面向长序列自主作业的非对称 Actor-Critic 强化学习方法

瞿宇珂[1]，任君凯[1*]，罗嘉威[1]，倪子淇[2]，卢惠民[1]，叶益聪[2]

（1.国防科技大学 智能科学学院, 湖南 长沙 410073；2.国防科技大学 空天科学学院, 湖南 长沙 410073）

摘　要：在医疗、服务、实验室、工业生产线等典型场景中，机器人需要具备较强的长序列自主作业能力。针对长序列自主作业任务学习难度大与奖励设计复杂等挑战，提出了一种非对称 Actor-Critic 强化学习方法。通过整合多个 Critic 网络协同训练单个 Actor 网络，引入生成对抗模仿学习为 Critic 网络生成内在奖励，降低了长序列任务学习难度。为进一步加速学习进程，设计了两阶段学习方法，利用模仿学习为强化学习提供高质量的预训练行为策略。在化学实验室中机器人长序列自主作业仿真结果表明，该方法能显著提高长序列操作技能的学习效率与行为策略的鲁棒性。

关键词：自主作业机器人；强化学习；非对称 Actor-Critic；长序列操作

中图分类号：TP249　　**文献标识码：**A　　**文章编号：**

Asymmetric Actor Critic Reinforcement Learning Method for Long-Sequence Autonomous Manipulation

QU Yuke[1], REN Junkai[1], LUO Jiawei[1], NI Ziqi[2], LU Huimin[1], Ye Yicong[2]*

(1. College of Intelligence Science and Technology, National Unibersity of Defense Technology, Changsha 410073, China; 2. College of Aerospace Science and Engineering, National Unibersity of Defense Technology, Changsha 410073, China)

Abstract: In typical scenarios such as healthcare, services, laboratories, and industrial production lines, robots need to have high long-sequence autonomous manipulation capabilities. An asymmetric Actor Critic reinforcement learning method is proposed to address the challenges of difficulty in learning long sequence autonomous manipulation tasks and complex reward design. By integrating multiple Critic networks to collaboratively train a single Actor network and introducing generative adversarial imitation learning to generate intrinsic rewards for Critic networks, the difficulty of learning long sequence tasks is reduced. To further accelerate the learning process, a two-stage learning approach was designed, utilizing imitation learning to provide a high-quality pre-trained behavior policy for reinforcement learning. The simulation results of long-sequence autonomous manipulation of robot in the chemistry laboratory show that this method can significantly improve the learning efficiency of long-sequence operation skills and the robustness of behavior policy.

Keywords: autonomous manipulation robot; reinforcement learning; asymmetric actor-critic; long-sequence operation

随着人工智能与机器人技术的快速发展，自主作业机器人逐渐引起了业界的广泛关注。由于其具有较高的执行效率与精度，不受体力限制且能够承担大量重复性、繁重、以及危险的任务，因此，自主作业机器人有望在医疗、服务、工业生产线等多个领域中解放人力，改善人类生活方式与生活质量。然而，在上述典型应用场景中，任务通常由多个环环相扣的步骤组成，要求机器人拥有较高的长序列自主作业能力，在精准执行每个子任务的同时，还需具备全局任务理解、动态调整与环境适应的能力[1][2]。因此，设计智能高效的长序列任务决策与规划策略，保证自主作业机器人在解决复杂实际任务时的作业效率与可靠性，具有重要的理论和实践意义。

近些年，模仿学习与强化学习在机器人自主作业领域取得了显著进展。在物体抓取与放置[3][4]、开关门[5]和积木堆叠[6]等任务中，机器人已经基本能够实现高精度的自主操作。然而，现有技术在面对长序列自主作业任务时仍面临诸多挑战。一方面，模仿学习依赖于预先采集的示教数据，缺乏有效的探索与泛化机制，因此在面对示教样本分布外的新状态时表现欠佳[7]；另一方面，虽然强化学习通过试错学习的方式，在一定程度上能够缓解上述问题，但在学习长序列操作技能时，由于探索空间庞大且缺乏先验知识，训练成本较高[8]。此外，强化学习在解决长时序问

收稿日期：2024-12-15

基金项目：国家自然科学基金资助项目(U1913202, U22A2059, 62203460)

第一作者：瞿宇珂（1999—），男，湖南岳阳人，博士研究生，E-mail：quyuke999@163.com；

* **通信作者：**任君凯（1991—），男，河北石家庄人，副研究员，博士，E-mail：jk.ren@nudt.edu.cn；

题时还面临稀疏奖励等挑战。尽管已有研究采用生成对抗模仿学习（Generative Adversarial Imitation Learning, GAIL）[9]学习内在奖励函数，或利用大语言模型推断环境中的奖励函数[10]来尝试解决这一问题。然而，鉴于技能序列的冗长性与复杂性，设计或学习有效的奖励函数依然十分困难。

为了克服上述挑战，研究者设计了分层学习框架，并且在经典模仿学习与强化学习的基础上提出了分层模仿学习[11]与分层强化学习[12]方法，将长序列任务根据其任务特点与控制频率分解为上下两层进行学习。在下层，长序列自主作业任务被拆解为多个简单的子任务进行学习，机器人能学习到多个操作子技能；在上层，通过设计子技能调度规划器，根据当前状态以及其他相关指标来调度下层子技能，通过将下层子技能顺序组合，能够完成原本复杂的长序列自主作业任务，从而降低了任务的学习难度，并提高了技能的可解释性和可迁移性。然而，分层学习在实际运用过程中，也存在一些挑战：首先，由于包含多个子技能模型，分层学习大大增加了模型复杂度[13]；其次，大量的子技能会给上层规划器带来更大的规划压力，甚至导致技能调度错误的情况。尤其是多个子技能对应的状态信息可能具有一定的相似度，容易严重影响技能调度的稳定性[14]；此外，由于每个子技能都是单独训练的，缺乏对技能间连接关系的考虑，这可能导致在实际应用中技能之间存在“序列交接”失误的问题[15]，进而影响整体任务的成功率。

