DOI: 10.13382/j.jemi.B2407549

机器人大模型发展与挑战*

邓 鹏[1,2] 唐文涛[1,2] 罗 静[1,2]
(1.荆楚理工学院新能源学院 荆门 448000;2.荆楚理工学院智能制造与先进技术应用研究所 荆门 448000)

摘 要:近年来预训练大模型的研究取得了显著成就,本文论述了预训练大模型在机器人技术中的应用。机器人中的传统深度学习模型是在为特定任务定制的小数据集上训练的,这限制了它们在不同应用中的适应性。相比之下,在互联网规模数据上预训练的大模型似乎具有优越的泛化能力,并且在某些情况下显示出一种探索能力,在训练数据中未出现的情况下可以找到 one-shot 解决方案。大模型具有增强机器人自主性任务的各个组成部分的潜力,从感知到决策和控制。本文研究了最近使用或建立大模型来解决机器人问题的论文,探讨了大模型如何有助于提高机器人在感知、决策和控制领域的能力,从而推动机器人大模型在更多领域实现应用落地。同时,讨论了阻碍大模型在机器人自主系统中应用的挑战,如机器人应用中的数据稀缺性、机器人自身的可变性、多模态表示的局限性和实时性能,并为未来的改进提供了机会和潜在的方法。

关键词:机器人;大语言模型;视觉语言模型;多模态;基础模型
中图分类号:TP242.6;TN98 **文献标识码**:A **国家标准学科分类代码**:510.8050

Robotic large model development and challenges

Deng Peng[1,2] Tang Wentao[1,2] Luo Jing[1,2]
(1. College of New Energy, Jingchu University of Technology, Jingmen 448000, China; 2. Institute of Intelligent Manufacturing and Advanced Technology Application, Jingchu University of Technology, Jingmen 448000, China)

Abstract: The research on pre-trained large models has made remarkable achievements in recent years, this paper reviews the application of pre-trained large model in robotics. Traditional deep learning models in robots were trained on small datasets customized for specific tasks, which limits their adaptability in different applications. In contrast, large models pre-trained on internet-scale data appear to have superior generalization capabilities and in some cases show an exploratory ability to find one-shot solutions where they are not present in the training data. The underlying model has the potential to enhance the various components of a robot's autonomous task, from perception to decision making and control. This paper examines recent papers that use or build large models to solve robotics problems, exploring how large models can help improve robots' capabilities in the areas of perception, decision making, and control, thereby promoting the application of large robot models in more fields. Meanwhile, the challenges that hinder the application of large models in robotic autonomous systems were discussed in this paper, such as data scarcity in robotic applications, the variability of robots themselves, the limitations of multimodal representations, and real-time performance, and provides opportunities and potential approaches for future improvements.

Keywords: robotics; large language models; visual-language models; multimodal; foundation models

收稿日期: 2024-05-24 Received Date: 2024-05-24
*基金项目:荆门市科技计划项目(2023YFYB040)、湖北省高等学校优秀中青年科技创新团队项目(T2021028)、荆楚理工学院教学研究项目(JX2023-014)资助

0 引　言

大模型是指具有大规模参数和复杂计算结构的机器学习模型。大模型在自然语言处理、图像识别、计算机视觉、金融科技和智能交通等领域都有广泛的应用。目前,大模型在视觉和语言处理方面已经取得了重大突破,例如 Transform 双向编码器表示[1](bidirectional encoder representations from transformers, BERT)、通用预训练模型[2](generative pre-train model3, GPT-3)、通用预训练模型[3](generative pre-train model4, GPT-4)、对比语言图像预训练[4](contrastive language-image pre-training, CLIP)、零样本文本到图像生成[5](zero-shot text-to-image generation, DALL-E)、路径大语言模型[6](pathways language model, PaLM-E)等。然而,大模型正以难以预料的潜力开启机器人领域的新可能性,如自动驾驶、家用机器人、工业机器人、辅助机器人、医疗机器人和多机器人协作系统。预训练的大型语言模型(large language models,LLM)、大型视觉语言模型(large vision-language models,VLM)、大型音频语言模型(large audio-language models,ALM)和大型视觉导航模型(large visual navigation models,VNM)可用于改进机器人执行中的各种任务。将大模型集成到机器人中是一个快速发展的领域,最近大量研究人员开始探索利用机器人领域中的这些大型模型进行感知、决策、规划和控制的方法。

大模型是在广泛和多样化的数据上进行预训练的,这在其他领域(如自然语言处理、计算机视觉和医疗检测[7-8])中得到了证明,可以显著扩展其适应性、泛化能力和整体性能。与机器人技术特别相关的是,多模态大模型可以将从各种传感器收集的多模态异构数据融合和对齐为机器人理解和推理所需的紧凑的同构表示形式[9]。这些学习到的表示形式有可能用于自主机器人的任何部分,包括感知、决策和控制。

将大模型集成到机器人系统中可以通过增强机器人感知环境和与环境交互的能力来实现上下文感知机器人系统。例如,在感知领域,已经发现 VLM 通过学习视觉和文本数据之间的关联来提供跨模态理解,帮助完成诸如零样本图像分类、零样本物体检测[10]和 3D 分类[11]等任务。另一个例子是,三维世界中的语言映射[12]可以通过将单词与三维环境中的特定对象、位置或动作相关联来增强机器人的空间意识。

在决策或规划领域,已经发现 LLM 和 VLM 可以帮助机器人进行高级规划[13]。机器人可以通过在操作、导航和交互中利用语言线索来执行更复杂的任务。例如,对于模仿学习[14]和强化学习[15]等机器人策略学习技术,大模型似乎提供了提高数据效率和增强上下文理解的可能性。特别是语言驱动的奖励可以通过提供形状奖励来指导强化学习智能体[16]。此外,研究人员还利用语言模型为策略学习技术提供反馈[17]。一些研究表明,VLM 模型的视觉问答能力可以在机器人中得到应用。例如,研究人员已经使用 VLM 来回答与视觉内容相关的问题,以帮助机器人完成任务[18]。

尽管大模型在视觉和语言处理方面具有变革性的能力,但模型在现实世界机器人任务中的泛化和微调仍然具有挑战性。本文研究了现有的关于大模型在机器人中应用的文献以及其中的方法和应用,探讨了大模型如何有助于提高机器人在感知、决策和控制领域的能力。本文讨论了阻碍大模型在机器人自主系统中应用的挑战,如机器人应用中的数据稀缺性、机器人自身的可变性、多模态表示的局限性和实时性能,并为未来的改进提供了机会和潜在的方法。

1 大模型技术背景

1.1 大语言模型

LLM 有数十亿个参数,并在数万亿个标记上进行训练。这种大规模使得通用预训练模型(generative pre-train model2, GPT-2)[19]和 BERT[1]等模型分别在 Winograd 模式挑战[20]和通用语言理解评估[21](general language understanding evaluation,GLUE)基准测试中实现了最先进的性能。它们的后继者包括 GPT-3[2]、开放高效的基础语言模型[22](open and efficient foundation language models, LLaMA)、PaLM-E[23]和通用语言模型[24](general language model 130 billion, GLM-130B),它们在参数数量(通常超过 1 000 亿个)、上下文窗口的大小(通常超过 1 000 个令牌)和训练数据集的大小(通常有 10 s TB 的文本)方面都有很大的增长。

1.2 视觉 transformer

视觉 transformer[25-27](vision transformer, ViT)是一种用于计算机视觉任务的 transformer 架构,包括图像分类、分割和目标检测。ViT 将图像视为称作标记的图像补丁序列。在图像标记化过程中,图像被分割成固定大小的小块。然后将这些小块平面化成一维向量,称为线性嵌入。为了捕捉图像斑块之间的空间关系,将位置信息添加到每个标记中。这个过程被称为位置嵌入。与位置编码结合的图像标记被馈送到 transformer 编码器中,自关注机制使模型能够捕获输入数据中的长期依赖关系和全局模式。在本文中,只关注那些具有大量参数的 ViT 模型。ViT 改进版本[28]将 ViT 模型按比例放大,有 2 B 个参数。此外,ViT 改进版本[29]有 4 B 个参数。ViT 改进版本[30]是一个具有 220 亿个参数的视觉 transformer 模型,

