软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(4):1715−1757 [doi: 10.13328/j.cnki.jos.007250] [CSTR: 32375.14.jos.007250]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向具身人工智能的物体目标导航综述*

陈铂垒, 康嘉绪, 钟 萍, 崔永正, 卢思怡, 杨昊楠, 王建新

(中南大学 计算机学院, 湖南 长沙 410083)

通信作者: 钟萍, E-mail: ping.zhong@csu.edu.cn

摘 要: 近年来随着计算机视觉和人工智能领域的不断发展, 具身人工智能 (embodied AI) 受到国内外学术界和工业界的广泛关注. 具身人工智能强调具身智能体通过与环境进行情景化的交互来主动获取物理世界的真实反馈, 并通过对反馈进行学习使具身智能体更加智能. 作为具身人工智能具体化的任务之一, 物体目标导航要求具身智能体在事先未知的、复杂且语义丰富的场景中搜寻并导航至指定的物体目标 (例如: 找到水槽). 物体目标导航在辅助人类日常活动的智能助手方面有着巨大的应用潜力, 是其他基于交互的具身智能研究的基础和前置任务. 系统地分类和梳理当前物体目标导航相关工作, 首先介绍环境表示和视觉自主探索相关知识, 从 3 种不同的角度对现有的物体目标导航方法进行分类和分析, 其次介绍两类更高层次的物体重排布任务, 描述逼真的室内仿真环境数据集、评价指标和通用的导航策略训练范式, 最后比较和分析现有的物体目标导航策略在不同数据集上的性能, 总结该领域所面临的挑战, 并对发展前景作出展望.

关键词: 物体目标导航; 具身人工智能; 视觉自主探索; 视觉物体重排布

中图法分类号: TP18

中文引用格式: 陈铂垒, 康嘉绪, 钟萍, 崔永正, 卢思怡, 杨昊楠, 王建新. 面向具身人工智能的物体目标导航综述. 软件学报, 2025, 36(4): 1715–1757. http://www.jos.org.cn/1000-9825/7250.htm

英文引用格式: Chen BL, Kang JX, Zhong P, Cui YZ, Lu SY, Yang HN, Wang JX. Survey on Object Goal Navigation for Embodied AI. Ruan Jian Xue Bao/Journal of Software, 2025, 36(4): 1715–1757 (in Chinese). http://www.jos.org.cn/1000-9825/7250.htm

Survey on Object Goal Navigation for Embodied AI

CHEN Bo-Lei, KANG Jia-Xu, ZHONG Ping, CUI Yong-Zheng, LU Si-Yi, YANG Hao-Nan, WANG Jian-Xin

(School of Computer Science and Engineering, Central South University, Changsha 410083, China)

Abstract: With the continuous development of computer vision and artificial intelligence (AI) in recent years, embodied AI has received widespread attention from academia and industry at home and abroad. Embodied AI emphasizes that an agent should actively obtain real feedback from the physical world by interacting with the environment in a contextualized way and make itself more intelligent through learning from the feedback. As one of the concrete tasks of embodied AI, object goal navigation requires an agent to search for and navigate to a specified object goal (e.g., find a sink) in a previously unknown, complex, and semantically rich scenario. Object goal navigation has great potential for applications in smart assistants that support daily human activities, serving as a fundamental and antecedent task for other interaction-based embodied AI research. This study systematically classifies current research on object goal navigation. Firstly, the knowledge related to environmental representation and autonomous visual exploration is introduced, and existing object goal navigation methods are classified and analyzed from three different perspectives. Secondly, two categories of higher-level object rearrangement tasks are introduced, with a description of datasets for realistic indoor environment simulation, evaluation metrics, and a generic training paradigm for navigation strategies. Finally, the performance of existing object goal navigation strategies is compared and analyzed on different datasets. The challenges in this field are summarized, and development trends are predicted.

Key words: object goal navigation; embodied AI; autonomous visual exploration; visual object rearrangement

* 基金项目: 国家自然科学基金 (62172443); 湖南省自然科学基金 (2022JJ30760); 长沙市自然科学基金 (kq2202107, kq2202108)
收稿时间: 2023-05-29; 修改时间: 2023-10-08, 2024-05-14; 采用时间: 2024-07-15; jos 在线出版时间: 2024-11-27
CNKI 网络首发时间: 2024-11-28

过去的 10 年里, 计算机视觉、深度学习乃至广泛的人工智能领域的发展取得了长足的进步, 推动了传统的被动学习范式向主动学习的转变[1], 促进了具身人工智能 (embodied AI) 的快速发展[2,3]. 在大规模逼真场景数据集[4–9]和高性能仿真器[10,11]的加持下, 具身人工智能领域着重研究“智能是如何从与环境交互的过程中产生的”[1], 鼓励具身智能体 (embodied agent) 以交互和探索的方式主动地学习, 并创造性地解决环境中具有挑战性的问题. 作为具身人工智能具体化的任务之一, 物体目标导航 (object goal navigation, ObjectNav)[12]要求具身智能体在事先未知场景中搜寻由自然语言指定的物体, 并在有限的时间预算内导航至物体附近 (例如导航到沙发附近, 如图 1 和图 2 所示). 从应用角度看, ObjectNav 在辅助人类日常活动的智能助手方面有着巨大的应用潜力. 从研究领域看, ObjectNav 促使具身智能体从与环境的交互和反馈中学习和进步, 是基于导航的下游具身任务 (例如视觉物体重排布[13]) 的基础, 能够为后续任务做无监督的准备.

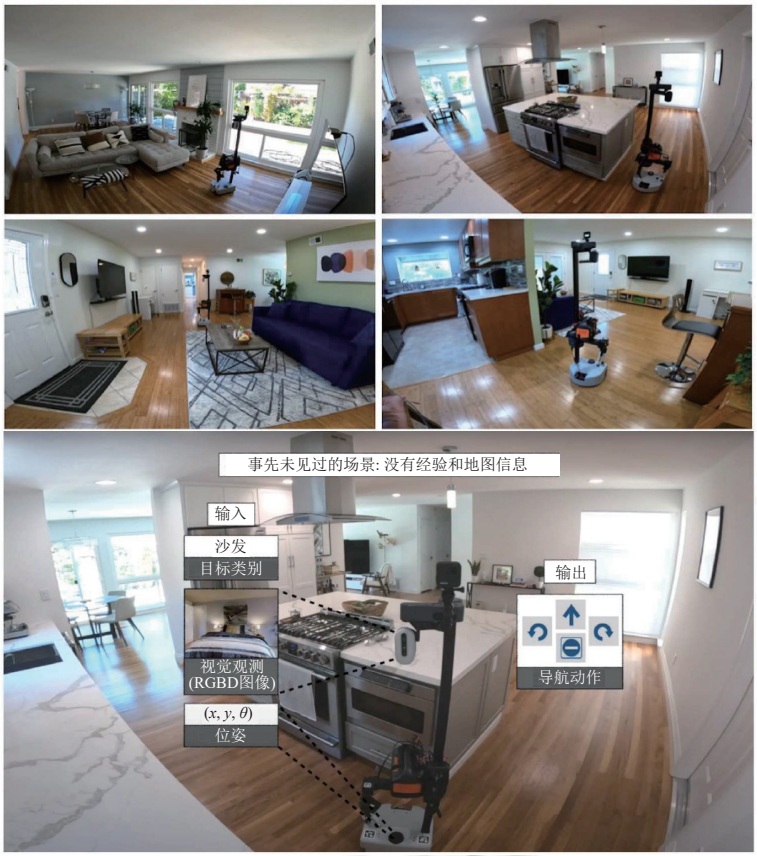


图 1　多样化的环境与 ObjectNav 任务的输入输出

在 ObjectNav 任务中[12], 具身智能体通常以目标类别、自身位姿和捕获的 RGBD 视觉观测为输入, 输出当前时刻的导航动作 (通常包括直行、左转、右转和停止), 如图 1 所示. 多样化且事先未知的环境所导致的巨大的观测空间使得 ObjectNav 任务极具挑战性. 此外, 具身智能体还需要应对视觉-语言匹配、运动控制和碰撞避免等问题. 实证研究表示[14], 人类擅长在未知的环境中执行 ObjectNav 任务, 在 MP3D 数据集上达到了 88.9% 的导航成功率. 具身智能体被期待像人类一样通过记忆环境中的空间和语义模式, 利用不同物体之间的从属或者共现 (co-occurrence) 关系, 高效地执行复杂的 ObjectNav 任务. 对于具身智能体而言, 想要高效地导航到物体目标至少需要具备 3 个方面的能力[15]: (1) 视觉-语言匹配能力. 具身智能体应当理解自然语言指令形式的物体目标, 并将指令与环境中的真实物体相匹配; (2) 复杂场景探索能力. 具身智能体应当高效地捕获有助于导航的场景布局和语义先验

信息, 避免无效的场景感知; (3) 高效的场景记忆能力. 具身智能体应当跟踪和记录已搜索过的区域, 避免重复和冗余的搜索. 从微观层面看, 具身智能体至少需要具备 5 种思维[16]: (1) 直接依赖视觉观测进行决策的思维; (2) 利用物体间的语义关系搜寻目标物体的搜索思维; (3) 高效地捕获环境布局的探索思维; (4) 根据物体的方位运动到目标位置的导航思维; (5) 运动过程中的避障思维. 图 3 根据不同的方法所强调的思维对后文介绍的物体目标导航策略进行了分类.

图 2 主流的物体目标导航示例

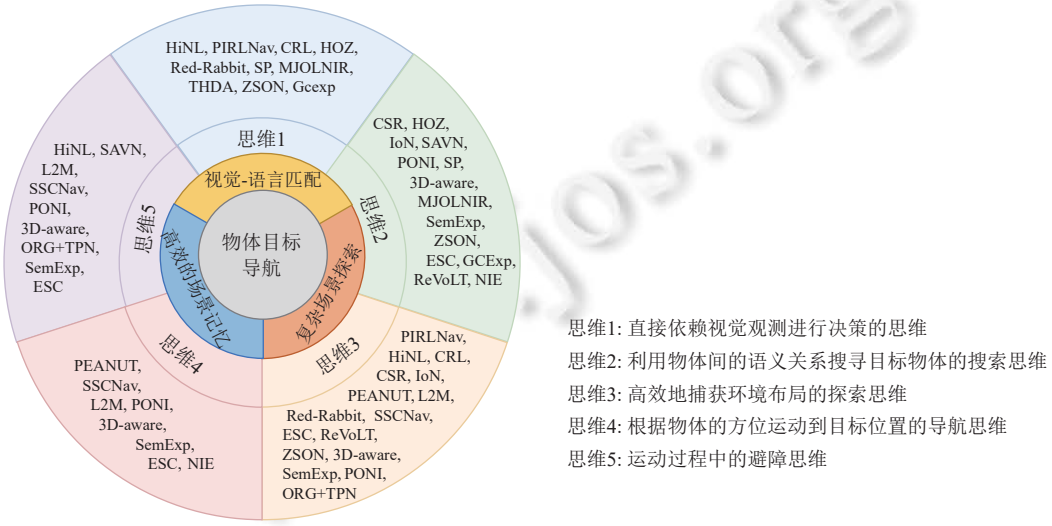


图 3 根据物体目标导航方法强调的思维分类

当前主流的 ObjectNav 策略往往是通过强调上述的一种或多种能力, 或者是通过强调上述的一种或者多种思维实现的. 现有的方法要么以端到端的方式从 RGB 视觉观测直接学习 ObjectNav 策略以应对上述挑战, 要么采用模块化的设计从 RGBD 视觉观测预测语义地图以更好地解决上述问题. 主流的方法按导航任务类型可以划分为单物体目标导航任务[17–24]和多物体目标导航任务[25–30], 按导航模型结构可以划分为端到端的导航策略[31–40]和模块化的导航策略[18–24,41,42]. 顾名思义, 单物体目标导航要求具身智能体搜寻并导航至环境中的一个物体实例, 多物体目标导航要求具身智能体有序或无序地搜寻并导航至多个不同类别的物体实例. 端到端的导航策略往往采用一个模型建模和学习多种能力或思维, 以视觉观测和定位等数据作为模型输入, 直接输出最终的导航动作. 模块化的导航策略往往依赖独立的语义建图、高层次语义探索和低层次导航等模块来分别建模各种能力或思维. 这些策略普遍采用模仿学习预训练和强化学习调优的范式进行训练, 通过行为克隆为强化学习设置一个合理的起点, 从而避免盲目地探索配置空间导致的冷启动问题[15]. 图 4 可视化了不同分类法则下物体目标导航方法所占比重.

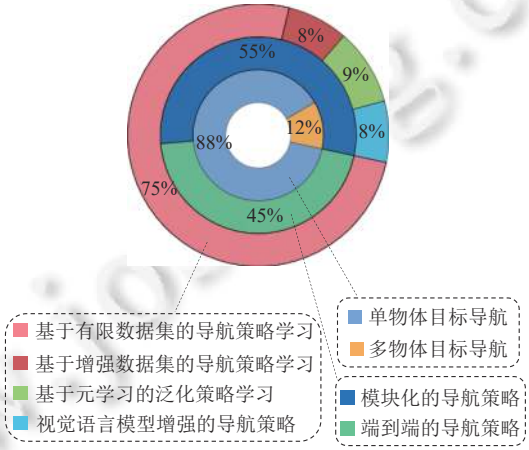


图 4 不同分类法则下物体目标导航方法所占比重

主流的方法按照发展进程可以分为如图 5 所示的 4 类. 早期, 得益于 MP3D[4]、Gibson[5]和 AI2-THOR[7]等逼真的室内场景数据集的发布, 基于有限数据集的导航策略学习方法[32,36,43–45]迅速发展起来. 新颖的视觉表示[22,23,32,34,36]和辅助任务[31]被提出用于增强导航策略的泛化能力. 此外, ObjectNav 还被解耦为“去哪里寻找目标”的高层次语义探索问题和“如何导航到目标”低层次导航问题[20,41], 并分而治之. 之后, 为了缓解有限的数据集导致的模型训练过拟合问题, 基于增强数据集的导航策略学习方法[46,47]被提出. 通过对现有数据集进行编辑和融合, 或者引入新的场景, 导航模型的性能和泛化能力被进一步提高. 最近, 考虑到现实世界中物体类别和场景的多样性, ObjectNav 策略被期望能够泛化到事先未见过的物体类别和环境, 因此基于元学习的泛化策略学习方法[39,40,48,49]被提出. 以基于梯度的元学习为基础, 具身智能体被赋予学习如何去学习的能力, 通过想象未见过的物体类别的特征和反复试错, 在测试的过程中逐渐学习和优化. 随着大规模视觉语言模型的发展[50,51], 强大的视觉语言对齐能力被用于增强 ObjectNav 的语义先验信息, 视觉语言模型增强的 ObjectNav 策略[52,53]被提出. 大规模视觉语言模型为零样本 (zero-shot) ObjectNav 的实现提供了可能, 即不需要或仅需要少量的训练就能实现对未见过的场景和物体目标的泛化.

然而, 高性能 ObjectNav 的实现并非易事, 研究显示目前最先进的 ObjectNav 策略在 MP3D 验证集上仅达到 40% 左右的成功率[41]. ObjectNav 的发展需要诸如环境表示、视觉自主探索[54]和强化学习等技术的支撑. 当然, ObjectNav 的发展也为物体重排布[13]和交互抓取等下游具身任务铺平了道路. 通过对相关领域国内外已发表的综述类文献进行调研, 我们发现现有的综述文献大都从较高的层面梳理具身人工智能领域发展现状[55,56]. 然而, 对于具身人工智能各个子领域的系统描述还有所欠缺, 诸如物体目标导航、视觉自主探索、视觉物体重排布和视觉语言导航等子领域. 虽然已有综述文献系统地梳理了视觉语言导航[3,57]和机器人自主探索[58–61]领域的关键技术和发展现状, 但是针对物体目标导航及其相关领域的总结还有待补充和完善. 本文将重点围绕物体目标导航展开介绍,

同时梳理面向具身人工智能的环境表示方法, 详细介绍和讨论视觉自主探索前置任务和物体重排布后置任务的最新进展. 图 5 所示的框架图自底向上展示了任务层次关系、方法分类和应用场景. 此外, 本文也将介绍用于 ObjectNav 策略学习和评估的数据集、评价指标和训练范式.

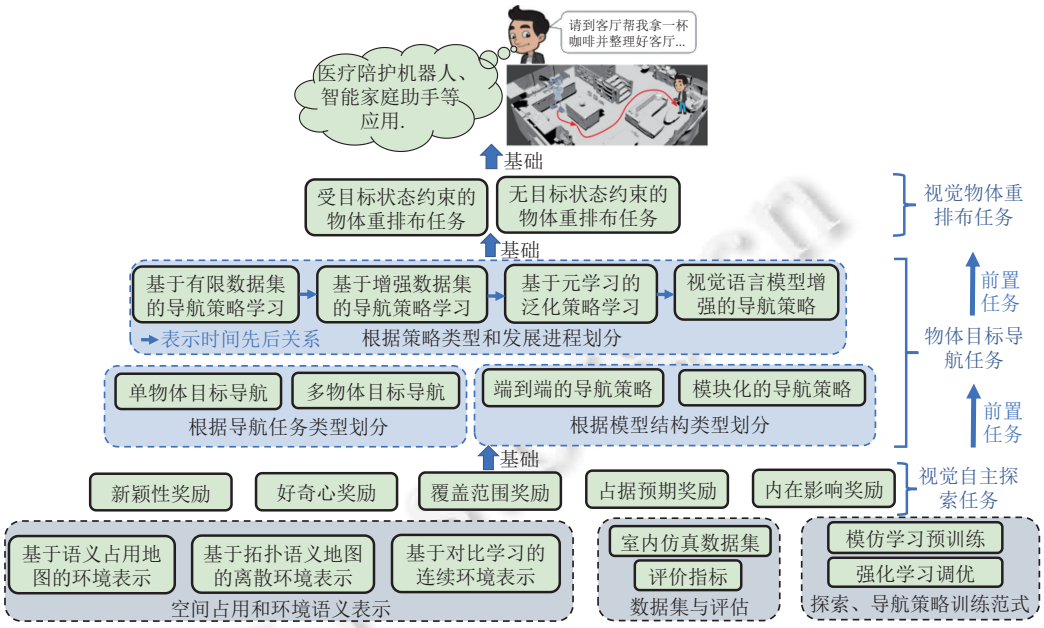


图 5 面向具身人工智能的物体目标导航综述架构图

本文第 1 节对物体目标导航的相关背景和视觉自主探索前置任务进行介绍. 第 2 节对面向具身人工智能的物体目标导航方法进行分类、分析和总结. 第 3 节对面向具身人工智能的物体重排布方法进行分析. 第 4 节对数据集、评价指标和策略训练范式进行总结. 第 5 节对现有方法的性能进行分析. 最后, 在第 6 节对面临的挑战和发展前景进行分析, 并总结全文.

1 相关背景

在本节中, 我们将介绍物体目标导航所采用的环境表示和视觉自主探索前置任务. 第 1.1 节中介绍主流的空间占用和环境语义表示方法; 第 1.2 节介绍视觉自主探索任务的定义、分类以及其与物体目标导航的关系.

1.1 空间占用和环境语义表示

视觉导航和探索通常是长序列决策任务, 运动决策不仅基于当前的视觉观测, 还依赖于过去的经验和环境记忆, 包括过去的动作、视觉观测和神经网络隐藏状态等信息. 现有的大多数方法要么采用循环神经网络隐式地编码过去的经验和记忆[32,34,36−38], 要么采用地图显式地记录环境的空间结构和语义模式[18−23]. 然而, 长短期记忆 (long and short term memory, LSTM) 网络和门控循环单元 (gate recurrent unit, GRU) 等循环神经网络被证明在捕获运动轨迹中的长期依赖方面是低效的[62,63], 不利于处理和执行长动作序列. 此外, 隐式的记忆存储要求网络模型分配大量的参数维护经验和记忆, 这不仅增加了导航策略模型训练和推理的负担, 还将导致策略学习与基本导航任务偏离. 为了缓解此类问题, 一方面, 基于 Transformer[64]的方法提出利用自注意力机制来捕获历史导航轨迹中不同时间步的长距离依赖关系[35,48,65,66], 为将来的动作预测提供时间序列特征. 另一方面, 基于地图的环境表示方法被提出显式地存储和维护环境中的低维和高维特征, 具体分为 3 类: (1) 基于语义占用地图的环境表示[67−70]; (2) 基于拓扑语义地图的离散环境表示[71−77]; (3) 基于对比学习的连续环境表示 (continuous environment representation, CER)[78,79].

如图 6 所示, 二维或三维语义地图不仅指示了空间占用情况, 还为具身智能体提供了有助于导航策略学习的空间位置信息和对应位置的语义类别信息. 首先对 RGB 图像进行语义分割, 然后结合相机的内参、外参和深度图像生成三维语义点云. 通过对三维语义点云进行拼接、融合和进一步的维护, 构建三维语义占用地图. 如图 7 所示, 将三维语义点云沿 Z 轴向 XY 坐标平面投影, 则生成二维语义占用地图[67]. 如果不考虑语义信息, 则仅生成对应的占用地图. 还有一类方法将通过视觉神经网络学习到的特征或者视觉神经网络的隐藏层状态存入占据栅格中构建语义占据地图[68−70], 此类地图被验证除了包含语义信息外还包含物体目标的轮廓和纹理特征. 基于语义占用地图的环境表示经过神经网络的处理, 能够为视觉导航和探索提供环境的空间线索、语义关系信息甚至物体的轮廓和纹理特征.

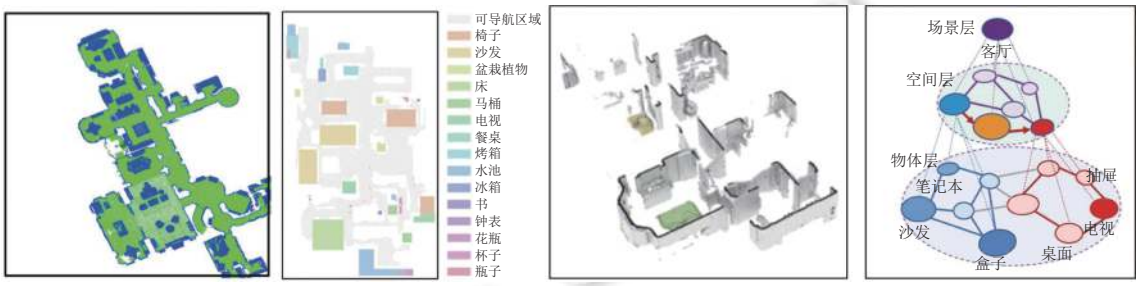


图 6 基于语义占用地图的环境表示和基于拓扑语义地图的离散环境表示

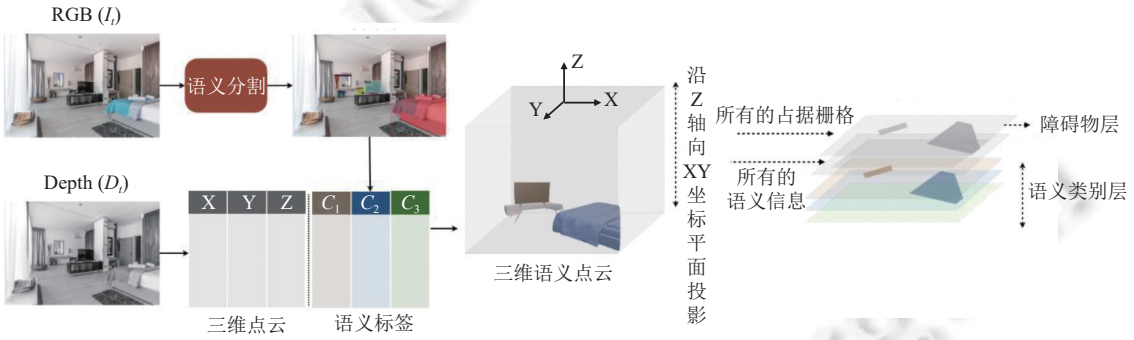


图 7 二维语义占用地图环境表示构建

拓扑语义地图属于离散的高维环境表示, 对环境中的高层次语义和关系进行表征, 相对于语义占用地图更加的精简和抽象. 拓扑语义地图的节点可以表示为物体的一组属性, 也可以表示为由神经网络编码的高维视觉图像特征. 拓扑语义地图的边可以表示为节点之间的空间和语义关系. 图 6 给出了分层次拓扑语义地图的一个示例[32], 将一个客厅场景划分为子空间, 子空间及其空间连通性构成了空间层. 对于每一个子空间中的视觉观测, 通过对视觉图像进行物体目标检测, 将物体的属性 (例如颜色) 和类别嵌入作为节点特征, 将节点之间的空间关系 (例如欧氏距离) 作为边, 构成了物体层离散的语义拓扑图. 这里的离散, 即拓扑图所包含的特征通常是人为指定的, 通常具有强烈的归纳偏差. 例如根据某些特征, “电视”在人类看来是“电视”, 但是在具身智能体看来未必是“电视”, “鼠标在书桌上”并不意味着“鼠标在书桌的边缘上”. 这些归纳偏差的存在是因为人类的生活环境是语义丰富且复杂多样的, 物体实例之间的关系无法通过人为指定的离散的关系来详尽地描述. 环境中的物体和它们之间的关系构成了一个庞大的特征和关系空间, 其中的部分特征和关系对于视觉导航和探索是非常重要的.