因此，本文针对机器人长序列自主作业过程中学习难度大与奖励设计难的挑战问题，提出了一种面向长序列自主作业的非对称 Actor-Critic 强化学习方法。具体而言，该方法将复杂的长序列自主作业任务分解为若干子任务，旨在降低任务的学习难度以及奖励函数设计的复杂性。在此基础上，该方法构建了一个包含多 GAIL 模型和评价器（Critic）的网络架构，用于联合训练一个执行器（Actor）网络。进一步地，该方法设计了一种两阶段的训练方法，利用模仿学习初始化 Actor 网络的权重参数，以减少后续强化学习过程中的无效探索，从而有效提升学习效率。最终，通过多个机器人长序列自主作业的仿真实验，本文验证了所提方法的有效性。本文的研究聚焦于机器人自主作业中的长序列上肢操作任务，在这些任务中，机器人需要完成一系列灵巧操作，包括抓取、放置、打开抽屉和柜门等任务。由于前置任务的完成度会直接影响后续技能的执行，因此技能之间存在较强的关联性。本文主要贡献如下：

1)提出了一种非对称 Actor-Critic 强化学习方法，该方法将一个复杂的长序列自主作业任务分解为多个简单的子任务进行学习，每个子技能都配备了独立的 GAIL 网络和 Critic 网络，用于联合训练同一个执行器 Actor 网络。该方法降低了奖励函数设计的复杂性，简化了 GAIL 网络和 Critic 网络的学习过程，帮助 Critic 网络对 Actor 网络的指导更加准确，从而使智能体学习到更加高效鲁棒的操作技能。

2)设计了一种两阶段学习框架，一阶段通过模仿学习训练初始行为策略，以较少的训练代价为强化学习的 Actor 网络提供先验知识。二阶段，通过 GAIL 辅助下的奖励函数引导强化学习智能体学习全局最优策略。该方法可以有效提高 Actor 网络的学习效率。

3)在 Unity 引擎中搭建了长序列自主作业任务场景进行仿真实验，以评估本文方法。实验结果显示，该方法能够成功学会多种长序列操作技能用于解决不同的长序列自主作业任务，并展现出良好的鲁棒性和执行效率。

1　长序列自主作业任务分析与建模

1.1　长序列自主作业任务分析

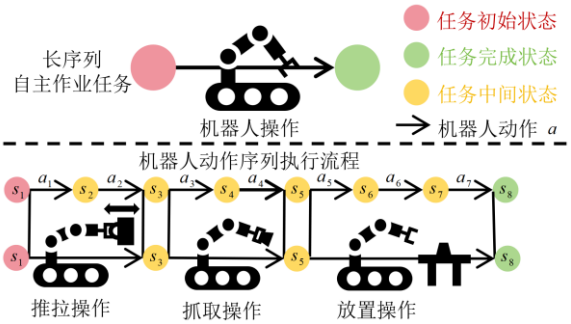


图 1　机器人执行长序列自主作业任务示意图

Fig.1　Schematic diagram of robot executing long sequence operation tasks

如图 1 所示，在执行长序列自主作业任务时，任务的长时性和顺序依赖性凸显，要求机器人执行一系列较长的动作序列，这些动作可归纳为不同的操作行为，如推拉、抓取、放置等。这些操作间存在着“序列交接”问题[15]，即前一操作的结束状态直接决定了后一操作的起始条件，对后续操作的成功执行具有决定性影响。因此，机器人不仅需要精准掌握每个独立操作的执行技巧，

还必须深入理解并有效学习各操作间的交接过程，以确保整个任务序列的连贯性和最终的成功率。

在强化学习的理论框架内，可以将长序列自主作业任务表示为马尔可夫决策过程（Markov Decision Processes，MDPs）。一个典型的 MDP 模型全面涵盖了状态空间、动作空间、奖励函数以及状态转移概率矩阵。值得注意的是，在深度强化学习中，通过训练神经网络能够隐式地学习环境中的状态转移规律，从而替代了显式构建状态转移矩阵的需求。鉴于此，本节将聚焦于状态空间、动作空间以及奖励函数的设计，这些要素在强化学习系统中扮演着至关重要的角色。

1.2 环境与智能体

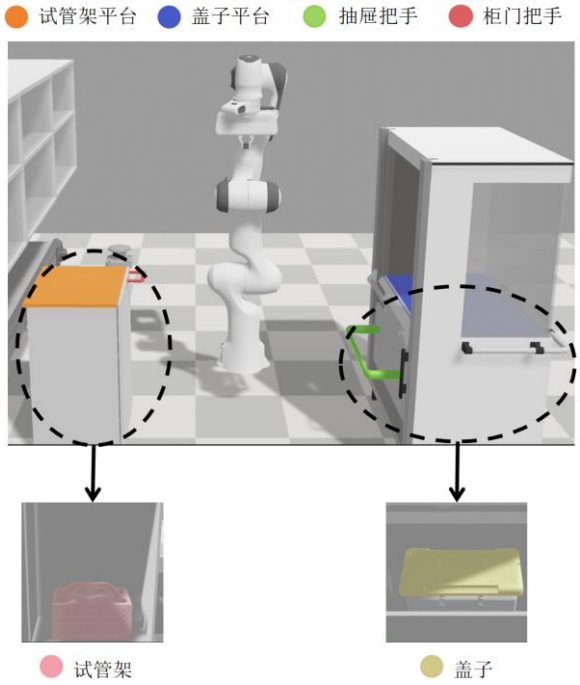


图 2 化学实验室仿真环境[14]

Fig.2 Chemical laboratory simulation environment

在介绍马尔可夫决策过程（MDP）的关键要素之前，本文首先将对环境与智能体进行简要说明。如图 2 所示，本文的研究场景设定在一个化学实验室中，旨在学习能完成长序列自主作业任务的操作技能。在这个环境中，共有五个可操作的目标物体、两个不可操作的目标物体以及一个智能体。关于目标物体的分类细节，请参见表 1。智能体为 Franka 机械臂，通过控制其末端执行器来执行一系列操作。