用于 PaLM-E 和 PaLI-X[31]（multilingual language-image model evaluated），并帮助完成机器人任务。

改进降噪功能检测框[32]（detr with improved denoising anchor boxes，DINO）是一种用于训练 ViT 的自监督学习方法。DINO 是一种没有标签的知识蒸馏形式。神经网络架构由 ViT 或残差网络[33]（residual network，ResNet）主干和包含多层感知（Multi-Layer Perception，MLP）层的投影头组成。DINO 改进版本[34]提供了多种预训练的视觉模型，这些模型在[34]中引入的语言视觉数据（language visual data，LVD-142M）数据集上使用不同的 ViT 进行训练。分割一切模型[35]（segment anything model，SAM）提供零样本提示图像分割。

1.3 多模态视觉-语言模型

多模态是指一个模型能够接受输入的不同“模态”，例如图像、文本或音频信号。VLM 是一种同时接受图像和文本的多模态模型。在机器人应用中常用的 VLM 是对比语言图像预训练[4]（contrastive language image pre-training，CLIP）。CLIP 提供了一种比较文本描述和图像之间相似性的方法。CLIP 使用互联网规模的图像-文本对数据来捕获图像和文本之间的语义信息。CLIP 模型架构包含一个文本编码器[19]和一个图像编码器（ViT 的改进版本），它们被联合训练以最大化图像和文本嵌入的余弦相似度。引导语言图像预训练[36]（bootstrapping language-image pre-training，BLIP）侧重于多模态学习，在预训练过程中对 3 个目标进行联合优化。CLIP 改进版本[37]旨在构建对齐良好且基于实例的文本-图像点智能体。它使用跨模态对比目标学习语义和实例级对齐的点云表示。细粒度交互式语言图像预训练[38]（fine-grained interactive language-image pre-training，FILIP）侧重于在多模态学习中实现更精细的对齐。它集成了一个跨模态后期交互机制，该机制利用了视觉和文本标记之间的最大相似性。细粒度语言图像预训练[39]（fine-grained language-image pre-training，FLIP）提出了一种简单高效的 CLIP 训练方法。FLIP 算法在训练过程中随机掩码并去除大量图像补丁。该方法旨在提高 CLIP 的训练效率同时保持其性能。

1.4 具身多模态语言模型

具身智能体是一种与虚拟或物理世界互动的 AI 系统，包括虚拟助手或机器人。具身语言模型是将真实世界的传感器和驱动模式纳入预训练大语言模型的基础模型。典型的视觉语言模型是在一般的视觉语言任务上训练的，比如图像字幕或视觉问答。PaLM-E[6]是一种多模态语言模型，它不仅在互联网规模的通用视觉语言数据上进行了训练，同时也在具体的机器人数据上进行了训练。PaLM-E 是由 PaLM 和一个 ViT[30]构建的。ViT 将图像转换成一系列嵌入向量，这些嵌入向量通过仿射变换投射到语言嵌入空间。整个模型是端到端的训练，从预训练的 LLM 和 ViT 模型开始。

2 国外研究进展

2.1 机器人感知大模型

与周围环境交互的机器人接收不同形式的原始感官信息，如图像、视频、音频和语言。这种高维数据对于机器人在环境中理解、推理和互动至关重要。当前的大模型，包括那些已经在视觉和自然语言处理（natural language processing，NLP）领域开发的模型，是将这些高维输入转换为抽象的、结构化的表示工具，可以更容易地解释和操作。特别是多模态模型使机器人能够将不同的感官输入整合到包含语义、空间、时间和功能信息的统一表示中。

1）目标检测

零样本物体检测允许机器人识别和定位他们以前从未遇到过的物体。GLIP[40]（ground language-image pre-training）通过将目标检测重新定义为短语映射，将目标检测与短语映射相结合。在此框架中，检测模型的输入不仅包括图像，还包括描述检测任务的所有潜在类别的文本输入。最近，部分分割语言图像预训练[41]（part segmentation language-image pre-training，PartSLIP）证明了 GLIP 可以用于 3D 物体的 low-shot 头部分割。PartSLIP 从多个视图中渲染对象的 3D 点云，并结合这些视图中的 2D 边界框来检测对象部件。

开放世界语言视觉 transformer[42]（open world language-ViT，OWL-ViT）是一个开放词汇的对象检测器。OWL-ViT 使用视觉 transform 架构，具有对比图像-文本预训练和端到端的微调。与 GLIP 不同，GLIP 将检测定义为单个文本查询的短语映射问题，并限制了可能的对象类别的数量，而 OWL-ViT 可以处理多个基于文本或图像驱动的查询。OWL-ViT 已被应用于机器人学习中，例如在文献[43]中作为开放词汇对象检测器来寻找“感兴趣的实体”（例如花瓶或抽屉把手），并最终定义值图以优化操作轨迹。改进对齐 DINO[44]将 DINO[32]与基于实际的预训练相结合，通过融合视觉和语言，将闭集 DINO 模型扩展为开集检测。Grounding DINO 在开集目标检测方面优于 GLIP。

零样本 3D 分类器可以使机器人在没有明确训练数据的情况下对环境中的物体进行分类。大模型是执行 3D 分类的有力候选。CLIP 改进版本[45]通过将点云与文本对齐，将 CLIP 对 2D 图像的预训练知识转化为对 3D 点云的理解。作者建议将每个点投影到一系列预定义的图像平面上以生成深度图。然后，使用 CLIP 视觉编码器

对点云的多视图特征进行编码,并用自然语言预测每个视图的标签。BERT 改进版本[46]使用基于 transformer 的架构从点云中提取特征,将 BERT 的概念推广到三维点云。与 CLIP 改进版本将匹配点云和文本的任务转换为图像-文本对齐不同,语言、图像和点云的统一表示[47-48](unified representation of language, images, and point clouds, ULIP)是用于 3D 理解的语言,图像和点云的统一表示。它通过使用对象三元组(图像、文本、点云)进行预训练来实现这一点。该模型使用形状网络[49](shape network55, ShapeNet55)中的少量自动合成三元组进行训练,形状网络是一个大规模的 3D 模型存储库。ULIP 使用 CLIP 作为视觉语言模型。ULIP[47-48]表明,使用统一的 ULIP 多模态表示可以提高 PointBERT 等模型的识别能力。

2) 语义分割

语义分割将图像中的每个像素划分为语义类。这提供了关于图像中对象边界和位置的细粒度信息,并使嵌入的智能体能够在更细粒度的级别上理解环境并与之交互。一些研究探讨了诸如 CLIP 之类的大模型如何增强语义分割任务的泛化性和灵活性。

语言分割(language segmentation, LSeg)是一种语言驱动的语义分割模型[50],它将语义相似的标签与嵌入空间中的相似区域相关联。LSeg 使用基于 CLIP 架构的文本编码器来计算文本嵌入,使用基于密集预测转换器(dense prediction transformer,DPT)底层架构的图像编码器[51]。与 CLIP 类似,LSeg 使用文本和图像嵌入创建联合嵌入空间。SAM[35]引入了一个提示式输入分段框架,该框架由提示式分段的任务定义、分段基础模型和数据引擎组成。SAM 采用来自掩码自动编码器[52](mask auto-encoder, MAE)的预训练视觉 transformer 作为图像编码器,同时使用来自改进 CLIP[53]的文本编码器用于稀疏输入(点、框和文本),并使用单独的密集输入编码器用于掩码。

更快 SAM 的改进版本[54]和移动 SAM 的改进版本[55]在更快的推理速度下实现了与 SAM 相当的性能。TAM[56](track anything model)结合了 SAM 和一种先进的视频对象分割(video object segmentation,VOS)模型[57],实现了交互式视频对象跟踪和分割。任意 3D[58]采用了一系列视觉语言模型和 SAM 来将物体提升到 3D 的级别。它使用 BLIP[36]生成文本描述,同时使用 SAM 从视觉输入中提取感兴趣的对象。任意 3D 将提取的对象提升到使用文本到图像扩散模型的神经辐射场[59](neural radiance field, NeRF)表示中,使其能够集成到 3D 场景中。

