] VORA S, LANG A H, HELOU B, et al. Pointpainting: sequential fusion for 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 4604 - 4612.
[61] HUANG T, LIU Z, CHEN X, et al. Epnet: enhancing point features with image semantics for 3d object detection[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2020: 35 - 52.
[62] LIANG M, YANG B, WANG S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2018: 641 - 656.
[63] LIU Z, HUANG T, LI B, et al. EPNet + +: cascade bi-directional fusion for multi-modal 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(7): 8324 - 8341.
[64] WANG C, MA C, ZHU M, et al. Pointaugmenting: cross-modal augmentation for 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11794 - 11803.
[65] QI C R, LIU W, WU C, et al. Frustum pointnets for 3D object detection from rgb-d data[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 918 - 927.
[66] PANG S, MORRIS D, RADHA H. CLOCs: camera - LiDAR object candidates fusion for 3D object detection[C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10386 - 10393.
[67] 宁大海,郑晟.可见光和红外图像决策级融合目标检测算法[J].红外技术,2023,45(3):282 - 291.
NING Dahai, ZHENG Sheng. An object detection algorithm based on decision-level fusion of visible and infrared images[J]. Infrared Technology , 2023, 45(3): 282 - 291. (in Chinese)
[68] RONECKER M P, ZHU Y. Deep Q-network based decision making for autonomous driving[C]// 2019 3rd International Conference on Robotics and Automation Sciences (ICRAS). IEEE, 2019: 154 - 160.
[69] PENG B, SUN Q, LI S E, et al. End-to-end autonomous driving through dueling double deep Q-network[J]. Automotive Innovation, 2021, 4(3): 328 - 337.
[70] CUI Y, DONG X, LI D, et al. An end-to-end deep reinforcement learning method for UAV autonomous motion planning[C]// 2022 7th International Conference on Robotics and Automation Engineering (ICRAE). IEEE, 2022: 100 - 104.
[71] LI Y, GAO J, CHEN Y, et al. Deep reinforcement learning-based end-to-end navigation of mobile robots with reward shaping [C]//2024 IEEE 22nd International Conference on Industrial Informatics (INDIN). IEEE, 2024: 1 - 6.
[72] HUANG P, ZHU L, ZHANG Z, et al. An end-to-end learning-based row-following system for an agricultural robot in structured apple orchards[J]. Mathematical Problems in Engineering, 2021, 2021(1): 6221119.
[73] LIU B, JIANG Y, ZHANG X, et al. LLM + P: empowering large language models with optimal planning proficiency[J]. arXiv Preprint, arXiv:2304.11477, 2023.
[74] CHEN Y, CUI W, CHEN Y, et al. Robogpt: an intelligent agent of making embodied long-term decisions for daily instruction tasks[J]. arXiv Preprint, arXiv:2311.15649, 2023.
[75] HRIDHAR M, THOMASON J, GORDON D, et al. Alfred: a benchmark for interpreting grounded instructions for everyday tasks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10740 - 10749.
[76] LIU Y, PALMIERI L, KOCH S, et al. Delta: decomposed efficient long-term robot task planning using large language models [J]. arXiv Preprint, arXiv:2404.03275, 2024.
[77] SONG C H, WU J, WASHINGTON C, et al. Llm-planner: few-shot grounded planning for embodied agents with large language models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 2998 - 3009.
[78] LIN K, AGIA C, MIGIMATSU T, et al. Text2motion: from natural language instructions to feasible plans[J]. Autonomous Robots, 2023, 47(8):1345 - 1365.
[79] RANA K, HAVILAND J, GARG S, et al. Sayplan: grounding large language models using 3d scene graphs for scalable robot task planning[J]. arXiv Preprint, arXiv:2307.06135, 2023.
[80] CHEN Y, ARKIN J, DAWSON C, et al. Autotamp: autoregressive task and motion planning with llms as translators and checkers[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6695 - 6702.
[81] MAO Q, LIAO Z, YUAN J, et al. Multimodal tactile sensing fused with vision for dexterous robotic housekeeping[J]. Nature Communications, 2024, 15(1): 6871.
[82] YUAN W, DONG S, ADELSON E H. Gelsight: high-resolution robot tactile sensors for estimating geometry and force[J]. Sensors, 2017, 17(12): 2762.
[83] JIANG D, LI G, SUN Y, et al. Manipulator grabbing position detection with information fusion of color image and depth image using deep learning[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12: 10809 - 10822.
[84] 吴培良,李瑶,牛明月,等.一种视/触觉融合的柔性物体抓取力估计方法[J].机器人,2024,46(5):544 - 553.
WU Peiliang, LI Yao, NIU Mingyue, et al. A visual-tactile fusion method for estimating the grasping force on flexible objects