现有的大量的方法以静态图像为基础构建拓扑场景关系图[71−73], 这些方法没有识别和跟踪场景随时间变化的机制, 因此不适合动态导航任务. 为了缓解这一问题, 有的方法从视频中创建拓扑场景图[74,75], 能够从时间维度捕获信息. 然而, 这些方法依赖于预先录制的视频, 因此不适用于需要与环境进行交互学习的具身导航任务. 最近, 一

些工作尝试将基于知识图谱的常识知识注入拓扑场景图中[76,77], 构建更加鲁棒的离散拓扑场景图. 但是, 为了探索更加适合具身导航任务的环境表征, 研究者们逐渐开始研究连续环境表示的构建方法, 基于对比学习的连续环境表示应运而生[78,79]. 如图 8 所示, 基于对比学习的连续环境表示方法[78]将成对的语义关系嵌入到一个潜在的特征空间, 鼓励具身智能体以探索的方式逐渐构建一个鲁棒的、全面的环境表示. 首先, Fast R-CNN[80]被用于基于 RGB 图像检测物体, 物体的属性和成对的物体之间的关系被 CER 编码器编码为固定长度的向量, 这样的特征向量被期望能够描述连续的、多维度的语义关系. InfoNCE 损失 $\mathcal{L}_{\text{contrast}}$ [81]被用来作为对比损失训练 CER 编码器, 促使特征空间中相似的特征相互靠近, 不同的特征相互远离:

$$\mathcal{L}_{\text{contrast}}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\dfrac{\text{sim}\left(\tilde{z}_i^1,\tilde{z}_i^2\right)}{\tau}\right)}{\sum_{j,k=1}^{N}\exp\left(\dfrac{sim\left(\tilde{z}_j^1,\tilde{z}_k^2\right)}{\tau}\right)} \tag{1}$$

其中, $\left(\tilde{z}_i^1,\tilde{z}_i^2\right)$ 表示成对的正样本, $\left(\tilde{z}_j^1,\tilde{z}_k^2\right)$ 表示成对的负样本, $sim(\cdot)$ 表示点乘操作, τ 表示 Softmax 温度缩放参数.

如图 9 所示, 通过联合优化一个基于强化学习的探索策略和一个视觉表示模型, Du 等人[79]提出了好奇心驱动的表示学习方法, 促使具身智能体在探索环境的过程中主动地学习环境表示. 具体来说, 探索策略和视觉表示学习相互博弈, 视觉表示模型的优化目标是最小化表示学习的目标损失 $\mathcal{L}_{\text{rep}}$:

$$\min_{\phi}\ \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\mathcal{L}_{\text{rep}}\left(M_\phi,\boldsymbol{x}\right)\right] \tag{2}$$

探索策略的奖励函数被设置为最大化表示学习的目标损失:

$$\max_{\theta}\ \mathbb{E}_{\boldsymbol{x}\sim\pi_\theta}\left[\sum_{t=0}^{T}\mathcal{L}_{\text{rep}}\left(M_\phi,\boldsymbol{x}\right)\right] \tag{3}$$

其中, M_ϕ 为表示学习模型, $\boldsymbol{x}$ 为从数据分布 p_{data} 中的采样, π_θ 表示探索策略. 因此, 探索策略被训练以最大化表示学习模型的错误, 具身智能体在这样的过程中被激励去探索充满不确定性的环境. 随着探索策略提供越来越难学习的数据, 学习的环境表示也变得越来越全面和鲁棒.

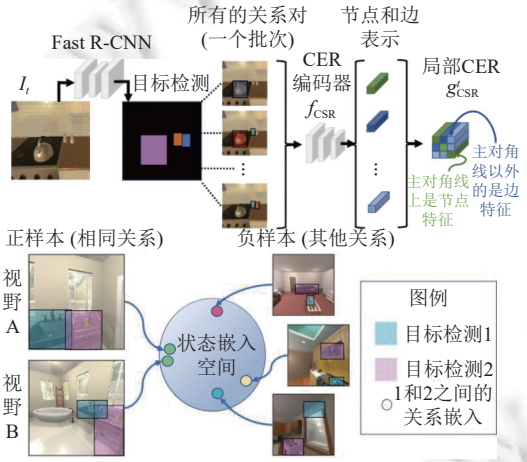


图 8 基于目标检测和对比学习的连续环境表示

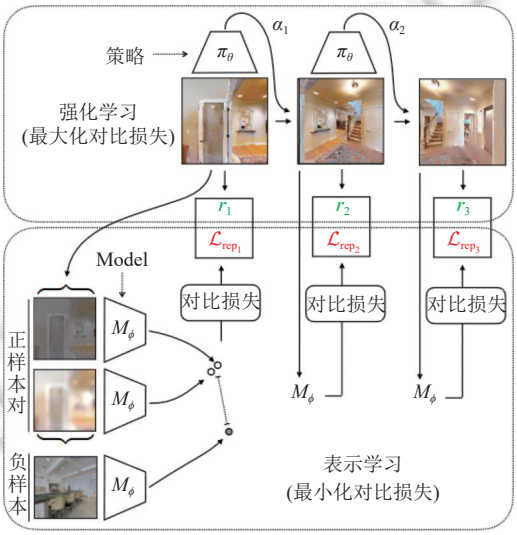


图 9 基于对比学习的对抗式连续环境表示

1.2 视觉自主探索任务与策略

视觉自主探索是具身智能体不可或缺的能力之一, 它使得具身智能体无需依赖人类的部署即可熟悉和适应复

杂多样的场景, 允许具身智能体为未知环境中的下游任务做无监督的准备[82]. 在视觉自主探索任务中, 具身智能体置身于一个事先未知的环境中, 能够捕获自我为中心的局部视觉观测和自身位姿状态. 在没有全局地图和任务目标的情况下, 具身智能体必须在探索策略的引导下, 在开放的场景中自主探索和收集空间和语义信息, 同时显式或隐式地构建环境表示. 如图 10 所示, 当接收到明确任务目标时 (例如去客厅拿一杯咖啡), 具身智能体能够凭借自主探索过程中建立的环境表示和先验知识, 解决诸如物体目标导航等下游任务.

图 10 视觉自主探索与下游任务示例

为了充分完整地探索环境, 具身智能体不仅需要同时定位与地图构建 (simultaneous localization and mapping, SLAM) 和分层次路径规划等模块的协同工作, 关键还需要一个鲁棒的探索策略持续地提供长期的探索目标. 如图 11 所示的探索框架[83], 在每个时间步 t, 具身智能体从环境获得视觉观测 s_t 和位姿传感器数据 x_t, SLAM 神经网络率先接收这些数据并建立环境表示:

$$m_t, \hat{x}_t = f_{\text{SLAM}}(s_t, x_t) \tag{4}$$

其中, m_t 表示部分的基于地图的环境表示, $\hat{x}_t$ 为具身智能体在 m_t 中的位姿估计, 全局探索策略依据现有地图表示 m_t 和具身智能体位姿估计 $\hat{x}_t$ 选择合适的长期导航目标 g_t^l:

$$g_t^l = \pi_G(m_t, \hat{x}_t) \tag{5}$$

其次, 长期导航目标 g_t^l、位姿估计 $\hat{x}_t$ 和地图 m_t 被整合和输入到路径规划器 f_{plan} 中:

$$trj\left(\hat{x}_t, g_t^l\right) = f_{\text{plan}}\left(m_t, \hat{x}_t, g_t^l\right) \tag{6}$$

输出一条引导探索的全局导航路径 (蓝色). 分层次路径规划策略根据全局导航路径生成航路点作为短期探索目标 g_t^s, 进一步结合视觉观测 s_t, 利用局部导航策略输出具体的导航动作:

$$g_t^s = waypoint\left(trj\left(\hat{x}_t, g_t^l\right), \hat{x}_t\right) \tag{7}$$

$$a_t = \pi_L(g_t^s, \hat{x}_t) \tag{8}$$

其中, a_t 为具身智能体经过一次视觉自主探索迭代后执行的导航动作.

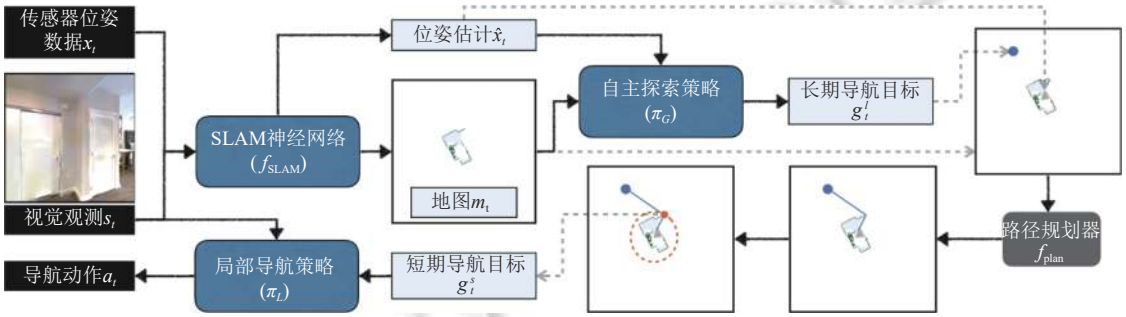


图 11 视觉自主探索架构示意图

现有的视觉自主探索方法基于强化学习框架, 试图赋予具身智能体某种奖励以鼓励具身智能体在环境中的探索行为. 本文基于 Ramakrishnan 等人[54]和 Bigazzi 等人[84]的工作, 梳理和总结了 5 种现有的探索策略, 其中包括新颖性奖励 (novelty reward)、好奇心奖励 (curiosity reward)、覆盖范围奖励 (coverage reward)、占用预期奖励 (occupancy anticipation reward) 和内在影响奖励 (intrinsic impact reward), 如图 5 所示.

新颖性奖励: 鼓励具身智能体探索先前未访问过的状态[85−89]. 在探索过程中, 每个状态 $s\in S$ 被分配一个访问计数 $n(s)$, 其中 S 为状态空间, 新颖性奖励与当前状态下访问频率的平方根成反比:

$$r_t \propto \frac{1}{\sqrt{n(s_t)}} \tag{9}$$

最具代表性的是文献 [85] 中的工作, 研究者将 3D 环境离散为 2D 网格作为状态空间 S, 其中每个单元格被视为一个离散的独立状态 s, 并根据公式 (9) 分配新颖性奖励. 早期, 研究者基于这种新颖性奖励定义了基于马尔可夫决策过程 (Markov decision process, MDP) 的最优值函数的置信区间, 并将其应用于传统的强化学习 (reinforcement learning, RL)[86]. 最近的工作将这一思想扩展到函数近似, Bellemare 等人[87]的工作通过引进伪计数的概念, 结合复杂的密度模型[88]或直接通过哈希函数[89]将高维度连续空间离散化, 用来辅助表格 RL.

好奇心奖励: 鼓励具身智能体探索难以准确预测的状态[90−92], 尽管这个状态已被访问多次. 基于动力学 (dynamic) 的好奇心奖励已被证明在大规模场景探索任务中表现良好[90,91]. 具身智能体首先通过学习一个前向动力学模型 (forward dynamic model) F 来预测具身智能体在当前状态 s_t 下, 采取的动作 a_t 对环境的影响:

$$\hat{s}_{t+1} = F(s_t, a_t) \tag{10}$$

然后, 在每个时间步对好奇心奖励 r_t 进行计算:

$$r_t \propto \| \hat{s}_{t+1} - s_{t+1} \|_2^2 \tag{11}$$

其中, $\hat{s}_{t+1}$ 和 s_{t+1} 分别表示预测的 $t+1$ 时刻的状态和 $t+1$ 时刻的真实状态. 以最小化前向动力学预测损失 $J(\theta)$ 为目标, 前向动力学模型 F 被在线优化:

$$J(\theta) = \min_\theta \| F_\theta(s_t, a_t) - s_{t+1} \|_2^2 \tag{12}$$

因此, 如果前向动力学模型 F 能够准确无误地预测状态的转变, 也就意味着具身智能体已经掌握了环境的演化规律, 好奇心奖励会鼓励具身智能体移动并探索新的状态. 在 Pathak 等人[90]的基于好奇心奖励的工作中, 图像特征被作为前向动力学模型的状态表示, 是对具身智能体的视觉观测状态最直观的表征. Ramakrishnan 等人[54]的工作没有延续这一方法, 他们使用了门控循环单元 GRU 隐藏状态计算好奇心奖励, 有助于缓解具身智能体有限的局部环境观测导致的偏差. 然而, 具身智能体所处的环境的演化可能存在固有的随机性, 好奇心奖励会诱导具身智能体重复访问部分随机状态[90,91]以获取高额积累奖励, 具身智能体永远无法准确预测并探索这样的场景. 为此, Burda 等人[91]通过将一个固定的、参数随机的神经网络定义为奖励预测函数, 提出了随机网络蒸馏方法来打破这一僵局. 此外, 基于探索分歧的方法也被用于解决这一问题[92], 与基于动力学的方法只采用一个前向动力学模型不同, 该方法通过学习一组前向预测模型并将它们的预测分歧作为鼓励探索新状态的内在动机.

覆盖奖励: 旨在引导具身智能体在短期内捕获尽可能多的感兴趣的事物[83,93]. 因此, 基于覆盖范围的奖励最大化在每个时间步收集的信息, 例如最大化物体和地标的数量, 或者最大化观测区域的面积和体积. 值得一提的是, 基于覆盖范围的奖励与新颖性奖励是截然不同的, 新颖性奖励鼓励访问所有地点而不是观察所有环境. 对于覆盖奖励, 在特定位置捕获的信息量取决于周围的 3D 结构和具身智能体的观测位姿. 基于覆盖奖励, Chaplot 等人[83]提出了基于 RL 的探索策略, 将覆盖范围奖励定义为:

$$r_t \propto AS_t - AS_{t-1} \tag{13}$$

其中, AS_t 表示在 t 时刻探索的体素数量. 该奖励鼓励具身智能体调整位姿, 优先考虑环境中未观测过的部分. 所以具身智能体不会因为重新访问相同的区域而获得奖励[93]. 然而在大规模环境中, 覆盖奖励容易面临稀疏奖励导致极端情况[54]. 例如, 具身智能体从一个大型环境中心位置开始探索, 当探索完成一半后, 将没有足够的奖励信号引导具身智能体探索另一半空间. 为了克服这一局限, Ramakrishnan 等人[54]结合新颖性奖励设计了一个覆盖奖励的变体, 称为平滑覆盖奖励, 通过对探索过的区域进行访问计数, 允许具身智能体在不常访问的区域中导航以走出稀疏奖励区域. 基于覆盖奖励的探索策略[93]通过将模仿学习和强化学习相结合来优化区域覆盖奖励目标, 以学习更好的泛化策略.

占用预期奖励: 旨在最大限度地提高环境重建的准确性[94−100]. 具体来说, 具身智能体被鼓励积极地探索最有

助于重建环境中未见过的区域的位置, 并最小化环境重建的误差. 占用预期奖励通常被建模为环境重建质量:

$$r_t \propto -d\left(V(P),\hat{V}_t(P)\right) \tag{14}$$

其中, P 表示相机位姿, $V(P)$ 表示环境重建的真实结果, $\hat{V}_t(P)$ 表示具身智能体在时间步 t 的实际环境重建, d 表示距离函数. 不同于好奇心奖励引导具身智能体探索不确定性区域, 占用预期奖励更支持具身智能体访问那些“有把握”的区域. 早期的环境重建局限于体素重建[94,95], 后来通过引入语义的概念扩展到 3D 语义重建, 要求具身智能体检测环境中的物体级概念. 早期, 基于占用预期奖励的探索方法[96]被提出引导具身智能体准确地对场景和物体进行像素级重建. Ramakrishnan 等人[95]的工作使用 pix2pix 模型[97]增强重建质量, 展示了探索策略的泛化能力, 能够广泛地转移至其他任务. 此外, 基于占用预期奖励的探索也可以表述为基于信息增益的探索[98−100], 其目标是通过探索并重建工作空间来降低环境的模糊性, 明确工作空间的占用情况.

内在影响奖励: 鼓励具身智能体对其内部环境表示的改变[101,102]. 这种改变意味着具身智能体在探索过程中对环境的认知的改变, 也就是内在影响. 在时间步 t, 内部环境表示的改变被度量为两个连续状态编码 $\phi(s_t)$ 和 $\phi(s_{t+1})$ 的 l_2 范数:

$$r_t = \| \phi(s_t) - \phi(s_{t+1}) \|_2 \tag{15}$$

然而直接使用上述内在影响奖励可能会导致自主探索陷入死循环困境, 即具身智能体陷入一个周期性循环, 虽然内部环境表示的改变很大, 但是无法逃离局部区域. 为了克服这个问题, Raileanu 等人[101]使用状态访问计数 $N(s_t)$ 来辅助离散空间中内在影响奖励的度量, 一定程度上缓解了这个问题. 但是在连续空间中, 访问计数的概念并不直接适用. 最近, Bigazzi 等人[102]提出了内在影响奖励的变体, 分别使用网格法 (grid) 和密度模型估计 (density model estimation, DME), 在连续空间中实现伪计数 $\hat{N}(s_t)$, 将内在影响奖励变更为:

$$r_t = \| \phi(s_t) - \phi(s_{t+1}) \|_2 \Big/ \sqrt{\hat{N}(s_t)} \tag{16}$$

其中, $\phi(s_t)$ 和 $\phi(s_{t+1})$ 分别表示相邻时间步状态 s_t 和 s_{t+1} 的编码, $\hat{N}(s_t)$ 是估计的伪访问计数, 伪计数权重的引入, 有效缓解了具身智能体产生周期性的“高影响、低探索”问题.

在 ObjectNav 任务中, 具身智能体通过自主探索工作空间来捕获环境的先验知识, 基于对环境的认知推理、搜寻并导航至物体目标. 因此自主探索是 ObjectNav 的前置任务, 对于 ObjectNav 任务至关重要, 为 ObjectNav 任务提供基础特征. 当然, 过度地探索环境不是 ObjectNav 任务的本意, ObjectNav 通常被视为一个权衡探索 (exploration) 和利用 (exploitation) 的问题, 通过在探索和利用之间切换, 高效地捕捉有助于物体目标定位的信息, 并合理地利用它们导航至目标.

2 面向具身人工智能的物体目标导航

近年来, 以计算机视觉、深度学习和强化学习等技术为基础, 环境表示学习和视觉自主探索领域取得了丰硕的成果. 人们期望具身智能体能够基于语义丰富的环境表示进一步执行有目的性的探索和搜索任务, 例如物体目标导航任务. 本节将从任务类型 (第 2.1 节)、模型结构 (第 2.2 节) 和发展进程 (第 2.3 节) 这 3 个不同的角度分别梳理和分析面向具身人工智能的物体目标导航方法. 此外, 本节在表 1 中总结了现有的物体目标导航策略的优势和局限性.

表 1 物体目标导航策略的优势和局限性总结

方法	年份	发表渠道	优势	局限性
SAVN[39]	2019	CVPR	首次提出采用元学习技术提高导航策略的泛化性能	缺少有效的长期记忆, 易导致重复探索
DD-PPO[103]	2020	ICLR	近乎完美地解决了点导航问题, 为物体导航提供了一种高效的局部运动规划器	缺少环境表示能力, 无法有效完成物体导航任务
Li等人[104]	2020	CVPR	将物体导航过程解耦为多种可迁移的元能力, 提升智能体在未知环境中的自主学习能力	缺少有效的长期记忆, 缺少对物体导航所需基本能力和思维的讨论

表 1 物体目标导航策略的优势和局限性总结 (续)

方法	年份	发表渠道	优势	局限性
ORG+TPN[36]	2020	ECCV	综合利用视觉观测中的局部和全局空间特征学习视觉表示, 能够帮助具身智能体脱离局部陷阱	缺少有效的长期记忆, 易导致重复探索
SemExp[45]	2020	NeurIPS	首次提出基于语义地图环境表示的物体导航策略和模块化的物体导航策略	动作空间大, 采样效率低, 易导致重复探索
SSCNav[18]	2020	ICRA	通过想象未知区域中的场景先验信息及其置信度来增强物体目标导航决策	缺少有效的长期记忆, 易陷入局部探索区域; 动作空间较大, 采样效率低
MJOLNIR[42]	2020	CoRL	利用环境中普遍存在的物体间语义和空间关系, 引导智能体逐步探索目标物体	导航策略的泛化性能受限于外部知识和环境布局差异影响
THDA[46]	2021	ICCV	通过数据增强提高训练场景的多样性和复杂性, 从而提高导航策略的泛化能力	依赖大量的训练场景和算力
Red-Rabbit[31]	2021	ICCV	提出利用辅助任务和奖励函数以端到端的方式训练通用物体导航策略	缺少有效的长期记忆, 易导致重复探索
Mayo等人[34]	2021	CVPR	提出一种空间注意力机制, 引导智能体关注物体之间的关系和目标物体的方位	缺少有效的长期记忆, 易导致重复探索
NIE[105]	2021	ICCV	提出交互式导航策略, 支持具身智能体通过改变环境状态实现更高效的导航	缺少有效的长期记忆, 特征融合机制较为粗糙
HOZ[32]	2021	ICCV	利用由粗到细的分层次场景先验引导智能体逐步搜寻物体	不同层次的拓扑场景先验之间的融合机制较为粗糙
GCExp[106]	2021	ROMAN	提出利用高层次的房间类别信息引导智能体寻找目标物体	导航性能受限于场景分类和拓扑图构建的准确性
L2M[19]	2021	ICLR	利用未观察区域中的语义类别不确定性来确定长期导航目标, 具备高效的物体搜索能力	动作空间较大, 采样效率低, 导航性能受限于语义地图构建的准确性
VTNET[65]	2021	ICLR	综合利用视觉观测中的局部和全局空间特征学习视觉表示, 以增强导航策略网络	缺少有效的长期记忆, 视觉表示的质量受限于物体检测模型
DOA[107]	2021	ACM MM	采用有向物体注意力图指导智能体学习物体之间的注意力关系, 引导智能体关注正确的物体目标	拓扑图的构建导致细粒度场景先验信息的损失, 不利于短期规划
Habitat-Web[14]	2022	CVPR	基于大量的专家演示和模仿学习技术提高物体导航性能	需要人工收集大量的专家演示
ZSON[53]	2022	NeurIPS	通过对CLIP模型进行微调, 利用CLIP中内化的丰富的视觉-语言特征提高导航泛化性能	缺少有效的长期记忆, 易导致重复探索
OVRL[108]	2022	ICLR	利用知识蒸馏技术学习隐式的场景表示, 为物体导航提供丰富的视觉特征	缺少有效的长期记忆, 易导致重复探索
ProcTHOR[47]	2022	NeurIPS	通过数据增强提高训练场景的多样性和复杂性, 从而提高导航策略的泛化能力	依赖大量的训练场景和算力
Li等人[49]	2022	Complex & Intelligent Systems	基于元学习技术, 充分利用视觉观测的上下文内容, 提高导航策略对未知环境的泛化性能	缺少有效的长期记忆, 易导致重复探索
OMT[66]	2022	ICRA	利用Transformer存储历史信息	采样效率低, 长期记忆不可避免的丢失
TransNav[48]	2022	Computational Design and Engineering	从空间和时间维度分别充分利用空间视觉特征和时序视觉特征, 增强导航策略的性能	对长序列的视觉观测进行建模, 导致计算量的急剧增加
GMAN[40]	2022	ECCV	基于生成式元对抗网络技术生成未知环境中的陌生物体的视觉特征, 从而泛化到新的物体目标	缺少有效的长期记忆, 易导致重复探索
Zhu等人[21]	2022	IROS	提出基于监督学习, 根据语义地图直接预测未知区域中距离智能体最近的物体位置	需要额外收集训练数据, 泛化性能易受环境布局影响
PONI[20]	2022	CVPR	提出基于监督学习, 预测未探索区域的潜在信息增益和目标物体存在的可能性	需要收集大量的地图训练数据, 泛化性能易受环境布局影响
Al-Halah等人[109]	2022	CVPR	利用模块化迁移学习提出一种适应多种目标模态的语义视觉导航策略	没有对多模态融合进行充分研究, 导航效率和成功率较低

表 1 物体目标导航策略的优势和局限性总结 (续)

方法	年份	发表渠道	优势	局限性
Campari等人[110]	2022	CVPR	增量地学习可重用的高维环境特征, 在导航过程中有效地获取和回忆历史知识	在复杂和未知环境中, 抽象模型的性能受限于特征提取模块
Stubborn[111]	2022	IROS	提出了一种多维度的障碍物地图和一种不依赖环境语义的启发式探索策略	动作空间过于简化, 容易导致无效的导航步骤
Min等人[23]	2022	IROS	利用自主探索和对比学习技术以具身的方式训练语义分割模型并构建环境表示, 提高导航效率	需要额外收集训练数据, 对复杂环境的泛化性能有待商榷
3D-aware[22]	2023	CVPR	提出了基于3D语义地图的场景表示和一种角向引导的探索策略, 采用3D场景先验引导智能体识别目标物体	3D语义地图的更新和维护效率低, 容易导致导航步骤的浪费
Dang等人[112]	2023	ICCV	提出 “搜索”和“导航”两种导航思维, 不同的导航阶段灵活地使用不同的思维	忽略了对探索、避障等重要思维的建模和讨论
Li等人[35]	2023	IEEE RAL	基于Transformer存储历史信息, 实现当前视觉观测与历史记忆的交互	对长序列的视觉观测进行建模, 导致计算量的急剧增加
PIRLNav[15]	2023	CVPR	深入研究了将模仿学习预训练和强化学习微调相结合的训练范式	导航模型设计过于简单, 对复杂场景的泛化性能未知
L-sTDE[38]	2023	CVPR	利用环境中的物体布局, 通过因果推断消除已知环境和未知环境之间的特征误差	导航性能受环境布局信息的制约, 在不同环境间的泛化性能未知
HiNL[37]	2023	CVPR	抑制了重复的历史记忆对导航策略的负面影响, 使智能体能快速响应环境变化	简单地将历史记忆编码为向量, 容易造成历史信息的丢失和遗忘
PEANUT[41]	2023	ICCV	采用监督学习, 根据语义地图直接预测未知区域中物体目标的潜在位置	需要收集大量的地图训练数据, 泛化性能易受环境布局影响
ESC[113]	2023	ICML	通过大语言模型将常识性知识引入导航策略实现零样本学习, 提高了在未知环境中的泛化性能	导航性能受限于提示符的设计和语义地图构建的准确性
ReVoLT[114]	2023	arXiv	使用多层空间-语义拓扑图构建场景表示, 利用分层次的场景先验增强导航策略	损失了细粒度的场景先验信息, 不利于短期规划
Chen等人[115]	2023	IEEE TCSVT	提出一种基于对比学习的连续环境表示方法和一种多步前向规划方法, 增强了智能体的环境理解和决策能力	连续环境表示的性能受限于语义地图构建的精度
SHRL[116]	2023	IEEE TMM	提出一种基于技能的分层强化学习导航框架, 调度不同的技能解决导航过程中的不同子问题	缺乏对避障、导航等思维的讨论, 基于向量的场景表示容易造成历史信息的丢失和遗忘

2.1 单物体与多物体目标导航任务

想象一下, 你走到一个家庭机器人助手面前问它: “能去看看我的笔记本电脑在不在桌子上吗? 如果在, 请拿来给我.” 为了成功地完成这项任务, 机器人助手需要具备视觉感知、语言理解、情景记忆、推理和规划和导航等广泛的技能. 作为中间环节, 搜索和导航到一个指定物体 (寻找笔记本电脑) 的能力是不可或缺的, 这也是本文讨论的重点. 物体目标导航按照任务类型可以分为单目标导航和多目标导航, 定义分别如下.