3) 3D 场景和对象表示

场景表示允许机器人理解周围环境,促进空间推理,并提供上下文感知。语言驱动的场景表示将文本描述与视觉场景对齐,使机器人能够将单词与对象、位置和关系联系起来。在本节中,研究了最近使用基础模型来增强场景表示的工作。

(1) 3D 场景中的语言映射

语言映射是指将环境的几何表征与语义表征相结合。一种可以为智能体提供强几何先验的表示是隐式表示。隐式表示的一个例子是 NeRF[59-61]。NeRF 从一组从不同视点捕获的 2D 图像(不需要明确的深度信息)中创建高质量的场景和对象的 3D 重建。NeRF 神经网络将相机姿态作为输入,并预测场景的 3D 几何形状以及颜色和强度。Kerr 等[62]提出了语言嵌入辐射场(language-embedded radiance fields,LERF),将 CLIP 嵌入到密集的多尺度 3D 场中。这将生成环境的 3D 表示,可以通过查询生成语义相关性图。LERF 模型以三维位置(x, y, z),观察方向(φ,θ)和比例因子作为输入,输出 RGB 值,密度(σ)以及 DINO[32]和 CLIP 特征。在 CLIP 场域[63]中,隐式场景表示 $g(x,y,z):\Re^3 \rightarrow \Re^d$ 通过解码 d 维潜在向量到不同模态特定输出来训练。该模型从预训练的图像模型中提取信息,方法是将像素标签反向投影到 3D 空间,并训练输出头来预测来自开放词汇对象检测器、CLIP 视觉表示和使用对比损失的单一编码实例标签的语义标签。

另一个相关的工作是视觉语言映射[64](visual language maps, VLMaps),它将像素嵌入从 LSeg 投影到自上而下网格图中的网格单元。该方法不需要训练,而是直接将像素嵌入反投影到网格单元中,并在重叠区域中取平均值。语义抽象[65](semantic abstraction,SemAbs)提出了另一种通过解耦视觉语义推理和 3D 推理来理解 3D 场景的方法。在 SemAbs 中,给定场景的 RGB-D 图像,语义感知的 2D VLM 为每个查询对象提取 2D 相关性图,而语义抽象的 3D 模块使用相关性图预测每个对象的 3D 占用。

目前的 VLM 可以对 2D 图像进行推理,然而,它们并没有建立在 3D 世界的基础上。建立三维 VLM 大模型的主要挑战是三维数据的稀缺性。特别是,与语言描述相匹配的 3D 数据很少。规避这个问题的一个策略是利用在大规模数据上训练的 2D 模型来监督 3D 模型。例如,特征 NeRF[66]的作者提出通过神经渲染将 2D 视觉基础模型(即 DINO 或 latent diffusion)提取到 3D 空间来学习 3D 语义表示。

(2) 对象表示

在学习对象之间的对应关系时,如何促进对象操纵的能力,特别是在技能从已知类别的训练对象转移到测试时的新对象实例或新对象类别时。传统上,对象对应关系已经学习使用强大的监督,如关键点和关键帧。神

经描述符字段[67](neural descriptor fields,NDF)通过利用占用网络的分层激活来消除密集标注的需要。然而,这种方法仍然需要对每个目标对象类别进行许多训练形状。其他工作已经开始直接从预训练视觉模型的图像特征构建对象表示。

机器人操作特征域[68](feature fields for robotic manipulation,F3RM)建立在 NDF 的基础上,开发支持寻找相应对象区域的场景表示。F3RM 使用与 NDF 相似的特征表示相对于物体的 6 自由度姿态(例如,抓住杯子的把手)。对象之间的对应关系也可以直接从 DINO 特征中提取[69],而无需训练。该方法首先利用多视图提取两个对象的密集 ViT 特征图;通过计算特征图上的周期性距离度量[70],可以找到两个对象上的相似区域。有了二维的 patch 对应关系,物体之间的 7 维刚体变换(即一个 SO(3)位姿、一个平移量和一个尺度标量)可以与 RANSAC 和 Umeyama 的方法[71]一起求解。

4) 可操作性学习

可操作性是指对象、环境或实体为智能体提供特定功能或交互的潜力。它们可以包括推、拉、坐或抓等动作。检测可操作性弥合了感知和动作之间的差距。

可操作性融合[72]综合了复杂的相互作用,例如关节手与给定物体的相互作用。给定 RGB 图像,可操作性融合旨在生成用于手-物交互(hand-object interaction,HOI)的人手图像。作者提出了一种基于大规模预训练扩散模型的两步生成方法,该模型基于在哪里交互(布局)和如何交互(内容)。视觉-机器人桥[73](vision-robotic bridge,VRB)在人类行为的网络视频上训练视觉提供模型。特别是它估计可能的位置和方式。

2.2 机器人决策大模型

在机器人的决策和规划领域内,LLM 和 VLM 可能会成为增强机器人能力的有价值的工具。VLM 也可能对这一领域做出贡献。VLM 专注于可视化数据的分析。这种视觉理解是机器人明智决策和复杂任务执行的关键组成部分。机器人现在可以利用自然语言线索来提高它们在涉及操作、导航和交互的任务中的表现。视觉语言目标条件策略学习,无论是通过模仿学习还是强化学习,都有望使用基础模型进行改进。本节强调 LLM 和 VLM 在机器人决策中的潜在贡献。

1) 策略学习

在语言条件模仿学习中,学习到一个目标条件策略 $\pi_\theta(\alpha_t \mid s_t, l)$,输出基于当前状态 $s_t \in S$ 和语言指令 $l \in L$ 的动作 $\alpha_t \in A$,损失函数定义为最大似然目标条件模仿目标:

$$s_{GCIL} = E_{(\boldsymbol{\tau},l)\sim D} \sum_{t=0}^{|\boldsymbol{\tau}|} \log \pi_\theta(\alpha_t \mid s_t, l) \tag{1}$$

其中,$\boldsymbol{D}$ 是语言标注的演示数据集 $\boldsymbol{D} = \{t_i\}_i^N$。演示可以表示为轨迹-图像序列、RGB-D 体素观测等。语言指令与演示配对,用作训练数据集。每个语言标注论证 $\boldsymbol{\tau}_i$ 由 $\boldsymbol{\tau}_i = \{(s_1, l_1, a_1), (s_2, l_2, a_2), \cdots\}$ 组成。在测试时,给机器人一系列指令,语言条件视觉运动策略 π_θ 在给定指令的每个时间步长下提供闭环动作。

由于通过配对演示和语言教学来生成语言标注数据是一个代价较高的过程,播放监督的潜在运动规划[74](play-supervised latent motor plans, Play-LMP)的作者建议从远程操作的游戏数据中学习。此外,还学习了目标条件策略来解码推理的规划,以执行用户指定的任务。在后续工作[75]中,作者提出了多上下文模仿(multi-context imitation,MCIL),它在非结构化数据上使用语言条件模仿学习。多上下文模仿框架是基于重新标记的模仿学习和标记的指令跟随。MCIL 假定可以访问多个上下文模拟数据集。解决语言条件模仿学习中数据标注挑战的另一种方法包括利用大模型通过标记演示来提供反馈。在文献[76]中,作者提出使用预训练的大模型来提供反馈。为了将训练好的策略部署到新任务或新环境中,使用随机生成的指令来执行策略,并且预训练的大模型通过标记演示来提供反馈。CLIP 传输[77]也提出了基于视觉操作的语言条件模仿学习,将 CLIP 的语义理解与 Transporter 的空间精度相结合[78]。这个端到端框架解决了语言指定的操作任务,而不需要任何对象姿态或实例分割的显式表示。CLIP 传输基于精确空间推理的语义概念,但它仅限于二维观察和行动空间。为了解决这一限制,感知者-行动者[79](perceiver actor, PerAct)的作者提出用 3D 体素来表示观察和动作空间,并利用体素补丁的 3D 结构进行有效的语言条件行为克隆,并使用 transformer 来模仿少数演示中的 6 自由度机器人操作任务。

