arXiv:2301.07137v1 [cs.RO] 17 Jan 2023

异构多机器人强化学习

Matteo Bettini
剑桥大学，英国剑桥
mb2389@cl.cam.ac.uk

Ajay Shankar 剑桥
大学，英国剑桥
as3233@cl.cam.ac.uk

Amanda Prorok 剑
桥大学，英国剑桥
asp45@cl.cam.ac.uk

摘要

合作多机器人任务可以从机器人在物理特征和行为特征上的异质性中获益。尽管如此，传统的多智能体强化学习（MARL）框架缺乏显式适应策略异构性的能力，通常还会限制智能体共享神经网络参数。这种强制同质性限制了模型在受益于异质行为的任务中的应用。本文阐明了异质性在MARL 策略中的作用。为此，我们提出异构图神经网络近端策略优化（HetGPPO），这是一种训练异质 MARL 策略的新范式，利用图神经网络实现可微的智能体间通信。HetGPPO 使相互通信的智能体能够学习异质行为，同时支持在部分可观测环境中进行完全去中心化训练。除此之外，我们还从分类学角度进行综述，揭示了比以往所识别的更多异构类别。为说明模型的必要性，我们对同质模型可用于模拟异质行为的技术进行了特征分析，并展示了这种“表观异构性”在真实世界条件下的脆弱性。通过仿真和真实世界实验，我们表明：（i）当同质方法因异构需求过强而失效时，HetGPPO 能够成功；（ii）当同质方法能够学习表观异质行为时，HetGPPO 对训练噪声和部署噪声均具有更强的韧性。

关键词

异构性、多智能体强化学习、多机器人系统

1 引言

用于应对复杂协作任务的多机器人系统，往往可以借助异质的物理和/或行为特征来完成使命。这类异构系统已应用于灾害应对 [37], 协作建图 [8], 农业 [26], 和包裹运输 [22]等领域。然而，为这些任务综合出最优去中心化策略在计算上可能非常困难，其计算复杂度通常会随智能体数量呈指数级增长 [5]。尽管存在元启发式算法 [9],等更快速且可扩展的解决方案，但它们无法保证最优性。多智能体强化学习（MARL） [61] 可作为一种可扩展的方法，用于为这些问题寻找近似最优解。然而，缺乏智能体间通信的 MARL 算法难以直接应用于真实世界的机器人问题，因为单个智能体普遍存在部分可观测性。通信是克服这种部分可观测性并实现协作的关键。

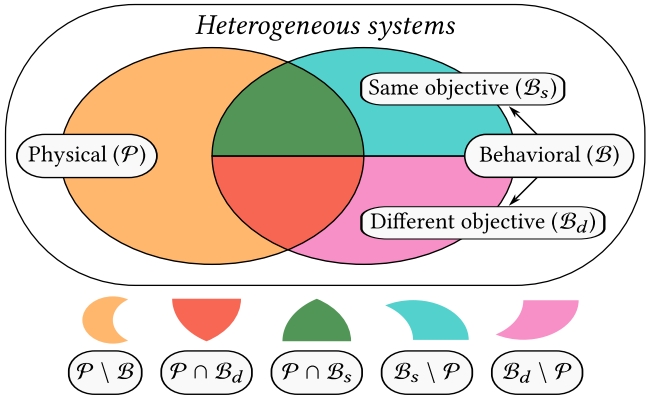


图 1：异构多机器人/智能体系统的分类体系。上方：三种异质性类别（$\mathcal{P}$、$\mathcal{B}_s$、$\mathcal{B}_d$）。下方：五种互斥的异质性子类。每个异构系统都属于其中一种子类。

我们的研究涉及异质多机器人强化学习，这一范式处于MARL（具备智能体间通信）与异构多机器人系统的交界处。

大多数协作式多智能体强化学习研究都要求智能体共享策略神经网络参数，以提高训练样本效率 [24, 44, 53]。这会导致智能体的模型完全相同，因而是同质的。虽然这有助于加快训练，但可能会阻碍在需要异质行为的场景中进行学习。克服这种强加同质性的经典方法，是将一个唯一整数（例如智能体的索引）作为每个智能体观测的一部分 [19, 24]。这样，智能体便能共享同一策略，同时表现出表面上不同的行为。尽管这一方案应用广泛，但仍有许多缺点 [12]。

我们致力于学习真正异质的去中心化多智能体强化学习策略。针对不同目标进行优化时，学习异质策略是常见做法 [33], ，但将这一范式应用于目标相同的场景的研究却十分有限。现有解决方案寥寥无几，且针对特定任务量身定制，因此未能对多智能体强化学习中的异构性进行更广泛的研究与分类。此外，这些方案仅限于无噪声电子游戏类多智能体强化学习基准 [30, 45],，未考虑具备智能体间通信的现实世界多机器人任务。因此，我们需要一个框架，使通信中的多智能体强化学习智能体能够实现真正的异构性，并学习可在（现实世界）异质多机器人系统中以去中心化方式运行的策略。

在这项工作中，我们提出了异质图神经网络近端策略优化（HetGPPO），这是一种用于异构多智能体强化学习的范式，可克服上述问题。HetGPPO 是一个用于训练异构多智能体强化学习策略的框架

它利用图神经网络（GNN）实现可微分的智能体间通信。我们的架构使异构智能体能够进行学习，同时仅以局部通信和局部观测为条件。这使得 HetGPPO 能够以去中心化方式进行训练，符合分散式训练、分散式执行（DTDE）范式 [25]。