单物体目标导航: 要求具身智能体在一个事先未知的环境中导航至由一个特定标签指定的物体目标. 对于一个在事先未知的环境中以随机的坐标和位姿初始化的具身智能体, 类别层次的 (class-level) 物体目标导航要求其探索并导航至一个物体类别的任何实例 (例如“找到一把椅子”), 实例层次的 (instance-level) 物体目标导航[33]要求其探索并导航至具有某种属性的特定实例 (例如“找到一把木质的棕色的椅子”).

多物体目标导航: 要求具身智能体探索环境并按照指定的语义标签序列搜索并导航至多个不同的物体目标, 是对单物体目标导航任务的泛化.

虽然基于地图的记忆结构对于具身智能体来说不一定是最优的, 但它们具有较强的空间结构建模和可解释性优势, 被研究人员广泛地采用, 也已被证明在各种导航任务中优于基于神经网络的隐式记忆结构[32,34,36−38]. 基于这个观点, Wani 等人[25]提出了一个统一的框架 MultiON, 采用如图 12 所示的导航框架重点研究了基于地图的环境表示在单目标和多目标导航中的应用, 验证了简单的语义地图环境表示在提升导航性能方面优于一个更复杂的基

于神经网络隐式记忆. 该框架允许通过更改物体目标的数量来调整导航的难度, 验证了具身智能体的导航性能随着任务的复杂性增加 (物体目标的增多) 而急剧下降. 该框架采用 Actor-Critic 强化学习架构, 通过 one-hot 编码来表示多个物体类别标签, 同时将视觉观测和占据地图作为数据输入, 经过简单的特征拼接并输入到 GRU 网络中, 最终输出导航动作的概率分布和状态值.

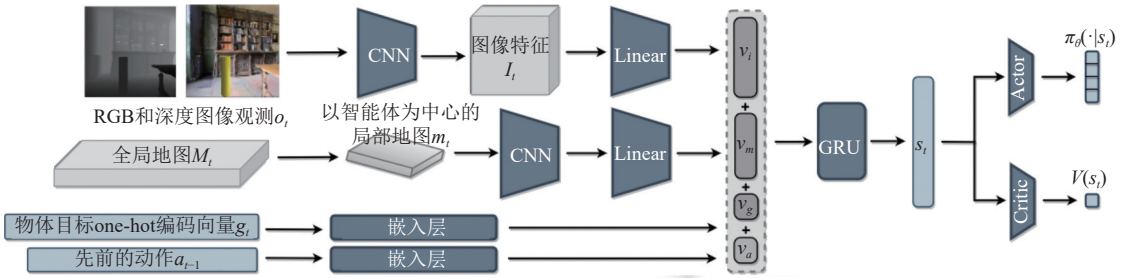


图 12 多 (单) 物体目标导航框架 (MultiON)

基于语义地图环境表示, Raychaudhuri 等人[26]最近采用模块化的设计, 通过单独开发物体检测模块、地图构建模块、探索模块和导航模块系统地研究了物体目标导航任务, 并提出了一个新的名为 MultiON 2.0 的大规模数据集. 类似地, Sadek 等人[27]提出了一种模块化的混合导航方法, 重点研究了多物体目标导航策略从仿真到现实的迁移. Sadek 等人将多物体目标导航问题分解为两个层次: (1) 采用监督学习和强化学习训练的深度神经网络分别解决探索、语义地图构建和目标检索问题; (2) 一旦确定了目标位置, 再将经典的 SLAM 技术和符号路径规划器相结合解决点导航问题. 最近, Zeng 等人[28]将基于语义占用地图的环境表示和基于拓扑语义地图的离散环境表示相结合, 相应提出了数据驱动的策略函数和知识驱动的策略函数, 在 Gibson 和 MP3D 数据集上均取得了优异的多物体目标导航性能. 不同于上述的环境表示方式, Marza 等人[29]最近提出采用两个神经网络学习神经隐式表示 (neural implicit representation) 来动态学习并构建场景上下文信息, 为多物体目标导航任务提供了新的范式. 上述物体目标导航策略[26−29]均采用不同的环境表示方法来准确辨别和存储环境中的有效信息, 提高具身智能体的场景探索和场景记忆能力.

现有的绝大多数导航方法研究集中在基于固定相机的问题定义上. 考虑到人类通常通过环顾四周来导航至一个或多个物体目标, 而具身智能体所搭载的相机的姿态是固定的、视野是有限的, 这很大程度上限制了 ObjectNav 的性能. 为了缓解这一问题, Chen 等人[30]提出了一个面向探索的主动相机观测策略, 将相机的转向运动定义为强化学习框架下的马尔可夫决策过程, 初步解决了两个问题: 1) 如何在复杂环境中学习一个好的相机控制策略, 2) 如何将相机控制策略与导航策略进行协调. 面向多物体目标导航, Chen 等人提出了如图 13 所示的基于地图环境表示和 Actor-Critic 强化学习框架导航方法. 为了便于学习主动相机观测策略, 一个启发式模块被提出用于提供专家经验, 在训练之初提供监督信号, 鼓励相机观测视野朝向包含大量潜在信息的未观测的区域. 随后, 主动相机观测策略在强化学习中进行微调, 以近似人类主动环顾四周的探索行为. 值得一提的是, 现有的导航策略的动作输出被作为主动相机观测策略的输入, 用于学习两种策略的协调机制. 主动相机观测策略有利于具身智能体充分地捕捉与导航目标相关的场景信息, 提高其在复杂环境中的探索能力.

2.2 端到端的物体目标导航策略

端到端的物体目标导航策略通常采用一个神经网络来建模导航所需的几种能力或思维. 神经网络以原始图像数据为输入, 直接输出导航动作. 人类和动物都能够很好地在新的环境中适应导航, 这种适应能力是很自然的, 但也是很巧妙的, 需要我们在新的视觉观察和过去的经验之间找到相似之处. 这种适应性归功于人类有能力对新的视觉信息进行分类, 并智能地关注与导航最相关的语义线索. 受此启发, Mayo 等人[34]提出了一种端到端的用于编码物体语义信息和空间信息的视觉注意力概率模型, 如图 14 所示. 该方法重点强调用于物体目标导航的视觉语言匹配能力和高效的场景记忆能力. 该模型由 3 种用于导航的注意力机制组成: 考虑图像中物体目标

信息的目标注意力机制, 考虑到具身智能体的上一个动作的动作注意力机制和考虑先前所有步骤的记忆注意力机制. 视觉图像特征和 3 种注意力机制的融合被作为注意力嵌入, 通过 LSTM 网络预测导航动作. 值得一提的是, 该方法具有良好的可解释性, 目标注意力揭示了物体目标在图像中可能存在的位置, 动作注意力揭示了下一步向右转更有助于观测到物体目标, 记忆注意力赋予了“冰箱”更高的注意力, 因为它与“烤面包机”在空间和语义方面有一定的关联.

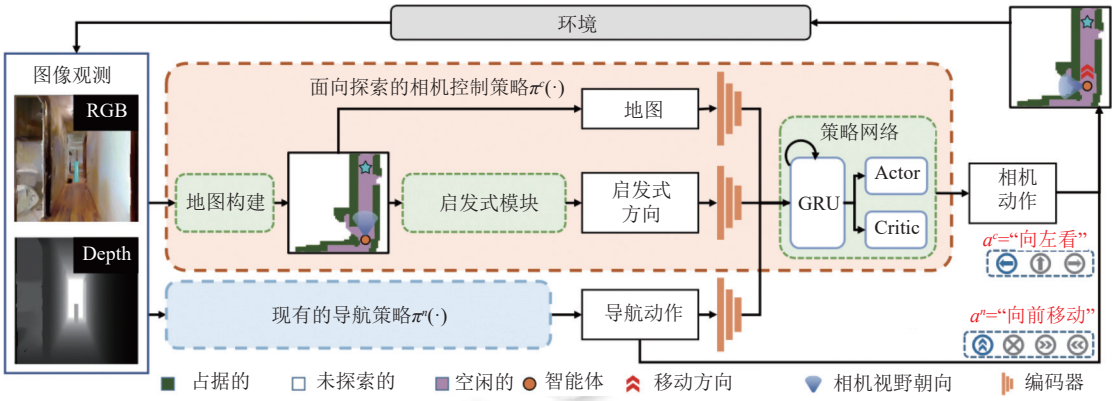


图 13 面向探索的主动相机观测策略和现有导航策略的结合

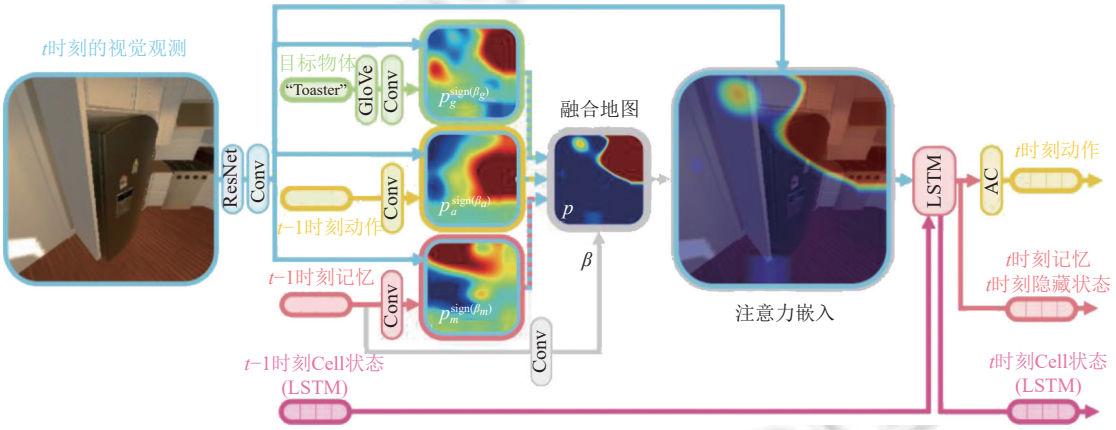


图 14 端到端的视觉注意力概率模型 (具身智能体被要求导航至冰箱附近的烤面包机)

然而, 由于 LSTM 和 GRU 等循环神经网络被证明在捕获运动轨迹中的长期依赖方面是低效的, 不利于处理和执行长动作序列. 因此, 基于 Transformer 的端到端的方法被提出, 通过利用多头注意力机制来捕获历史导航轨迹中不同时间步的长距离依赖关系, 图 15 所示的模型采用了 Actor-Critic 强化学习架构, 以 RGB 图像、深度图像和任务观测作为输入, 预测导航动作. 然而, 考虑到 Transformer 模型的训练往往依赖大量的数据, 在强化学习框架下直接利用 Transformer 结构是非常不稳定的. Li 等人[35]进一步提出了一个辅助任务来预测下一个航路点, 有利于引导强化学习和对环境的表示学习. 同图 14 所示的方法类似, Li 等人[35] 的工作也重点强调用于物体目标导航的视觉语言匹配能力和高效的场景记忆能力. 类似地, 基于辅助任务的方法也存在于 Ye 等人[31]的工作中. 该工作证明了辅助任务通过最小化有效的 RNN 维数来避免模型过拟合, 一个高性能具身智能体必须通过学习平滑的、低维的循环特征来实现长期的连续规划.

除了采用注意力机制和 Transformer 编码器来隐式地维护导航过程中的历史记忆, Du 等人[36]提出了一种物体关系拓扑图 (object representation graph, ORG), 在基于图像的目标检测过程中提取场景中成对的物体间的关系, 以重点强调用于物体目标导航的高效的场景记忆能力. 如图 16 所示, ORG 被作为局部拓扑环境表示, 与当前视觉特

征相结合构成了兼顾全局和局部场景先验的强大视觉表示. 为了防止具身智能体陷入局部死锁状态中, 模型构建了一个记忆增强的试探策略网络 (tentative policy network, TPN), 在训练阶段通过模仿学习克隆专家轨迹, 学习逃出死锁状态的动作. 在推理阶段, 即使具身智能体无法获得外部专家的监督, 试探策略网络首先检测具身智能体是否陷入死锁状态, 然后通过历史经验中的死锁状态-动作数据对, 指导具身智能体做出逃离死锁状态的动作尝试. 最近, Du 等人[37]在 ORG+TPN 的基础上重点研究了导航过程中的历史信息对当前决策的影响, 提出了一种历史启发的导航策略学习框架, 在 AI-THOR 数据集上取得了最佳的物体目标导航性能.

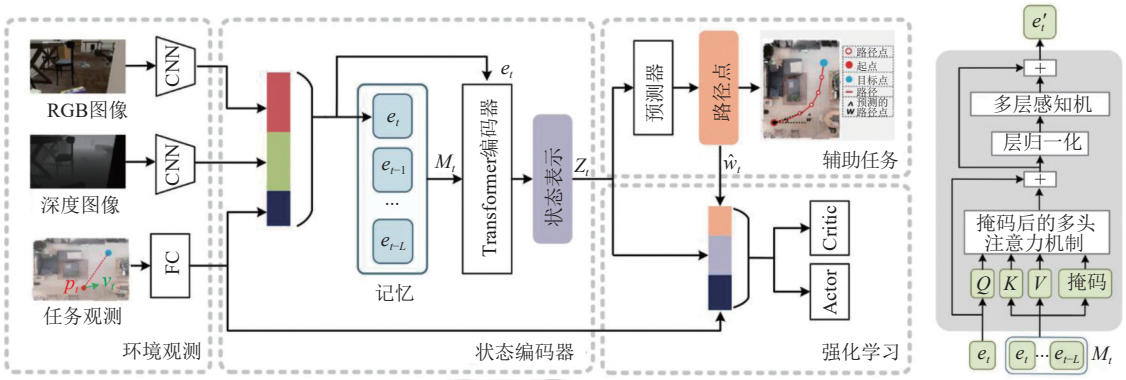


图 15 基于 Transformer 和特定辅助任务的端到端的物体目标导航策略

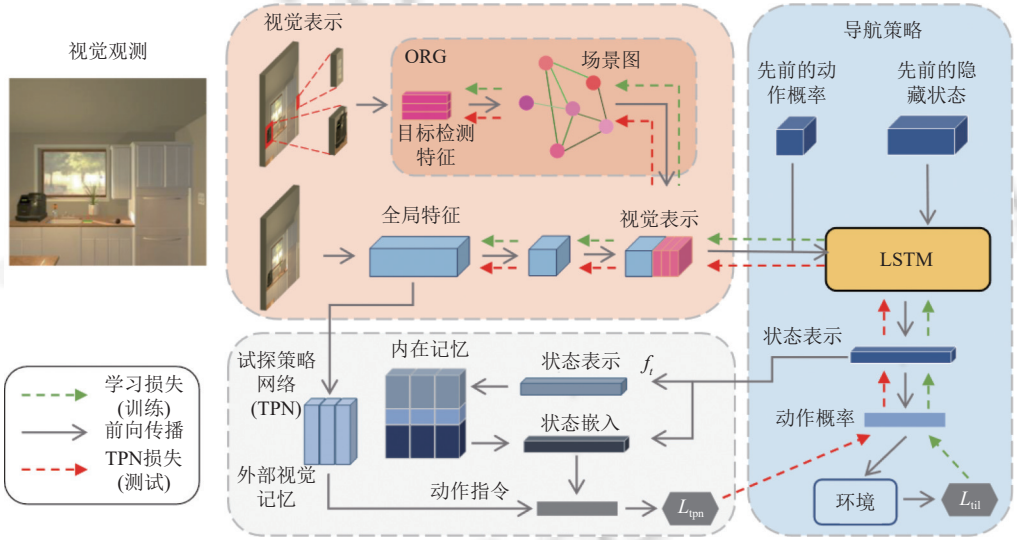


图 16 基于拓扑场景图的模块化导航框架 (ORG+TPN)

与 Du 等人[36]的工作类似, Zhang 等人[32]基于视觉图像维护了一个分层次的拓扑场景图来维护空间和语义模式, 并提出了一种在线学习机制, 根据实时的视觉观测更新场景图. 其中, 每个场景层的节点对应一个特定的场景 (例如客厅), 每个场景节点由一组空间层的节点组成, 空间层拓扑图是通过匹配相关房间级别的拓扑图并融合构建的, 每个空间节点有一组空间和语义相关的物体节点组成. Zhang 等人提出了如图 17 所示的端到端的物体目标导航框架, 分别采用不同的图卷积网络实现场景图不同层次的消息传递, 提取有利于导航的空间和语义线索. 其中, ResNet18 和 Fast R-CNN 模型分别被用来提取视觉图像特征和执行物体检测, LSTM 被用于沿时间维度维护导航过程中的历史特征. 分层次的拓扑场景图提供的由粗到细的环境布局和场景先验提高了具身智能体的视觉-语言匹配、场景探索和场景记忆能力. 最近, Zhang 等人[38]在他们的工作 (HOZ) 的基础上从因果推理的角度重点研

究了训练和测试环境之间的差异对物体目标导航性能的影响. 他们提出的基于布局的软总直接效应 (layout-based soft total direct effect, L-sTDE) 框架大幅提升了 HOZ 的导航性能.

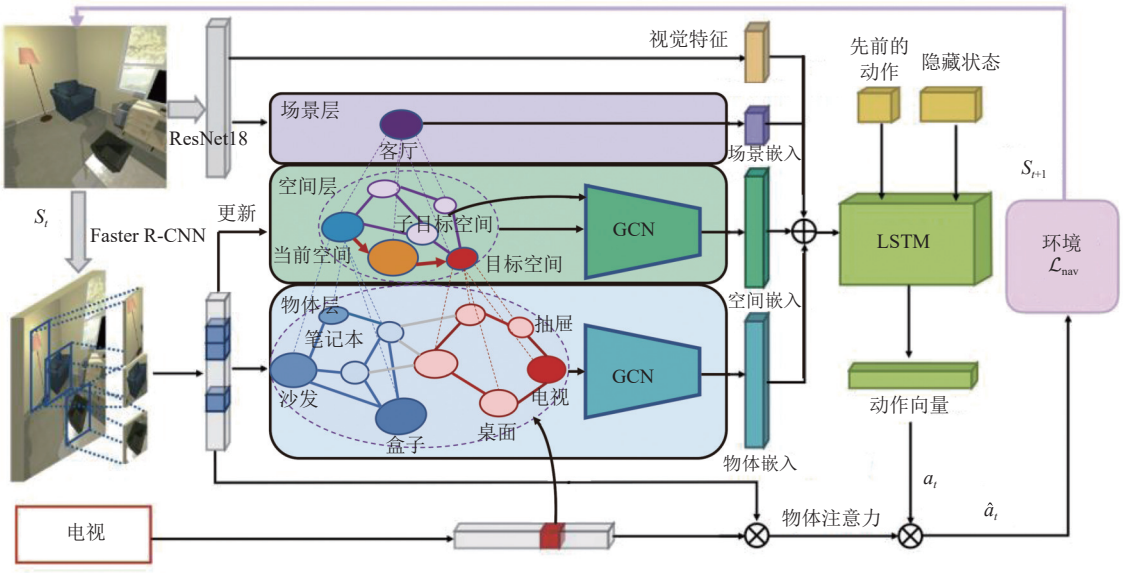


图 17 基于分层次拓扑场景图的端到端物体目标导航框架 (HOZ)

在物体目标导航问题早期的研究中, 基于元学习 (meta learning) 的端到端的导航策略被提出赋予具身智能体学习如何学习的能力, 进而泛化到事先未见过的场景. 如图 18 所示, 基于自监督交互损失 $\mathcal{L}_{\text{int}}^{\phi}$, Wortsman 等人[39]最早提出了名为 SVAN 的自适应 ObjectNav 策略. 在训练过程中, 交互梯度和导航梯度通过网络反向传播, 同时导航梯度被用来更新自监督交互损失的参数. 在推理过程中, 交互损失 $\mathcal{L}_{\text{int}}^{\phi}$ 的参数保持固定, 而网络的其余部分则使用交互梯度进行更新. 这种方法促使具身智能体在执行任务的同时不断地学习, 只是在不同的学习阶段, 学习的内容和学习的方式有所不同. 基于元学习的导航策略使具身智能体的场景探索、场景记忆和视觉-语言匹配能力在不同的学习阶段中逐步提高. 最近, Zhang 等人[40]提出了如图 19 所示的端到端的生成式元对抗网络 (generative meta-adversarial network, GMAN), 包括一个特征生成器和一个元鉴别器, 致力于提高 ObjectNav 策略对于事先未见过的物体类别的泛化能力. 特征生成器基于目标物体的语义类别嵌入想象并生成未见过的目标的初始特征. 元鉴别器对生成器进行监督, 使其能够进一步学习新环境的背景特征, 逐步调整生成的特征以接近目标物体的真实特征. 经过调整的特征作为更具体的目标物体特征表示, 指导具身智能体完成 ObjectNav 任务. 值得注意的是, 不同于上述的类别层次的物体目标导航策略, Li 等人[33]通过将导航过程划分为导航阶段和物体实例定位阶段, 提出了一种适用于实例层次的物体目标导航的端到端的 ObjectNav 策略.

2.3 模块化的物体目标导航策略

从 ObjectNav 策略所强调的能力来看, 端到端的 ObjectNav 策略通常直接将视觉观察映射为低级导航动作, 重点强调视觉语言匹配能力和高效的场景记忆能力. 其优点在于采用一个模型综合地学习 ObjectNav 所需的多种能力或者思维, 策略学习的流程相对简单, 易于部署. 但是, 寄希望于单一模型学习导航所需的多种技能是困难的, 需要规模巨大的训练数据和算力. 一方面, 端到端的 ObjectNav 策略所采用的神经网络被视为一个黑盒模型, 可解释性差. 另一方面, 端到端的方法将复杂的场景先验隐式地存储于模型内部, 增加了导航策略的负担, 导致模型的复杂度较高. 为了克服这些弊端, 近年来模块化的 ObjectNav 策略陆续被提出, 与端到端的方法形成了强有力的竞争[17]. 模块化的策略分别采用不同的模块建模导航所需的几种能力或思维, 重点强调高效的探索和场景记忆能力, 这些模块一般包括:

(1) 视觉感知模块: 主要负责对具身智能体的视觉观察进行编码, 或者通过语义分割获取环境语义信息.

(2) 语义映射模块: 整合视觉特征、深度信息和语义信息构建环境表示, 例如占用语义地图和拓扑语义地图.

(3) 动作决策模块: 根据具身智能体的视觉观测信息和环境表示, 基于传统的分层次导航策略或基于强化学习的导航策略进行路径规划和运动决策.