利用真实机器人的语言条件模仿学习来部署机器人策略学习技术面临着持续的挑战。这些模型依赖于端到端学习,其中策略将像素或体素映射到动作。为了提高策略的鲁棒性和适应性,可以采用数据增强和领域自适应等技术使策略对分布转移具有更强的鲁棒性。收集增强压缩训练[14](collect, augment, compress, train, CACTI)是一种新的框架,旨在使用稳定融合[80]等基础模型增强机器人学习的可扩展性。CACTI 引入了数据收集、数据增强、视觉表示学习和模仿策略训练 4 个阶段。在数据增强阶段,CACTI 采用 Stable Diffusion[80]等视觉生成模型,通过场景和布局变化来增强数据,从而增强视觉多样性。CACTI 经过训练,可以在模拟和现实世界的厨房环境中进行多任务和多场景机器人操作。

强化学习(reinforcement learning,RL)是一系列方法,使机器人能够通过优化奖励函数与环境相互作用来优化策略。在 RL 问题中,使用从与环境的交互中收

集的样本数据最大化策略的预期回报。在自适应智能体[81](adaptive agent, AdA)中,作者提出了一个强化学习模型,该模型是一个对各种任务进行预训练的智能体,旨在通过使用快速的上下文学习反馈来快速适应开放式的三维问题。这项工作考虑了导航、协作和分工任务。Palo 等[15]提出了一种通过集成 LLM 和 VLM 来创建更统一的强化学习框架的方法。这项工作考虑了机器人操作任务。他们的方法解决了与探索、经验重用和转移、技能调度以及从观察中学习相关的核心强化学习挑战。

2) 语言-图像价值学习

在价值学习中,目的是构建一个价值函数,使不同模式的目标保持一致,并由于价值函数的递归性质而保持时间一致性。R3M[82](reusable representation for robotic manipulations)使用各种人类视频数据集(如 Ego4D)为机器人操作提供预训练的视觉表示,并可作为机器人操作任务中策略学习的冻结感知模块。在 Franka Emika Panda 的手臂上演示了 R3M 的预训练视觉表征,并实现了不同的下游操作任务。与 R3M 类似,VIP[83](value-implicit pretraining)采用时间对比学习来捕获视频中的时间依赖性,但不需要视频语言校准。VIP 是一种自监督方法,用于从视频中学习视觉目标条件值函数和表示。VIP 学习基于视觉目标的下游任务奖励,并可用于零样本奖励规范。

语言-图像价值学习[84](language-image value learning, LIV)是一种以控制为中心的视觉语言表示。LIV 通过学习多模态视觉语言值函数和使用语言对齐视频的表示来推广先前的工作 VIP。对于指定为语言目标或图像目标的任务,训练一个编码通用值函数的多模态表示。LIV 还专注于机器人操作任务。LIV 是一种以控制为中心的基于大型人类视频数据集(如 EPICKITCHENS[85])预训练的视觉语言表征。Nair 等[86]从离线数据中学习语言条件奖励,并在模型预测控制期间使用该方法来完成语言指定的任务。

价值函数可用于帮助从 LLM 获得的映射语义信息到机器人运行的物理环境。通过利用价值函数,机器人可以将 LLM 处理的信息与周围的特定位置和物体联系起来。Ahn[87]等通过学习研究了大语言模型与物理世界的整合。Inner Monologue 研究了基础环境反馈提供给 LLM 的作用[88],从而与环境闭合回路。通过利用感知模型集合(例如,场景描述符和成功检测器)以及预训练的语言条件机器人技能,反馈用于具有大型语言模型的机器人规划。文本到行为[89](text to motion, Text2Motion)是一种基于语言的长视界机器人操作规划框架。与 SayCan 和 Inner Monologue 类似,Text2Motion 在每个时间步计算与每个技能相关的分数(S_{LMM})。任务规划问题是通过最大化给定语言指令和初始状态的技能序列的可能性来找到一个技能序列。Mahmoudieh[90]提出了使用 CLIP 进行构造奖励的方法。这项工作考虑了机器人操作任务。该模型利用 CLIP 对目标文本描述的场景中的地面物体与空间关系规则配对,通过使用原始像素作为输入来塑造奖励。Mees[91]提出了分层通用语言条件策略。这项工作考虑了机器人操作任务,只需要用语言标注总数据的 1%。

3) 任务规划

LLM 可用于为执行复杂的长视机器人任务提供上层任务规划。如上所述,SayCan[87]使用 LLM 在语言中进行上层任务规划,尽管使用学习值函数将这些指令置于环境中。

时间逻辑对于在机器人系统中施加时间规范是有用的。Chen 等[92]提出了从自然语言(natural language, NL)到时间逻辑(temporal language, TL)的翻译。创建具有 28 k 个 NL-TL 对的数据集,并使用该数据集对 T5[93]模型进行微调。LLM 通常用于规划任务子目标。这项工作考虑了机器人导航任务。在文献[94]中,不是直接进行任务规划,而是执行从自然语言任务描述到中间任务表示的几次翻译。经典的任务规划需要广泛的领域知识,搜索空间大[95-96]。LLM 可用于生成完成高级任务所需的任务序列。在编程提示[97](program prompt, ProgPrompt)中,作者介绍了一种使用 LLM 直接生成动作序列的输入方法,无需额外的领域知识。对 LLM 的输入包括动作、环境中的对象和可执行的示例程序。虚拟家庭[98]被用作演示的模拟器。代码生成策略[99]探讨了使用代码编写 LLM 来基于自然语言命令生成机器人策略代码。这项工作考虑了机器人操作和导航任务,使用的是来自 Everyday Robots 的真实世界的移动机械臂机器人。

4) 机器人 transformer

可以通过提供结合感知、决策和动作生成的集成框架,用于机器人的端到端控制。Xiao 等[100]使用真实世界图像的监督视觉预训练证明了自监督的有效性,直接从像素输入学习运动控制任务,该工作的重点是机器人操作任务。同样,Buceker 等[101]研究了在真实世界机器人任务的各种野外视频上使用自监督视觉预训练。这项工作考虑了机器人操作任务。

基于 Transformer 的策略模型的另一个例子是在 RT-1(robotics transformer1)上的工作[102],其中作者演示了一个显示可伸缩性属性的模型。为了训练模型,作者使用了超过 13 万个真实世界机器人经验的大型数据集,包括 700 多个任务,这些数据集是由 13 个机器人组合在 17 个月内收集的。RT-1 接收图像和自然语言指令作为输入和输出离散的位置和机械臂动作。

后续工作中,称为 RT-2[103](robotics transformer2),展示了一个视觉-语言-动作(vision-language-action,VLA)模型,该模型有效地利用这些数据生成机器人控制的广义动作。通过大量的实验表明,利用 VLM 有助于增强视觉和语义概念的泛化,并使机器人能够响应所谓的思维链提示输入,其中智能体执行更复杂的多阶段语义推理。在 RT-X[104](robotics transformer-X)中,作者提供了许多标准化数据格式的数据集和模型,从而可以探索在机器人操作背景下训练大型交叉机器人模型的可能性。特别是,作者从 21 个机构合作收集的 22 个不同机器人中收集了一个数据集,展示了 527 项技能(包括 160 266 项任务)。有了这个统一的数据集,RT-X 证明了基于 RT-1 和 RT-2 的模型在这种多智能体、多样化的数据上进行训练可以在机器人领域之间表现出较高的性能,并通过利用其他平台的经验提高了多个机器人的能力。

其他工作研究了机器人控制的预训练 transformer,使用来自多个机器人的自监督轨迹数据进行训练。例如,感知-行动因果 transformer[105] 是一种生成式 transformer 架构,它从具有自我监督的机器人数据中构建表征。这项工作考虑了机器人导航任务。该领域的另一项工作是使用控制 transformer 的自监督多任务预训练[106],它引入了与控制变压器相关的自监督多任务,提供了为顺序决策任务量身定制的预训练微调方法。