我们首先在第 2 节提出异构系统分类法，旨在根据异构性的来源对这类系统进行分类。我们在第 3 节中运用这一分类法，对多机器人系统和多智能体强化学习领域的相关工作进行归类。第 4 节阐述本文所研究的多智能体强化学习问题。第 5 节介绍 HetGPPO 及其同质对应方法 GPPO。为说明策略异构性的必要性，我们归纳并定义了同质模型用于模拟异质行为的技术（第 6 节）。通过示例场景，我们展示了这些技术的工作方式，以及与真正的异构模型相比，它们在真实世界条件下为何可能表现脆弱。最后，第 7 节对我们的框架在模拟环境和真实世界的多机器人协作场景中进行了评估。结果表明：（i）当同质方法因异构需求强烈而失效时，HetGPPO 能够成功；（ii）当同质方法能够学会表面上异质的行为时，HetGPPO 对训练噪声和部署噪声都具有更强的韧性。此外，我们的真实机器人实验表明，异构策略本质上更能适应真实世界条件。

本文展示了异质多智能体强化学习应用于真实世界多机器人系统的强大能力。我们的主要贡献如下：

(1) 提出一种异构系统分类法，统一归类多机器人和多智能体领域的研究；(2) 探讨同质模型用于模拟异质行为的行为类型判定 技术，并通过实证证据说明这些技术在部署时的脆弱性；(3) 提出 HetGPPO，这是一种能够以去中心化方式学习异构通信策略的多智能体强化学习模型；(4) 通过仿真和真实世界实验，详细评估异构策略与同质策略在若干合作多机器人任务中的性能和韧性。

2 异构系统分类法

尽管已有大量工作试图推动对异构系统的研究（参见 [1]及其中的参考文献），机器人与学习领域仍缺乏一套共同认可且结构清晰的异构系统分类体系。为了恰当刻画异构性（多样性）领域的相关工作，我们提出了一种异构系统分类法，如图1所示。根据该分类法，系统异构性分为两类：物理 （P）和行为（B）。

当团队中的至少一个组成部分（即智能体、机器人）在硬件或物理约束方面与其他部分不同时，我们就认为该团队是物理上 （P）异质的。也就是说，团队成员可能配备不同的传感器、执行器，具有不同的运动约束等。这些物理差异可能导致能力不同。例如，一架小型无人机可能能够飞行并进行激进机动，但其电池续航时间可能比一台体型较大、速度较慢的地面机器人短。在学习情境中，这种异质性可能导致不同的观测空间和动作空间，例如机器人配备了不同的传感器或执行器时。

当团队中至少有一个组成部分在软件或行为模型方面与其他部分不同，该团队就被视为行为异质（B）。也就是说，两个行为异质的智能体在观察相同输入时，可能产生不同的策略输出。例如，两架物理同质的无人机可能协同执行场地监测：其中一架无人机可以从远处勘察，并引导另一架前往需要近距离检查的区域。行为异质性分为两类：相同目标（B_s）和不同目标（B_d）。在 B_s 异质系统中，智能体通过异质行为优化相同的目标函数。在多智能体强化学习中，这意味着它们共享相同的（全局或局部）奖励函数。B_s 异质系统通常代表合作场景 [11]。不过，它们也可以用于建模这样的对抗场景：具有相同目标的智能体争夺有限资源 [7]。在 B_d 异质系统中，智能体通过异质行为优化不同的目标函数。在多智能体强化学习中，这意味着它们具有不同的局部奖励函数，或者具有由这些局部函数组合而成的全局奖励。B_d 异质系统通常代表非合作或对抗场景 [33]。不过，它们也可以用于建模合作场景，在这类场景中，智能体针对一项高阶任务优化不同的子函数 [12]。例如，在合作搜索与救援场景中，一台机器人可能只负责清除碎片，而其他机器人则负责在没有杂物的区域搜索。

物理异构性与行为异构性并非互斥。因此，前文介绍的三种异构类别（P、B_s、B_d）界定了一个系统可能归属的五种异质性子类：

- $\mathrm{P} \setminus \mathrm{B}$：智能体在物理上不同，但共享相同的行为模型。
- $\mathrm{P} \cap \mathrm{B}_d$：智能体在物理上不同，且行为模型和目标各不相同。
- $\mathrm{P} \cap \mathrm{B}_s$：智能体在物理上不同，且行为模型各不相同，但共享相同目标。
- $\mathrm{B}_s \setminus \mathrm{P}$：智能体在物理上同质，且共享相同目标，但行为模型各不相同。
- $\mathrm{B}_d \setminus \mathrm{P}$：智能体在物理上同质，但行为模型和目标各不相同。

虽然该分类体系旨在对异质系统进行分类，但并不试图衡量异构性的程度。此外，它属于高层次分类，并未考虑动态物理异质性，例如不同的电池电量或硬件老化 [50]。

3 相关工作

本节回顾异构多机器人/智能体系统领域当前的最新进展。我们根据表1中的分类体系对相关工作进行分类。

3.1 多机器人系统中的异质性

关于异质机器人的核心文献通常着重于开发能够发挥团队

表1：按照第2节中的分类体系对异构多机器人/智能体系统领域的相关工作进行分类。

异质性类别	多机器人系统	MARL
$\mathrm{P} \setminus \mathrm{B}$	[8]	[55],[54]
$\mathrm{P} \cap \mathrm{B}_d$	[37]	[33],[12]
$\mathrm{P} \cap \mathrm{B}_s$	[17],[40],[18],[36],[39], [43],[42],[46],[28],[34], [35],[10],[14]	[49]
$\mathrm{B}_s \setminus \mathrm{P}$	[1],[4],[2],[3],[31], [52],[56]	[57],[11],[58]
$\mathrm{B}_d \setminus \mathrm{P}$	[23],[47]	[33],[12]