为了验证学习效果，本文设计了两种不同的长序列任务，在每个任务中智能体需要交互的对象不同，与对象的交互逻辑也有所差异。具体任务细节如下：

任务一：智能体需要打开抽屉，取出抽屉里的盖子，并将其放置于相应的平台上，最后末端执行器需要回到复位点。若抽屉未打开，智能体将无法取出盖子。

任务二：智能体需要打开绕单轴旋转的柜门，然后从柜子中取出试管架，并将其放置在指定的试管架平台上，最后末端执行器需要回到复位点。若柜门未打开，智能体将无法取出试管架。

以上两种任务均存在“序列交接”问题。如果前置操作未能成功完成或完成状态不佳（例如，抽屉仅打开一半），将对后续操作产生不良影响。

表 1 环境中目标物体的分类

Tab.1 Classification of target objects in the environment

	可操作目标物体	不可操作目标物体
目标物品类别		
目标物体名称	抽屉把手，柜门把手，试管架，盖子	试管架平台，盖子平台

1.3 状态空间设计

智能体的状态空间 S 包含了三类信息：智能体的本体信息 G 、目标物体信息 M 以及部分特定的环境信息 F 。智能体的本体信息 G 主要通过以下两个方面进行描述：末端执行器的速度 V ，以及末端夹爪的开合状态 $gripper_state$ 。对于目标物体信息 M ，采用相对于末端执行器的极坐标 m 进行表示。如图 3 所示，本文采用了一种特殊的表示方法：相对极坐标 m 中的角度 θ 用目标物体与末端执行器之间的归一化相对位置向量 $\vec{e}_{norm}=(p_x,p_y,p_z)$ 来替代，而极坐标的半径 ρ 则代表目标物体与末端执行器之间的距离。环境信息则包含一些难以通过相对位置量化的信息，例如抽屉的开启程度 d_{norm}（通过距离归一化值表示）或柜门的开启程度 θ_{norm}（通过角度归一化表示）。表 2 展现了每个任务的状态空间组成。

$$S=\{G,M,F\} \tag{1}$$

$$G=\{V,gripper_state\} \tag{2}$$

$$M=\{m_1,m_2,\cdots,m_n\} \tag{3}$$

$$m_i=\{p_x,p_y,p_z,\rho\} \tag{4}$$

$$F=\{d_{norm},\theta_{norm},\cdots\} \tag{5}$$

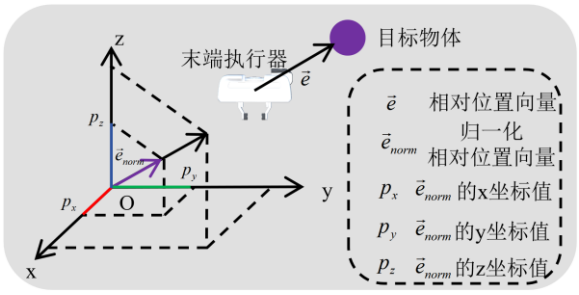


图 3 极坐标 m 中角度 θ 的特殊表示方法示意图

Fig.3 Schematic diagram of special representation method for angle θ in polar coordinate m

表 2　每个任务的状态空间组成

Tab.2　Composition of state space for each task

任务 \ 状态空间	智能体本体信息	目标物体信息	环境信息
任务 1	末端执行器的速度和末端夹爪的开合状态	抽屉把手、盖子和盖子平台平台相对于末端执行器的极坐标	复位点相对于末端执行器的极坐标，抽屉的开启程度
任务 2	末端执行器的速度和末端夹爪的开合状态	柜门把手、试管架和试管架平台相对于末端执行器的极坐标	复位点相对于末端执行器的极坐标，柜门的开启程度

1.4　动作空间设计

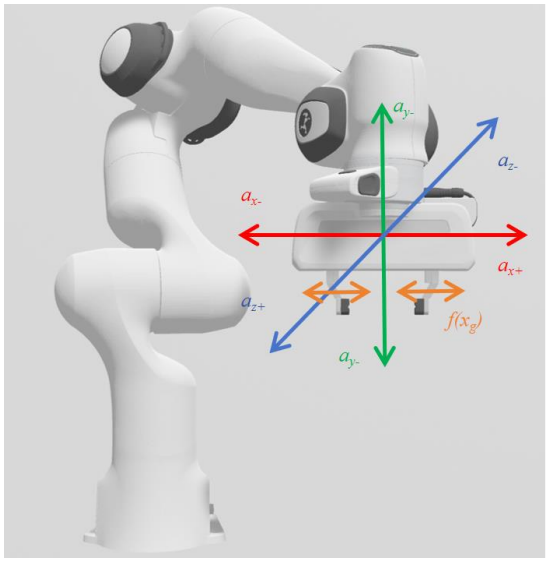


图 4　智能体的四维连续动作空间

Fig.4　The four-dimensional continuous action space of agent

本文采用四维连续动作空间 $A=(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$，具体设计如图 4 所示，包含两个方面：一是末端执行器的加速度，由 α_1（对应 $a_{x\pm}$）、α_2（对应 $a_{y\pm}$）和 α_3（对应 $a_{z\pm}$）三个参数表示；二是末端执行器夹爪的开合，该状态由 a_4（对应 x_g）来控制。由于夹爪的控制界面有限，它只提供打开和关闭的二进制命令。为了弥补这一点，引入了一个函数 $f(x_g)$ 来解释夹爪状态的切换过程。这种关系显示为：

$$f(x_g)=\begin{cases}\text{打开} & if\ x_g\in(0.9,1.0]\\ \text{保持现状} & if\ x_g\in[-0.9,0.9]\\ \text{关闭} & if\ x_g\in[-1.0,-0.9)\end{cases}\quad(6)$$

当 $x_g\in(-0.9,0.9)$ 时，夹爪保持当前状态，当且仅当 x_g 超出(-0.9,0.9)的区域时，夹持器的状态将切换为打开或关闭。

1.5　奖励函数设计

在强化学习中,奖励函数的设计是否合理，会直接影响智能体学习到的行为策略。通过设计能够准确反映任务本质的奖励函数，不仅能显著增进学习效果，还能提升任务完成的成功率。如式(7)所示，本文的的奖励函数 R 包括内在奖励 r_i 和外在奖励 r_e 两个部分。