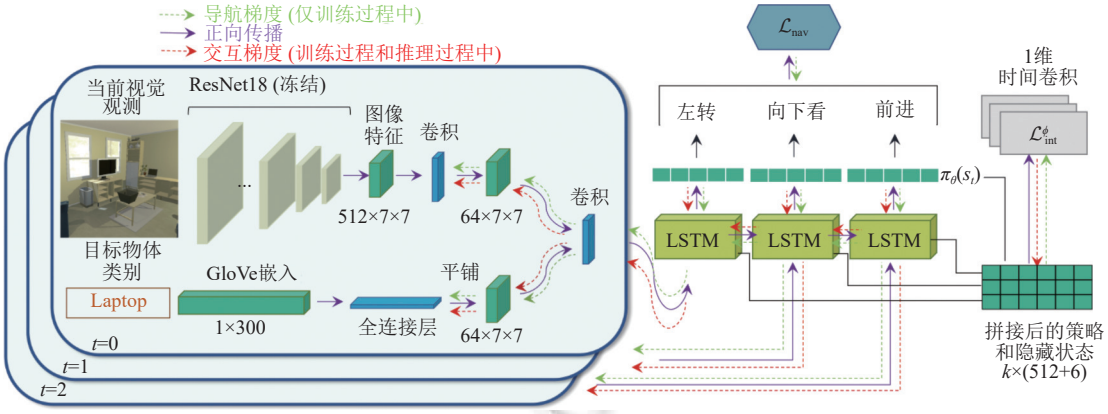


图 18 基于元学习的自适应 ObjectNav 策略 (SVAN)

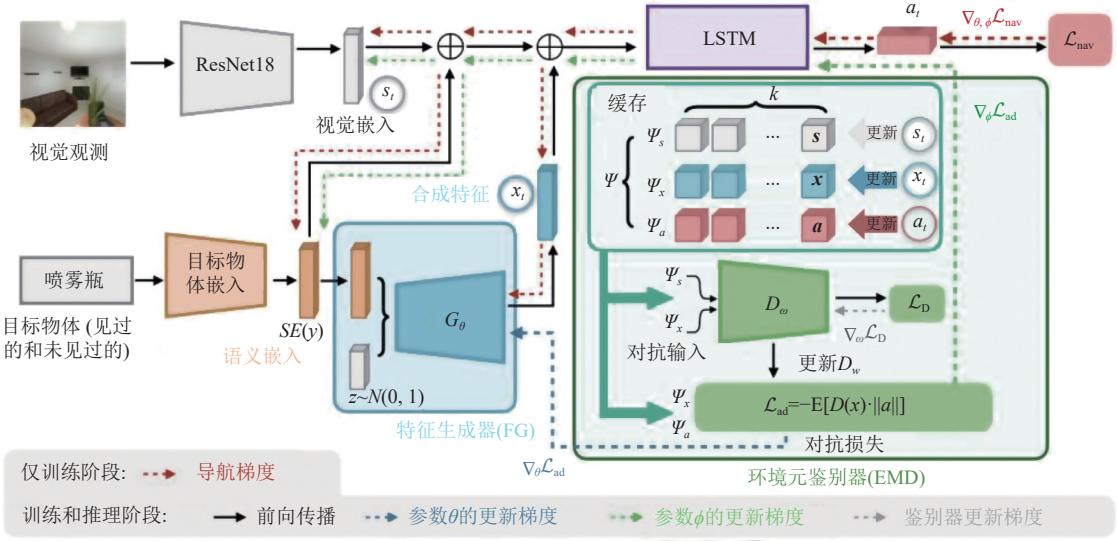


图 19 用于 ObjectNav 的生成式元对抗网络 (GMAN)

基于 2D 占用语义地图环境表示, SSCNav[18]将 ObjectNav 解耦为标准的视觉感知、语义映射和动作决策模块, 并提出一种基于语义地图补全的、置信度增强的 ObjectNav 策略. 如图 20 所示, 首先对 RGB 视觉观测进行语义分割, 结合深度图像生成语义地图环境表示. 考虑到基于帧数有限的图像构建的局部地图是残缺的, 一个语义地图补全网络被用来辅助预测视野之外的语义占用情况和语义关系, 弥补传感器视野的限制. 进一步地, 一个语义置信度预测网络被用来预测补全区域的不确定性, 指示预测结果的可信程度. 最后, 该导航框架将物体目标的语义类别嵌入、预测补全的局部地图和置信度得分拼接起来, 输入到导航网络, 用于预测局部的导航动作. SSCNav[18]采用基于语义地图环境表示, 重点强调高效的探索能力和场景记忆能力.

类似地, L2M[19]提出了一种空间预测策略, 促使具身智能体主动想象和利用视野之外的语义线索. L2M 同样利用语义类别的不确定性来确定长期导航目标, 不同于 SSCNav, L2M 通过集成多个模型的分歧来估计模型的不

确定性, 并进一步平衡探索和利用, 以提高搜寻物体目标的性能.

虽然局部语义地图有助于取得良好的导航性能, 但由于缺少全局的上下文场景先验, 此类方法容易陷入局部陷阱, 面临泛化能力差等问题. 基于全局语义地图表示, Ramakrishnan 等人[20]通过将 ObjectNav 问题解耦为两个子问题: “去哪里寻找目标”和“如何导航到目标”, 进而提出了一种名为 PONI 的导航策略. 得益于点目标导航技术的发展, “如何导航到目标”这一问题已经被近乎完美地解决. 因此“去哪里寻找目标”成为了实现 ObjectNav 的关键问题. PONI[20]是一种典型的权衡探索与利用的 ObjectNav 算法, 凸显具身智能体的探索能力和场景记忆能力. 如图 21 所示, 在整个导航过程中, PONI 利用视觉观测和位姿信息累积构建已探索区域的全局语义地图. 为了在未知的区域中搜索物体目标位置, PONI 提出了一个基于区域势函数和物体势函数的 encoder-decoder 架构的模块化导航策略. 区域势函数指示地图中未知区域所包含的潜在信息量, 指导具身智能体在搜索物体目标的过程中充分探索未知环境. 物体势函数则指示出未探索区域中存在物体目标的可能性, 指导具身智能体导航至具体的目标物体. PONI 的提出为模块化导航方法提供了新的范式, 基于全局语义地图表示的 ObjectNav 方法被研究者们进一步拓展. 不同于 PONI 所采用的势函数, Zhai 等人[41]提出了一种名为 PEANUT 的 ObjectNav 方法, 通过直接从全局语义地图预测物体目标的位置来学习环境布局中的空间和语义规律. Zhu 等人[21]则通过直接预测全局地图中不同位置到物体目标的距离, 指导具身智能体选择合适的中期导航目标. 然而, 考虑到具身智能体通常置身于 3D 场景中, 构建 2D 语义地图环境表示难免会丢失细粒度的 3D 空间信息. Zhang 等人[22]在 PONI 的基础上, 提出了角向引导探索和类别意识识别两个子策略, 为解决 ObjectNav 问题构建了第 1 个 3D 框架. 即便 3D 环境表示的构建和维护相对于 2D 环境表示需要大量的计算资源, Zhang 等人通过实验验证了 3D 场景表示的使用能够大大提升 ObjectNav 性能并降低模型训练的成本. 3D 场景表示使具身智能体能够同时捕捉场景中细粒度 3D 空间和语义信息, 使其具备更高效的场景记忆和视觉-语言匹配能力.

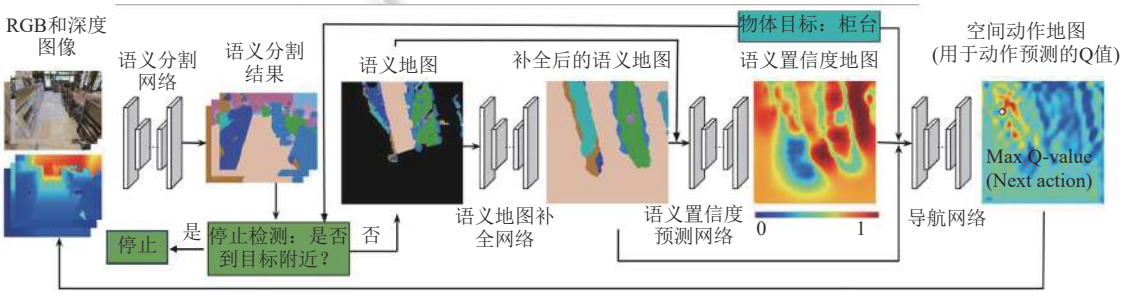


图 20 基于局部语义地图的模块化物体目标导航框架 (SSCNav)

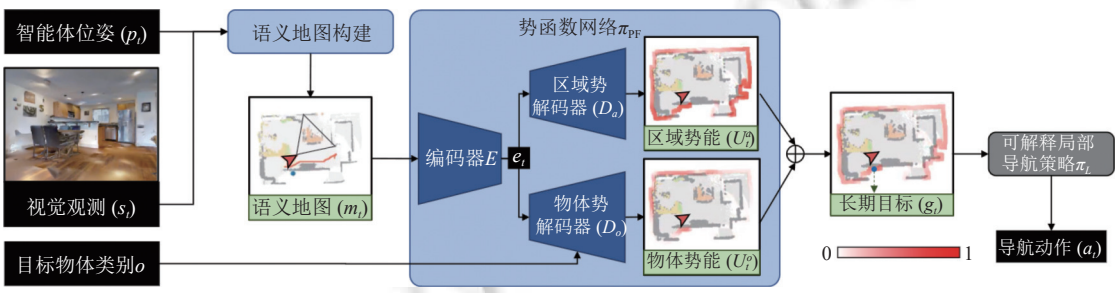
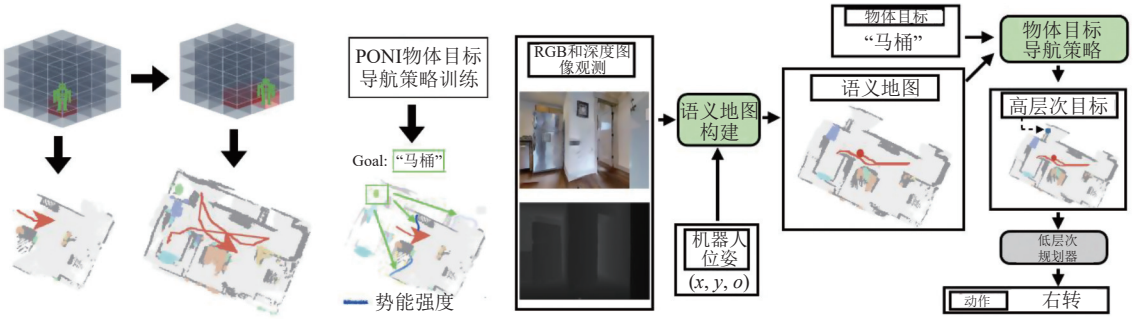


图 21 探索和导航相结合的模块化物体目标导航框架 (PONI)

虽然上述基于语义地图表示的模块化导航策略取得了优异的性能, 但是大都依赖人工标注的语义信息. 在实际应用中, 具身智能体被要求在没有语义注释的环境中也应达到相同的导航效果, 而无需将环境中的物体语义预先注释一遍. 为此, Min 等人[23]提出了一种自监督的 ObjectNav 算法, 利用自主探索和对比学习技术以具身的方式训练语义分割模型和构建环境表示 (如图 22(a) 所示), 并鼓励具身智能体基于 PONI 算法纯粹地从自己标记的语义地图中

学习 ObjectNav 策略 (如图 22(b) 所示). 具身的训练方式使智能体在与环境交互的过程中提高场景探索、场景记忆和视觉语言匹配能力. 值得一提的是, Min 等人率先进行了真实场景的机器人实验, 验证了所提出的方法的可行性.



(a) 具身场景表示学习 (b) ObjectNav策略

图 22 基于自主探索和对比学习技术的自监督 ObjectNav 策略流程

除了采用基于占用语义地图的环境表示, 基于拓扑场景图的模块化 ObjectNav 算法被提出改善对未知环境的泛化能力. 人类往往依靠在成长过程中积累的物体间的先验知识辅助未知环境中的探索和目标搜寻. 例如当我们需要寻找杯子时, 我们可能会到咖啡机旁边或者橱柜里去寻找. 受此启发, Yang 等人[24]提出通过构建知识图谱将语义先验知识整合到 ObjectNav 策略学习网络中. 考虑到在 ObjectNav 任务中直接搜索较小的目标物体是具有挑战性的, MJOLNIR 方法[42]通过构建场景拓扑图学习场景中物体之间的关系, 利用场景上下文有效地对目标进行分层次搜索. MJOLNIR 将与目标物体存在空间或语义关系的物体定义为父物体, 当具身智能体检测到父物体时会更倾向于在父物体周围寻找目标物体以达到分层次搜索的目的. MJOLNIR 强调了物体之间的共现关系, 促使具身智能体学习并利用更高效的场景记忆, 以搜寻目标物体. Pal 等人[42]通过实验证明, 基于语义拓扑图的分层次的导航算法的性能优于直接搜寻目标物体的导航算法. 上述两种方法都基于外部数据集 Visual Genome[117]提取语义关系并构建场景拓扑图, 同时强调了视觉语言匹配、探索和场景记忆能力. 然而, 外部数据集通常与任务场景存在差距, 不能完全覆盖所有的物体-物体关系. 为了缓解这个问题, Du 等人[36]提出了一种无需外部数据集的物体关系拓扑图 ORG, 在基于图像的目标检测过程中直接提取场景中成对的物体间的关系. 不同之处在于, Du 等人[36]提出的是一种端到端的导航策略, 如图 16 所示.

2.4 物体目标导航策略发展进程梳理与分析

过去 5 年的时间里, 真实室内场景数据集和高性能仿真器的发布促进了物体目标导航技术的快速发展. 在 2017–2019 年这段时间里, 大规模真实室内真实场景数据集陆续公开, 高性能仿真器陆续发布. 在此之前, Gupta 等人[118]也曾基于预先收集的小规模室内场景图像, 针对视觉导航任务展开研究. 大规模真实场景数据和仿真器的结合使用, 允许具身智能体身临其境, 以交互的方式在不断地试错的过程中学习导航技能. 本文按照发展时间线将现有的主流物体目标导航策略分为 4 个类别分别介绍.

2.4.1 基于有限数据集的导航策略学习

所谓有限数据集, 指的是 Matterport3D[4]、Gibson[5]和 AI2-THOR[7]等室内真实场景数据集, 与后文中的增强数据相区分. 有限的数据集为具身导航策略学习提供了丰富的场景和视觉图像观测, 但是无法涵盖真实世界中人类建造的所有的场景、房间或物体类型. 所以基于有限数据的策略学习难免具有一定的局限性, 甚至会导致策略过度拟合数据集, 难以泛化到未见过的场景. 早期, Mousavian 等人[43]基于较小规模的 active vision dataset 数据集[119], 提出采用语义分割和物体检测掩码作为神经网络的输入学习导航策略. 这项工作探索了结合真实环境图像和仿真图像联合训练导航策略的方法, 致力于降低将导航策略从仿真迁移到真实环境的难度. Wu 等人[44]基于 House3D 数据集, 提出了一种称为贝叶斯关系记忆的基于概率的拓扑场景表示, 致力于提高视觉语义导航智能体在未见过的环境中的场景探索、场景记忆和视觉-语言匹配能力.

MP3D 和 Gibson 数据集的发布和流行, 促使 Chaplot 等人[45]率先提出了基于语义地图环境表示的模块化物体

目标导航策略, 是一项具有里程碑意义的工作. 以此为基础, 在此之后的 2–3 年的时间里, SSCNav、L2M、PONI 和 PEANUT 等基于语义地图环境表示的方法被陆续提出, 基于高性能仿真器渲染的 RGB 图像、深度图像和语义分割, 充分地挖掘和利用了大规模真实场景数据集的空间结构和语义信息, 大幅度提高了导航的成功率等性能指标. 几乎与 Chaplot 等人[45]同时, Du 等人[36]基于 AI2-THOR 数据集和仿真器, 提出了基于离散拓扑环境表示的端到端目标导航策略. 相比于 MP3D 和 Gibson, AI2-THOR 数据集只提供 RGB 图像, 因此拓扑场景图和循环记忆网络被用来从时间和空间维度构建和维护有利于导航的环境特征. 之后, Zhang 等人[32]基于 AI2-THOR 数据集进一步提出了分层次的拓扑场景表示, 几乎将拓扑环境表示的优势发挥到了极致. 虽然上述方法中的策略学习受限于固定的场景, 基于有限的数据集的物体目标导航策略层出不穷, 经久不衰, 贯穿整个发展进程.

2.4.2 基于增强数据集的泛化策略学习

Maksymets 等人[46]通过实验验证了有限的数据集所导致的模型训练过拟合问题, 即经过大规模训练的具身智能体在训练环境中达到 94% 的成功率, 在测试环境中的成功率仅为 8%. 具体的原因是具身智能体过度拟合了训练环境的布局, 不需要探索环境就可以沿着最短路径导航至物体目标, 但是过度拟合的环境特征无法应对测试集中未见过的场景. Maksymets 等人提出了名为寻宝数据增强 (treasure hunt data augmentation, THDA) 方法, 通过在 MP3D 场景中插入新的 3D 物体来增强训练场景的复杂度, 以提高具身智能体在复杂多样的环境中的场景探索、场景记忆和视觉-语言匹配能力. Ramrakhya 等人[14]提出通过收集人类专家演示和模仿学习来增强物体目标导航性能的方法, 命名为 Habitat-Web. Ramrakhya 等人发现, 经过模仿学习训练的具身智能体从人类专家演示中学习了有效的物体搜索行为, 即窥视房间、检查角落里的小物体和通过转弯获得全景视野图像等技巧. 这些技巧很难通过强化学习突出地展示出来, 即使通过先进的强化学习技术诱导这些行为也需要繁琐的奖励工程.

最近, Deitke 等人[47]提出了名为 ProcTHOR 的数据增强方法, 通过随机采样任意多样化的、交互式的、可定制的和高性能虚拟环境来训练和评估具身智能体. 该方法以采样的方式对现有的场景数据集进行融合, 通过各个数据集的优势缓解模型训练过拟合问题. 该文还通过实验证明了 ProcTHOR 方法强大的零样本学习能力, 即通过 ProcTHOR 预训练的模型不需要下游任务的微调, 能够击败下游任务中最先进的方法. 考虑到重新构建一个新的、多演化的数据消耗大量的时间和经济成本, 上述基于增强数据集的泛化策略通常通过修改或者融合现有的数据集来进一步提升导航策略的性能, 缓解对训练环境布局的过度拟合.

2.4.3 基于元学习的泛化策略学习

考虑到现实环境的复杂多样, 采用数据集增强手段来改善物体目标导航性能仍然具有局限性. 对于人类而言, 学习本质上是一种持续的现象. 当人们学习一项新的任务时, 思维的训练和推理之间没有明显的区别, 因为我们会在执行任务的推理过程中不断地修正思维. 学习如何学习对于我们是一项关键的能力, 使我们能够毫不费力地适应新的环境和工作. 然而这与机器学习中的传统设置形成对比, 在机器学习中, 经过训练的模型在推理过程中被冻结. 在物体目标导航的早期研究中, 基于元学习的泛化策略学习方法就曾被提出, 赋予具身智能体学习如何去学习的能力. 2019 年, Wortsman 等人[39]基于模型不可知的元学习 (model agnostic meta-learning, MAML) 和强化学习提出了一种自适应视觉导航 (self-adaptive visual navigation, SAVN) 方法, 它可以在没有任何显式监督的情况下学习如何适应新的环境, 并泛化到没有见过的场景.

2022 年, Zhou 等人[48]基于 Transformer 和元强化学习提出了名为 TransNav 的物体目标导航框架, 考虑了导航过程中局部和全局视觉信息, 以及时间序列特征. Li 等人[49]在 SAVN 的基础上, 将离散的拓扑环境表示与元强化学习相结合, 弥补了具身智能体在训练场景和未见过的场景之间的导航性能差距. 这些基于元学习的泛化策略改善了具身智能体的无效探索行为, 加快模型的收敛速度. 上述的 3 种方法试图将导航策略泛化到未见过的环境中, 使其在训练场景中观察到的物体类别上取得合理的导航性能. 最近, Zhang 等人[40]提出了用于应对以未见过的物体类别为目标物体的生成式元对抗网络, 使具身智能体通过综合目标物体和环境的特征来“想象”没有见过的物体, 并导航至该物体. 基于元学习的泛化策略学习使具身智能体主动地学习和适应未知环境的空间和语义模式, 通过增强智能体的视觉-语言匹配、场景探索和场景记忆能力提高的导航效率.

2.4.4 视觉语言模型增强的导航策略

考虑到世界上不同的国家和地区的房屋结构不同, 房屋中的物品各色各样, 物体目标导航策略必须在不进行

额外微调的情况下应对各色各样的物体, 才能被广泛应用. 近期, CLIP[120]等视觉语言模型在针对任意物体的图像分类方面表现出了令人印象深刻的性能, 能够对训练中未明确看到的物体类别进行分类. 在物体目标导航任务中, 具身智能体必须在未见过的环境中找到通过文本标签指定的任意目标物体, 本质上可以建模为面向自然语言和视觉图像的多模态匹配任务. 基于这一观点, Gadre 等人[52]率先提出了基于 CLIP 的零样本 ObjectNav 策略 CoW, 以提高具身智能体的视觉-语言匹配能力. Gadre 等人在 Habitat 和 RoboTHOR 模拟器中评估 CoW 策略, 发现基于 CLIP 的物体定位和经典的探索策略, 在不需要额外的训练的情况下, 通常在成功、效率和鲁棒性方面优于基于学习的方法.

随后, 为了充分利用大规模视觉语言的表征和推理能力, Majumdar 等人[53]提出首先学习一个图像目标导航策略, 其中导航目标由一幅图像指定, 具身智能体被要求在真实场景中搜索并导航至该图像的拍摄位置. 通过将目标图像和物体类别的文本标签分别编码到多模态语义嵌入空间, CLIP 等大规模视觉语言模型能够被用于目标图像和物体类别之间的匹配, 因此具身智能体可以被指示寻找用自然语言描述的物体目标. 最近, Zhou 等人[113]也做了类似的研究, 提出了一种新的零样本物体导航方法, 称为软常识约束探索 (exploration with soft commonsense constraints, ESC), 将视觉语言预训练模型中的常识知识转移到真实世界中的物体目标导航, 而无需通过任何导航经验或任何其他视觉环境的训练. 常识知识的引入使具身智能体克服仿真和真实场景之间的差异, 使其具备更强大的视觉-语言匹配能力, 从而提高物体目标导航效率.

3 面向具身人工智能的视觉物体重排布

近年来, 场景表示、视觉自主探索和物体目标导航相关技术取得了显著的进步, 人们开始期望具身智能体完成如图 23 所示的更复杂、更加切合实际的任务. Weihs 等人[13]于 2020 年提出了一项新的前沿挑战——视觉物体重排布 (visual object rearrangement) 任务. 给定一个目标场景状态, 视觉物体重排布任务的目标是通过与环境交互, 将当前场景状态转变为目标状态. 例如房间中的椅子发生了移动, 具身智能体需要检测出椅子位置的变化, 并通过移动椅子将房间状态恢复为变动之前的目标状态. 显然, 视觉物体重排布任务要求具身智能体反复导航至不同的目标物体, 是 ObjectNav 的后置任务. 目前, 视觉物体重排布任务可以被归纳为两类: 受目标状态约束的物体重排布任务和无目标状态约束的物体重排布任务.

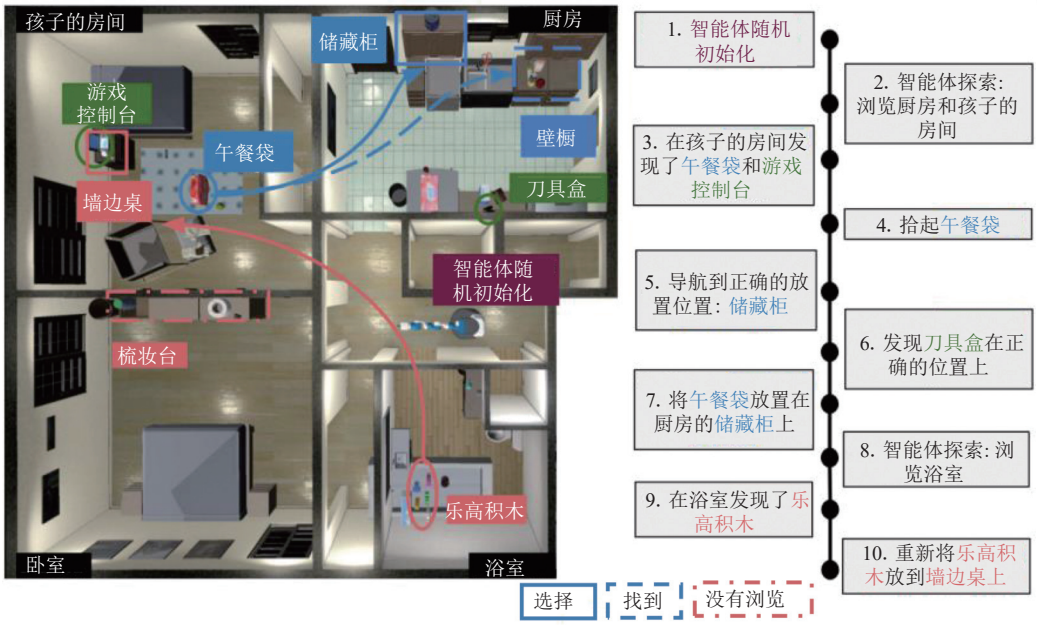


图 23 面向具身人工智能的视觉物体重排布任务

3.1 受目标状态约束的物体重排布任务

Weihs 等人[13]基于 AI2-THOR 环境构建了 RoomR 数据集, 并于 2020 年首次提出了物体重排布任务基准模型. Weihs 等人将物体重排布任务分为两个步骤: 预排布 (walkthrough) 和重排布 (unshuffle). 在预排布步骤中, 具身智能体充分地探索场景的目标状态并构建场景记忆. 在重排布步骤中, 场景目标状态被打乱, 具身智能体依靠预排布阶段构建的场景记忆识别当前状态中错位的物体并将其恢复至目标状态中相应的位置, 如图 24 所示. 顾名思义, 目标状态已知是指在重排布步骤开始之前, 具身智能体在预排布步骤中已经探索过目标状态并获取了场景记忆. 在 2021 年重排布挑战赛中, 物体重排布任务挑战被划分为两个赛道: 一阶段物体重排布任务和两阶段物体重排布任务.