物理异质性优势的协调算法。因此，这些研究属于 $\mathrm{P} \cap \mathrm{B}_s$ 类。这种多样性可以表现为不同的传感器探测范围 [42], 和不同的感知能力 [46], ，或不同的最大速度 [28]。这些差异可用于解决各种问题，例如多机器人覆盖 [28, 42, 46] 和异构任务分配 [40, 43],，并采用能够应对机器人能力 [17] 或环境 [18, 36, 39] 中的不确定性的鲁棒方案。在主动采样和建图 [34, 35], 领域，传感器异构性也受到关注，因为异构的计算资源会影响任务执行 [10]。最后，$\mathrm{P} \cap \mathrm{B}_s$ 多样性也已在更复杂的问题中得到研究，例如异质轨迹规划 [14]。

有趣的是，这种没有行为多样性的物理多样性（$\mathrm{P} \setminus \mathrm{B}$）往往会成为问题的约束。这一异质性类别中的研究试图在行为层面调和机器人的物理异质性，以便将同质化解决方案应用于当前问题。异构多机器人同步定位与建图就是一个应用示例：来自配备多样化传感器的不同机器人的扫描数据需要相互匹配，才能构建同质的共享地图 [8]。

针对物理同质机器人的行为异质性是一个探索较少但前景广阔的研究方向 [1]。该领域的研究大多解决协作问题，利用 $\mathrm{B}_s \setminus \mathrm{P}$ 异质性。Balch [2, 3] 和 Li等人 [31] 的早期研究侧重于利用强化学习为多机器人团队学习行为专门化。博弈论自主赛车 [52, 56]构成了 $\mathrm{B}_s \setminus \mathrm{P}$ 异质性的对抗性场景。请注意，当所有参与者都采用对称的纳什均衡策略时，博弈论控制器不会表现出异质行为 [38]。然而，当团队中的部分机器人采用传统的模型预测控制器时，就会出现异质性。

相反，不同目标下的异质行为（B_d）也已在协作机器人任务中得到研究，例如，将全局任务划分为由相同机器人组成的群体所承担的子任务（$\mathrm{B}_d \setminus \mathrm{P}$） [23, 47]。如果不同子群的机器人在物理上也存在差异，则可利用这些差异来应对复杂的多机器人任务，例如灾后协作建图 [37], ，从而形成 $\mathrm{P} \cap \mathrm{B}_d$ 异质性。

本小节讨论的所有研究都聚焦于某一特定异质性类别和问题，并针对该设定提出专门的解决方案。在很大程度上，这些方法采用传统的控制理论方法。相比之下，我们的工作提出了一种基于学习的框架，用于合成可通信的多智能体/机器人策略，并且适用于任何异质性类别。

3.2 多智能体强化学习中的异构性

近年来，多智能体强化学习作为解决多机器人问题的有效技术，日益受到关注 [61]。借助多智能体强化学习，可以为困难的协作问题构建高效的分布式多智能体控制器 [5]。针对物理同质智能体的同质策略（共享参数）在多智能体强化学习中十分常见 [20, 24, 29, 44, 53]，并构成了相关研究文献的核心。为了模拟异质行为，一种常见做法是在每个智能体的观测空间中加入一个唯一索引，用以表示该智能体的类型 [19, 24]。在这种情况下，智能体共享同一个同质多模态策略，并以唯一的常量索引为条件。我们将在第 6 节深入阐述并讨论这种方法的局限性。多智能体强化学习中的 $\mathrm{P} \setminus \mathrm{B}$异质性，侧重于为物理上不同的智能体利用参数共享和同质训练的优势。实现方式包括将异质观测空间映射为同质的定长编码 [55], ，或通过填充并将智能体索引纳入观测 [54]。

大多数异质多智能体强化学习文献都属于 B异质性类别。可以通过多种技术，为物理同质智能体学习不同的行为角色，例如，依据潜在表示对智能体策略进行条件化 [57], 分解并聚类动作空间 [58], ，或通过内在奖励最大化智能体轨迹与其角色之间的互信息 [11]。上述所有工作都考虑了目标相同的物理同质智能体，因此利用了 $\mathrm{B}_s \setminus \mathrm{P}$ 异质性。此外，它们没有使用智能体间通信，因此在高度部分可观测的协作问题中的应用受到限制。考虑物理属性不同的机器人时，必须将异质动作空间或观测空间纳入考量。例如，可以针对每种物理属性不同的机器人类型，采用一个专用的图神经网络层 [49]，以此对具有通信能力的 $\mathrm{P} \cap \mathrm{B}_s$ 异质性进行建模。对于某些机器人类型预先已知的任务，这种方法或许适用，但它会阻碍物理同质智能体学习异质行为。

具有不同目标的行为异质性（B_d）源于不同的智能体奖励函数，正如第2节所述。MAD- DPG [33] 采用这种范式，通过集中式训练学习各自独立（而非共享）的行动者与评论家网络。他们在混合的合作-竞争任务中测试了这一方法。这些任务同时考虑了物理同质和物理属性不同的智能体（即最大速度不同的智能体）。因此，MADDPG 利用了 $\mathrm{B}_d \setminus \mathrm{P}$ 和 $\mathrm{P} \cap \mathrm{B}_d$异质性类别。 [12],也研究了相同的异质性类别，并提出在物理同质且共享相同奖励函数的智能体子群之间共享参数。然而，这种方法会阻止目标相同的物理同质智能体采用不同的行为角色来完成任务。

本节讨论的大多数研究都针对某一特定异构性子类中的问题提出解决方案。

虽然少数研究可能适用于多个类别 [11, 33, 57],，但它们采用集中式训练方法，且未考虑智能体间通信。这是使多智能体强化学习适用于多机器人问题所需的两个关键特征。

4 问题表述

现在，我们对本文所研究的多机器人多智能体强化学习问题进行表述。为此，我们首先介绍部分可观测马尔可夫决策过程（POMDP）的多智能体扩展 [27]。

部分可观测马尔可夫博弈。部分可观测马尔可夫博弈（POMG）定义为一个元组

$$\left\langle \mathcal{V}, \mathcal{S}, O, \{\sigma_i\}_{i\in\mathcal{V}}, \mathcal{A}, \{\mathcal{R}_i\}_{i\in\mathcal{V}}, \mathcal{T}, \gamma \right\rangle,$$