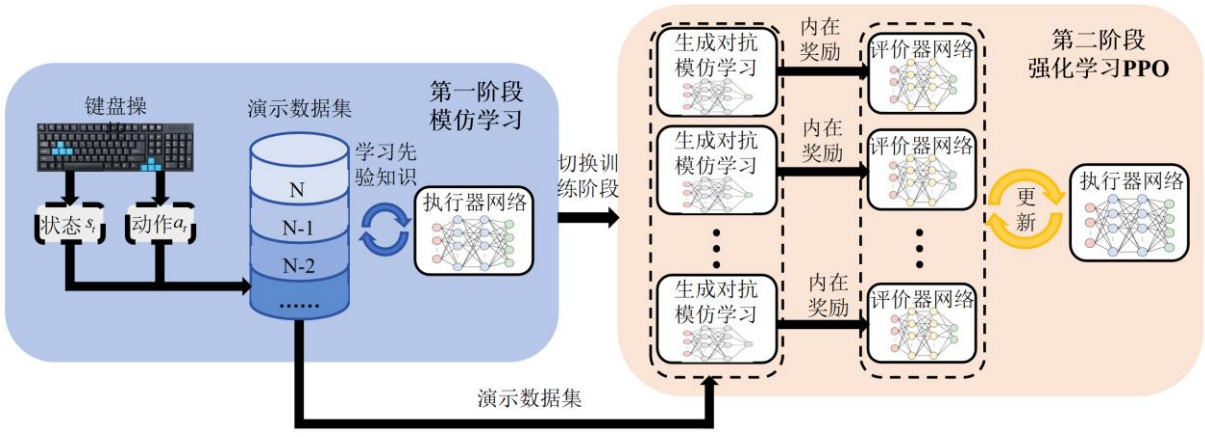


图 5　两阶段学习方法

Fig.5　Two stage learning method

$$R = r_i + r_e \tag{7}$$

其中，内在奖励 r_i 则由GAIL模块产生，GAIL模块用演示轨迹和代理轨迹来训练鉴别器，以便它能够评估代理的当前状态 s_t。内在奖励是根据鉴别器 *discriminator* 的分数产生的，如下式所示：

$$r_i = -ln(1 - discriminator(s_t)) \tag{8}$$

外在奖励 r_e 来自于智能体与环境的交互，每个长序列自主作业任务会被划分为不同的子任务进行学习，针对子任务设计不同的奖励函数，奖励函数设置如式(9)所示，r_{step} 是单步奖励，智能体的每步决策都会得到单步奖励 r_{step}，r_{step} 根据子任务完成的进程而变大。$r_{success}$ 是成功奖励，智能体在完成子任务时会得到成功奖励 $r_{success}$。

$$r_e = \begin{cases} r_{step} & \text{随着子任务进程增加} \\ r_{success} & \text{当子任务完成时} \end{cases} \tag{9}$$

2 方法框架

如图 5 所示，本文采用了一种两阶段的方法来训练智能体的 Actor 网络。在首个阶段，利用模仿学习方法来初始化 Actor 网络的权重参数，旨在为后续的强化学习过程奠定一个良好的起点。进入第二阶段，采用了近端策略优化（Proximal Policy Optimization, PPO）算法对 Actor 网络进行训练。为了应对长序列自主作业任务在训练过程中存在的挑战，设计了一个非对称 Actor-Critic 强化学习方法。通过上述方法，智能体最终能够成功且稳定地完成多个长序列自主作业任务。

2.1 模仿学习

为了减少强化学习过程中的无效探索，加速智能体的学习，本文引入模仿学习，为智能体 Actor 网络注入先验知识。首先，采用人工示教的方式，引导智能体完成任务演示。接着，利用这些示教过程中产生的状态 s_t 与动作 a_t 序列，构建出用于模仿学习的演示数据集。在模仿学习的阶段，通过最小化演示动作与 Actor 网络输出动作之间的均方误差损失，来实现 Actor 网络对先验知识的学习。值得注意的是，在训练过程中会避免只使用模仿学习。原因在于，本文方法的演示数据集相对较小，这可能导致策略模型迅速过拟合数据集，一个过拟合的模型并不是一个理想的学习起点。因此，在模仿学习的同时，本方法也会进行强化学习，以确保网络的潜在的泛化性。此外，在 PPO 算法中 Actor 网络会输出动作所遵循的高斯分布的均值 μ 和方差 σ，然后根据这个高斯分布来选择执行的动作 a_t。然而，如果将该动作 a_t 用于模仿学习，会导致方差 σ 迅速下降，从而降低 Actor 网络的贪婪程度。这不利于训练初期智能体对环境的探索。因此，本方法仅采用动作均值 μ 来进行模仿学习。

2.2 近端策略优化算法

近端策略优化（proximal policy optimization，PPO）算法是一种基于策略梯度的强化学习算法，通过引入一种简单而有效的裁剪机制，限制了 Actor 网络更新步长的大小，从而避免了因大幅度更新而导致的训练不稳定问题。PPO 算法在 Actor-Critic 框架下初始化两个神经网络——Actor 网络和 Critic 网络。Actor 网络负责根据当前状态输出动作概率分布，而 Critic 网络则评估当前状态的价值。算法利用 Critic 网络的输出值（即状态价值预估）与实际的奖励信息，进一步计算出优势函数 $\bar{A}_t$，该函数量化了采取特定动作相较于平均行为所展现出的优势程度。为了确保更新后的新 Actor 网络不过度偏离旧 Actor 网络，算法通过裁剪概率比 $r_t(\theta)$ 对 Actor 网络更新幅度进行控制，其数学表示如下式所示。

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta old}(a_t \mid s_t)} \tag{10}$$

$$L(\theta) = -\frac{1}{T}\sum_{t=1}^{T} min(r_t(\theta)\bar{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\bar{A}_t) \tag{11}$$