图 24 物体重排布任务中的当前状态和目标状态示例

两阶段物体重排布任务是指预排布步骤和重排布步骤依次进行, 在重排布步骤中, 具身智能体依靠在预排布步骤中构建的场景记忆执行重排布任务. 如图 25 所示, Weihs 等人[13]提出的基准模型分别利用在预排布步骤中构建的显式和隐式场景记忆识别错位物体和预测导航动作. 具体来说, 在预排布步骤中, 模型通过构建显式的 2D 地图环境表示记录目标状态, 以便在重排布步骤中具身智能体将当前状态与基于地图的环境表示相匹配, 检测错位物体. 同时, 具身智能体的视觉感知状态被隐式地存储在 LSTM 网络中, 指导具身智能体在重排布步骤中的导航和交互.

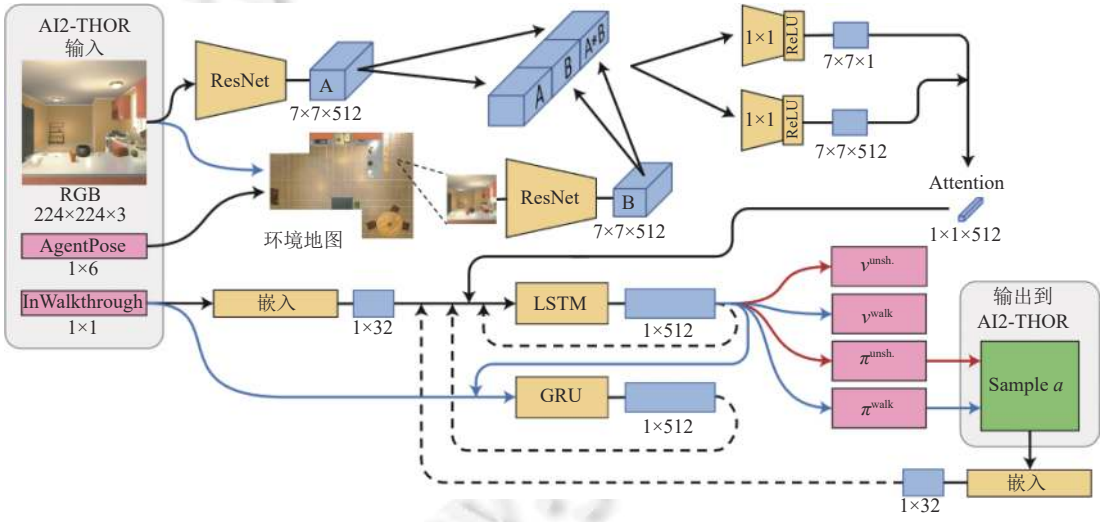


图 25 基于环境地图的二阶段物体重排布算法框架

与上述过程不同, 一阶段物体重排布任务规定具身智能体能在执行重排布步骤时, 能够直接访问目标状态, 而不经历预排布步骤的探索过程. 因此, 一阶段物体重排布任务相对容易一些. EmbCLIP[121]是目前一阶段物体重排布任务中效果最好的方法之一. EmbCLIP 不依赖语义映射模块, 而是采用基于 CLIP 大规模视觉语言模型的端到端的算法框架, 提升对目标状态和当前状态的理解能力, 能够更准确识别错位物体. EmbCLIP 获得 2021 年一阶段物体重排布挑战赛的第 1 名.

基于 Weihs 等人[13]的工作, Trabucco 等人[122]提出了基于 3D 语义地图表示和语义搜索的物体重排布算法框架. 如图 26 所示, 该方法基于当前状态和目标状态分别构建两个 3D 语义地图, 通过采用匈牙利算法匹配两个语义地图中对应位置的物体差异来执行错位物体检测. 该方法基于高精度的 3D 语义地图, 使用语义搜索策略准确定位错位物体并进行重排布, 保证了物体重排布任务的效率. 上述两种方法均为模块化的物体重排布策略, 实证研究表明, 基于占用地图环境表示的模块化策略优于端到端的策略的性能, 特别是在两阶段物体重排布任务中.

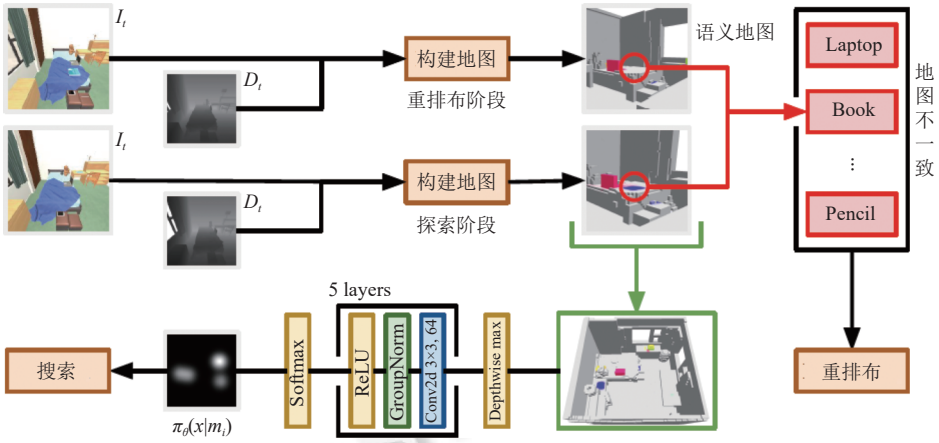


图 26 基于 3D 语义地图的两阶段物体重排布算法框架

3.2 无目标状态约束的物体重排布任务

然而, 在实际应用过程中, 物体重排布问题的求解往往不具备执行预排布步骤的条件, 因此物体重排布任务的目标状态是未知的. 为了缓解这个问题, 基于先验知识的物体重排布方法被提出, 允许具身智能体根据人类常识信息执行物体错位检测, 并进行物体重排布. 由于缺少对目标状态的说明, 基于先验知识的物体重排布任务要求具身智能体具有更强的自主推理能力, 能够适应更广泛的任务和环境. 以图 27 所示场景为例, 我们期望具身智能体在目标状态未知的情况下检测并拾起掉落到地板上的遥控器, 进一步推理放置遥控器的合适位置并执行重排布, 而不要求具身智能体在整理客厅之前探索整个客厅.

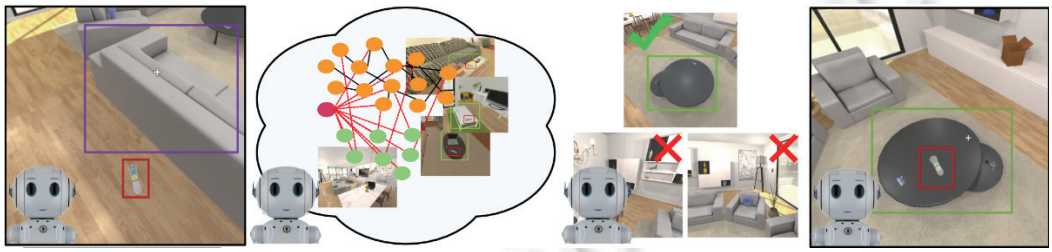


图 27 基于先验知识的物体重排布示例

自主重排布过程需要同时考虑 3 个子任务: 1) 探索: 具身智能体在环境中自主探索, 识别环境中存在的物体; 2) 错位检测: 具身智能体根据常识性知识自主识别错位物体并推理合适的放置位置; 3) 重排布: 通过 ObjectNav 和环境交互将检测到的错位物体放置到正确的位置. Sarch 等人[123]提出基于先验知识的物体重排布策略 TIDEE, 利用自然语言中表达的先验知识和从视觉观察中提取的空间-空间、物体-空间、物体-物体关系图识别错位物体, 并将错位物体重排布到合适的状态. 如图 28 所示, 该策略利用联合记忆图网络 (Memex) 构建可靠的任务场景上下文, 推断错位物体的合适位置, 最后通过视觉搜索网络指导具身智能体执行 ObjectNav 任务, 并重排布错位物体.

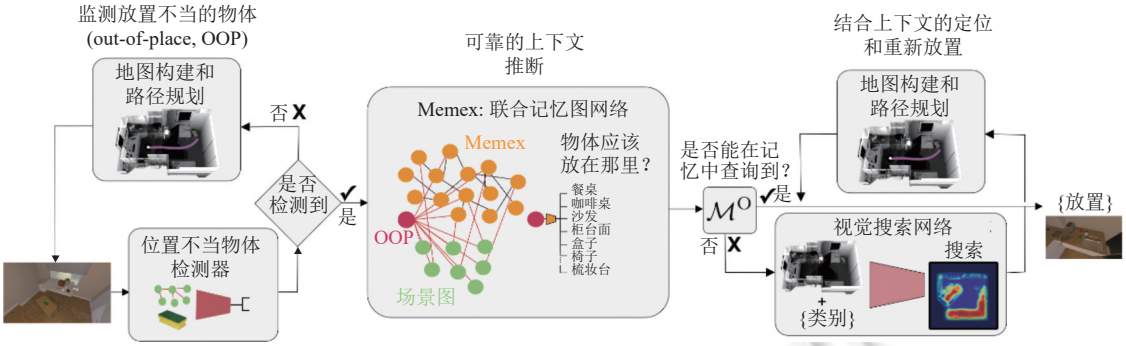


图 28 基于先验知识的物体重排布框架 (TIDEE)

然而, 在较大的任务场景中, 通过维护一个庞大的物体场景关系图推理每一个错位物体的合适位置需要大量计算成本, 并且会降低重排布的效率. 针对这个问题, Kant 等人[124]基于部分可观测场景提出新的重排布策略, 仅在部分观察的场景中识别错误物体, 并推理错位物体在当前场景中的正确容器. 当具身智能体捕获到新局部观测时, 继续重复这个过程, 进而维护一个物体-容器对列表. 在重排布任务过程中, 具身智能体基于大规模语言模型提取物体排布的先验知识, 并结合物体-容器列表找出放置错位物体的最适合的容器.

4 数据集、评价标准和策略学习

具身人工智能技术的快速涌现和发展离不开大规模逼真场景数据集和高性能仿真器. 同时, 合理的评价标准和高效的训练范式也是不可或缺的. 第 4.1 节着重介绍视觉自主探索、物体目标导航和视觉物体重排布任务普遍使用的逼真室内场景数据集, 第 4.2 节总结了研究领域内公认的评价标准, 第 4.3 节介绍当前主流的基于模仿学习和强化学习的导航策略学习范式.

4.1 数据集

与传统的基于预先收集的图像、文本或视频数据集进行学习的互联网人工智能 (Internet artificial intelligence) 不同, 具身人工智能强调具身智能体在与环境交互的过程中学习技能, 依赖于与环境交互的过程中捕获的位姿、视觉图像、环境布局等元数据. 另一方面, 早期的图像数据集存在数据量、数据种类和场景类型有限等缺陷, 不适应物体目标导航等任务的具身交互要求. 近年来, Matterport3D (MP3D)[4], Gibson[5], HM3D[6], AI2-THOR[7], iGibson[8,9]等数据集的陆续发布缓解了这些问题. 本节首先总结视觉自主探索和 ObjectNav 任务通用数据集 (见表 2), 然后对视觉物体重排布任务的专用数据集展开介绍.

表 2 视觉自主探索和 ObjectNav 通用数据集属性比较

数据集	发布时间	场景数目	物体数目	场景覆盖面积 (km^2)	可导航面积 (km^2)	支持交互	物理引擎
Matterport3D[4]	2017	90	40	101.82	30.22	否	—
Gibson[5]	2018	572	—	217.99	81.84	否	Pybullet
HM3D[6]	2021	1 000	—	365.42	112.50	是	—
AI2-THOR[7]	2019	120	3 578	—	—	是	Unity
iGibson 1.0[8]	2021	15	570	—	—	是	Pybullet
iGibson 2.0[9]	2021	15	1 217	—	—	是	Pybullet
ProcTHOR[47]	2022	—	1 633	—	—	是	Unity

4.1.1 视觉自主探索和物体目标导航数据集

(1) Matterport3D

为了弥补传统图像数据集的缺陷, Chang 等人[4]于 2017 年发布了 Matterport3D (MP3D) 数据集. MP3D 数据

集是一个大型 RGBD 数据集, 包含来自 90 个室内场景的 10 800 个全景视图的 194 400 张 RGBD 图像和相应的位姿. 与传统图像数据集不同, MP3D 数据集中的图像数据覆盖了场景的全部范围, MP3D 通过在环境中全面覆盖视点, 扫描每一个视点周围 360 度的全景视觉观测. MP3D 数据集同时提供了环境的 3D 纹理和语义信息, 语义信息包括涉及 40 个物体类别的 50 811 个实例级物体语义注释. 研究者通常采用 Habitat 仿真器[10]对 MP3D 数据集进行渲染, 开展具身人工智能技术研究.

(2) Gibson

Gibson 数据集由 Xia 等人[5]于 2018 年发布, 包括 572 个完整建筑中的 1 447 层空间, 总覆盖面积为 211 km^2. Gibson 数据集涵盖多种类型的场景, 包括住宅、办公室、酒店、博物馆、医院、建筑工地等. 每个空间都包含一组 RGB 全景图, 深度图像, 语义信息和 3D 纹理数据. Gibson 数据集是采用 3D 扫描重建的方法从真实环境中构建的, 做到了对真实世界的高质量的模拟重建.

(3) HM3D

HM3D 由 Ramakrishnan 等人[6]于 2021 年发布, 相较于其他 3D 场景数据集, 其优势主要表现在大规模、完整性, 视觉保真度这 3 方面. HM3D 是一个超大型的图像数据集, 包含真实世界中的 1 000 个建筑的 3D 纹理网格, 场景覆盖范围达到 112.5 km^2, 数据集规模比 MP3D, Gibson 数据集大 1.4–3.7 倍. 该数据集包含超过 10 600 个房间, 分布于大约 1 920 个建筑楼层, 覆盖多层住宅, 办公室, 餐厅和商店等多个场景. HM3D 使用密集视点均匀采样的方法弥补了 3D 重建过程中出现缝隙或者黑洞的情况, 提供了更完整的 3D 重建.

(4) AI2-THOR

AI2-THOR[7]本身不仅是一个数据集, 更是一个综合的视觉 AI 研究框架, 由 Kolve 等人于 2019 年发布. 与 MP3D 等数据集相比, 其最大优越性在于 AI2-THOR 支持具身智能体与物体进行多种类型的交互. AI2-THOR 包括 3 758 个可交互物体实例, 例如打开或关闭冰箱, 面包切片并使用烤箱进行烘焙等. 同时, AI2-THOR 提供的逼真的物体和场景是通过专业艺术家建模搭建的, 与真实世界近乎相同. AI2-THOR 囊括了 iTHOR, RoboTHOR, ProcTHOR-10K 和 ArchitecTHOR 等场景数据集. 其中, iTHOR 是原始数据集, 由包含卧室、浴室、厨房、客厅等场景在内的 120 个房间组成.

(5) ProTHOR

ProcTHOR 是 Deitke 等人[47]在 AI2-THOR 基础上提出的一种程序化场景生成框架. ProcTHOR 包含了 108 种室内常见物体以及 1 633 个可供交互的物体实例, 并且支持对物体材质, 空间位置, 房间布局, 甚至光照条件进行修改, 以此将仿真场景数量拓宽到一个极大的数量级. 相较于使用 Matterport3D、Gibson 和 HM3D 等基于 3D 扫描数据生成的环境, ProcTHOR 生成的场景既支持不同的物体状态 (打开、关闭和损坏等), 又支持机械臂与物体之间的多种交互操作, 实现了跨导航、交互、操作多种任务的智能体训练. 除了为具身智能体训练提供仿真环境, ProcTHOR 生成了 10 000 个不同规模和布局的室内场景并收集了对应的数据, 提出了目前最大的交互式家庭环境数据集——ProcTHOR-10K 数据集.

(6) iGibson 1.0 和 iGibson 2.0

iGibson 是一种新型的仿真模拟环境, 由 Shen 等人[8]于 2021 年发布. iGibson 既拥有庞大的图像数据集和逼真的 3D 环境, 又具有近乎真实世界的完整的动作交互过程. iGibson 1.0[8] 拥有 15 个支持完全互动的场景, 覆盖 108 个房间, 高度重建了真实世界的室内环境. 同时, 通过物理引擎对环境中的物体模型进行渲染, 具有真实材料质感和随视角光线动态变化的视觉属性. iGibson 1.0 优秀的人机交互界面为具身人工智能技术研究提供了便捷. iGibson 2.0[9]在 iGibson 1.0 的基础上丰富了物体的状态, 包括温度、湿度、清洁度、切片状态等, 将交互推广到更广泛的领域. 在交互任务方面, iGibson 进一步提出一系列谓词逻辑函数, 描述物体的状态变化, 例如温度变化, 清洁度变化. 此外 iGibson 2.0 在 iGibson 1.0 的基础上设计了一个新的虚拟现实界面, 支持用户与具身智能体的交互.

4.1.2 视觉物体重排布数据集

(1) RoomR 数据集

RoomR 数据集[13]是 Weihs 等人基于 AI2-THOR 仿真环境面专门为物体重排布任务设计的, 发布于 2021 年.

RoomR 数据集包括 6 000 个不同的物体重排布任务设置, 涉及 120 个不同场景中的 72 种不同类型的物体类别. 每组数据由房间初始状态、具身智能体的起始位置和目标状态组成. 在每一个独特的物体重排布任务设置中, 可移动的物体和可以打开但不可移动的物体的分布是随机和均匀的. 在生成初始状态和目标状态的过程中, 可打开但不可移动的物体是否打开和打开程度以及可移动物体的位置被随机地设置. 同时, RoomR 数据集保证被打乱的物体不会隐藏在容器中, 降低了物体重排布任务的难度. 在整个数据集中, 共有 1 895 个可拾取物体实例和 1 262 个可打开但不可拾取物体实例. 平均每个房间分别为 15.7 个可拾取物体实例和 10.5 个可打开但不可拾取物体实例.

(2) Habitat 2.0

Habitat 2.0 数据集[11]由 Szot 等人于 2021 年发布, 包含基于人工设计的 ReplicaCAD 数据集构建的交互式 3D 仿真环境. ReplicaCAD 数据集不仅包括 111 个布局独特的公寓场景和 92 种可交互物体, 还提供任务场景的动态参数、语义类别和表面注释. Habitat 2.0 仅支持具身智能体与刚体的动作交互. 与其他环境相比, 虽然 Habitat 2.0 在可交互类型数量方面并不占优势, 但 Habitat 2.0 有着其他数据集所不可比拟的运算速度. 在 Habitat 2.0 环境中, Fetch 机器人能够以每秒 1 200 步的速度进行交互. 作为参考, 每秒 30 步的交互速度被认为是实时交互, 其他模拟器通常能达到每秒 10–400 步. 因此, Habitat 2.0 的仿真速度达到了实时交互水平的 40 倍.

(3) HouseKeep

HouseKeep[124]是用于评估家庭环境中的具身智能体性能的基准任务, 由 Kant 等人于 2022 年发布. HouseKeep 数据集包括整洁和不整洁的室内场景, 包含 105 个房间和涵盖 268 个类别的 1 799 个物体. 通过将场景的目标状态打乱, 成对的目标状态和打乱后的状态被用于视觉物体重排布任务. HouseKeep 没有明确指示哪些物体需要重排布, 具身智能体需要依靠外部的先验知识来识别哪些物体发生错位, 并从先验知识中学习如何将错位物体重排布到正确位置. RoomR、Habitat 2.0 和 HouseKeep 这 3 种数据集的对比表 3 所示.

表 3 视觉物体重排布任务数据集对比

数据集	发布时间	场景数目	房间数目	物体类别	物体数目	重排布任务设置	物理引擎
RoomR[13]	2021	—	120	72	3 157	6 000	Unity
Habitat 2.0[11]	2021	111	—	92	—	—	Bullet
HouseKeep[124]	2022	14	105	268	1 799	—	—

4.2 评价指标

4.2.1 视觉自主探索评价指标

视觉自主探索任务的通用评价指标包括:

(1) 探索面积 (exploration area or volume, EA): 探索任务结束时, 具身智能体在场景中遍历的面积, 单位: m^2.

(2) 探索率 (exploration ratio, ER): 探索任务结束时, 具身智能体遍历的面积或体积占工作空间总面积或体积的比值.

(3) 交并比 (intersection over union, IoU): 探索任务结束时, 重建的地图与环境基准地图的交并比, 用于衡量重建的地图与基准的差异.

(4) 地图准确率 (map accuracy, MA): 探索任务结束时, 重建的地图与环境基准地图相匹配的区域面积.

不同的自主探索奖励有各自不同的优势, 因此适合不同的下游任务[54]. 一个好的自主探索方法会访问有趣的地点, 并收集对各种下游任务有用的信息. 例如, 基于覆盖奖励的自主探索策略可能无法与场景中的物体充分交互, 从而导致在以物体为中心的任务上性能较差. 然而, 基于覆盖奖励的自主探索策略可能有利于三维重建任务, 短期内遍历大面积的工作空间. 因此, 视觉自主探索的部分评价指标与下游任务密切相关, 通常还包括:

(5) 搜索物体数量 (number of objects, Objs.): 探索任务结束时, 具身智能体在工作空间中搜寻到的物体数量占要求的总数量的比值.

(6) 搜索路标数量 (number of landmarks, Lands.): 探索任务结束时, 具身智能体在工作空间中搜寻到的路标数量占要求的总数量的比值.

(7) 视图定位精度 (view location accuracy, V.Loc.): 探索任务结束时, 具身智能体在工作空间中定位到指定视图的成功率.

(8) 导航准确率 (navigation accuracy, Nav.): 导航准确率采用路径长度加权成功率 (success weighted by path length, *SPL*) 来衡量. 这个标准在衡量成功率的同时, 还衡量导航至目标的效率:

$$SPL = \frac{1}{N}\sum_{i}^{N} S_i \frac{l_i}{\max(p_i, l_i)} \tag{17}$$

设 l_i 为第 i 次任务中从具身智能体的起始位置到目标的最短路径距离, p_i 为本次任务中具身智能体实际走过的路径长度. 假设 S_i 是第 i 次任务中成功的二进制指标, 若任务成功为 1, 任务失败为 0. N 为验证任务总数.

(9) 重建准确率 (reconstruction accuracy, Recon.): 探索任务结束时, 具身智能体构建的地图的准确率, 计算为:

$$Precision = \frac{TP}{TP + FP},$$

其中, 将重建的地图中占据情况重建正确的设置为“正阳性” (true positive, *TP*), 重建错误的设置为“假阳性” (false negative, *FP*).

4.2.2 物体目标导航评价指标

在每一次物体目标导航任务中, 当具身智能体执行停止动作后, 如果具身智能体距最近的目标类别物体实例的距离在某个阈值 d_s 内则认为单次物体目标导航任务是成功的. 衡量物体目标导航算法性能优劣的首要标准是成功率, 成功率指成功的任务数量占所有物体目标导航任务的比例, 计算方式如公式 (18) 所示:

$$SuccessRate\,(SR) = \frac{n}{N} \tag{18}$$

其中, n 为成功的任务次数, N 为全部的导航任务数量.

同时, 为了更全面的评价物体目标导航算法的性能优劣, 从导航效率、导航准确率等多个方面设置标准对物体目标导航算法的性能进行评价.

(1) 路径长度加权成功率 *SPL*: 定义与公式 (17) 相同.

(2) 距离成功的距离 (distance to success, *DTS*). 任务结束时具身智能体距离成功阈值边界的距离. 计算方法如下:

$$DTS = \max(\|x_T - G\|_2 - d_s, 0) \tag{19}$$

其中, $\|x_T - G\|_2$ 是任务结束时具身智能体距离目标位置的 L2 距离, d_s 是成功阈值.

(3) 目标进度 (goal progress). 具身智能体每次运动后与目标类别物体缩减的距离.

(4) 轨迹相似性 (normalized dynamic time warping, nDTW). 衡量具身智能体的真实路径与专家路径之间的相似度, 惩罚偏离专家路径的偏差动作.

(5) 导航误差 (navigation error, NE). 具身智能体最终位置与目标位置之间的平均距离, 单位为 m.

(6) 路径长度软加权成功率 (soft success weighted by path length, *SoftSPL*). *SoftSPL* 与 *SPL* 直接将失败任务计算为 0 不同, *SoftSPL* 衡量了具身智能体停止位置与目标位置之间的距离 (即使任务失败).

$$SoftSPL = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{d_{T_i}}{d_{\text{init}_i}}\right)\frac{l_i}{\max(p_i, l_i)} \tag{20}$$

其中, d_{init_i} 指具身智能体的初始位置与目标位置的测地距离, d_{T_i} 指具身智能体的最终位置与目标位置的测地距离.