其中，$\mathrm{V} = \{1, \ldots, n\}$ 表示智能体集合，S 是所有智能体共享的状态空间，$\mathrm{O} \equiv \mathrm{O}_1 \times \ldots \times \mathrm{O}_n$ 和 $\mathrm{A} \equiv \mathrm{A}_1 \times \ldots \times \mathrm{A}_n$ 分别是观测空间和动作空间，其中 $\mathrm{O}_i \subseteq \mathrm{S} \quad \forall i \in \mathrm{V}$。此外，$\{\sigma_i\}_{i\in\mathcal{V}}$ 和 $\{\mathrm{R}_i\}_{i\in\mathcal{V}}$ 分别是智能体的观测函数和奖励函数[1]，使得 $\sigma_i\colon \mathrm{S} \mapsto \mathrm{O}_i$，且 $\mathrm{R}_i\colon \mathrm{S} \times \mathrm{A} \times \mathrm{S} \mapsto \mathbb{R}$。T 是随机状态转移模型，定义为 $\mathrm{T} : \mathrm{S} \times \mathrm{A} \times \mathrm{S} \mapsto [0, 1]$。最后，$\gamma$ 是折扣因子。

我们将智能体组织成通信图 $\mathrm{G} = (\mathrm{V}, \mathrm{E})$。节点 $i \in \mathrm{V}$ 表示智能体，边 $e_{ij} \in \mathrm{E}$ 表示通信链路。边集合取决于最大智能体通信范围，并会随时间变化。每个智能体的通信邻域定义为 $\mathrm{N}_i \equiv \{v_j \mid e_{ij} \in \mathrm{E}\}$。

在每个时间步 t，每个智能体 i 都会获取一个观测 $o_i^t = \sigma_i(s^t) \in \mathrm{O}_i$，该观测是全局状态 $s^t \in \mathrm{S}$ 的一部分。此观测会传递给相邻智能体 N_i^t。随机策略 π_i 利用这些信息计算动作 $a_i^t \sim \pi_i(\cdot|o^t{}_{\mathcal{N}_i})$。随后，将智能体的动作 $\mathbf{a}^t = (a_1^t, \ldots\ldots, a_n^t) \in \mathrm{A}$ 与当前状态 s^t 一并输入转移模型，以获得下一状态 $s^{t+1} \sim \mathrm{T}(\cdot|s^t, \mathbf{a}^t)$。然后将奖励 $r_i^t = \mathrm{R}_i(s^t, \mathbf{a}^t, s^{t+1})$ 提供给智能体 i。

每个智能体的目标是在时域为 T、可能无限长[2]的回合中，最大化折扣奖励之和 $v_i^t = \sum_{k=0}^{T} \gamma^k r_i^{t+k}$。$v_i^t$ 称为回报。每个智能体都有一个价值函数 $V_i(o_{\mathcal{N}_i}) = \mathbb{E}_{\pi_i}\left[v_i^t | o^t{}_{\mathcal{N}_i} = o_{\mathcal{N}_i}\right]$，它表示从观测 $o_{\mathcal{N}_i}$ 出发并遵循策略 π_i 时的期望回报。该函数用于估计观测的“优劣”。在这项工作中，我们采用近端策略优化（PPO）演员-评论家算法 [48],，该算法使用神经网络和约束策略梯度更新来近似策略（演员）和价值函数（评论家）。

问题。学习异构策略 $\pi_i(o^t{}_{\mathcal{N}_i}; \theta_i)$ 和评论家网络 $V_i(o_{\mathcal{N}_i}; \theta_i)$，并以神经网络参数 θ_i 为条件，且每个智能体的参数各不相同。来自智能体邻域 N_i 的观测 $o^t{}_{\mathcal{N}_i}$ 通过可微通信信道获取，从而使学习天然具备去中心化能力。

我们的目标是明确异构性在多智能体强化学习策略中的作用。为此，我们构建了一个模型来解决上述问题描述，并通过实证支持的论述阐明同质策略的不足（以及它们所依赖的行为类型判定 技术）。

5 异质模型

我们介绍构成本研究所采用方法的两种多智能体强化学习模型：图神经网络近端策略优化（GPPO）及其异质对应模型 HetGPPO。

GPPO 建立在独立近端策略优化（IPPO）[13]的基础上。在 IPPO 中，每个智能体都会学习一个局部评论家 $V_i(o_i)$ 和一个演员 $\pi_i(o_i)$，且仅以自身观测为条件。评论家仅以局部观测而非完整状态 s 为条件，会在训练过程中引入非平稳性。这会导致其他智能体被视为环境的一部分，而未在评论家中显式建模。虽然这可能带来问题，但其优点是不需要在训练期间获取全局信息。此外，已有研究表明，在最先进的多智能体强化学习基准测试中，IPPO 的表现优于许多完全可观测的评论家模型 [13]。

GPPO 克服了 IPPO 的局限，同时保留了其优势。它采用 GNN 通信层，使智能体能够在邻域内共享信息，以协同合作并克服部分可观测性。因此，GPPO 的评论家 $V_i(o_{\mathcal{N}_i})$ 和演员 $\pi_i(o_{\mathcal{N}_i})$ 以局部通信邻域观测 $o_{\mathcal{N}_i}$ 为条件。这样有助于克服非平稳性，同时训练期间只需使用局部信息并进行通信。