θ 代表 Actor 网络参数，$r_t(\theta)$ 代表新旧 Actor 网络在状态 s_t 下选择动作 a_t 的概率比值，$L(\theta)$ 是 Actor 网络的损失函数，*clip* 是一个裁剪函数，用于将 $r_t(\theta)$ 的值限制在 $1-\varepsilon$ 和 $1+\varepsilon$ 的范围内，其中 ε 是一个人为设定的超参数。简单来说，当新旧 Actor 网络之间的差异超出预设阈值时，裁剪机制将介入调整，以确保 Actor 网络更新维持在一个合理的幅度范围之内。

2.3 非对称 Actor-Critic 强化学习

为了提高长序列操作任务的学习效率，本文提出了一种非对称 Actor-Critic 强化学习方法，使用多个 GAIL 模型和 Critic 共同训练一个 Actor 网络，具体结构如图 6 所示。

2.3.1 算法结构

本方法将复杂的长序列任务拆解为多个简单的子任务，以此降低长序列任务学习的难度。针对每个拆解后的子任务，本方法独立设计了奖励函数，并分配了特定的 Critic 网络来学习各子任务的状态价值函数。这些 Critic 网络均具备相同的网络架构，包括一层标准化层（Normalize）、三层隐藏层（每层包含 256 个神经元）和一层用

于输出价值函数的输出层组成。所有 Critic 网络共同训练一个 Actor 网络，Actor 网络的网络架构跟 Critic 网络大致一样，区别在于最后一层为四维的动作均值 μ 输出层，动作方差 σ 由一个可学习的四维参数进行表示。同时，为了进一步提升学习效果，每个子任务配置了独立的 GAIL 网络，用于生成子任务的内在奖励。每个 GAIL 网络也采用了相同的网络结构，包括一层标准化层（Normalize）、两层隐藏层（每层各含 128 个神经元），以及一层负责输出对当前状态评估值的输出层。

Actor 网络状态输入为长序列任务中所有的状态信息 s_t，而在构建 Critic 网络和 GAIL 网络的输入时，本方法设计了一种状态掩码机制。该机制根据子任务 ID，选择性地排除与当前子任务无关状态信息。其次，根据子任务 ID 将掩码状态 s_t^{mask} 输入相应的 Critic 网络和 GAIL 网络。基于上述操作，Critic 网络和 GAIL 网络能够专注于与当前子任务高度关联的状态空间，从而提高了学习的准确性和效率。

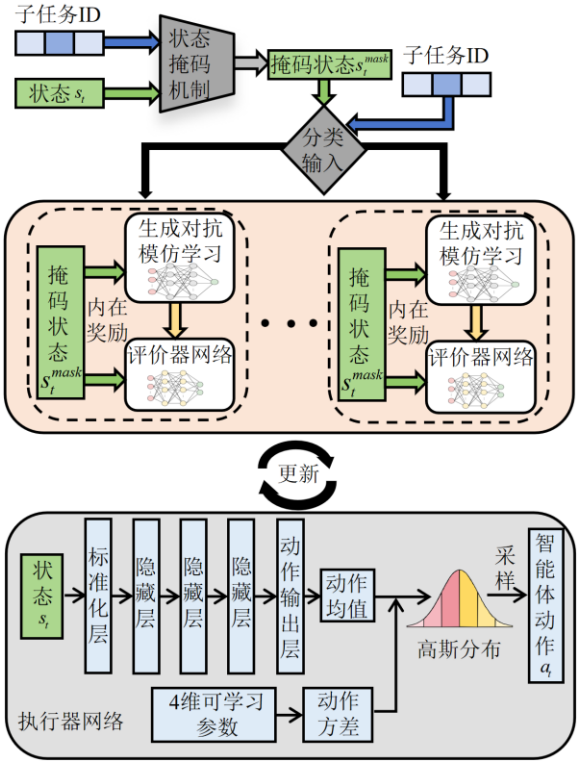


图 6　非对称 Actor-Critic 结构

Fig.6　Asymmetric Actor-Critic Structure

2.3.2　网络微调

一旦各个子任务的学习阶段完成，Actor 网络便成功掌握了多个的子技能。为了习得这些子技能间的平滑过渡行为，需对 Actor 网络实施进一步的微调。具体而言，在非对称 Actor-Critic 强化学习方法中加入了针对长序列自主作业任务的 Critic 网络和 GAIL 网络，用于训练 Actor 网络。鉴于 Actor 网络已具备各子技能的知识基础，无需构建高度复杂的奖励函数，仅需设计简洁的引导机制，促使 Actor 网络能够自主地将离散化的子技能融合为连贯的长序列操作技能。具体的奖励函数设计细节参见式(12)与式(13)，其中 r_{step} 为单步奖励，w_i 代表每个子任务完成前后的奖励值，$r_{success}$ 为成功奖励。

$$r_e = \begin{cases} r_{step} = w_1^{step} + w_2^{step} + \cdots + w_n^{step} \\ r_{success} \end{cases} \tag{12}$$

$$w_i^{step} = \begin{cases} 0 & \text{子任务}i\text{ 没有完成} \\ \alpha & \text{子任务}i\text{ 已经完成} \end{cases} \tag{13}$$

在推进长序列任务学习的同时，本方法持续维护对各个子任务的学习，旨在防止 Actor 网络在专注于长序列任务学习时，可能会对已掌握的知识造成潜在破坏。

3　实验与分析

为了验证本文提出算法的有效性，本节在 Unity 仿真平台上设计了如 1.2 节所描述的操作任务，并与文献[14]中的技能学习方法进行对比（后续统称对比方法），以证明本文算法的有效性和优势。对比方法是一种端到端的操作技能学习方法，其主体网络为传统的 Actor-Critic 架构。此外，本节对第一阶段的模仿学习展开消融实验，进一步验证了本文提出的两阶段学习方法的优势。

3.1　实验参数设置

表 3　算法部分超参数设置

Tab.3　Partial hyperparameter settings for algorithm

参数	值
Learning_rate	3e-4
Beta	1e-2
Epsilon	0.2
Lambd	0.95
Buffer_size	10,240
Batch_size	1,024