(7) 路径长度加权进度 (progress weighted by path length, *PPL*). 该评价标准多用于多物体目标导航任务中, 衡量了具身智能体寻找多个目标物体时的导航效率.

$$PPL = \bar{s} \cdot \bar{d} / \max\left(p, \bar{d}\right) \tag{21}$$

其中, $\bar{s}$ 指具身智能体成功找到的目标物体数量, $\bar{d} = \sum_{i=1}^{l} d_{i-1,i}$, $\bar{d}$ 为具身智能体开始位置到所有目标物体的总测地距离, $d_{i-1,i}$ 为第 $i-1$ 个和第 i 个目标之间的测地距离, l 为找到的目标物体数量, p 为具身智能体总路径长度.

路径加权成功率在成功率的基础上进一步评价了算法导航的效率, *SoftSPL* 在 *SPL* 的基础上提出了更宽松的评价标准, 导航结束时到目标物体的距离 (*DTS*) 和导航误差 (NE) 则评价了导航算法的准确率, 轨迹相似性则评价

了物体目标导航算法规划路径的质量, 这些标准均是物体目标导航算法领域内常用的评价标准.

4.2.3 视觉物体重排布评价指标

本文总结了视觉重排布挑战赛[13]中提出的 4 个评价指标.

(1) 成功率 (success rate, SR). 如果具身智能体恢复了目标状态中所有错位物体的位置则任务成功, 否则任务失败.

(2) 恢复率 (*fixed strict*). 该指标衡量了具身智能体恢复的错位物体的比例.

$$fixed\ strict = 1 - \frac{|M_{\text{end}}|}{|M_{\text{start}}|} \tag{22}$$

其中, $M_{\text{start}} = \{i \mid s_i^0 \not\approx s_i^*\}$, 表示在重排布阶段开始时错位物体的集合. $M_{\text{end}} = \{i \mid s_i \not\approx s_i^*\}$, 表示在重排布阶段结束时错位物体的集合. s_i^0, s_i^* 分别表示重排布阶段开始前和结束时物体 i 的状态.

(3) 错放率 (misplaced, *M*). 此指标等于任务结束时错位物体数除以任务开始时错位物体数.

$$M = \frac{|M_{\text{end}}|}{|M_{\text{start}}|} \tag{23}$$

如果具身智能体在重排布阶段结束时造成比任务开始时更多的错位物体, 则此衡量指标可能大于 1.

(4) 能量剩余率 (energy remaining, *E*). 该指标衡量了重排布任务结束后, 场景状态与目标状态之间的相似性.

$$E = \left(\sum_{i=1}^{n} D(s_i, s_i^*)\right) \Big/ \left(\sum_{i=1}^{n} D\left(s_i^0, s_i^*\right)\right) \tag{24}$$

其中, 能量函数 $D: S \times S \to [0,1]$, 当两个状态接近时, 能量函数递减到 0, 若两个状态近似相等, 则能量函数为 0. Energy remaining 定义为重排布任务结束时剩余的能量除以重排布开始时的总能量.

4.3 模仿学习预训练与强化学习调优

Ramrakhya 等人[14]通过人为控制具身智能体在 MP3D 数据集上完成物体目标导航, 收集了 80k 专家知识, 得到了 ObjectNav 任务的人类基线 (human baseline), 即 88.9% 的导航成功率, 这远高于当前最好的导航策略性能 (35.4% 的成功率). Ramrakhya 等人研究发现, 经过模仿学习训练的具身智能体能够从人类专家演示中学习有效的物体搜索行为, 即窥视房间、检查角落里的小物体和通过转弯获得全景视野图像等技巧. 这些技巧很难通过强化学习突出地展示出来, 即使通过先进的强化学习技术诱导这些行为也需要繁琐的奖励工程. 然而收集 80k 规模专家知识花费了大约 2 894 h 的操作时间和 50k 美元费用, 高昂的代价致使基于模仿学习的策略学习难以扩展.

此外, Ramrakhya 等人在研究中发现, 通过模仿学习训练的导航策略难以泛化至训练数据集以外的场景, 因为模仿学习强调的是对动作的克隆, 而不是强调搜寻并到达物体目标. 另一方面, 真实场景数据集和高性能仿真器的联合使用, 允许采用强化学习技术以交互和试错的方式学习具身导航策略. 然而, 强化学习往往需要细致的奖励工程, 奖励函数的设计需要权衡探索和利用 (exploration and exploitation). 为了鼓励具身智能体高效地导航至物体目标, 奖励函数的设计会不可避免地惩罚环境探索行为, 尽管环境探索是重要且不可或缺的. 此外, 采用强化学习技术学习的策略容易过度拟合小规模场景数据集, 损害导航策略的泛化性能.

主流的 ObjectNav 方法通常采用模仿学习预训练和强化学习调优相结合的方式来训练导航策略. 为了拓展模仿学习的可扩展性, 强化学习被用于对模仿学习预训练模型进行微调. Ramrakhya 等人[15]提出了名为 PIRLNAv 的导航策略学习范式, 采用模仿学习预训练为强化学习的 bootstrapping 提供一个合理的起点, 减少了强化学习的负担. 因为大多数随机采样的行动轨迹都不会产生积极的奖励, 在缺乏奖励的情况下从头开始学习通常是不切实际的. 虽然已经有大量的工作基于 Actor-Critic 强化学习框架研究了 IL 和 RL 相结合的训练范式, 但大多数方法注重通过模仿学习预训练学习策略 (Actor) 的参数, 而不更新值函数 (Critic) 的参数. 因此, 直接使用这些 IL 预训练策略权重来初始化一个新的 RL 策略, 通常会导致灾难性的失败, 因为在 RL 训练的早期会针对策略进行破坏性的策略更新. 为了克服这一挑战, Ramrakhya 等人提出了一个两阶段的学习范式, 首先只训练 Critic, 然后逐渐过渡到同时训练 Actor 和 Critic, 从而避免 RL 之初由未训练过的 Critic 所导致的破坏性参数更新. 不同的学习阶段, Critic 和 Actor 的学习率随步数的变化如图 29 所示.

图 30 说明了 IL 阶段采用不同数量的专家演示对 RL 微调的性能的影响. 显然, 当不采用 IL 时, 即 IL 专家演示数量为 0, 用于 ObjectNav 的 RL 是失败的, 即成功率近乎为 0. 深蓝色表示随着专家演示数量的增加, RL 后的 ObjectNav 成功率逐渐增加. 浅橘色表示随着 IL 数据集规模的扩大, RL 微调过程中获取的奖励先趋于平稳然后降低, 这意味着 IL 阶段的学习为 RL 提供了一个很好的起点. 这些研究结果表明, 通过有效地权衡 RL 调优后的性能和 IL 预训练数据集大小, 我们可以在不需要大量昂贵的专家演示的情况下获得更先进的 ObjectNav 性能.

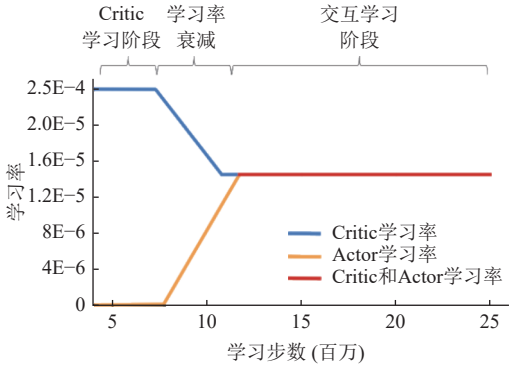


图 29 Critic 和 Actor 的学习率随步数的变化

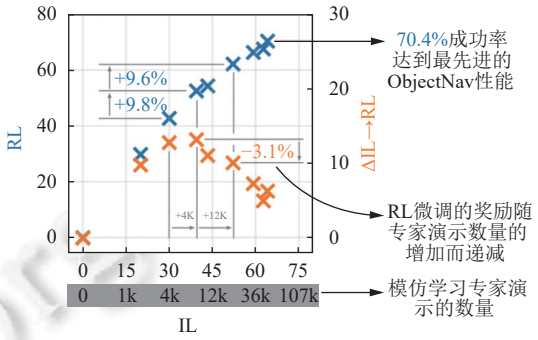


图 30 模仿学习专家演示数量对强化学习调优的性能影响

5 现有方法的性能比较及分析

基于第 4 节所述的数据集和评价指标, 本文总结了现有的视觉自主探索、ObjectNav 和视觉物体重排布策略的性能比较, 最优性能用粗体表示. 根据第 2 节所述的 ObjectNav 模型结构和策略划分, 对现有 ObjectNav 方法进行了细致的分类和梳理. 表 4 为视觉自主探索任务中的算法总结, 表 5 介绍了多目标 ObjectNav 算法性能对比, 表 6–表 8 按照模型架构, 分别总结了现有的 ObjectNav 策略在不同数据集上的实验结果. 表 9 分别介绍了现有的一阶段和二阶段视觉物体重排布任务的基线和最佳性能.

表 4 P3D 数据集上的视觉自主探索性能比较

方法	年份	EA↑	ER↑	IoU↑	MA↑	Objs.	Lands.	V.Loc.	Nav.	Recon.
随机动作	—	—	—	—	—	0.25	0.16	0.13	0.64	0.38
边界探索[125]	1997	—	—	—	—	0.57	**0.50**	**0.24**	**0.70**	**0.46**
新颖性奖励[85]	2020	—	—	—	—	**0.62**	0.35	0.22	0.68	0.45
好奇心奖励[90]	2020	157.27	—	0.34	109.79	0.31	0.21	0.15	0.64	0.40
覆盖奖励[83]	2020	158.83	—	0.36	116.58	0.56	0.35	0.22	0.68	0.44
占据预期奖励[54]	2020	147.33	—	0.42	126.86	0.53	0.40	0.21	0.69	0.45
内在影响奖励 (grid)[101]	2022	157.19	—	0.44	133.97	—	—	—	—	—
内在影响奖励 (DME)[102]	2022	**166.20**	—	0.43	133.27	—	—	—	—	—
ANS[83]	2020	73.28	0.52	—	—	—	—	—	—	—
OccAnt[126]	2020	—	—	0.34	100.30	—	—	—	—	—
S-ANS[127]	2022	84.40	—	—	—	—	—	—	—	—
UPEN[128]	2022	113.00	**0.68**	**0.46**	**141.60**	—	—	—	—	—

5.1 视觉自主探索性能比较及分析

为了对视觉自主探索策略进行总结, 探索面积 (EA)、探索率 (ER)、IoU 和地图准确率 (MA) 被用来评价策略的性能, 搜索的物体数量 (Objs.)、搜索的路标数量 (Lands.)、视图定位精度 (V.Loc.)、导航准确率 (Nav.) 和重建准确率 (Recon.) 被用来评估探索策略对下游任务的贡献. 表 4 中的第 1–2 行总结了视觉自主探索任务的基线. 其中随

机动作是指具身智能体在探索过程中随机选取探索动作, 不接受任何探索奖励的引导. 边界探索是指具身智能体总是优先探索最近的未知的区域. 所谓边界, 也就是地图中已知空间与未知空间的临界区域. 表 4 中的第 3–8 行总结了基于第 1.2 节中所述的 5 种探索奖励的自主探索性能. 其中 grid 和 DME 表示基于栅格计数和稠密模型估计分别实现的两种内在奖励影响. 从实验结果来看, 内在影响奖励在探索面积、探索率和地图准确率方面的优势更加明显. 不同类型的探索奖励对不同的下游任务的贡献不同, 例如新颖性奖励鼓励具身智能体搜索到更多的物体.

表 5　多物体目标导航性能对比

方法	视觉数据	视觉模块	年份	SR (%)↑	Progress↑	*SPL* (%)↑	*PPL*↑
MultiON[25]	RGBD	CNN	2020	22.0	40.0	17.0	30.0
Chen等人[30]	RGBD	—	2022	51.1	67.3	38.7	49.5
Marza等人[29]	RGBD	—	2023	**58.3**	69.4	**43.8**	**52.1**

表 6　MP3D 数据集上的 ObjectNav 性能比较

数据集划分	模型结构	方法	视觉数据	视觉模块	策略类型	年份	SR (%)↑	*SPL* (%)↑	*DTS* (m)↓	*SoftSPL* (%)↑
验证集	端到端	Random	—	—	有限数据集	—	0.005	0.005	8.05	—
		DD-PPO[103]	RGBD	ResNet	有限数据集	2020	8.0	1.8	6.90	—
		THDA[46]	RGBD	RedNet[129]	增强数据集	2021	28.4	11.0	5.59	18.9
		Red-Rabbit[31]	RGBD	ResNet18	有限数据集	2021	34.6	7.93	—	—
		Habitat-Web[14]	RGBD	ResNet18	有限数据集	2022	35.4	10.2	—	—
		ZSON[53]	RGB	ResNet50	视觉语言模型	2022	15.3	4.8	—	—
		OVRL[108]	RGBD	ResNet50	增强数据集	2022	28.6	7.4	—	—
	模块化	SemExp[45]	RGBD	Mask R-CNN[130]	有限数据集	2020	36.0	14.4	6.73	—
		SSCNav[18]	RGBD	ACNet[131]	有限数据集	2020	27.1	15.7	—	—
		GCExp[106]	RGBD	—	有限数据集	2021	**70.3**	**51.6**	**3.66**	**55.1**
		L2M[19]	RGBD	UNet	有限数据集	2021	39.1	17.0	**3.37**	22.1
		Zhu等人[21]	RGBD	Mask R-CNN[130]	有限数据集	2022	**43.6**	23.2	—	—
		PONI[20]	RGBD	Mask R-CNN[130]	有限数据集	2022	27.8	12.0	5.60	—
		Al-Halah等人[109]	RGB	ResNet9	增强数据集	2022	14.6	10.8	—	—
		Campari等人[110]	RGBD	RedNet[129]	有限数据集	2022	27.9	13.1	6.16	23.3
		3D-aware[22]	RGBD	RedNet[129]	有限数据集	2023	34.0	14.6	4.74	**30.5**
		PEANUT[41]	RGBD	PSPNet[132]	有限数据集	2023	40.5	15.8	—	—
		ESC[113]	RGBD	GLIP[133]	视觉语言模型	2023	36.1	**17.7**	—	24.0
		ReVoLT[114]	RGBD	YOLOv4	有限数据集	2023	**85.7**	**7.0**	**0.03**	—
测试集	端到端	DD-PPO[103]	RGBD	ResNet	有限数据集	2020	0.00	0.00	10.32	0.94
		Red-Rabbit[31]	RGBD	ResNet18	有限数据集	2021	23.67	6.22	9.14	12.14
		Habitat-Web[14]	RGBD	ResNet18	有限数据集	2022	**27.8**	**9.9**	—	—
		THDA[46]	RGBD	RedNet[129]	增强数据集	2021	21.08	8.75	9.20	16.96
		OVRL[108]	RGBD	ResNet50	增强数据集	2022	23.2	7.6	—	—
	模块化	SemExp[45]	RGBD	Mask R-CNN[130]	有限数据集	2020	17.85	7.07	8.82	14.50
		PONI[20]	RGBD	Mask R-CNN[130]	有限数据集	2022	20.01	8.82	**8.68**	**17.08**
		Stubborn[111]	RGBD	Mask R-CNN[130]	有限数据集	2022	23.7	9.8	—	—

表 4 中还总结了最近的 ANS[83]、OccAnt[126]、S-ANS[127]、UPEN[128]这 4 种方法的探索性能. ANS[83]于 2020 年率先提出了面向具身人工智能的视觉自主探索方法, 提出了模块化的主动视觉 SLAM 框架, 在 MP3D 数据集上达到了 73.28 m^2 的探索面积和 52% 的探索率. 后来, OccAnt[126]基于占用预期奖励针对自主探索精度展开研究, 在 MP3D 数据集上达到了 0.34 的 IoU 指标和 100.3 的地图准确率指标. 这一结果与表 4 中的占用预期奖励的实验数

据是匹配的, 占用预期奖励在重建精度准确率方面达到了相对较高的性能指标. S-ANS[127]通过引入归纳偏差, 从模型结构设计方面对 ANS 进行了改进, 其性能的提升主要体现在探索面积方面, 达到了 84.4 m^2 的探索面积. 最近, UPEN[128]通过学习一组前向预测模型并将它们的预测分歧作为鼓励探索新状态的内在动机, 在探索面积、探索率、IoU 和地图准确率方面都达到了最先进的水平.

表 7 Gibson 数据集和 HM3D 数据集上的 ObjectNav 性能比较

数据集	模型结构	方法	视觉数据	视觉模块	策略类型	年份	SR (%)↑	*SPL* (%)↑	*DTS* (m)↓
Gibson	端到端	DD-PPO[103]	RGBD	ResNet	有限数据集	2020	15.0	10.7	3.24
		ZSON[53]	RGB	ResNet50	视觉语言模型	2022	31.3	12.0	—
		Li等人[35]	RGBD	CNN	有限数据集	2023	**82.8**	**48.7**	—
	模块化	SemExp[45]	RGBD	Mask R-CNN[130]	有限数据集	2020	54.4	19.9	1.723
		PONI[20]	RGBD	Mask R-CNN[130]	有限数据集	2022	73.6	41.0	1.25
		Al-Halah等人[109]	RGB	ResNet9	增强数据集	2022	33.0	23.6	—
		Min等人[23]	RGBD	DeepLabv2[134]	有限数据集	2022	60.0	31.2	1.89
		3D-aware[22]	RGBD	RedNet[129]	有限数据集	2023	74.5	42.1	**1.16**
HM3D	端到端	DD-PPO[103]	RGBD	ResNet	有限数据集	2020	26.0	12.0	—
		Habitat-Web[14]	RGBD	ResNet18	有限数据集	2022	55.0	22.0	—
		OVRL[[108]	RGBD	ResNet50	增强数据集	2022	60.0	27.0	—
		ProcTHOR[47]	RGB	CNN	增强数据集	2022	54.0	32.0	—
		PIRLNav[15]	RGB	ResNet50	有限数据集	2023	62.2	28.7	—
	模块化	PEANUT[41]	RGBD	PSPNet[132]	有限数据集	2023	**64.0**	**33.0**	—
		ESC[113]	RGBD	GLIP[133]	视觉语言模型	2023	44.0	25.2	—

表 8 AI2-THOR 数据集上的 ObjectNav 性能比较

模型结构	方法	视觉数据	视觉模块	策略类型	年份	所有步数		步数大于等于5	
						SR (%)↑	*SPL* (%)↑	SR (%)↑	*SPL* (%)↑
端到端	SAVN[39]	RGB	ResNet18	元强化学习	2019	40.9	16.2	28.7	23.5
	Li等人[104]	RGB	—	元强化学习	2020	27.7	11.5	20.6	8.0
	ORG+TPN[36]	RGB	Faster R-CNN[135]	有限数据集	2020	69.3	39.4	60.7	38.6
	Ye等人[31]	RGBD	ResNet18	有限数据集	2021	74.0	46.0	—	—
	Mayo等人[34]	RGB	ResNet18	有限数据集	2021	46.2	17.9	32.6	16.0
	NIE[105]	RGBD	Mask R-CNN[130]	有限数据集	2021	80.0	31.3	—	—
	HOZ[32]	RGB	ResNet18	有限数据集	2021	70.6	40.0	62.8	39.2
	Li等人[49]	RGB	—	元强化学习	2022	71.0	19.6	62.0	24.2
	OMT[66]	RGB	ResNet50	有限数据集	2022	71.1	26.7	—	—
	TransNav[48]	RGB	Faster R-CNN[135]	元强化学习	2022	76.2	46.4	68.3	40.2
	GMAN[40]	RGB	ResNet18	元强化学习	2022	48.8	25.1	—	—
	L-sTDE[38]	RGB	Mask R-CNN[130]	有限数据集	2023	75.1	41.5	—	—
	HiNL[37]	RGB	ResNet	有限数据集	2023	**80.1**	**49.8**	**74.6**	**47.6**
模块化	MJOLNIR[42]	RGB	ResNet18	有限数据集	2020	65.3	21.1	50.0	20.9
	VTNET[65]	RGB	ResNet18	有限数据集	2021	72.2	44.9	63.4	44.0
	DOA[107]	RGB	DERT[136]	有限数据集	2021	74.3	40.3	67.9	40.4
	Dang等人[112]	RGB	ResNet18	有限数据集	2023	82.4	48.9	76.2	49.3
	ESC[113]	RGBD	GLIP[133]	视觉语言模型	2023	38.1	22.2	—	—
	Dang等人[16]	RGB	ResNet18	有限数据集	2023	**83.1**	**50.2**	**77.0**	**50.9**
	SHRL[116]	RGB	DERT[136]	有限数据集	2023	78.3	41.8	70.6	0.42

表 9 视觉物体重排布性能比较

类型	方法	视觉数据	视觉模块	年份	*fixed strict* (%)↑	SR (%)↑	*E* (%)↓	*M* (%)↓
一阶段	ResNet18, IL[13]	RGB	ResNet18	2021	6.00	3.0	111	109
	ResNet18+ANM, IL[13]	RGB	ResNet18	2021	9.00	3.0	105	104
	EmbCLIP[118]	RGB	CLIP[120]	2022	**17.00**	**8.00**	**89**	**88**
两阶段	ResNet18, PPO+IL[13]	RGB	ResNet18	2021	0.7	0.2	121	—
	ResNet18+ANM, PPO+IL[13]	RGB	ResNet18	2021	1.4	0.3	110	—
	CSR[78]	RGB	Fast R-CNN[80]	2022	1.9	0.4	117	—
	Trabucco等人[122]	RGB	—	2022	**16.56**	**4.70**	—	—
	TIDEE[123]	RGB	DERT[136]	2022	11.6	2.4	**93**	**94**
	Human[13]	—	—	2021	91.2	83.4	9	—

5.2 物体目标导航性能比较及分析

MultiON 为多目标导航任务提出了基准模型, 模型通过将视觉观测中的语义信息转化为 2D 语义地图指导具身智能体执行多物体目标导航任务. MultiON[25]提供了多种基准方案的对比, 包括不使用语义地图、使用先验地图 (oracle map)、使用具身智能体自主构建的全局地图 (ObjRecogMap) 等. 表 5 中列出了使用 ObjRecogMap 的实验结果, 成功率达到了 22.0%. Chen 等人[30]首次提出使用主动摄像机策略解决多目标物体导航问题, 相对于 MultiON 将成功率提高到 51.1%. 具体来说, Chen 等人通过将相机控制策略与导航策略相结合, 促使具身智能体自主调节相机视野方向, 以更全面的观察和探索环境. 最近, Marza 等人[29]通过采用一种新颖的神经隐式表示, 刷新了 SR、*SPL* 和 *PPL* 指标, 取得了最先进的多物体目标导航性能.

DD-PPO[103]是一种分散的分布式近端策略优化 (decentralized distributed proximal policy optimization) 算法. 在巨量计算资源和分布式强化学习的加持下, DD-PPO 近乎完美地解决了点导航问题. 但在 ObjectNav 任务中 DD-PPO 仅取得了 8.0% (MP3D 验证集) 和 0.0% (MP3D 测试集) 的成功率. 如表 6 中的第 1–2 行所示, DD-PPO 和随机导航动作 (random) 被选作 ObjectNav 基准. 相比于 Random、DD-PPO、THDA[46]、ZSON[53]和 OVRL[108], Red-Rabbit[31]通过添加辅助任务和探索奖励以端到端的方式学习一个更通用的 ObjectNav 策略, 在 MP3D 的验证集上, 将端到端的 ObjectNav 成功率提高到 34.6%. 值得注意的是, Red-Rabbit 只在有限数据集上学习导航策略, 只采用了卷积层更少的预训练的 ResNet18 作为视觉编码器. Habitat-Web[14]通过收集人类专家演示进行模仿学习, 将 ObjectNav 成功率提升至 35.4%. Red-Rabbit 和 Habitat-Web 在 MP3D 测试集上的性能指标也处于领先地位.

对于模块化的 ObjectNav 策略, Chaplot 等人[45]提出的 SemExp 策略通过构建基于语义地图的场景表示, 在 MP3D 验证集上, 将 ObjectNav 成功率提高到 36.0%, 相比 DD-PPO 提升了 4.5 倍. 但是, SemExp 采用了性能较强的预训练的 Mask R-CNN[130]作为视觉编码器. Stubborn[111]针对 SemExp 无法有效利用语义线索和容易陷入局部陷阱的问题, 维护了一种多维度的障碍物地图和一种不依赖语义的探索策略, 在 MP3D 测试集上, 相比于 SemExp 策略将成功率提高到 23.7%. SSCNav[18]是典型的基于局部语义地图场景表示的模块化 ObjectNav 策略, 采用了语义场景补全和语义置信度预测等模块, 在 MP3D 验证集上取得了 27.1% 的成功率和 15.7% 的 *SPL* 指标. L2M[19]同样利用语义类别的不确定性来确定长期导航目标, 不同之处在于, L2M 通过集成多个模型的分歧来估计模型的不确定性, 并进一步平衡探索和利用, 在 MP3D 验证集上将成功率和 *SPL* 指标分别提升到了 39.1% 和 17.0%, 同时 L2M 取得了最低的 *DTS* 指标.