GPPO 模型如图 2 所示。在每个时间步，每个智能体 i 都会观测环境，收集观测 o_i。这些观测包含绝对几何特征，例如智能体的位置 $\mathbf{p}_i \in \mathbb{R}^2$。非绝对特征会输入多层感知机（MLP）编码器，得到嵌入表示 z_i。绝对位置和智能体的速度 $\mathbf{v}_i \in \mathbb{R}^2$ 用于计算边特征 e_{ij}，即智能体 i 和 j 之间的相对特征。在这项工作中，我们使用相对位置 $\mathbf{p}^{ij} = \mathbf{p}_i - \mathbf{p}_j$ 和相对速度 $\mathbf{v}^{ij} = \mathbf{v}_i - \mathbf{v}_j$ 作为边特征 $e_{ij} = \mathbf{p}_{ij} \,||\, \mathbf{v}^{ij}$，其中 $||$ 表示拼接操作。该过程确保 GNN 的输出具有对 $\mathbb{R}^2$ 平移的不变性（即，如果整个团队在空间中发生平移，输出仍保持不变），从而帮助模型 [21]实现泛化。随后，边特征 e_{ij} 和智能体嵌入表示 z_i 将用于消息传递 GNN 内核：

$$h_i = \psi_{\theta_i}(z_i) + \bigoplus_{j\in\mathcal{N}_i} \phi_{\theta_i}(z_j \,||\, e_{ij}).$$

其中，ψ_{θ_i} 和 ϕ_{θ_i} 是两个多层感知机，由智能体参数 θ_i[3] 参数化；$\bigoplus$ 是一个聚合算子（例如求和）。随后，图神经网络的输出 h_i 被输入两个不同的多层感知机解码器，分别输出动作 $a_i \sim \pi_i(\cdot|o_{\mathcal{N}_i})$ 和价值 $V_i(o_{\mathcal{N}_i})$。与 IPPO 类似，GPPO 使用参数共享来提高样本效率。因此 $\theta_1 = \ldots = \theta_n$。参数共享使智能体能够从集体经验中获益，从而缩短训练时间。另一方面，它要求采用集中式训练，并限制智能体策略必须完全相同（即同质）。

HetGPPO 移除了 GPPO 的参数共享约束，因此允许智能体策略各自分化，$\theta_1 \neq \ldots \neq \theta_n$。然而，在 GNN 通信的背景下，不共享参数会产生深远影响：GNN 的置换等变性

[1] 请注意，虽然我们使用局部智能体奖励函数 $\mathcal{R}_i$ 来构建问题（以实现对 $\mathcal{B}_d$ 异构性的学习），但我们的实验均采用全局奖励函数 $\mathcal{R} = \mathcal{R}_1 = \ldots = \mathcal{R}_n$，该函数无法分解为局部子函数。此奖励编码了一个全局协作目标，从而产生 $\mathcal{B}_s$ 异构性。

[2] γ^k 表示 γ 的 k 次方，而非时间步上标。

[3] 其中，θ_i 表示智能体 i 所有神经网络层的参数。

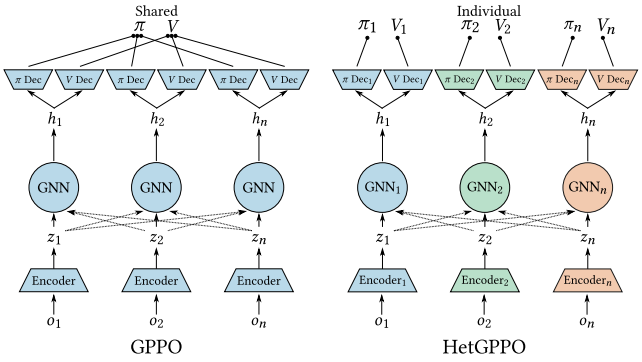


图 2：GPPO 和 HetGPPO 的架构：采用通信智能体的多智能体强化学习模型。每个智能体将其观测输入编码器，然后使用具有平移不变性的消息传递图神经网络聚合从邻居接收的消息，并更新其隐藏状态 h_i。随后将 h_i 用作策略解码器和值解码器（Dec）的输入。HetGPPO 等同于不采用参数共享的 GPPO。

[59] 不再成立，因为智能体现在会学习不同的消息编码和解释策略。因此，GNN 必须针对给定团队输入的所有可能置换，分别学习不同的团队输出，而不是只学习一个输出。这可能会降低泛化能力和样本效率。另一方面，梯度会通过通信邻域反向传播，使智能体能够从局部交互中集体学习。

HetGPPO 的结构（如图 2 所示）支持去中心化训练、去中心化执行（*DTDE*）。这是因为 GPPO 的评论家网络不以全局信息为条件。GPPO 使用参数共享，而 HetGPPO 不再需要参数共享，因此能够在任何仅支持智能体间通信的环境中进行训练。值得注意的是，通过实现一种临时机制来实现去中心化参数共享（例如通过分布式优化 [60]），GPPO 也可以采用去中心化方式进行训练。

我们使用 PyTorch [41] 实现 HetGPPO 和 GPPO，并采用 RLlib [32] 框架进行训练。代码可在此处获取[4]。仿真在自定义场景中使用 VMAS 仿真器 [6], 执行，可通过此链接获取[5]。

6 行为类型判定

上文介绍的 HetGPPO 使我们能够学习真正的异构策略。与直觉相反，使用 GPPO 这样的表面上同质模型，也可以学习异质行为。这使智能体能够模拟异质行为，同时利用参数共享带来的样本效率优势。共享模型可以包含不同的行为类型，而这些行为类型会由输入观测的特定组合激活。例如，如果两个机器人正将一个包裹运往目的地，模型可以识别某个智能体是否位于后方（距离目标更远），并为其分配一种不同于位于

[4]https://github.com/proroklab/HetGPPO [5]
https://github.com/proroklab/VectorizedMultiAgentSimulator