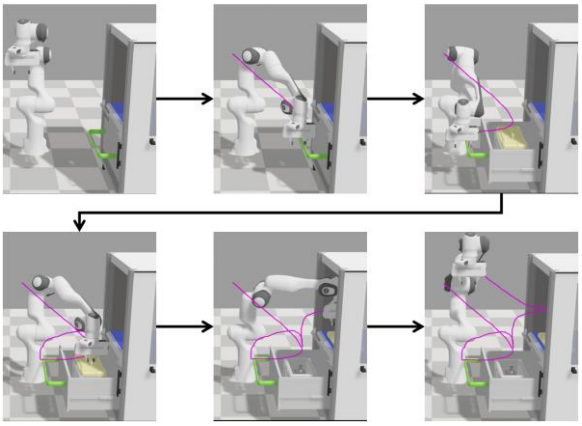
仿真环境中更新频率为 1000hz，相应的 Franka 机械臂的控制频率为 1000hz，智能体决策频率为 25hz，即每 0.04 秒就会进行一次动作决策，并且在两次决策之间，智能体会沿用上一次决策的动作进行控制。为了实验的公开性，表 3 将呈现部分算法超参数。其中，Learning_rate 为学习率；Beta 决定了算法的“贪婪程度”，Beta 越大，动作方差下降越慢或上升的越快；Epsilon

决定了新旧 Actor 网络之间更新的步长上限，等价于 2.2 节中的参数ε；Lambd 为广义优势估计（Generalized Advantage Estimation, GAE）中的正则化参数[16]；Buffer_size 为在更新网络前每个子任务应采集的样本量，Bacth_size 为更新网络时每个子任务的样本批次。

3.2 对比实验

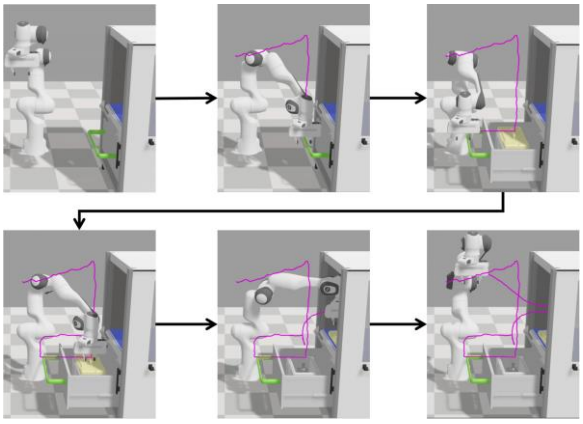
3.2.1 关于运动轨迹的对比实验

任务一和任务二的具体描述如 1.2 节所示。基于非对称 Actor-Critic 结构，任务一可被拆分成两个子任务进行学习：（1）拉动抽屉：智能体需要拉开关闭着的抽屉。（2）抓取物体并放置：智能体需要从抽屉抓取盖子，将其放在相应的平台上后复位末端执行器。基于非对称 Actor-Critic 结构，任务二同样可被拆分成两个子任务进行学习：（1）开启门扉：智能体需要打开绕着单轴旋转的柜门。（2）抓取物体并放置：智能体需要从柜子里抓取试管架，将其放在柜子上后复位末端执行器。



（a）本文方法的运动轨迹

（a）The motion trajectory of our method

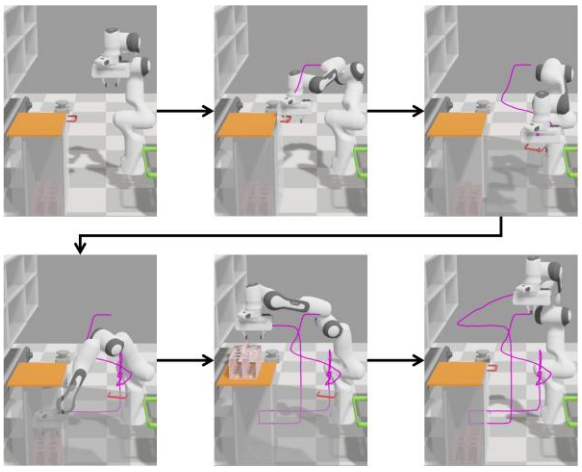


（b）对比方法[14]的运动轨迹

（b）The motion trajectory of Comparison method

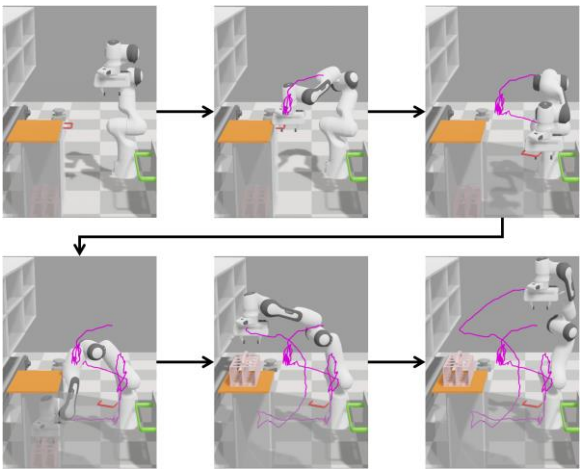
图 7 任务一中不同训练方法的运动轨迹对比图

Fig.7 Comparison of Motion Trajectory of Different Training Methods in Task 1



（a）本文方法的运动轨迹

（a）The motion trajectory of our method



（b）对比方法[14]的运动轨迹

（b）The motion trajectory of Comparison method

图 8 任务二中不同训练方法的运动轨迹对比图

Fig.8 Comparison of Motion Trajectory of Different Training Methods in Task 2

图 7 和图 8 是在不同训练方法下，智能体完成长序列自主作业任务的过程，其中紫色曲线为末端执行器的轨迹。可以看出，本文方法训练出的智能体，其末端执行器运动轨迹更加平滑。