3 国内研究进展

随着大模型技术的飞速发展,国内在机器人大模型研究方面虽然刚刚起步,但也取得了一些进展[107]。在机器人大模型研究方面,国内企业和研究机构在算法应用、算力和数据等方面取得了一些突破。例如,基于 Transformer 架构的预训练大模型在自然语言处理、计算机视觉等领域展现出强大的能力,为机器人智能化提供了坚实的基础,虽然与国际顶尖水平仍存在一定差距,但已能够满足部分应用场景的需求。国内企业智元机器人发布“远征”与“灵犀”两大系列共 5 款商用人形机器人产品,在交互服务、柔性智造、特种作业、科研教育及数据采集等场景提出了一些解决方案。除企业外,一批人形机器人实验室、创新中心,也在凝聚产学研各界优势力量,建设人形机器人开源社区,提升关键技术的供给能力[108]。

表 1 特定于机器人的大模型以及模型的结构、大小、预训练任务、推理速度和硬件信息

Table 1 Large models specific to the robot and the model's structure, size, pre-training tasks, reasoning speed, and hardware information

模型	Backbone	参数大小	预训练任务	推理速度	硬件
RoboCat[109]	仅解码器的 Transformer	1.18 B	操作任务	10~20 Hz	
Gato[110]	仅解码器的 Transformer	1.2 B	通用智能体	20 Hz	16×16 TPU v3
PaLM-E-562B[6]	仅解码器的 Transformer	562 B	语言子任务+控制策略	5~6 Hz	多 TPU 云服务器
ViNT[111]	Efficient+仅解码器的 Transformer	31 M	视觉导航	4 Hz	2×4090, 3×Titan Xp, 4×P100, 8×1080Ti, 8×V100, 8×A100
VPT[112]	卷积+ResNet	0.5 B	Minecraft 中的具身智能体	20 Hz	720 V100 GPU
RT-1[102]	EfficientNet+TokenLearner+仅解码器的 Transformer	35 M	真实世界机器人任务	3 Hz	
RT-2[103]	PaLI-X	55 B	真实世界机器人任务	1~3 Hz	多 TPU 云服务器
RT-2-X[104]	ViT + 语言模型 UL2	55 B	真实世界机器人	1~3 Hz	多 TPU 云服务器
LIV[84]	CLIP		奖励学习	15 Hz	8 个英伟达 V100 GPUs
SMART[106]	仅解码器的 Transformer	11 M	双向动力学预测和控制	1 Hz	8 个英伟达 V100 GPUs
COMPASS[113]	3D-Reset 编码器	20 M	对比损失	30 Hz	8 个英伟达 V100 GPUs
PACT[105]	仅解码器的 Transformer	12 M	动作预测	50 Hz	8 个英伟达 V100 GPUs

4 挑战和未来方向

在本节中,将研究把大模型集成到机器人设置中的挑战。还探讨了解决其中一些挑战的潜在方法。

4.1 克服机器人训练大模型中的数据稀缺性

一个主要的挑战是,与大模型训练的互联网规模的文本和图像数据相比,机器人特定的数据很少。讨论了克服数据稀缺性的各种技术。例如,为了扩大机器人学习的规模,最近的一些研究建议使用游戏数据而不是专家数据来进行模仿学习。另一种技术是使用填充技术进行数据增强。

1)使用非结构化游戏数据和视频扩展机器人学习

语言条件学习,如语言条件行为克隆,或语言条件功

能学习需要访问大型带标注的数据集。为了扩大学习规模,在 Play-LMP[74] 中,作者建议使用远程操作的人类提供的游戏数据,而不是完全标注的专家演示数据。游戏数据是非结构化、没有标签的,收集成本低,但内容丰富。收集游戏数据不需要场景分段、任务分段或重置到初始状态。此外,在观看人类游戏进行远程模仿学习[114](long-horizon imitation learning by watching human play, MimicPlay)中,目标条件轨迹生成模型是基于人类游戏数据训练的。游戏数据包括人类用手与环境互动的未标记视频序列。最近的一些研究,如文献[91]表明,训练用于机器人操作任务的视觉语言提供模型需要比例非常小(少于 1%)的语言标注数据。

2) 使用填充进行数据增强

收集机器人数据需要机器人与真实的物理世界进行交互。这种数据收集过程可能会带来巨大的成本和潜在的安全问题。解决这一挑战的一种方法是使用生成式人工智能,如文本到图像扩散模型来进行数据增强。例如,ROSIE[115](scaling robot learning with semantic imagined experience)提出了一种基于扩散的数据增强方法。给定一个机器人操作数据集,它们使用填充在文本指导下创建各种看不见的物体、背景和干扰物。这些方法面临的一个重要挑战是开发能够生成足够的语义和视觉上多样化的数据的填充策略,同时确保这些数据在物理上是可行和准确的。例如,使用填充来修改机器人夹爪内物体的图像可能会导致图像具有物理上不真实的抓取,从而导致下游训练性能不佳。对生成基础模型的进一步研究不仅可以评估视觉质量,还可以评估物理真实感,这可能会提高这些方法的通用性。

3) 克服训练 3D 大模型的 3D 数据稀缺

目前,多模态视觉和语言模型可以分析 2D 图像,但它们缺乏与 3D 世界的连接,其中包括 3D 空间关系,3D 规划,3D 特征等。开发基于 3D VLM 模型的主要障碍在于 3D 数据的稀缺性,特别是与语言描述配对的数据。如上文所述,语言驱动的感知任务,如语言驱动的 3D 场景表示,需要访问 3D 数据或带有相机内参矩阵的多视图图像,这些数据类型不容易获得。未来需要创建新的数据集或数据生成方法来克服 3D 领域的数据稀缺性。

4) 通过高保真仿真生成合成数据

通过游戏引擎进行高保真仿真可以提供有效的数据收集手段,特别是解决机器人的多模态和 3D 感知任务。例如,多模态传感器数据和地面实况标签[116](multi-modal sensor data and ground truth labels, TartanAir)是机器人导航任务的数据集,在文献[117]中收集了移动物体、变化的光线和各种天气条件。通过在仿真中收集数据,可以获得多模态传感器数据和精确的地面真值标签,如立体 RGB 图像、深度图像、分割、光流、相机姿态和 LiDAR 点云。搭建了大量具有各种风格和场景的环境,涵盖了具有挑战性的视点和多种运动模式,这些都是使用物理数据收集平台难以实现的。

5) 使用 VLM 的数据增强

可以使用 VLM 进行数据增强。在语言条件控制的数据驱动指令增强[118](data-driven instruction augmentation for language-conditioned control, DIAL)中,引入了用于语言条件控制的数据驱动指令增强。DIAL 使用 VLM 标记离线数据集,用于语言条件策略学习。DIAL 使用 VLM 执行指令扩充,以减少重新标记离线控制数据集。DIAL 包括 3 个步骤:(1)在带有标注的小型机器人操作轨迹数据集上对 VLM(如 CLIP)进行对比微调;(2)通过使用微调后的 VLM 对更大的轨迹数据集进行标注的相关性评分来生成新的指令标签;(3)在原始和重新标注的数据集上使用行为克隆来训练语言条件策略。

4.2　实时性

在机器人上部署大模型的另一个瓶颈是这些模型的推理时间较长。在表 1 中,展示了其中一些模型的推理时间。由表中结果可以看出,为了机器人系统的可靠实时部署,一些模型的推理时间仍然需要改进。实时性是任何机器人系统的基本要求,提高大模型的计算效率需要进行更多的研究。

4.3　多模态表示的局限性

多模态交互假设模态是可标记的,并且可以在不丢失信息的情况下标准化为输入序列。多模态模型提供了多种模态之间的信息共享,是多模态 transformer 的一种变体,在每对输入之间具有跨模态关注。在多模态表征学习中,假设跨模态交互和不同模态之间的异质性维度都可以通过简单的嵌入来捕获。在多模态表征学习领域,单一多模态模型能否适应所有模态的问题仍然是一个公开的挑战。