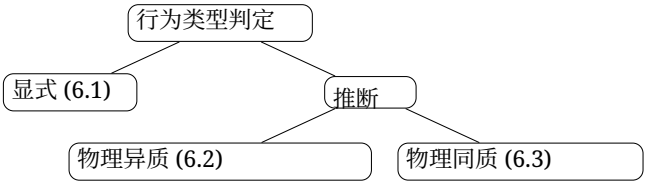


图 3：行为类型判定的不同形式。同质策略利用类型识别来区分不同智能体，并模拟异质行为。

前方智能体的行为类型。输入观测为模型向智能体分配行为类型提供了条件。

我们将这一识别过程称为类型识别。图 3 展示了行为类型化技术的分类，我们将在以下各小节中进行介绍。请注意，行为类型位于连续的行为空间中（而非离散集合的一部分） [2]。

6.1 显式行为类型识别

最常见的行为类型判定形式，是将智能体的索引 i 显式地作为观测的一部分输入。这种做法已在多智能体强化学习文献中得到广泛应用 [12, 16, 24, 54]。然而，这要求模型学习一种多模态策略，并根据这个整数索引切换模式。这可能导致智能体的策略出现不连续性，而且已被证明性能并非最优 [12]。

定义 *6.1*（显式行为类型识别）。当共享式分散多智能体强化学习策略能够根据与输入拼接的一个常数值对智能体进行类型识别，且该值因智能体而异时，即为显式行为类型识别。

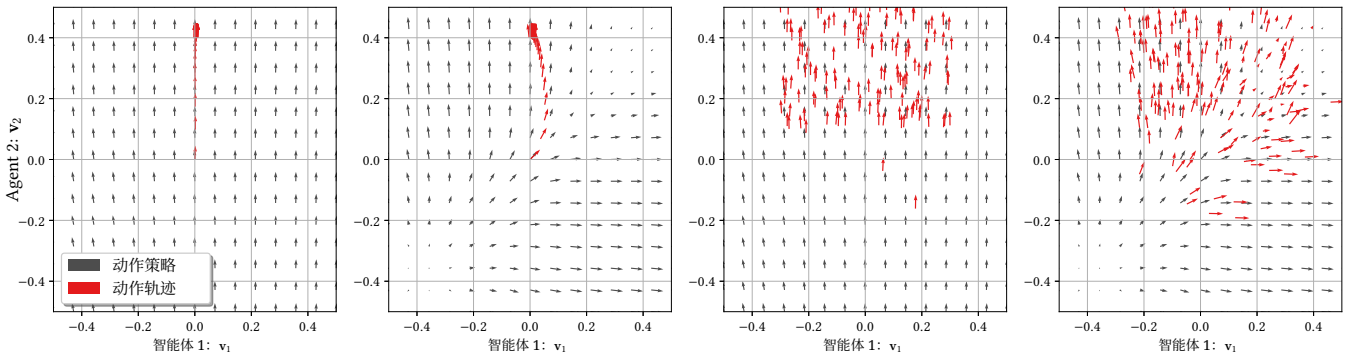
即使没有显式索引，共享策略仍可能模拟异质行为 [15]。我们将这种现象称为推断出的行为类型判定。对于物理异质和物理相同的智能体，都可能出现推断类型识别。

6.2 物理异质智能体的推断行为类型判定

我们首先通过一个案例研究，介绍物理异质智能体的推断行为类型判定。

定义 *6.2*（物理异质智能体的推断行为类型判定）。当共享式分散多智能体强化学习策略能够通过观测对 P 个异构智能体进行类型识别时，就会出现物理异质智能体的推断行为类型判定。

场景 A（图 4）。考虑两台质量不同的机器人，$m_1 > m_2$ 它们位于一维工作空间中的随机位置。机器人观测自身的位置 $\mathbf{p}_i \in \mathbb{R}$ 和速度 $\mathbf{v}_i \in \mathbb{R}$，并通过通信共享这些信息。它们的动作是施加力 $\mathbf{f}_i \in \mathbb{R}$。团队共同获得奖励，目标是在尽量减少能量消耗的同时，使团队中的最大速度达到最大。显然，此情形下的最优策略是质量较大的机器人完全不动，而较轻的机器人以最大速度移动。显然，这些行为是异质的，因为当两个智能体接收到相同观测时，$\mathbf{f}_1 \neq \mathbf{f}_2$。



（a）异质　（b）同质　(c) 含噪声的异质　(d) 含噪声的同质

图 4：场景 A 中学到的策略，以向量场（灰色）和环境中的滚动采样（红色）表示。(a) 和 (b) 不受部署噪声影响。(c) 和 (d) 的观测受到 ± 0.3 均匀噪声的影响。在这些图中，箭头表示团队动作向量 $\vec{\mathbf{f}}(\bar{\mathbf{v}}) = [\mathbf{f}_1(\bar{\mathbf{v}}), \mathbf{f}_2(\bar{\mathbf{v}})]$ 关于观测 $\bar{\mathbf{v}} = [\mathbf{v}_1, \mathbf{v}_2]$ 的函数。滚动采样始终从原点 ($\bar{\mathbf{v}} = [0, 0]$) 开始。我们可以看到，表示同质策略的向量场被迫对两个输入的置换保持不变，因此沿 $\mathbf{v}_1 = \mathbf{v}_2$ 对称。这使其在存在噪声时变得脆弱（d）：噪声会使观测落入平面中错误的区域，而对称性又会在此处强制采用次优策略（水平箭头）。