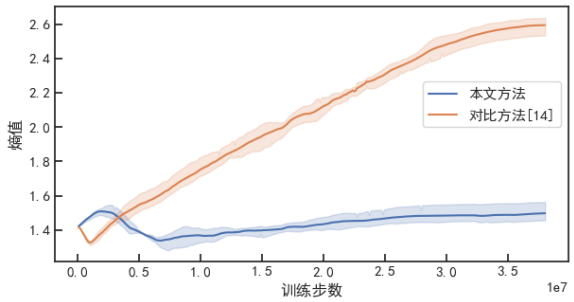


图 9 任务一中不同方法训练过程的 Actor 网络熵值曲线

Fig.9 Entropy curves of Actor networks trained using different methods in Task 1

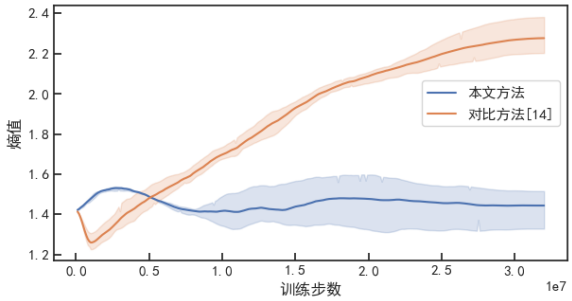
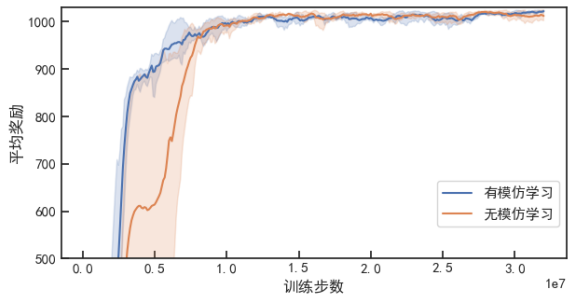


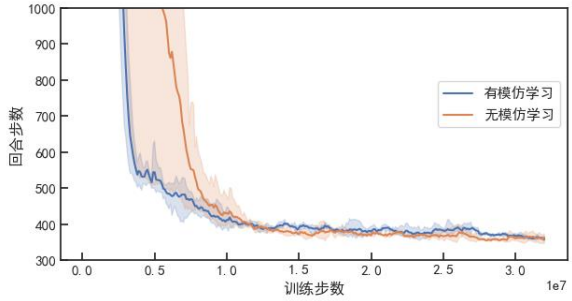
图 10 任务二中不同方法训练过程的 Actor 网络熵值曲线

Fig.10 Entropy curves of Actor networks trained using different methods in Task 2



（a）任务一中子任务学习的奖励曲线

（a）The reward curve for subtask learning in Task 1

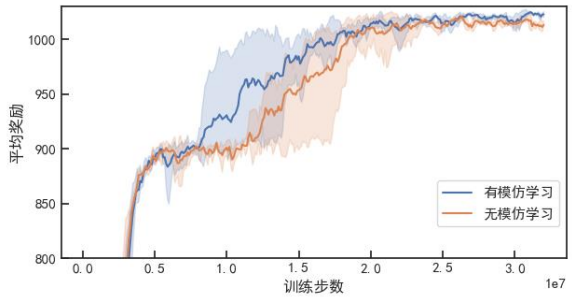


（b）任务一中子任务学习的回合步数曲线

（b）Round step curve of subtask learning in Task 1

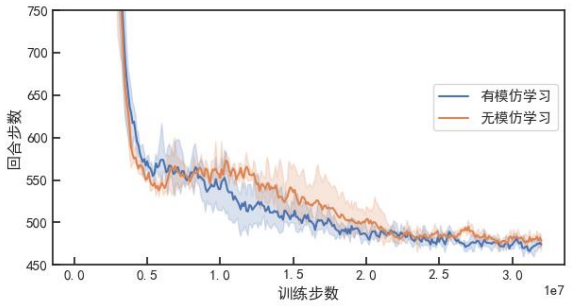
图 11 任务一中关于模仿学习的消融实验

Fig.11 Ablation experiment on imitation learning in Task 1



（a）任务二中子任务学习的奖励曲线

（a）The reward curve for subtask learning in Task 2



（b）任务二中子任务学习的回合步数曲线

（b）Round step curve of subtask learning in Task 2

图 12 任务二中关于模仿学习的消融实验

Fig.12 Ablation experiment on imitation learning in Task 2

图 9 和图 10 是在不同训练方法下，Actor 网络动作不确定性的熵值，熵值越大代表动作方差 σ 越大。值得注意的是，在任务一中，子任务学习阶段完成后，Actor 网络需要进行参数微调，才能学会子技能之间的过渡行为。而在任务二中，Actor 网络即使不经过参数微调，也顺利掌握了子技能之间的过渡行为。因此，任务一的完整训练步数比任务二多 6×10^6 步。从图 9 和图 10 可以明显发现，本文方法的动作熵值更低，这说明 Actor 网络的动作方差较小，进一步解释了末端执行器的运动轨迹更加平滑。

3.2.2 关于任务完成情况的对比实验

表 4 和表 5 描述了不同方法下的任务成功率和平均回合步数，每个数据都进行了 200 次实验。回合步数为智能体在一个回合内的决策步数，在一个回合中智能体的决策步数到达上限或者智能体完成子任务时，该回合将会结束，因此回合步数越低代表任务完成越快，训练效果越好。β 代表位置噪声，通过高斯分布拟合出一种动态变化的偏移量施加在物体的实际位置上，β 值越大，表示物体实际位置越不准确，状态空间中的相对极坐标也越不准确。随着位置噪声 β 的增大，本文方法仍然能保持较高的成功率。并且在相近成功率的情况下，本文方法也能以较短的步数完成任务。由此可见本文方法能够高效且鲁棒的完成长序列自主作业任务。

3.3 消融实验

图 11 展示了任务一中关于模仿学习的消融实验。其中，图 11 (a)为同时学习多个子任务时的平均奖励曲线，奖励越大代表训练效果越好。图 11 (b)为训练过程中智能体的平均回合步数曲线。同理，图 12 展示了任务二中关于模仿学习的消融实验。从曲线图可以看出，有模仿学习的平均奖励曲线和平均回合步数曲线皆优于没有模仿学习的平均奖励曲线和平均回合步数曲线，证明了两阶段的学习方法可以提高算法的收敛速度。