此外,当模态和文本之间的配对数据可用时,可以将该模态直接嵌入文本中。在机器人应用中,有一些模式没有足够的数据可用,并且能够将它们与其他模式对齐,它们需要首先转换为其他模式,然后使用。例如,在苏格拉底模型[119]中,每种模态,无论是视觉的还是听觉的,最初都被翻译成语言,之后语言模型试图对这些模态做出响应。

4.4　机器人自身的可变性

另一个挑战是机器人自身的灵活可变性。机器人平台本质上是多样化的,具有不同的物理特性、配置和功能。机器人所处的现实世界环境也是多种多样、不确定的。由于所有这些可变性,机器人解决方案通常针对具有特定布局、环境和特定任务对象的特定机器人平台进行定制。这些解决方案不能在各种场景、环境或任务中

普遍化。因此,为了构建通用的预训练机器人大模型,一个关键因素是构建预训练任务不可知、交叉体现和开放式的大模型,并捕获不同的机器人数据。在 ROSIE[120] 中,通过在语义文本指导下对各种看不见的物体、背景和干扰物进行绘画,为机器人学习生成了一个多样化的数据集。

5 结 论

通过对最近文献的研究,本文调查了大模型在机器人技术中的各种应用。本文深入研究了这些模型如何增强机器人在感知、决策、规划以及控制等领域的能力。本文还讨论了关于大模型的泛化、零样本能力、多模态能力和可扩展性,这些特征具有改变机器人技术的潜力。然而,在机器人应用中整合大模型时,必须认识到未来研究中必须解决的挑战和潜在风险。机器人应用中的数据稀缺性、机器人真实的高度可变性、多模态表示的局限性和实时性能仍然是需要未来研究的重要问题。本文已经深入研究了其中的一些挑战,并讨论了改进的潜在途径。

参考文献

[1] JACOB D, MING W C, KENTON L, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. ArXiv preprint arXiv: 1810.04805, 2018.

[2] TOM B, BENJAMIN M, NICK R, et al. Language models are few-shot learners [C]. Conference and Workshop on Neural Information Processing Systems, 2020: 1877-1901.

[3] OpenAI. GPT-4 technical report [R]. ArXiv preprint arXiv:2303.08774,2023.

[4] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]. International Conference on Machine Learning,PMLR, 2021: 8748-8763.

[5] ADITYA R, MIKHAIL P, GABRIEL G, et al. Zero-shot textto-image generation[C]. International Conference on Machine Learning, 2021: 8821-8831.

[6] DANNY D, FEI X, MEHDI S M. PaLM-E: An embodied multimodal language model[C]. International Conference on Machine Learning, 2023: 8469-8488.

[7] ANIMESH G. PlaTe: Visually-grounded planning with transformers in procedural tasks[C]. IEEE Robotics and Automation Letters, 2022, 7(2):4924-4930.

[8] QIU J, LI L, SUN J, et al. Large ai models in health informatics: Applications, challenges, and the future[J]. IEEE Journal of Biomedical and Health Informatics, 2023,DOI: 10.1109/JBHI.2023.3316750.

[9] JIANKAI S, CHUANYANG Z, ENZE X, et al. Reasoning with foundation models: Concepts, methodologies, and outlook [J]. In Zenodo preprint, 2023,DOI: 10.5281/zenodo.10298866.

[10] DINGYUAN Z, DINGKANG L, HONGCHENG Y, et all. SAM3D: Zero-shot 3D object detection via segment anything model [J]. ArXiv preprint arXiv: 2306.02245, 2023.

[11] YINING H, HAOYU Z, PEIHAO C, et al. 3D-LLM: Injecting the 3D world into large language models [C]. International Conference on Neural Information Processing Systems, 2023: 20482-20494.

[12] WILLIAM C, SIYI H, RAJAT T, et al. Leveraging large language models for robot 3D scene understanding[J]. ArXiv preprint arXiv:2209.05629, 2022.

[13] SHERRY Y, OFIR N, YILUN D, et al. Foundation models for decision making: Problems, methods, and opportunities [J]. ArXiv preprint arXiv: 2303.04129, 2023.

[14] ZHAO M, HOMANGA B, VINCENT M, et al. CACTI: A framework for scalable multi-task multi-scene visual imitation learning [J]. ArXiv preprint arXiv: 2212.05711, 2022.

[15] 陈潇磊,尤波,李佳钰,等. 基于驾驶员模型的六足机器人自主/协同决策[J]. 仪器仪表学报, 2023, 4(4): 91-100.
CHEN X L, YOU B, LI J Y, et al. Autonomous/collaborative decision making of hexapod robots based on driver model [J]. Journal of Instrumentation, 2023, 4(4):91-100.

[16] MINAE K, SANG M X, KALESHA B, et al. Reward design with language models [C]. International Conference on Learning Representations, 2023: 1002-1016.

[17] XIDONG F, YICHENG L, ZIYAN W, et al. ChessGPT: Bridging policy learning and language modeling [J]. ArXiv preprint arXiv:2306.09200, 2023.

[18] YIFAN D, JUNYI L, TIANYI T, et al. Zero-shot visual question answering with language model feedback [J]. ArXiv preprint arXiv:2305.17006, 2023.

[19] ALEC R, JEFFREY W, REWON C, et al. Language models are unsupervised multitask learners[R]. OpenAI Blog, 2019.

[20] HECTOR L, ERNEST D, LEORA M. The Winograd schema challenge[R]. In KR, 2012.

[21] ALEX W, AMANPREET S, JULIAN M, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding [J]. ArXiv preprint arXiv:

1804. 07461, 2018.

[22] HUGO T, THIBAUT L, GAUTIER I, et al. LLaMA: Open and efficient foundation language models [J]. ArXiv preprint arXiv:2302. 13971, 2023.

[23] AAKANKSHA C, SHARAN N, JACOB D, et al. PaLM: Scaling language modeling with pathways [J]. ArXiv preprint arXiv:2204. 02311, 2022.

[24] AOHAN Z, XIAO L, ZHENGXIAO D, et al. GLM-130B: An open bilingual pre-trained model [C]. International Conference on Learning Representations, 2023: 2003-2021.

[25] ALEXEY D, LUCAS B, ALEXANDER K, et al. An image is worth 16×16 words: Transformers for image recognition at scale [C]. International Conference on Learning Representations, 2021: 3444-3458.

[26] KAI H, YUNHE W, HANTING C, et al. A survey on vision transformer [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 87-110.

[27] 刘铁,段勇. 融合 CNN 和 Transformer 的机器人室内场景识别[J]. 电子测量与仪器学报, 2023, 37(5): 223-229.
LIU T, DUAN Y. Robot indoor scene recognition integrating CNN and transformer [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37 (5): 223-229.

[28] ZHAI X H, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022: 12104-12113.

[29] CHEN X, WAMG X, CHANGPINYO S, et al. PaLI: A jointly-scaled multilingual language-image model [C]. International Conference on Neural Information Processing Systems, 2022: 19827-19839.

[30] MOSTAFA D, JOSIP D, BASIL M, et al. Scaling vision transformers to 22 billion parameters[C]. International Conference on Machine Learning, 2023: 7877-7889.

[31] CHEN X, DJOLONGA J, PADLEWSKI P, et al. Pali-x: On scaling up a multilingual vision and language model[J]. ArXiv preprint arXiv:2305. 18565, 2023.

[32] MATHILDE C, HUGO T, ISHAN M, et al. Emerging properties in self-supervised vision transformers [C]. International Conference on Computer Vision, 2021: 7109-7123.

[33] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 770-778.

[34] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: Learning robust visual features without supervision [J]. ArXiv preprint arXiv: 2304. 07193, 2023.

[35] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything [C]. International Conference on Computer Vision, 2023: 4015-4026.

[36] LI JUN N, LI D X, XIONG C M, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation [C]. International Conference on Machine Learning, PMLR, 2022: 12888-12900.