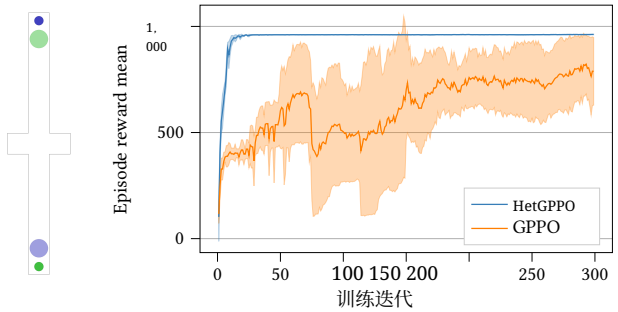
我们使用 GPPO 和 HetGPPO 在此场景中训练智能体。图 4 展示了各模型学习到的策略的图示。在这些图中，箭头表示团队动作向量 $\vec{\mathbf{f}}(\bar{\mathbf{v}}) = [\mathbf{f}_1, \mathbf{f}_2] = [0, 0]$ ($\vec{\mathbf{v}}) \mathbf{f}(\bar{\mathbf{v}})]$ 随观测 $\bar{\mathbf{v}}$ vv变化的情况。竖直箭头表示，当两个智能体都静止时，智能体 1 希望保持静止，而智能体 2 希望提高速度。我们用灰色矢量场绘制每一对观测对应的策略平均动作[6]，并用红色显示策略的一次滚动采样。在图 4a 中，我们看到 HetGPPO 能够学习到不依赖任何观测的最优策略。得益于基于物理特征推断的类型划分，GPPO（图 4b）出人意料地也能学习到产生最优滚动采样的策略。我们可以看到，由于同质性，GPPO 策略被迫关于 $\mathbf{v}_1 = \mathbf{v}_2$ 轴对称。因此，当 GPPO 智能体在 $[0, 0]$, 处生成时，它们会被迫采取相同的动作来提高各自的速度。然而，由于它们的 P 差异，这一动作会产生不同的速度，使滚动采样迅速偏离对称状态，从而产生最优行为。物理差异（即智能体质量不同）会导致相同动作产生不同观测（即智能体速度不同），这使同质模型能够学习到最优策略，同时表现出明显的异质行为——这是针对物理异质智能体的推断行为类型判定的一个例子。

基于物理特征推断的类型划分被证明是一种脆弱的解决方案。当在滚动采样期间注入加性均匀观测噪声时，我们观察到 HetGPPO 的滚动采样（图 4c ）完全不受影响，而 GPPO 的滚动采样（图 4d）偶尔会落在对角线的另一侧，产生与最优行为对称相反的行为，并导致重型智能体移动（水平箭头）。

6.3 物理同质智能体的推断出的行为类型划分

现在，我们通过一个案例研究来介绍物理同质智能体的推断出的行为类型划分。

[6]近端策略优化策略具有随机性，并输出一个关于 $\mathbf{f}_i$ 的分布。

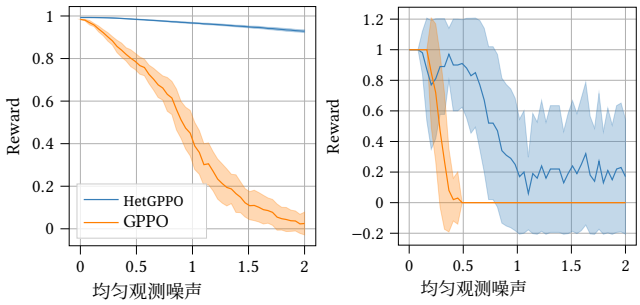


（a）场景　(b) 训练性能

图 5：场景 B。(a)：两台机器人（较大的圆圈）位于走廊两侧，需要互相礼让才能到达各自的目标（较小的圆圈）。(b) 场景 B 的训练曲线显示，异质模型能够立即解决该场景，而同质模型则需要约 300 次训练迭代，才能为物理上相同的智能体学习推断出的行为类型划分。我们绘制了 10 次不同运行的均值和标准差。每次迭代均基于 200 个回合进行。

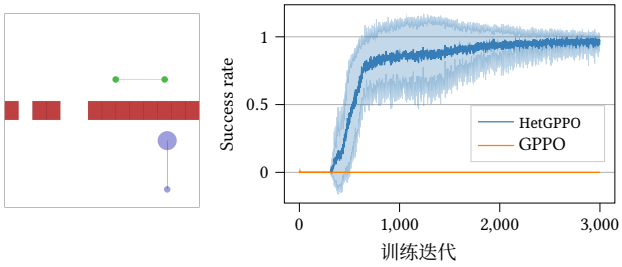
定义 *6.3*（物理同质智能体的推断行为类型判定）。当共享式分散多智能体强化学习策略能够通过物理同质智能体的观测对其进行类型识别时，就会出现物理同质智能体的推断行为类型判定。

场景 B（图 5）。现在考虑两台物理同质的机器人，它们分别从狭窄走廊的两端出发，如图 5a 所示。每台机器人都位于另一台机器人的目标前方。走廊仅够一台机器人通过，但中央有两个机器人大小的凹室。机器人会观测并通信各自的二维位置和速度，任务是在不发生碰撞的情况下到达各自的目标。因此，只有一台机器人向另一台让行，任务才能完成。



(a) 场景 A (b) 场景 B

图 6：在两种推断类型场景中存在部署噪声时，同质模型与异质模型的性能。奖励归一化至 0 到 1 之间。对所有观测施加均匀噪声，其单位与观测相同。我们报告了 100 次运行中归一化奖励的均值和标准差，噪声值范围为 0 到 2，共 50 个。



(a) 场景 (b) 训练性能

图 7：不同尺寸机器人在通道场景中的性能评估。在此，同质模型无法对物理异质智能体进行推断出的行为类型划分，因为$\mathcal{P}$ 异构性不会影响机器人的观测。因此，只有异质模型能够解决该任务。我们绘制了 4 次运行的平均成功率和标准差。每次迭代均基于 200 个经验回合进行。