表 4　任务一中不同方法的成功率和平均回合步数

Tab.4　The success rate and average number of turn steps of different methods in Task 1

位置噪声 β /cm	$\beta=\pm0.5$		$\beta=\pm1.0$		$\beta=\pm2.0$	
方法 \ 指标	成功率/%	平均回合步数	成功率/%	平均回合步数	成功率/%	平均回合步数
本文方法	**99.5**	**728**	**99.5**	**733**	**99.5**	**736**
对比方法[14]	**99.5**	866	**99.5**	873	**99.5**	883
位置噪声/cm	$\beta=\pm3.0$		$\beta=\pm4.0$		$\beta=\pm5.0$	
方法 \ 指标	成功率/%	平均回合步数	成功率/%	平均回合步数	成功率/%	平均回合步数
本文方法	**99.5**	**736**	**98.0**	**812**	**96.0**	**876**
对比方法[14]	97.5	983	94.0	1160	81.0	1747

表 5　任务二中不同方法的成功率和平均回合步数

Tab.5　The success rate and average number of turn steps of different methods in Task 2

位置噪声 β /cm	$\beta=\pm0.5$		$\beta=\pm1.0$		$\beta=\pm2.0$	
方法 \ 指标	成功率/%	平均回合步数	成功率/%	平均回合步数	成功率/%	平均回合步数
本文方法	**96.0**	**1465**	**96.0**	**1492**	**96.0**	**1458**
对比方法[14]	81.5	2240	81.0	2267	83.5	2161
位置噪声/cm	$\beta=\pm3.0$		$\beta=\pm4.0$		$\beta=\pm5.0$	
方法 \ 指标	成功率/%	平均回合步数	成功率/%	平均回合步数	成功率/%	平均回合步数
本文方法	**96.5**	**1512**	**96.0**	**1492**	**91.5**	**1561**
对比方法[14]	72.5	2413	77.5	2364	83.5	2189

4　结论

本文聚焦于机器人自主作业任务中存在的挑战，创新性地提出了一种非对称 Actor-Critic 强化学习方法。该方法将冗长复杂的长序列自主作业任务分解为多个简洁易处理的子任务进行学习。每个子任务均配备独立的 GAIL 网络和 Critic 网络。这种分解策略显著降低了 GAIL 网络和 Critic 网络学习难度，使 Critic 网络能为 Actor 网络提供更精确、更有针对性的反馈。此外，任务分解还能简化奖励函数的设计。为提升学习效率，本文引入了两阶段学习策略。通过模仿学习，为后续强化学习阶段奠定优质的学习起点，从而加速了整体学习进程。在实验部分，本文验证了该方法在多个自主作业任务下的有效性，并将其性能与基于传统的 Actor-Critic 架构的方法进行了对比。结果显示，本文方法能够高效学习长序列自主作业行为策略，在任务成功率、执行效率以及鲁棒性等方面展现出显著优势。在未来，我们计划将研究拓展至机器人多源感知领域，实现从多源感知输入到决策输出的端到端无缝映射。

参考文献

[1] BILLARD A and KRAGIC D. Trends and challenges in robot manipulation [J]. Science, 2019, 364(6466).

[2] Tee K P, Cheong S, Li J, et al. A framework for tool cognition in robots without prior tool learning or observation [J]. Nature Machine Intelligence, 2022, 4:

533-543.

[3] Zhang J, Gireesh N, Wang J, et al. GAMMA: Graspability-Aware Mobile MAnipulation Policy Learning based on Online Grasping Pose Fusion [C]// Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 1399-1405.

[4] FUJITA Y, UENISHI K, UMMADISINGU A, et al. Distributed reinforcement learning of targeted grasping with active vision for mobile manipulators [C]// Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 9712-9719.

[5] XIA F, LI C, Martín-Martín R, et al. ReLMoGen: Integrating Motion Generation in Reinforcement Learning for Mobile Manipulation [C]// Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021:4583-4590.

[6] Jia Z, Thumuluri V, Liu F, et al. Chain-of-thought predictive control[J]. arXiv preprint 2023, arXiv: 2304.00776.

[7] HUANG X, BATRA D, RAI A, et al. Skill Transformer: A Monolithic Policy for Mobile Manipulation [C]// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 10818-10828.

[8] THOR M and MANOONPONG P. Versatile modular neural locomotion control with fast learning [J]. Nature Machine Intelligence, 2022, 4: 169-179.

[9] LI C, VIASTELICA M, BLAES S, et al. Learning agile skills via adversarial imitation of rough partial demonstrations[C]// Proceedings of Conference on Robot Learning (CoRL). PMLR, 2023: 342-352.

[10] TRIANTAFYLLIDIS E, CHRISTIANOS F, LI Z. Intrinsic language-guided exploration for complex long-horizon robotic manipulation tasks[C]// Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 7493-7500.

[11] ZHANG D, LI Q, ZHENG Y, et al. Explainable hierarchical imitation learning for robotic drink pouring [J] IEEE Transactions on Automation Science and Engineering, 2022, 19(4): 3871-3887.

[12] LI C, XIA F, Martín-Martín R, et al. Hrl4in: Hierarchical reinforcement learning for interactive navigation with mobile manipulators [C]// Proceedings of Conference on Robot Learning. PMLR, 2020: 603-616.

[13] JIANG R, CHENG X, SANG H, et al. GTHSL: A Goal-Task-Driven Hierarchical Sharing Learning Method to Learn Long-Horizon Tasks Autonomously [J]. IEEE Transactions on Industrial Electronics, 2024: 1-12.

[14] TRIANTAFYLLIDIS E, ACERO F, LIU Z, et al. Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network ROMAN [J]. Nature Machine Intelligence, 2023, 5(9): 991-1005.

[15] GU J, CHAPLOT D S, SU H, et al. Multi-skill mobile manipulation for object rearrangement [C]// Proceedings of The Eleventh International Conference on Learning Representations (ICLR). 2023.

[16] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[J]. arXiv preprint 2015, arXiv: 1506.02438.