[J]. ROBOT, 2024, 46(5): 544 – 553. (in Chinese)
[85] LI S, YU H, DING W, et al. Visual-tactile fusion for transparent object grasping in complex backgrounds[J]. IEEE Transactions on Robotics, 2023, 39(5): 3838 – 3856.
[86] LI H, ZHANG Y, ZHU J, et al. See, hear, and feel: smart sensory fusion for robotic manipulation[J]. arXiv Preprint, arXiv:2212.03858, 2022.
[87] YUAN Y, CHE H, QIN Y, et al. Robot synesthesia: in-hand manipulation with visuotactile sensing[C]// 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6558 – 6565.
[88] GUO H, MIAO Z, JI J C, et al. An effective collaboration evolutionary algorithm for multi-robot task allocation and scheduling in a smart farm[J]. Knowledge-Based Systems, 2024, 289: 111474.
[89] GUO H, SANG H, ZHANG B, et al. An effective metaheuristic with a differential flight strategy for the distributed permutation flowshop scheduling problem with sequence-dependent setup times[J]. Knowledge-Based Systems, 2022, 242: 108328.
[90] MIAO Z, GUO H, PAN Q, et al. A reinforcement learning-enhanced multi-objective iterated greedy algorithm for weeding-robot operation scheduling problems[J]. Expert Systems with Applications, 2025, 263: 125760.
[91] DAI L L, PAN Q K, MIAO Z H, et al. Multi-objective multi-picking-robot task allocation: mathematical model and discrete artificial bee colony algorithm[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 25(6): 6061 – 6073.
[92] 王猛, 赵博, 刘阳春, 等. 同种农机机群动态作业任务分配方法[J]. 农业工程学报, 2021, 37(9): 199 – 210.
WANG Meng, ZHAO Bo, LIU Yangchun, et al. Dynamic task allocation method for the same type agricultural machinery group[J]. Transactions of the CSAE, 2021, 37(9): 199 – 210. (in Chinese)
[93] VÁZQUEZ-ARELLANO M, GRIEPENTROG H W, REISER D, et al. 3-D imaging systems for agricultural applications—a review[J]. Sensors, 2016, 16(5): 618.
[94] WANG C, GLENN N F. Integrating LiDAR intensity and elevation data for terrain characterization in a forested area[J]. IEEE Geoscience and Remote Sensing Letters, 2009, 6(3): 463 – 466.
[95] BARROWS C, BULANON D M. Development of a low-cost multispectral camera for aerial crop monitoring[J]. Journal of Unmanned Vehicle Systems, 2017, 5(4): 192 – 200.
[96] HIRAOKA N, SUDA T, HIRAI K, et al. Improved measurement of soil moisture and groundwater level using ultrasonic waves [J]. Japanese Journal of Applied Physics, 2011, 50(7S): 07HC19.
[97] TAO Y, ZHOU J. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking[J]. Computers and Electronics in Agriculture, 2017, 142: 388 – 396.
[98] LI Z, YUAN X, WANG C. A review on structural development and recognition-localization methods for end-effector of fruit-vegetable picking robots[J]. International Journal of Advanced Robotic Systems, 2022, 19(3): 17298806221104906.
[99] XIAO X, WANG Y, ZHOU B, et al. Flexible hand claw picking method for citrus-picking robot based on target fruit recognition[J]. Agriculture, 2024, 14(8): 1227.
[100] SHAMMI S A, MENG Q. Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling[J]. Ecological Indicators, 2021, 121: 107124.
[101] SHAH D U, REYNOLDS T P S, RAMAGE M H. The strength of plants: theory and experimental methods to measure the mechanical properties of stems[J]. Journal of Experimental Botany, 2017, 68(16): 4497 – 4516.
[102] WANG P, TANG Y, LUO F, et al. Weed25: a deep learning dataset for weed identification[J]. Frontiers in Plant Science, 2022, 13: 1053329.
[103] LI S, XU H, JI Y, et al. Development of a following agricultural machinery automatic navigation system[J]. Computers and Electronics in Agriculture, 2019, 158: 335 – 344.
[104] HOSSAIN M S, SALAM A. Text-to-3D scene generation using semantic parsing and spatial knowledge with rule based system [J]. Int. J. Comp. Sci. Issues (IJCSI), 2017, 14: 37 – 41.
[105] CAI J, DENG W, GUANG H, et al. A survey on data-driven scenario generation for automated vehicle testing[J]. Machines, 2022, 10(11): 1101.
[106] FRIDMAN R, ABECASIS A, KASTEN Y, et al. Scenescape: text-driven consistent scene generation[J]. Advances in Neural Information Processing Systems, 2023, 36: 39897 – 39914.
[107] YIN J, SHEN J, CHEN R, et al. Is-fusion: instance-scene collaborative fusion for multimodal 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 14905 – 14915.
[108] TIBONI G, ARNDT K, KYRKI V. DROPO: Sim-to-real transfer with offline domain randomization[J]. Robotics and Autonomous Systems, 2023, 166: 104432.
[109] HO D, RAO K, XU Z, et al. Retinagan: an object-aware approach to sim-to-real transfer[C]// 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 10920 – 10926.
[110] 王纪章,王涛,付益辉,等.数字孪生技术在农业生产管控中的应用现状与展望[J].农业机械学报,2024,55(9):1 – 20.
WANG Jizhang, WANG Tao, FU Yihui, et al. Application status and development prospects of digital twin technology in agricultural production process control[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(9): 1 – 20. (in Chinese)