[37] ZENG Y H, JIANG CH H, MAO J G, et al. CLIP2: Contrastive language-image-point pretraining from real-world point cloud data [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 2022-2041.

[38] YAO L W, HUANG R, HOU L, et al. FILIP: Fine-grained interactive language-image pre-training [C]. Proceedings of International Conference on Learning Representations, 2022: 3021-3041.

[39] LI Y H, FAN H Q, HU R H, et al. Scaling language-image pre-training via masking[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 19021-19042.

[40] LI L H, ZHANG P, ZHANG H, et al. Grounded language-image pre-training [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022: 10965-10975.

[41] LIU M H, ZHU Y H, CAI H, et al. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023: 21736-21746.

[42] MINDERER M, GRITSENKO A, STONE A, et al. Simple open vocabulary object detection with vision transformers [C]. European Conference on Computer Vision, 2022: 728-755.

[43] HUANG W L, WANG CH, ZHANG R H, et al. VoxPoser: Composable 3D value maps for robotic manipulation with language models[C]. Conference on Robot Learning, 2023: 2320-2340.

[44] LIU SH L, ZENG ZH Y, REN T H, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection [J]. ArXiv preprint arXiv: 2303. 05499, 2023.

[45] ZHANG R R, GUO Z Y, ZHANG W, et al. PointCLIP:

Point cloud understanding by CLIP [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 8552-8562.

[46] YU X M, TANG L L, RAO Y M, et al. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling[C]. Conference on Computer Vision and Pattern Recognition, 2022: 19313-19322.

[47] XUE L, GAO M F, XING CH, et al. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 1179-1189.

[48] XUE L, YU N, ZHANG SH, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 27091-27101.

[49] CHANG A X, FUNKHOUSER T, GUIBAS L, et al. Shapenet: An information-rich 3d model repository[J]. ArXiv preprint arXiv:1512.03012, 2015.

[50] LI B, WEINBERGER K Q, BELONGIE S, et al. Language-driven semantic segmentation [C]. International Conference on Machine Learning, 2022: 3200-3228.

[51] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 12179-12188.

[52] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 16000-16009.

[53] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]. International Conference on Machine Learning, 2021: 8748-8763.

[54] ZHAO X, DING W CH, AN Y Q, et al. Fast segment anything[J]. ArXiv preprint arXiv:2306.12156, 2023.

[55] ZHANG CH N, HAN D SH, QIAO Y, et al. Faster segment anything: Towards lightweight sam for mobile applications [J]. ArXiv preprint arXiv: 2306.14289, 2023.

[56] YANG J Y, GAO M Q, LI ZH, et al. Track anything: segment anything meets videos [J]. ArXiv preprint arXiv:2304.11968, 2023.

[57] CHENG H K, SCHWING A G. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model[C]. European Conference on Computer Vision, 2022: 640-658.

[58] SHEN Q H, YANG X Y, WANG X CH. Anything-3D: Towards single-view anything reconstruction in the wild[J]. ArXiv preprint arXiv:2304.10261, 2023.

[59] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis [J]. Communications of the ACM, 2021: 65(1):99-106.

[60] SUN J K, XU Y, DING M Y, et al. NeRF-Loc: Transformer based object localization within neural radiance fields [J]. IEEE Robotics and Automation Letters, 2023: 8(8):5244-5250.

[61] SUN J K, QIU J N, ZHENG CH Y, et al. Aria-NeRF: Multimodal egocentric view synthesis[J]. ArXiv preprint arXiv:2311.06455, 2023.

[62] KERR J, KIM C M, GOLDBERG K, et al. LERF: Language embedded radiance fields [C]. International Conference on Computer Vision, 2023: 19729-19739.

[63] SHAFIULLAH N M M, PAXTON C, PINTO L, et al. CLIP-Fields: Weakly supervised semantic fields for robotic memory [C]. Robotics: Science and Systems, 2023: 40032-40052.

[64] HUANG C, MEES O, ZENG A, et al. Visual language maps for robot navigation [C]. In 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023: 10608-10615.

[65] HA H, SONG SH R. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models[C]. Conference on Robot Learning, 2022: 19222-19248.

[66] YE J L, WANG N Y, WANG X L. Feature NeRF: Learning generalizable nerfs by distilling pre-trained vision foundation models [J]. ArXiv preprint arXiv: 2303.12786, 2023.

[67] SIMEONOV A, DU Y, TAGLIASACCHI A, et al. Neural Descriptor Fields: SE (3)-equivariant object representations for manipulation[C]. IEEE International Conference on Robotics and Automation, 2022: 30292-30210.

[68] SHEN W, YANG G, YU A, et al. Distilled feature fields enable few-shot manipulation[C]. Conference on Robot Learning, 2023: 3094-3129.

[69] GOODWIN W, HAVOUTIS I, POSNER I. You only look at one: Category-level object representations for pose estimation from a single example [J]. ArXiv preprint arXiv:2305.12626, 2023.

[70] GOODWIN W, VAZE S, HAVOUTIS I, ET AL. Zero-shot category-level object pose estimation[C]. European Conference on Computer Vision, 2022: 21291-21310.

[71] UMEYAMA S. Least-squares estimation of transformation

parameters between two point patterns [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1991: 13(4):376-380.

[72] YE Y E, LI X T, GUPTA A, et al. Affordance diffusion: Synthesizing hand-object interactions [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 22479-22489.

[73] BAHL S, MENDONCA R, CHEN L L, et al. Affordances from human videos as a versatile representation for robotics [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 13778-13790.

[74] LYNCH C, KHANSARI M, XIAO T, ET AL. Learning latent plans from play [C]. Conference on Robot Learning, 2020: 1113-1132.

[75] LYNCH C, SERMANET P. Language conditioned imitation learning over unstructured data[J]. Robotics: Science and Systems, 2021: 3920-3939.

[76] GE Y, MACALUSO A, LI L E, et al. Policy adaptation from foundation model feedback[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 19059-19069.

[77] SHRIDHAR M, MANUELLI L, FOX D. CLIPort: What and where pathways for robotic manipulation [C]. Conference on Robot Learning, 2022: 894-906.

[78] ZENG A, FLORENCE P, TOMPSON J, et al. Transporter networks: Rearranging the visual world for robotic manipulation [C]. Conference on Robot Learning, 2020: 726-747.

[79] SHRIDHAR M, MANUELLI L, FOX D. Perceiver-actor: A multi-task transformer for robotic manipulation [C]. Conference on Robot Learning, 2023: 785-799.

[80] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 3029-3045.

[81] TEAM A A, BAUER J, BAUMLI K, et al. Human-timescale adaptation in an open-ended task space[C]. International Conference on Machine Learning, 2023: 9821-9840.

[82] NAIR S, RAJESWARAN A, KUMAR V, et al. R3M: A universal visual representation for robot manipulation[J]. ArXiv preprint arXiv:2203.12601, 2022.

[83] MA Y J, SODHANI S, JAYARAMAN D, et al. VIP: Towards universal visual reward and representation via value-implicit pre-training[C]. International Conference on Learning Representations, 2023: 29012-29033.

[84] MA Y J, KUMAR V, ZHANG A, et al. LIV: Language-image representations and rewards for robotic control[C]. International Conference on Machine Learning, 2023: 23301-23320.

[85] DAMEN D, DOUGHTY H, FARINELLA G M, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset [C]. European Conference on Computer Vision, 2018: 720-736.

[86] NAIR S, MITCHELL E, CHEN K, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation[C]. Conference on Robot Learning, 2022: 1303-1315.

[87] AHN M, BROHAN A, BROWN N, et al. Do as I can, not as I say: Grounding language in robotic affordances[C]. Conference on Robot Learning, 2023: 287-318.

[88] HUANG W L, XIA F, XIAO T, et al. Inner monologue: Embodied reasoning through planning with language models[J]. ArXiv preprint arXiv:2207.05608, 2022.

[89] LIN K, AGIA C, MIGIMATSU T, et al. Text2Motion: From natural language instructions to feasible plans[J]. Autonomous Robots, 2023, 47(8):1345-1365.