PONI[20]是典型的基于全局语义地图环境表示的模块化 ObjectNav 策略, PONI 相比对 SSCNav 在成功率方面提升了 0.7%, 但是 *SPL* 指标降低了 3.7%. 考虑到全局环境表示为 ObjectNav 提供更多的空间和语义线索, 3D-Aware[22]、PEANUT[41]和 ESC[113]沿用了 PONI 的框架, 将成功率和 *SPL* 指标分别提升到 40.5% 和 15.8%. 值得一提的是, 3D-aware 是第 1 个引入 3D 环境表示的 ObjectNav 策略. ESC 利用 GLIP[133]中的视觉编码器和大规模视觉语言模型的推理能力, 实现了零样本 (zero-shot) ObjectNav, 在 MP3D 验证集上达到了 36.1% 的成功率和最高的 17.7% 的 *SPL* 指标. 通过对表 6 中的实验数据进行总结, 不难发现模块化的 ObjectNav 策略在 MP3D 验证集上的

性能优于端到端的 ObjectNav 策略. 相反, 在 MP3D 测试集上, 端到端的 ObjectNav 策略的性能优于模块化的 ObjectNav 策略. 表 6 中的红色数据表示 GCExp[106]和 ReVoLT[114]策略的训练或评估只采用 MP3D 的部分场景, 各类方法所取得的最佳性能以粗体数据显示. 因此出现成功率、*SPL* 指标、*DTS* 指标和 *SoftSPL* 指标偏高或偏低的情况.

由于 Gibson 数据集相对于 MP3D 数据集具有较低的复杂度, 因此各种 ObjectNav 策略在 Gibson 数据集上的性能指标更高, 如表 7 所示. 同样以 DD-PPO 为基准, Li 等人[35]提出了基于 Transformer 的端到端的 ObjectNav 策略, 一方面利用多头注意力机制来捕获历史导航轨迹中不同时间步的长距离依赖关系, 另一方面采用辅助任务改善模型的训练, 在 Gibson 数据集上达到了最高的 82.8% 的成功率和最高的 48.7% 的 *SPL* 指标. 值得注意的是, Li 等人[35]没有采用任何预训练的视觉编码器. 对于模块化 ObjectNav 策略, 相比于 SemExp[45]、Al-Halah 等人[109]和 Min 等人[23]的工作, PONI[20]和 3D-aware[22]分别取得 73.6% 和 74.5% 的成功率. 此外, PONI 和 3D-aware 也取得了较高的 *SPL* 指标, 3D-aware 达到了最低的 *DTS* 指标, 仅为 1.16 m. PONI[20]和 3D-aware[22] 分别采用了 Mask R-CNN 和 RedNet 作为视觉编码器.

类似于在 MP3D 验证集上的表现, PEANUT[41]作为最新的模块化 ObjectNav 策略, 在 HM3D 数据集上也取得了最高的 64.0% 的成功率和最高的 33.0% 的 *SPL* 指标. 以 DD-PPO[103] 为基准, Habitat-Web[14]、OVRL[108]、ProcTHOR[47]和 PIRLNav[15]这 4 种模块化的 ObjectNav 策略都提升了至少 1 倍的性能, 具体来说, 成功率提升了 28.0%–36.2%, *SPL* 指标提升了 10.0%–20.0%. 其中, PIRLNav 采用了独特的训练范式, 通过将 IL 和 RL 调优相结合, 取得了 62.2% 的次最优成功率. 值得注意的是, PIRLNav 只采用 RGB 图像作为导航策略的输入, 并没有利用深度图像数据.

由于 AI2-THOR 数据集相对于其他几种数据集包含较少的场景类型, 各种 ObjectNav 策略在 AI2-THOR 上取得了较为可观的性能. 基于 AI2-THOR 数据集的 ObjectNav 策略普遍报告了两类评价指标 ($L\geqslant 5$): (1) 导航步数大于等于 5 步情况下的 SR 和 *SPL* 指标; (2) 所有步数情况下的 SR 和 *SPL* 指标 (All).

最早的基于元强学习的泛化学习策略 SAVN[39]是基于 AI-THOR 数据集提出的, 致力于提高 ObjectNav 策略在未见过的场景中的泛化能力. 考虑所有步数情况时, SAVN 分别取得了 40.9% 的 SR 和 16.2% 的 *SPL* 指标, 考虑步数大于等于 5 的情况时, SAVN 分别取得了 28.7% 的 SR 和 23.5% 的 *SPL* 指标. Li 等人[49]最新的工作延续了 SAVN 的思路, 通过将层次化语义信息与元学习相结合, 弥补了已知环境与未知环境的泛化性能差距. 与基于 SAVN 的思路不同, GMAN[40]提出了一种新颖的生成式元对抗网络, 致力于提高 ObjectNav 策略对于未见过的物体目标的泛化能力, 在 AI2-THOR 上的性能相较于 SAVN 有所提高. Du 等人[36]提出了一种分层强化学习方法, 引入了一个目标关系图 (ORG) 和一个试探策略网络 (TPN), 成功率达到了 69.3% (All) 和 60.7% ($L\geqslant 5$). 在最新的工作中, L-sTDE[38]通过提出一种基于布局的软总直接效应框架大幅提升了 HOZ 的导航性能, 取得了 75.1% (All) 的成功率指标和 41.5% 的 *SPL* 指标. HiNL[37]在 ORG+TPN 的基础上提出了一种历史启发的导航策略学习框架, 在 AI-THOR 数据集上取得了最佳的物体目标导航性能.

考虑到人类在陌生环境中通常依赖历史经验搜寻目标物体, 研究者开始尝试将人类先验知识融合到导航策略学习网络中. 作为模块化的 ObjectNav 策略, MJOLNIR[42]通过从外部数据库提取描述场景先验的知识图谱, 学习环境中物体-物体之间的空间和语义关系, 在 AI2-THOR 上取得了 65.3% (All) 和 50.0% ($L\geqslant 5$) 的成功率. HOZ[32]作为一种端到端的 ObjectNav 策略, 则进一步考虑了空间-空间, 空间-物体, 物体-物体多层次空间和语义关系, 通过维护一个分层次语义关系图, 引导具身智能体进行由粗到细的进行分层次目标物体探索. HOZ 的成功率相比于 MJOLNIR 提升了 5.3% (All) 和 12.8% ($L\geqslant 5$). DOA[107]提出了一种有向物体注意力图来指导具身智能体显式地学习物体之间的注意力关系并关注正确的物体, 取得了优异的 SR 和 *SPL* 指标.

利用历史经验辅助当前时刻的动作决策, 是提高导航效率和避免重复探索的重要手段. 考虑到 Transformer 在捕获长距离依赖特征方面的优势, VTNET[65], OMT[66], TransNav[48]均基于 Transformer 提出了利用经验信息的 ObjectNav 策略. 通过综合利用视觉观察的局部空间特征和全局空间特征, VTNET 学习更强大的视觉表示. OMT 结合目标物体类别关注历史经验中有价值的信息, 从而指导具身智能体在没有先验知识的情况下在室内环境中完

成导航. TransNav 考虑了导航过程中局部和全局视觉信息, 以及时间序列特征. 考虑所有步数情况时, 这 3 种方法的在 AI2-THOR 上的成功率均超过了 70%, 考虑步数大于等于 5 的情况时, 这 3 种方法的在 AI2-THOR 上的成功率均超过了 60%.

Dang 等人[112]早期的工作将 ObjectNav 过程分解为“搜索”和“导航”两种思维, 提出在不同的导航阶段分别灵活使用不同的思维策略. 最近, Dang 等人[16]进一步将 ObjectNav 解耦为 5 种元思维: 感知思维, 搜索思维, 导航思维, 探索思维和避障思维, 通过多重思维协作模块来促进思维之间的相互协作. 如表 8 所示, Dang 等人的最新工作[16]在 AI2-THOR 上取得了最先进的性能. 具体来说, 考虑所有步数情况时, 分别取得了 83.1% 的 SR 和 50.2% 的 *SPL* 指标, 考虑步数大于等于 5 的情况时, 分别取得了 77.0% 的 SR 和 50.9% 的 *SPL* 指标.

从导航策略提取视觉特征的角度来看, Ye 等人[31]和 NIE[105]两种端到端的方法均采用了 RGBD 视觉观测作为模型输入, 取得了较好的导航性能. 不同于 Ye 等人[31]采用的基于 ResNet18 的视觉编码器, NIE[105]采用了 Mask-RCNN 作为视觉编码器. 然而, 同样作为端到端的方法的 HiNL[37]只采用了 RGB 图像和基于 ResNet 的视觉编码器, 在 AI-THOR 数据集上取得了更好的导航性能, 这体现了 HiNL 的历史启发的导航策略学习框架的优势. 在模块化的导航策略中, Dang 等人[16]同样只采用了 RGB 图像和基于 ResNet18 的视觉编码器, 在 AI-THOR 数据集上取得了最佳的物体目标导航性能, 这体现了通过构建多重思维协作模块来促进思维之间的相互协作的重要性.

5.3 视觉物体重排布性能比较及分析

视觉物体重排布任务由 Weihs 等人[13]于 2021 年提出, 他们同时发布了 RoomR 数据集和一阶段、两阶段物体重排布基线方法. 表 9 中的列出了两种具有代表性的基线方法——ResNet18, IL 和 ResNet18+ANM, IL. ResNet18, IL 表示采用 ResNet18 进行图像特征提取, 采用纯粹地通过模仿学习来训练具身智能体解决一阶段和两阶段物体重排布任务. ResNet18+ANM, IL 首先基于 AI-THOR 数据集预训练一个 ANS 的变体用于构建语义地图, 然后冻结 ANS 并通过模仿学习来训练具身智能体. 从各项指标来看, 基线方法在一阶段和两阶段任务上的 *fixed strict* 指标分别仅有 6%–9% 和 0.7%–1.4%, 成功率分别仅有 3% 和 0.2%–0.3%. 由此可见, 两阶段任务相较于一阶段任务难度更大. 此外, 较高的 *E* 和 *M* 指标也反映基线方法的重排布错误率较高, 基本不能够解决物体重排布问题. 表 9 的最后一行列出了由人类完成重排布任务的各项数据指标.

针对两阶段重排布任务, Trabucco 等人[122]分别基于目标状态和当前状态构建精准的 3D 语义地图, 用于错位物体检测, 在 RoomR 数据集上的 *fixed strict* 指标达到了 16.56%, 成功率达到了 4.7%, 相较于 ResNet18+ANM IL 基线分别提高了 15.86% 和 4.3%. EmbCLIP[121]利用 CLIP 模型学习视觉表示, 提升对目标状态和当前状态的理解能力, 更容易找到需要重排布的目标物体, 在一阶段任务中取得了当前最佳性能. TIDDE[119]则无需提前学习目标状态, 基于先验知识在探索过程中实时检测错位物体, 并及时进行修正错误. 在两阶段任务中, TIDDE 的性能仍明显优于基线模型, 证明了 TIDDE 具备一定的自适应能力.

相对于 ObjectNav 而言, 视觉物体重排布是一个更年轻、更前沿、更具有实际应用价值的研究方向, 正受到越来越多研究者的关注. 从表 9 列出的实验数据来看, EmbCLIP 所取得的 17% *fixed strict* 指标和 8% 的 SR 指标与人类取得的性能相差甚远, 视觉物体重排布相关技术的发展道阻且长.

6 总结与前景展望

本文讨论了物体目标导航对于具身人工智能发展的重要性, 广泛地回顾了现有的物体目标导航方法, 并从多个角度对其进行梳理和分析. 本文还描述了物体目标导航相关的前置和后置任务, 在介绍性的层面上讨论了这些问题. 回顾本文, 得知物体目标导航及其相关领域已经取得的巨大进展, 并衍生了新的研究课题和可以扩展的方向.

随着大规模视觉语言模型的兴起, 物体目标导航方法开始朝着明确引入外部知识的方向发展, 整合外部知识和人类偏好不仅有利于提高导航性能, 还可以提高人工智能的可解释性和可信度. 然而, 现有的物体目标导航研究普遍聚焦于学习导航策略, 但具身导航的“最后一公里问题”是如何与物体交互, 还没有被很好地研究和讨论, 比如“拿起一把勺子”. 此外, 用于导航策略学习的场景也缺乏多样性. 现有的绝大多数场景数据集的构建基于对美国房屋的图像扫描, 这些场景不包括仓库和医院等有意义的场景.

对于物体目标导航未来可能的发展方向, 我们着重讨论以下几个方面.

(1) 基于外部知识和大模型的物体目标导航: 一方面, 外部常识知识 (例如维基百科中对一般房屋和物体目标的描述) 的引入能够进一步提高物体目标导航的可解释性和可信度[24,42]. 另一方面, 以 ChatGPT 和 CLIP 为代表的大模型的兴起和流行为物体目标导航相关领域的发展提供了更多的可能性[52,53,113,115,121,137–143]. 作为首个基于 CLIP 的零样本物体目标导航方法, CoW[52]充分利用 CLIP 视觉编码器中内化的丰富视觉特征来显式地推断目标物体的方位. 一些方法[53,121]通过对 CLIP 的视觉-语言编码器进行微调, 增强具身智能体对于新环境和新物体目标的泛化能力. 考虑到人类在导航时不是被动地接受所有的视觉刺激, 而是主动调节并选择性地处理与当前任务相关的视觉特征. 基于这一观点, Eftekhar 等人[138]提出了一种代码本模块, 选择性地从 CLIP 视觉编码器中抽取有利于具身导航的视觉表示. 因此, 如何高效地从大模型中提取特定于导航任务的视觉特征并加以合理利用, 是一项有意义的研究课题.

基于大语言模型的物体目标导航方法[140–142]通常采用精心设计的提示来激活大语言模型的规划和决策能力. 得益于大语言模型的常识知识和强大的推理能力, 具身智能体能够在未知的场景中寻找任何物体. 尽管大语言模型能够在样本数据匮乏的情况下泛化到新的导航任务, 但是仍然存在许多技术和理论挑战[143]: 例如如何基于文本、图像和其他传感器数据进行多模态导航决策, 解决大语言模型与物理世界脱节的问题; 如何减少基于大语言模型的具身智能体的决策延迟. 未来, 大语言模型将持续助力具身人工智能的发展与变革[144].

(2) 人机物三元融合的物体目标导航: 现有的绝大多数物体目标导航策略都是针对静态环境设计的, 主要考虑了具身智能体 (机器人) 和物体两类元素. 在未来, 具身智能体的导航可能与人类的生活密切相关. 由于人类的运动和彼此之间的交互, 现实中的环境往往是动态的, 并且包含大量的、复杂的人机物三元交互. 因此, 具身智能体必须在导航的过程中遵从社会意识, 在不侵犯人类安全空间的前提下舒适地导航. 尽管已有最新的工作对此展开研究[145–147], 人机物三元融合的物体目标导航仍是一个开放的、有待深入研究的课题.

(3) 从仿真到现实的转移: 目前, 绝大多数物体目标导航及其相关任务的研究都是基于仿真环境开展的. 一方面, 仿真器渲染的视觉图像与现实环境中的视觉观测存在光照、纹理等方面的差异. ObjectNav 策略对仿真图像的过度拟合导致其难以泛化到现实环境中. 虽然已有工作采用元强化学习来改善具身智能体的视觉泛化能力[148,149], 但是距离实际应用还有很大的差距. 另一方面, 现有的 ObjectNav 方法普遍假设具身智能体的动作空间和状态空间是离散的, 以降低该任务的研究难度. 然而, 假设具身智能体只具有前进、左转、右转、停止几个离散动作, 不能保证其在复杂的现实环境中连续、高效地运动.

(4) 可解释性导航策略: 现有的物体目标导航方法往往采用深度神经网络模型建模导航所需的多种能力或思维, 通常是抽象和难以解释的. 尽管最近有工作试图对具身智能体的导航行为进行解释[150,151], 但是如何设计 ObjectNav 策略以具体化探索、导航、避障等思维还有待进一步研究.

(5) 物体目标导航的“最后一公里”问题: 当前的物体目标导航相关研究停留在“导航到目标物体附近”阶段, 还没有充分涉及与目标物体的交互. 如何使具身智能体学会与环境中的物体进行交互是物体目标导航发展的下一个目标. 比如我们可能会让机器人去厨房拿一个勺子, 前提是我们要让机器人学会拿起勺子. 最近, 在大语言模型的加持下, 具身智能的倡导者们针对机器人导航和抓取任务提出了相应的具身大模型[152,153]. 未来, 具身大模型可能在解决物体目标导航和抓取问题方面大显身手.

(6) 学习环境的多样性: 现有的绝大多数场景数据集是基于美国的房屋构建的, 而且并没有包含仓库或医院等有意义的、具身智能体应用广泛应用的场景. 由于大多数数据集来自国外的建筑, 考虑到不同的文化差异, 基于中式建筑的数据集有必要被构建和发布. 尽管新的场景数据集被陆续提出以促进物体目标导航的发展[154], 但是多样化的学习环境仍然需要被进一步考虑和深入研究.

(7) 隐私保护: 在物体目标导航任务训练和推理的过程中, 具身智能体可以观察和存储敏感信息, 这些信息可能会被泄露或滥用, 因此导航过程中有效的隐私保护是至关重要的. 因此, 可以将联邦学习[155]和差分隐私[139]等相关领域的技术与 ObjectNav 任务进行交叉, 以保护训练和推理过程中的隐私.

References:

[1] Deitke M, Batra D, Bisk Y, *et al*. Retrospectives on the embodied AI workshop. arXiv:2210.06849, 2022.

[2] Liu HP, Guo D, Sun FC, Zhang XY. Morphology-based embodied intelligence: Historical retrospect and research progress. Acta Automatica Sinica, 2023, 49(6): 1131–1154 (in Chinese with English abstract). [doi: 10.16383/j.aas.c220564]

[3] Sima SL, Huang Y, He KJ, An D, Yuan H, Wang L. Recent advances in vision-and-language navigation. Acta Automatica Sinica, 2023, 49(1): 1–14 (in Chinese with English abstract). [doi: 10.16383/j.aas.c210352]

[4] Chang A, Dai A, Funkhouser T, Halber M, Niebner M, Savva M, Song SR, Zeng A, Zhang YD. Matterport3D: Learning from RGB-D data in indoor environments. In: Proc. of the 2017 Int'l Conf. on 3D Vision (3DV). Qingdao: IEEE, 2017. 667−676. [doi: 10.1109/3DV.2017.00081]

[5] Xia F, Zamir AR, He ZY, Sax A, Malik J, Savarese S. Gibson env: Real-world perception for embodied agents. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9068–9079. [doi: 10.1109/CVPR.2018.00945]

[6] Ramakrishnan SK, Gokaslan A, Wijmans E, Maksymets O, Clegg A, Turner J, Undersander E, Galuba W, Westbury A, Chang A, Savva M, Zhao YL, Batra D. Habitat-Matterport 3D dataset (HM3D): 1 000 large-scale 3D environments for embodied AI. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. NIPS, 2021.

[7] Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, Deitke M, Ehsani K, Gordon D, Zhu YK, Kembhavi A, Gupta A, Farhadi A. AI2-THOR: An interactive 3D environment for visual AI. arXiv:1712.05474, 2017.

[8] Shen BK, Xia F, Li CS, Martín-Martín R, Fan LX, Wang GZ, Pérez-D'Arpino C, Buch S, Srivastava S, Tchapmi L, Tchapmi M, Vainio K, Wong J, Fei-Fei L, Savarese S. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In: Proc. of the 2021 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Prague: IEEE, 2021. 7520–7527. [doi: 10.1109/IROS51168.2021.9636667]

[9] Li CS, Xia F, Martín-Martín R, Lingelbach M, Srivastava S, Shen BK, Vainio KE, Gokmen C, Dharan G, Jain T, Kurenkov A, Liu KR, Gweon H, Wu JJ, Fei-Fei L, Savares S. iGibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In: Proc. of the 5th Conf. on Robot Learning. London: PMLR, 2022. 455–465.

[10] Savva M, Kadian A, Maksymets O, Savva M, Kadian A, Maksymets O, Zhao YL, Wijmans E, Jain B, Straub J, Liu J, Koltun V, Malik J, Parikh D, Batra Dhruv. Habitat: A platform for embodied AI research. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9338–9346. [doi: 10.1109/ICCV.2019.00943]

[11] Szot A, Clegg A, Undersander E, Wijmans E, Zhao YL, Turner J, Maestre N, Mukadam M, Chaplot D, Maksymets O, Gokaslan A, Vondrus V, Dharur S, Meier F, Galuba W, Chang A, Kira Z, Koltun V, Malik J, Savva M, Batra D. Habitat 2.0: Training home assistants to rearrange their habitat. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 20.

[12] Batra D, Gokaslan A, Kembhavi A, Maksymets O, Mottaghi R, Savva M, Toshev A, Wijmans E. ObjectNav revisited: On evaluation of embodied agents navigating to objects. arXiv:2006.13171, 2020.

[13] Weihs L, Deitke M, Kembhavi A, Mottaghi R. Visual room rearrangement. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5918–5927. [doi: 10.1109/CVPR46437.2021.00586]

[14] Ramrakhya R, Undersander E, Batra D, Das A. Habitat-Web: Learning embodied object-search strategies from human demonstrations at scale. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5163–5173. [doi: 10.1109/CVPR52688.2022.00511]

[15] Ramrakhya R, Batra D, Wijmans E, Das A. PIRLNav: Pretraining with imitation and RL finetuning for ObjectNav. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 17896–17906. [doi: 10.1109/CVPR52729.2023.01716]

[16] Dang RH, Chen L, Wang LY, He ZT, Liu CJ, Chen QJ. Multiple thinking achieving meta-ability decoupling for object navigation. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: PMLR, 2023. 6855–6872.

[17] Gervet T, Chintala S, Batra D, Malik J, Chaplot DS. Navigating to objects in the real world. Science Robotics, 2023, 8(79): eadf6991. [doi: 10.1126/scirobotics.adf6991]

[18] Liang YQ, Chen BY, Song SR. SSCNav: Confidence-aware semantic scene completion for visual semantic navigation. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation. Xi'an: IEEE, 2021. 13194–13200. [doi: 10.1109/ICRA48506.2021.9560925]

[19] Georgakis G, Bucher B, Schmeckpeper K, Singh S, Daniilidis K. Learning to map for active semantic goal navigation. In: Proc. of the 10th Int'l Conf. on Learning Representations. ICLR, 2022.

[20] Ramakrishnan SK, Chaplot DS, Al-Halah Z, Malik J, Grauman K. PONI: Potential functions for objectgoal navigation with interaction-

free learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 18868–18878. [doi: 10.1109/CVPR52688.2022.01832]

[21] Zhu MZ, Zhao BL, Kong T. Navigating to objects in unseen environments by distance prediction. In: Proc. of the 2022 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Kyoto: IEEE, 2022. 10571–10578. [doi: 10.1109/IROS47612.2022.9981766]

[22] Zhang JZ, Dai L, Meng FP, Fan QN, Chen XL, Xu K, Wang H. 3D-aware object goal navigation via simultaneous exploration and identification. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 6672–6682. [doi: 10.1109/CVPR52729.2023.00645]

[23] Min SY, Tsai YHH, Ding W, Farhadi A, Salakhutdinov R, Bisk Y, Zhang J. Object goal navigation with end-to-end self-supervision. arXiv:2212.05923, 2022.

[24] Yang W, Wang XL, Farhadi A, Gupta A, Mottaghi R. Visual semantic navigation using scene priors. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019.

[25] Wani S, Patel S, Jain U, Chang AX, Savva M. MultiON: Benchmarking semantic map memory using multi-object navigation. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 813.

[26] Raychaudhuri S, Campari T, Jain U, Savva M, Chang AX. Reduce, reuse, recycle: Modular multi-object navigation. arXiv:2304.03696, 2023.

[27] Sadek A, Bono G, Chidlovskii B, Baskurt A, Wolf C. Multi-object navigation in real environments using hybrid policies. In: Proc. of the 2023 IEEE Int'l Conf. on Robotics and Automation (ICRA). London: IEEE, 2023. 4085–4091. [doi: 10.1109/ICRA48891.2023.10161030]

[28] Zeng HT, Song XH, Jiang SQ. Multi-object navigation using potential target position policy function. IEEE Trans. on Image Processing, 2023, 32: 2608–2619. [doi: 10.1109/TIP.2023.3263110]

[29] Marza P, Matignon L, Simonin O, Wolf C. Multi-object navigation with dynamically learned neural implicit representations. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 10970–10981. [doi: 10.1109/ICCV51070.2023.01010]

[30] Chen PH, Ji DY, Lin KY, Hu WW, Huang WB, Li TH, Tan MK, Gan C. Learning active camera for multi-object navigation. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2078.

[31] Ye J, Batra D, Das A, Wijmans E. Auxiliary tasks and exploration enable objectgoal navigation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 16097–16106. [doi: 10.1109/ICCV48922.2021.01581]

[32] Zhang SX, Song XH, Bai YB, Li WJ, Chu YK, Jiang SQ. Hierarchical object-to-zone graph for object navigation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 15110–15120. [doi: 10.1109/ICCV48922.2021.01485]

[33] Li WJ, Song XH, Bai YB, Zhang SX, Jiang SQ. ION: Instance-level object navigation. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 4343–4352. [doi: 10.1145/3474085.3475575]

[34] Mayo B, Hazan T, Tal A. Visual navigation with spatial attention. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16893–16902. [doi: 10.1109/CVPR46437.2021.01662]

[35] Li WY, Hong RX, Shen JW, Yuan L, Lu Y. Transformer memory for interactive visual navigation in cluttered environments. IEEE Robotics and Automation Letters, 2023, 8(3): 1731–1738. [doi: 10.1109/LRA.2023.3241803]

[36] Du HM, Yu X, Zheng L. Learning object relation graph and tentative policy for visual navigation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 19–34. [doi: 10.1007/978-3-030-58571-6_2]

[37] Du HM, Li LC, Huang Z, Yu X. Object-goal visual navigation via effective exploration of relations among historical navigation states. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 2563–2573. [doi: 10.1109/CVPR52729.2023.00252]

[38] Zhang SX, Song XH, Li WJ, Bai YB, Yu XY, Jiang SQ. Layout-based causal inference for object navigation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 10792–10802. [doi: 10.1109/CVPR52729.2023.01039]

[39] Wortsman M, Ehsani K, Rastegari M, Farhadi A, Mottaghi R. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6743–6752. [doi: 10.1109/CVPR.2019.00691]

[40] Zhang SX, Li WJ, Song XH, Bai YB, Jiang SQ. Generative meta-adversarial network for unseen object navigation. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 301–320. [doi: 10.1007/978-3-031-19842-7_18]

[41] Zhai A, Wang SL. PEANUT: Predicting and navigating to unseen targets. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 10892–10901. [doi: 10.1109/ICCV51070.2023.01003]

[42] Pal A, Qiu YD, Christensen H. Learning hierarchical relationships for object-goal navigation. In: Proc. of the 2020 Conf. on Robot Learning. Cambridge: PMLR, 2021. 517–528.