[111] CHANG A, DAI A, FUNKHOUSER T, et al. Matterport3d: learning from rgb-d data in indoor environments[J]. arXiv Preprint, arXiv:1709.06158, 2017.
[112] SAVVA M, KADIAN A, MAKSYMETS O, et al. Habitat: a platform for embodied ai research[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9339 - 9347.
[113] MAKOVIYCHUK V, WAWRZYNIAK L, GUO Y, et al. Isaac gym: high performance GPU-based physics simulation for robot learning[J]. arXiv Preprint, arXiv:2108.10470, 2021.
[114] REID M, YAMADA Y, GU S S. Can wikipedia help offline reinforcement learning? [J]. arXiv Preprint, arXiv:2201.12122, 2022.
[115] REED S, ZOLNA K, PARISOTTO E, et al. A generalist agent[J]. arXiv Preprint, arXiv:2205.06175, 2022.
[116] BOUSMALIS K, VEZZANI G, RAO D, et al. Robocat: a self-improving generalist agent for robotic manipulation[J]. arXiv Preprint, arXiv:2306.11706,2023.
[117] XIAO T, RADOSAVOVIC I, DARRELL T, et al. Masked visual pre-training for motor control[J]. arXiv Preprint, arXiv:2203.06173, 2022.
[118] SUN Y, MA S, MADAAN R, et al. Smart: self-supervised multi-task pretraining with control transformers[C]// International Conference on Learning Representations, 2023: 1 - 21.
[119] LIU F, LIU H, GROVER A, et al. Masked autoencoding for scalable and generalizable decision making[J]. Advances in Neural Information Processing Systems, 2022, 35: 12608 - 12618.
[120] KARAMCHETI S, NAIR S, CHEN A S, et al. Language-driven representation learning for robotics[J]. arXiv Preprint, arXiv:2302.12766, 2023.
[121] RADOSAVOVIC I, SHI B, FU L, et al. Robot learning with sensorimotor pre-training[C]// Conference on Robot Learning. PMLR, 2023: 683 - 693.
[122] NUNES L, MARCUZZI R, CHEN X, et al. Segcontrast: 3D point cloud feature representation learning through self-supervised segment discrimination[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2116 - 2123.
[123] LU J, BATRA D, PARIKH D, et al. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]// Advances in Neural Information Processing Systems, 2019, 32: 13 - 23.
[124] CHEN Y C, LI L, YU L, et al. Uniter: universal image-text representation learning[C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 104 - 120.
[125] HONG Y, WU Q, QI Y, et al. Vln bert: a recurrent vision-and-language bert for navigation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021: 1643 - 1653.
[126] NAIR S, RAJESWARAN A, KUMAR V, et al. R3M: a universal visual representation for robot manipulation[C]// Proceedings of The 6th Conference on Robot Learning, 2022: 892 - 909.
[127] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 1998.
[128] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529 - 533.
[129] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv Preprint, arXiv:1707.06347, 2017.
[130] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv Preprint, arXiv:1509.02971, 2015.
[131] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]// International Conference on Machine Learning. PMLR, 2018: 1861 - 1870.
[132] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]// International Conference on Machine Learning. PMLR, 2018: 1587 - 1596.
[133] HWANGBO J, LEE J, DOSOVITSKIY A, et al. Learning agile and dynamic motor skills for legged robots[J]. Science Robotics, 2019, 4(26): eaau5872.
[134] 马楠,曹姗姗,白涛,等.农业复杂场景下多机器人协同 SLAM 研究进展与展望[J].智慧农业(中英文),2024,6(6):23 - 43.
MA Nan, CAO Shanshan, BAI Tao, et al. Research progress and prospect of multi-robot collaborative SLAM in complex agricultural scenarios[J]. Smart Agriculture, 2024, 6(6): 23 - 43. (in Chinese)
[135] 周俊,谢文松,伍星,等.基于时频脊线提取与改进稀疏分量分析的 RV 减速器复合故障盲分离[J].农业工程学报,2023,39(22):18 - 26.
ZHOU Jun, XIE Wensong, WU Xing, et al. Blind fault separation of RV reducers using time-frequency ridge extraction and improved sparse component analysis[J]. Transactions of the CSAE, 2023, 39(22): 18 - 26. (in Chinese)
[136] 翟长远,杨硕,王秀,等.农机装备智能测控技术研究现状与展望[J].农业机械学报,2022,53(4):1 - 20.
ZHAI Changyuan, YANG Shuo, WANG Xiu, et al. Status and prospect of intelligent measurement and control technology for agricultural equipment[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(4): 1 - 20. (in Chinese)