[90] MAHMOUDIEH P, PATHAK D, DARRELL T. Zero-shot reward specification via grounded natural language [C]. International Conference on Machine Learning, 2022: 14743-14752.

[91] MEES O, BORJA-DIAZ J, BURGARD W. Grounding language with visual affordances over unstructured data[C]. IEEE International Conference on Robotics and Automation, 2023: 11576-11582.

[92] CHEN Y CH, GANDHI R, ZHANG Y, et al. NL2TL: Transforming natural languages to temporal logics using large language models [J]. ArXiv preprint arXiv: 2305.07766, 2023.

[93] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. The Journal of Machine Learning Research, 2020: 21(1):5485-5551.

[94] CHEN Y CH, ARKIN J, DAWSON C, et al. Autotamp: Autoregressive task and motion planning with llms as translators and checkers [C]. 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6695-6702.

[95] HUANG J N, XIE S R, SUN J K, et al. Learning a decision module by imitating driver's control behaviors[C]. Conference on Robot Learning, 2021: 1-10.

[96] SUN J K, SUN H, HAN T, et al. Neuro-symbolic program search for autonomous driving decision module design [C]. Conference on Robot Learning, 2021: 21-30.

[97] SINGH I, BLUKIS V, MOUSAVIAN A, et al. ProgPrompt: Generating situated robot task plans using large language models [C]. IEEE International Conference on Robotics and Automation, 2023: 11523-11530.

[98] PUIG X, RA K, BOBEN M, et al. Virtual Home: Simulating household activities via programs[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8494-8502.

[99] LIANG J, HUANG W, XIA F, et al. Code as Policies: Language model programs for embodied control [C]. IEEE International Conference on Robotics and Automation, 2023: 9493-9500.

[100] XIAO T T, RADOSAVOVIC I, DARRELL T, et al. Masked visual pre-training for motor control[J]. ArXiv preprint arXiv:2203.06173, 2022.

[101] BUCKER A, FIGUEREDO L, HADDADIN S, et al. LATTE: Language trajectory transformer [C]. IEEE International Conference on Robotics and Automation, 2023: 72877294.

[102] BROHAN A, BROWN N, CARBAJAL J, et al. RT-1: Robotics transformer for real-world control at scale[C]. Robotics: Science and Systems, 2023: 2039-2056.

[103] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: Vision-language action models transfer web knowledge to robotic control [C]. Conference on Robot Learning, 2023: 8921-8943.

[104] O'NEILL A, REHMAN A, GUPTA A, et al. Open X-Embodiment: Robotic learning datasets and RT-X models[J]. ArXiv preprint arXiv:2310.08864, 2023.

[105] BONATTI R, VEMPRALA S, MA S, et al. Pact: Perception-action causal transformer for autoregressive robotics pre-training[C]. 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023: 3621-3627.

[106] SUN Y, MA S, MADAAN R, et al. SMART: Self-supervised multi-task pretraining with control transformers[C]. International Conference on Learning Representations, 2023: 3291-3312.

[107] 中国财经. 2024 世界机器人大会闭幕:头部企业加速构建产业生态行业发展仍需跨越鸿沟[EB/OL]. (2024-08-26). [2024-08-27]. http://www.ce.cn/cysc/newmain/yc/jsxw/202408/26/t20240826_39116945.shtml.
China Finance. 2024 world robot conference concludes: Leading enterprises accelerate the construction of industrial ecosystem, industry development still needs to cross the chasm [EB/OL]. (2024-08-26). [2024-08-27]. http://www.ce.cn/cysc/newmain/yc/jsxw/202408/26/t20240826_39116945.shtml.

[108] 前瞻产业研究院,华为云,首钢基金 CANPLUS. 2024 年前瞻中国 AI 大模型场景应用趋势蓝皮书[EB/OL]. (2024-07-30). [2024-08-27]. http://www.ecconsortium.org/Uploads/file/20240816/1723799167645192.pdf
Qianzhan Industry Research Institute, Huawei Cloud, Shougang Fund CANPLUS. Blue book on the trend of application of large model scenario of AI in China in 2024 [EB/OL]. (2024-07-30). [2024-08-27]. http://www.ecconsortium.org/Uploads/file/20240816/1723799167645192.pdf.

[109] BOUSMALIS K, VEZZANI G, RAO D, et al. RoboCat: A self-improving foundation agent for robotic manipulation [J]. ArXiv preprint arXiv: 2306.11706, 2023.

[110] REED S, ZOLNA K, PARISOTTO E, et al. A generalist agent [J]. ArXiv preprint arXiv: 2205.06175, 2022.

[111] SHAH D, SRIDHAR A, DASHORA N, et al. ViNT: A foundation model for visual navigation [J]. ArXiv preprint arXiv:2306.14846, 2023.

[112] BAKER B, AKKAYA I, ZHOKOV P, et al. Video PreTraining (VPT): Learning to act by watching unlabeled online videos [C]. Neural Information Processing Systems, 2022: 4382-4408.

[113] MA S, VEMPRALA S, WANG W, et al. Compass: Contrastive multimodal pretraining for autonomous systems[C]. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022: 1000-1007.

[114] WANG CH, FAN L X, SUN J K, et al. MimicPlay: Long-horizon imitation learning by watching human play[C]. Conference on Robot Learning, 2023: 8263-8282.

[115] YU T, XIAO T, STONE A, et al. Scaling robot learning with semantically imagined experience [J]. ArXiv preprint arXiv:2302.11550, 2023.

[116] WANG W SH, ZHU D L, WANG X W, et al. Tartanair: A dataset to push the limits of visual slam[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 4909-4916.

[117] SHAH S, DEY D, LOVETT C, et al. Airsim: High-fidelity visual and physical simulation for autonomous vehicles[C]. Field and Service Robotics: Results of the 11th International Conference. Springer International Publishing, 2018: 621-635.

[118] XIAO T, CHAN H, SERMANET P, et al. Robotic skill acquisition via instruction augmentation with vision-language models [J]. ArXiv preprint arXiv: 2211.11736, 2022.

[119] ZENG A, ATTARIAN M, ICHTER B, et al. Socratic models: Composing zero-shot multimodal reasoning with language [J]. ArXiv preprint arXiv: 2204.00598, 2022.

[120] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. ArXiv preprint arXiv:2302.13971, 2023.

作者简介

邓鹏(通信作者),分别在 2005 年和 2008 年于内蒙古科技大学获得学士学位和硕士学位,现为荆楚理工学院高级实验师,主要研究方向为机器人工程、机电控制。

E-mail: dengpeng@ jcut. edu. cn

Deng Peng (Correspondence author) received his B. Sc. degree and M. Sc. degree from Inner Mongolia University of Science and Technology in 2005 and 2008 respectively. Now he is a senior experimenter in Jingchu University of Technology. His main research interests include robot engineering and electromechanical control.

唐文涛,2005 年于山东理工大学获得学士学位,2009 年于山东师范大学获得硕士学位,2022 年于武汉大学获得博士学位,现为荆楚理工学院教授,主要研究方向为智能制造与先进技术。

E-mail: twt@ whu. edu. cn

Tang Wentao received his B. Sc. from Shandong University of Technology in 2005, M. Sc. degree from Shandong Normal University in 2009 and Ph. D. degree from Wuhan University in 2022. Now he is a professor in Jingchu University of Technology. His main research interests include intelligent manufacturing and advanced technology.

罗静,2012 年于湖北科技学院获得学士学位,2015 年于武汉大学获得硕士学位,2020 年于华中师范大学获得博士学位,现为荆楚理工学院讲师,主要研究方向为机器人工程、混沌控制、电机控制。

E-mail: luojing@ jcut. edu. cn

Luo Jing received his B. Sc. degree from Hubei University of Science and Technology in 2012, M. Sc. degree from Wuhan University in 2015 and Ph. D. degree from Central China Normal University in 2020, respectively. Now he is a lecturer in Jingchu University of Technology. His main research interests include robot engineering, chaos control and motor control.