[43] Mousavian A, Toshev A, Fišer M, Košecká J, Wahid A, Davidson J. Visual representations for semantic target driven navigation. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation. Montreal: IEEE, 2019. 8846–8852. [doi: 10.1109/ICRA.2019.8793493]

[44] Wu Y, Wu YX, Tamar A, Russell S, Gkioxari G, Tian YD. Bayesian relational memory for semantic visual navigation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2769–2779. [doi: 10.1109/ICCV.2019.00286]

[45] Chaplot DS, Gandhi D, Gupta A, Salakhutdinov R. Object goal navigation using goal-oriented semantic exploration. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 357.

[46] Maksymets O, Cartillier V, Gokaslan A, Wijmans E, Galuba W, Lee S, Batra D. THDA: Treasure hunt data augmentation for semantic navigation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 15354–15363. [doi: 10.1109/ICCV48922.2021.01509]

[47] Deitke M, Vander Bilt E, Herrasti A, Weihs L, Salvador J, Ehsani K, Han W, Kolve E, Farhadi A, Kembhavi A, Mottaghi R. ProcTHOR: Large-scale embodied AI using procedural generation. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 433.

[48] Zhou K, Zhang HY, Li F. TransNav: Spatial sequential Transformer network for visual navigation. Journal of Computational Design and Engineering, 2022, 9(5): 1866–1878. [doi: 10.1093/jcde/qwac084]

[49] Li F, Guo C, Zhang HY, Luo BH. Context vector-based visual mapless navigation in indoor using hierarchical semantic information and meta-learning. Complex & Intelligent Systems, 2023, 9(2): 2031–2041. [doi: 10.1007/s40747-022-00902-7]

[50] Yin J, Zhang ZD, Gao YH, Yang ZW, Li L, Xiao M, Sun YQ, Yan CG. Survey on vision-language pre-training. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2000–2023 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[51] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[52] Gadre SY, Wortsman M, Ilharco G, Schmidt L, Song SR. CoWs on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 23171–23181. [doi: 10.1109/CVPR52729.2023.02219]

[53] Majumdar A, Aggarwal G, Devnani B, Hoffman J, Batra D. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2343.

[54] Ramakrishnan SK, Jayaraman D, Grauman K. An exploration of embodied visual exploration. Int'l Journal of Computer Vision, 2021, 129(5): 1616–1649. [doi: 10.1007/s11263-021-01437-z]

[55] Zhang TY, Hu XG, Xiao J, Zhang GF. A survey of visual navigation: From geometry to embodied AI. Engineering Applications of Artificial Intelligence, 2022, 114: 105036. [doi: 10.1016/j.engappai.2022.105036]

[56] Duan JF, Yu S, Tan HL, Zhu HY, Tan C. A survey of embodied AI: From simulators to research tasks. IEEE Trans. on Emerging Topics in Computational Intelligence, 2022, 6(2): 230–244. [doi: 10.1109/TETCI.2022.3141105]

[57] Gu J, Stefani E, Wu Q, Thomason J, Wang X. Vision-and-language navigation: A survey of tasks, methods, and future directions. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin: ACL, 2022. 7606–7623. [doi: 10.18653/v1/2022.acl-long.524]

[58] Cao C, Zhu H, Ren Z, Choset H, Zhang J. Representation granularity enables time-efficient autonomous exploration in large, complex worlds. Science Robotics, 2023, 8(80): eadf0970. [doi: 10.1126/scirobotics.adf0970]

[59] Garaffa LC, Basso M, Konzen AA, de Freitas EP. Reinforcement learning for mobile robotics exploration: A survey. IEEE Trans. on Neural Networks and Learning Systems, 2023, 34(8): 3796–3810. [doi: 10.1109/TNNLS.2021.3124466]

[60] Wang L, Qi Y, He BB, Zhang YJ, Xu YC. Survey of autonomous exploration algorithms for robots. Journal of Computer Applications, 2023, 43(S1): 314–322 (in Chinese with English abstract). [doi: 10.11772/j.issn.1001-9081.2022111706]

[61] Zhang SY, Zhang XB, Yuan J, Fang YC. A survey on coverage and exploration path planning with multi-rotor micro aerial vehicles. Control and Decision, 2022, 37(3): 513–529 (in Chinese with English abstract). [doi: 10.13195/j.kzyjc.2021.1751]

[62] Fang K, Toshev A, Fei-Fei L, Savarese S. Scene memory Transformer for embodied agents in long-horizon tasks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 538–547. [doi: 10.1109/CVPR.2019.00063]

[63] Fortunato M, Tan M, Faulkner R, Hansen S, Badia AP, Buttimore G, Deck C, Leibo JZ, Blundell C. Generalization of reinforcement learners with working and episodic memory. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1117.

[64] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[65] Du HM, Yu X, Zheng L. VTNet: Visual Transformer network for object goal navigation. In: Proc. of the 9th Int'l Conf. on Learning Representations. ICLR, 2021.

[66] Fukushima R, Ota K, Kanezaki A, Sasaki Y, Yoshiyasu Y. Object memory Transformer for object goal navigation. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation. Philadelphia: IEEE, 2022. 11288–11294. [doi: 10.1109/ICRA46639.2022.9812027]

[67] Georgakis G, Schmeckpeper K, Wanchoo K, Dan S, Miltsakaki E, Roth D, Daniilidis K. Cross-modal map learning for vision and language navigation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15439–15449. [doi: 10.1109/CVPR52688.2022.01502]

[68] Henriques JF, Vedaldi A. MapNet: An allocentric spatial memory for mapping environments. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8476–8484. [doi: 10.1109/CVPR.2018.00884]

[69] Cartillier V, Ren ZL, Jain N, Lee S, Essa I, Batra D. Semantic MapNet: Building allocentric semantic maps and representations from egocentric views. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 964–972. [doi: 10.1609/aaai.v35i2.16180]

[70] Chen PH, Ji DY, Lin KY, Zeng RH, Li TH, Tan MK, Gan C. Weakly-supervised multi-granularity map learning for vision-and-language navigation. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2764.

[71] Xu DF, Zhu YK, Choy CB, Fei-Fei L. Scene graph generation by iterative message passing. In: Proc. of the 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3097–3106. [doi: 10.1109/CVPR.2017.330]

[72] Yang JW, Lu JS, Lee S, Batra D, Parikh D. Graph R-CNN for scene graph generation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 690–706. [doi: 10.1007/978-3-030-01246-5_41]

[73] Zellers R, Yatskar M, Thomson S, Choi Y. Neural motifs: Scene graph parsing with global context. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5831–5840. [doi: 10.1109/CVPR.2018.00611]

[74] Ost J, Mannan F, Thuerey N, Knodt J, Heide F. Neural scene graphs for dynamic scenes. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2855–2864. [doi: 10.1109/CVPR46437.2021.00288]

[75] Tsai YHH, Divvala S, Morency LP, Salakhutdinov R, Farhadi A. Video relationship reasoning using gated spatio-temporal energy graph. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10416–10425. [doi: 10.1109/CVPR.2019.01067]

[76] Giuliari F, Skenderi G, Cristani M, Wang YM, Del Bue A. Spatial commonsense graph for object localisation in partial scenes. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 19496–19505. [doi: 10.1109/CVPR52688.2022.01891]

[77] Gao C, Chen JY, Liu S, Wang LT, Zhang Q, Wu Q. Room-and-object aware knowledge reasoning for remote embodied referring expression. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3063–3072. [doi: 10.1109/CVPR46437.2021.00308]

[78] Gadre SY, Ehsani K, Song SR, Mottaghi R. Continuous scene representations for embodied AI. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 14829–14839. [doi: 10.1109/CVPR52688.2022.01443]

[79] Du YL, Gan C, Isola P. Curious representation learning for embodied intelligence. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 10388–10397. [doi: 10.1109/ICCV48922.2021.01024]

[80] Girshick R. Fast R-CNN. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1440–1448. [doi: 10.1109/ICCV.2015.169]

[81] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

[82] Zhu H, Kapoor R, Min SY, Han W, Li JT, Geng KW, Neubig G, Bisk Y, Kembhavi A, Weihs L. EXCALIBUR: Encouraging and evaluating embodied exploration. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14931–14942. [doi: 10.1109/CVPR52729.2023.01434]

[83] Chaplot DS, Gandhi D, Gupta S, Gupta A, Salakhutdinov R. Learning to explore using active neural SLAM. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: ICLR, 2019.

[84] Bigazzi R, Cornia M, Cascianelli S, Baraldi L, Cucchiara R. Embodied agents for efficient exploration and smart scene description. In: Proc. of the 2023 IEEE Int'l Conf. on Robotics and Automation (ICRA). London: IEEE, 2023. 6057–6064. [doi: 10.1109/ICRA48891.2023.10160668]

[85] Savinov N, Raichuk A, Vincent D, Marinier R, Pollefeys M, Lillicrap T, Gelly S. Episodic curiosity through reachability. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019.

[86] Strehl AL, Littman ML. An analysis of model-based interval estimation for Markov decision processes. Journal of Computer and System Sciences, 2008, 74(8): 1309–1331. [doi: 10.1016/j.jcss.2007.08.009]

[87] Bellemare MG, Srinivasan S, Ostrovski G, Schaul T, Saxton D, Munos R. Unifying count-based exploration and intrinsic motivation. In:

Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 1479–1487.

[88] Ostrovski G, Bellemare MG, van den Oord A, Munos R. Count-based exploration with neural density models. In: Proc. of the 34th Int'l Conf. Machine Learning. Sydney: IEEE, 2017. 2721–2730.

[89] Tang HR, Houthooft R, Foote D, Stooke A, Chen X, Duan Y, Schulman J, De Turck F, Abbeel P. #Exploration: A study of count-based exploration for deep reinforcement learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 2750–2759.

[90] Pathak D, Agrawal P, Efros AA, Darrell T. Curiosity-driven exploration by self-supervised prediction. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR, 2017. 2778–2787.

[91] Burda Y, Edwards H, Pathak D, Storkey A, Darrell T, Efros AA. Large-scale study of curiosity-driven learning. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019.

[92] Pathak D, Gandhi D, Gupta A. Self-supervised exploration via disagreement. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5062–5071.

[93] Chen T, Gupta S, Gupta A. Learning exploration policies for navigation. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019.

[94] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A. The kinetics human action video dataset. arXiv:1705.06950, 2017.

[95] Ramakrishnan SK, Jayaraman D, Grauman K. Emergence of exploratory look-around behaviors through active observation completion. Science Robotics, 2019, 4(30): eaaw6326. [doi: 10.1126/scirobotics.aaw6326]

[96] Jayaraman D, Grauman K. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1238–1247. [doi: 10.1109/CVPR.2018.00135]

[97] Isola P, Zhu JY, Zhou TH, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5967–5976. [doi: 10.1109/CVPR.2017.632]

[98] Cassandra AR, Kaelbling LP, Kurien JA. Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation. In: Proc. of the 1996 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Osaka: IEEE, 1996. 963–972. [doi: 10.1109/IROS.1996.571080]

[99] Burgard W, Stachniss C, Grisetti G. Information gain-based exploration using rao-blackwellized particle filters. In: Proc. of the 2005 Int'l Conf. on Robotics: Science and Systems. Cambridge: The MIT Press, 2005. 65–72.

[100] Sun Y, Gomez F, Schmidhuber J. Planning to be surprised: Optimal Bayesian exploration in dynamic environments. In: Proc. of the 4th Int'l Conf. on Artificial General Intelligence. Mountain View: Springer, 2011. 41–51. [doi: 10.1007/978-3-642-22887-2_5]

[101] Raileanu R, Rocktäschel. RIDE: Rewarding impact-driven exploration for procedurally-generated environments. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: ICLR, 2020.

[102] Bigazzi R, Landi F, Cascianelli S, Baraldi L, Cornia M, Cucchiara R. Focus on impact: Indoor exploration with intrinsic motivation. IEEE Robotics and Automation Letters, 2022, 7(2): 2985–2992. [doi: 10.1109/LRA.2022.3145971]

[103] Wijmans E, Kadian A, Morcos A, Lee S, Essa I, Parikh D, Savva M, Batra D. DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: ICLR, 2020.

[104] Li JC, Wang X, Tang SL, Shi HZ, Wu F, Zhuang YT, Wang WY. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12120–12129. [doi: 10.1109/CVPR42600.2020.01214]

[105] Zeng KH, Weihs L, Farhadi A, Mottaghi R. Pushing it out of the way: Interactive visual navigation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9863–9872. [doi: 10.1109/CVPR46437.2021.00974]

[106] Kumar G, Shankar NS, Didwania H, Roychoudhury RD, Bhowmick B, Krishna KM. GCExp: Goal-conditioned exploration for object goal navigation. In: Proc. of the 30th IEEE Int'l Conf. on Robot & Human Interactive Communication. Vancouver: IEEE, 2021. 123–130. [doi: 10.1109/RO-MAN50785.2021.9515530]

[107] Dang RH, Shi ZF, Wang LY, He ZT, Liu CJ, Chen QJ. Unbiased directed object attention graph for object navigation. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 3617–3627. [doi: 10.1145/3503161.3547852]

[108] Yadav K, Ramrakhya R, Majumdar A, Berges VP, Kuhar S, Batra D, Baevski A, Maksymets O. Offline visual representation learning for embodied navigation. arXiv:2204.13226, 2022.

[109] Al-Halah Z, Ramakrishnan SK, Grauman K. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17010–17020. [doi: 10.1109/CVPR52688.2022.01652]

[110] Campari T, Lamanna L, Traverso P, Serafini L, Ballan L. Online learning of reusable abstract models for object goal navigation. In:

Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 14850–14859. [doi: 10.1109/CVPR52688.2022.01445]

[111] Luo HK, Yue A, Hong ZW, Agrawal P. Stubborn: A strong baseline for indoor object navigation. In: Proc. of the 2022 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Kyoto: IEEE, 2022. 3287–3293. [doi: 10.1109/IROS47612.2022.9981646]

[112] Dang RH, Wang LY, He ZT, Su S, Tang JG, Liu CJ, Chen QJ. Search for or navigate to? Dual adaptive thinking for object navigation. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 8216–8225. [doi: 10.1109/ICCV51070.2023.00758]

[113] Zhou KW, Zheng KZ, Pryor C, *et al*. ESC: Exploration with soft commonsense constraints for zero-shot object navigation. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: PMLR, 2023. 42829–42842.

[114] Liu JJ, Guo JF, Meng ZH, Xue JT. ReVoLT: Relational reasoning and Voronoi local graph planning for target-driven navigation. arXiv:2301.02382, 2023.

[115] Chen BL, Kang JX, Zhong P, Cui YZ, Lu SY, Liang YX, Wang JX. Think holistically, act down-to-earth: A semantic navigation strategy with continuous environmental representation and multi-step forward planning. IEEE Trans. on Circuits and Systems for Video Technology, 2024, 34(5): 3860–3875. [doi: 10.1109/TCSVT.2023.3324380]

[116] Wang S, Wu ZH, Hu XB, Lin YS, Lv K. Skill-based hierarchical reinforcement learning for target visual navigation. IEEE Trans. on Multimedia, 2023, 25: 8920–8932. [doi: 10.1109/TMM.2023.3243618]

[117] Krishna R, Zhu YK, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS, Fei-Fei L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123(1): 32–73. [doi: 10.1007/s11263-016-0981-7]

[118] Gupta S, Davidson J, Levine S, Sukthankar R, Malik J. Cognitive mapping and planning for visual navigation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7272–7281. [doi: 10.1109/CVPR.2017.769]

[119] Ammirato P, Poirson P, Park E, Košecká J, Berg AC. A dataset for developing and benchmarking active vision. In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation. Singapore: IEEE, 2017. 1378–1385. [doi: 10.1109/ICRA.2017.7989164]

[120] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[121] Khandelwal A, Weihs L, Mottaghi R, Kembhavi A. Simple but effective: CLIP embeddings for embodied AI. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 14809–14818. [doi: 10.1109/CVPR52688.2022.01441]

[122] Trabucco B, Sigurdsson GA, Piramuthu R, Sukhatme GS, Salakhutdinov R. A simple approach for visual room rearrangement: 3D mapping and semantic search. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: ICLR, 2023.

[123] Sarch G, Fang ZY, Harley AW, Schydlo P, Tarr MJ, Gupta S, Fragkiadaki K. TIDEE: Tidying up novel rooms using visuo-semantic commonsense priors. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 480–496. [doi: 10.1007/978-3-031-19842-7_28]

[124] Kant Y, Ramachandran A, Yenamandra S, Gilitschenski I, Batra D, Szot A, Agrawal H. HouseKeep: Tidying virtual households using commonsense reasoning. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 355–373. [doi: 10.1007/978-3-031-19842-7_21]

[125] Yamauchi B. A frontier-based approach for autonomous exploration. In: Proc. of the 1997 IEEE Int'l Symp. on Computational Intelligence in Robotics and Automation: Towards New Computational Principles for Robotics and Automation. Monterey: IEEE, 1997. 146–151. [doi: 10.1109/CIRA.1997.613851]

[126] Ramakrishnan SK, Al-Halah Z, Grauman K. Occupancy anticipation for efficient exploration and navigation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 400–418. [doi: 10.1007/978-3-030-58558-7_24]

[127] Liu S, Okatani T. Symmetry-aware neural architecture for embodied visual exploration. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17221–17230. [doi: 10.1109/CVPR52688.2022.01673]

[128] Georgakis G, Bucher B, Arapin A, Schmeckpeper K, Matni N, Daniilidis K. Uncertainty-driven planner for exploration and navigation. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation. Philadelphia: IEEE, 2022. 11295–11302. [doi: 10.1109/ICRA46639.2022.9812423]

[129] Jiang JD, Zheng LA, Luo F, Zhang ZJ. RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation. arXiv:1806.01054, 2018.

[130] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[131] Hu XX, Yang KL, Fei L, Wang KW. ACNet: Attention based network to exploit complementary features for RGBD semantic

segmentation. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing (ICIP). Taipei: IEEE, 2019. 1440–1444. [doi: 10.1109/ICIP.2019.8803025]

[132] Zhao HS, Shi JP, Qi XJ, Wang XG, Jia JY. Pyramid scene parsing network. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239. [doi: 10.1109/CVPR.2017.660]

[133] Li LH, Zhang PC, Zhang HT, Yang JW, Li CY, Zhong YW, Wang LJ, Yuan L, Zhang L, Hwang JN, Chang KW, Gao JF. Grounded language-image pre-training. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10955–10965. [doi: 10.1109/CVPR52688.2022.01069]

[134] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

[135] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[136] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[137] Zhou GZ, Hong YC, Wu Q. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 7641–7649. [doi: 10.1609/aaai.v38i7.28597]

[138] Eftekhar A, Zeng KH, Duan JF, Farhadi A, Kembhavi A, Krishna R. Selective visual representations improve convergence and generalization for embodied AI. In: Proc. of the the 12th Int'l Conf. on Learning Representations. Vienna: ICLR, 2024.

[139] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Proc. of the 3rd Theory of Cryptography Conf. on Theory of Cryptography. New York: Springer, 2006. 265–284. [doi: 10.1007/11681878_14]

[140] Shah D, Equi MR, Osiński B, Xia F, Ichter B, Levine S. Navigation with large language models: Semantic guesswork as a heuristic for planning. In: Proc. of the 7th Conf. on Robot Learning. Atlanta: PMLR, 2023. 2683–2699.

[141] Song CH, Sadler BM, Wu JM, Chao WL, Washington C, Su Y. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 2986–2997. [doi: 10.1109/ICCV51070.2023.00280]

[142] Wu PY, Mu Y, Wu BX, Hou Y, Ma J, Zhang SH, Liu C. VoroNav: Voronoi-based zero-shot object navigation with large language model. arXiv:2401.02695, 2024.

[143] Tsai YHH, Dhar V, Li JL, Zhang BW, Zhang J. Multimodal large language model for visual navigation. arXiv:2310.08669, 2023.

[144] Xi ZH, Chen WX, Guo X, He W, Ding YW, Hong BY, Zhang M, Wang JZ, Jin SJ, Zhou EY, Zheng R, Fan XR, Wang X, Xiong LM, Zhou YH, Wang WR, Jiang CH, Zou YC, Liu XY, Yin ZY, Dou SH, Weng RX, Cheng WS, Zhang Q, Qin WJ, Zheng YY, Qiu XP, Huang XJ, Gui T. The rise and potential of large language model based agents: A survey. arXiv:2309.07864, 2023.

[145] Vuong AD, Nguyen TT, Vu MN, Huang BR, Nguyen D, Binh HTT, Vo T, Nguyen A. HabiCrowd: A high performance simulator for crowd-aware visual navigation. arXiv:2306.11377, 2023.

[146] Cancelli E, Campari T, Serafini L, Chang AX, Ballan L. Exploiting proximity-aware tasks for embodied social navigation. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 10923–10933. [doi: 10.1109/ICCV51070.2023.01006]

[147] Chen BL, Lu SY, Zhong P, Cui YZ, Liang YX, Wang JX. SemNav-HRO: A target-driven semantic navigation strategy with human-robot-object ternary fusion. Engineering Applications of Artificial Intelligence, 2024, 127: 107370. [doi: 10.1016/j.engappai.2023.107370]

[148] Luo Q, Sorokin M, Ha S. A few shot adaptation of visual navigation skills to new observations using meta-learning. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 13231–13237. [doi: 10.1109/ICRA48506.2021.9561056]

[149] Wang T, Wu ZK, Wang DL. Visual perception generalization for vision-and-language navigation via meta-learning. IEEE Trans. on Neural Networks and Learning Systems, 2023, 34(8): 5193–5199. [doi: 10.1109/TNNLS.2021.3122579]

[150] Dwivedi K, Roig G, Kembhavi A, Mottaghi R. What do navigation agents learn about their environment? In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10266–10275. [doi: 10.1109/CVPR52688.2022.01003]

[151] Yang ZJ, Majumdar A, Lee S. Behavioral analysis of vision-and-language navigation agents. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 2574–2582. [doi: 10.1109/CVPR52729.2023.00253]

[152] Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu TH, Huang WL, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I, Florence P. PaLM-E: An embodied multimodal language model. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: PMLR, 2023. 8469–8488.

[153] Huang WL, Wang C, Zhang RH, Li YZ, Wu JJ, Fei-Fei L. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In: Proc. of the 7th Conf. on Robot Learning. Atlanta: PMLR, 2023. 540–562.

[154] Khanna M, Mao YS, Jiang HX, Haresh S, Shacklett B, Batra D, Clegg A, Undersander E, Chang AX, Savva M. Habitat synthetic scenes dataset (HSSD-200): An analysis of 3D scene scale and realism tradeoffs for objectgoal navigation. arXiv:2306.11290, 2023.

[155] Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv:1610.05492, 2016.

附中文参考文献:

[2] 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. [doi: 10.16383/j.aas.c220564]

[3] 司马双霖, 黄岩, 何科技, 安东, 袁辉, 王亮. 视觉语言导航研究进展. 自动化学报, 2023, 49(1): 1–14. [doi: 10.16383/j.aas.c210352]

[50] 殷炯, 张哲东, 高宇涵, 杨智文, 李亮, 肖芒, 孙垚棋, 颜成钢. 视觉语言预训练综述. 软件学报, 2023, 34(5): 2000–2023. http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[51] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[60] 王乐, 齐尧, 何滨兵, 章永进, 徐友春. 机器人自主探索算法综述. 计算机应用, 2023, 43(S1): 314–322. [doi: 10.11772/j.issn.1001-9081.2022111706]

[61] 张世勇, 张雪波, 苑晶, 方勇纯. 旋翼无人机环境覆盖与探索规划方法综述. 控制与决策, 2022, 37(3): 513–529. [doi: 10.13195/j.kzyjc.2021.1751]

陈铂垒(1998－), 男, 博士生, 主要研究领域为物体目标导航, 视觉自主探索, 视觉物体重排布.

卢思怡(2000－), 女, 硕士生, 主要研究领域为人机交互, 物体目标导航.

康嘉绪(2001－), 男, 硕士生, 主要研究领域为物体目标导航, 视觉自主探索, 视觉物体重排布.

杨昊楠(2001－), 男, 硕士生, 主要研究领域为场景表示学习, 视觉物体重排布.

钟萍(1982－), 女, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为物联网, 人工智能, 移动互联网, 自主无人系统.

王建新(1969－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机算法与优化, 生物信息学, 网络优化理论.

崔永正(1998－), 男, 硕士生, 主要研究领域为机器人自主探索, 路径规划, 物体目标导航.