[137] 沈伟杰,肖茂华,宋新民,等. 基于连续小波变换的 CNN—SVM 农机滚动轴承故障诊断[J]. 中国农机化学报,2025,46(4):254 - 264.
SHEN Weijie, XIAO Maohua, SONG Xinmin, et al. Fault diagnosis of agricultural machinery rolling bearing based on CNN—SVM by continuous wavelet transform[J]. Journal of Chinese Agricultural Mechanization, 2025, 46(4): 254 - 264. (in Chinese)

[138] 邓文乾,赖颖杰,张世昂,等. 面向农田环境的农业多机器人协同技术研究进展[J]. 中国农机化学报,2024,45(10):289 - 297.
DENG Wenqian, LAI Yingjie, ZHANG Shi'ang, et al. Research progress of agricultural multi-robot collaboration technology for farmland environment[J]. Journal of Chinese Agricultural Mechanization, 2024, 45(10): 289 - 297. (in Chinese)

[139] WANG S, LU B. Detecting the weak damped oscillation signal in the agricultural machinery working environment by vibrational resonance in the duffing system[J]. Journal of Mechanical Science and Technology, 2022, 36(12): 5925 - 5937.

[140] 王培,张安琪,安晓飞,等. 深松作业质量多维度综合评价方法研究[J]. 农业机械学报,2025,56(4):81 - 86,97.
WANG Pei, ZHANG Anqi, AN Xiaofei, et al. Research on multi-dimensional comprehensive evaluation method of subsoiling operation quality[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(4): 81 - 86, 97. (in Chinese)

[141] 赵博,张巍朋,苑严伟,等. 农业装备运维与作业服务管理信息化技术研究进展[J]. 农业机械学报,2023,54(12):1 - 26.
ZHAO Bo, ZHANG Weipeng, YUAN Yanwei, et al. Research progress in information technology for agricultural equipment maintenance and operation service management[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(12):1 - 26. (in Chinese)

[142] 徐旻,张瑞瑞,陈立平,等. 智能化无人机植保作业关键技术及研究进展[J]. 智慧农业,2019,1(2):20 - 33.
XU Min, ZHANG Ruirui, CHEN Liping, et al. Key technology analysis and research progress of UAV intelligent plant protection[J]. Smart Agriculture, 2019, 1(2): 20 - 33. (in Chinese)

[143] 温昌凯,谢斌,李若晨,等. 拖拉机作业载荷数据平台设计与旋耕作业质量预测[J]. 农业机械学报,2020,51(8):372 - 381.
WEN Changkai, XIE Bin, LI Ruochen, et al. Construction of tractor working load data platform and prediction of rotary tillage quality[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(8): 372 - 381. (in Chinese)

[144] RAYHANA R, XIAO G, LIU Z. RFID sensing technologies for smart agriculture[J]. IEEE Instrumentation & Measurement Magazine, 2021, 24(3): 50 - 60.

[145] MORCHID A, EL ALAMI R, RAEZAH A A, et al. Applications of internet of things (IoT) and sensors technology to increase food security and agricultural sustainability: benefits and challenges[J]. Ain Shams Engineering Journal, 2024, 15(3): 102509.

[146] ZENG H, DHIMAN G, SHARMA A, et al. An IoT and blockchain-based approach for the smart water management system in agriculture[J]. Expert Systems, 2023, 40(4): e12892.

[147] ZHU Z, BAI Y, DAI W, et al. Quality of e-commerce agricultural products and the safety of the ecological environment of the origin based on 5G Internet of Things technology[J]. Environmental Technology & Innovation, 2021, 22: 101462.