同样，我们在此场景中使用 GPPO 和 HetGPPO 训练智能体。通过查看图 5b 中的训练奖励曲线，可以看到两种模型都能学会正确的行为（奖励 > 700）。GPPO 利用对物理同质智能体进行推断类型，并能够根据两台机器人的相对位置和速度动态分配“让行”角色。然而，我们观察到，学习行为类型判定需要全部 300 次训练迭代，而异质模型仅用 20 次迭代便学会了最优解。

6.4 行为类型判定的局限性

尽管同质模型可以通过行为类型判定来学习表面上异质的行为，但这并不能证明它是一种可靠且可扩展的解决方案。

在 [12] 中，研究表明，显式行为类型识别的性能会随着待学习类型数量的增加而下降。此外，作者通过实证表明，这种性能下降与共享同质模型的容量（即参数数量）无关。

推断索引也被证明是一种脆弱的解决方案。为了刻画这种脆弱性，我们在执行过程中注入观测噪声进行评估。结果如图 6 所示。我们报告了 100 次运行中归一化奖励的均值和标准差，噪声值取 0 到 2 之间的 50 个值。可以观察到，当噪声为 0 时，所有模型都从奖励为 1 的最优策略开始。随着噪声增加，我们观察到，同质模型要么几乎立即失去功能（如图 6b 中的场景 B），要么性能迅速下降（如图 6a 中的场景 A）。相比之下，异质策略能够容忍更高程度的噪声；即使在困难的走廊场景中，在较高噪声值下仍能有约 20% 的概率完成任务。

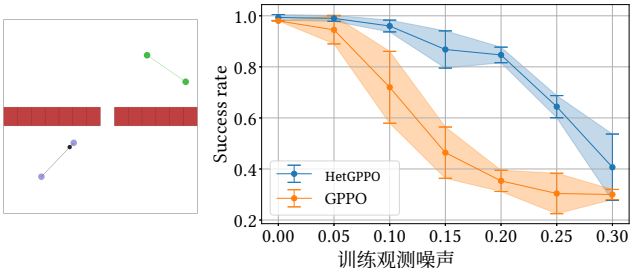
7 实验评估

现在，我们展示所提出的模型在模拟环境和真实环境中的一些评估结果。

性能评估。我们在一个模拟的二维任务上评估 HetGPPO，该任务需要异质行为。任务如图 7a 所示。这里，两个大小不同的机器人（蓝色圆形）通过两个转动关节与一根刚性连杆相连，需要在保持连杆与通道平行的同时穿过通道，然后与另一侧期望的目标位置（绿色圆形）对齐。通道由一个较大的缺口和一个较小的缺口组成，它们以随机的位置和顺序出现在墙上，但彼此间距始终相同。机器人团队以随机的顺序和位置出现在下侧，连杆始终与通道垂直。目标以随机位置水平地出现在上侧。每个机器人都会观测并传递自身速度、相对于每个缺口的位置，以及相对于目标中心的位置。塑形后的全局奖励由两个凸项组成。通过通道前，机器人因保持连杆与目标平行并将连杆中心移至通道中心而获得奖励。通过通道后，机器人因将连杆移至目标位置并使其方向符合期望而获得奖励。碰撞也会受到惩罚。

图 7b 展示了训练成功率（即每批次中完成任务的回合所占百分比）。异质模型能够学习到两种行为不同的策略：较大的机器人穿过较大的缺口，较小的机器人穿过较小的缺口，从而达到最优解。另一方面，由于不同的机器人尺寸所导致的 P 异质性不会影响机器人的观测，因此同质模型无法使用针对物理异质智能体的推断行为类型划分来分配这两种行为类型。采用同质策略的智能体始终无法穿过通道，因为它们会被不可避免的碰撞阻挡。

对训练噪声的鲁棒性。正如第 6 节所阐述的，同质模型能够学习异质行为。在本小节中，我们评估这一范式在训练期间存在观测噪声时的鲁棒性。我们考虑图 8a 所示的任务。该任务与图 7a 中的相同，不同之处在于，机器人现在在物理上同质，但连杆具有非对称质量（黑色圆圈），从而导致另一种类型的 P 异质性，



(a) 场景　　(b) 对训练噪声的鲁棒性

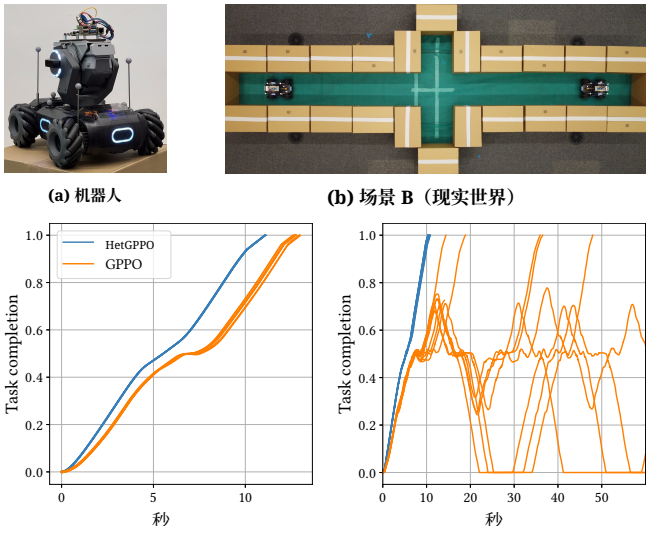
图 8：在带有非对称载荷的通道场景中，训练期间对均匀观测噪声的鲁棒性。随着噪声增加，异质模型能够保持较高的性能。我们针对 7 种不同的噪声值训练两个模型。对于每个噪声值，我们报告 5 次运行在经过 1000 次训练迭代后的成功率均值和标准差。每次训练迭代都基于 200 个经验回合进行。

这体现在速度观测中。通道是墙上的一个缺口，位置随机。智能体需要穿过缺口，同时使连杆保持与墙垂直并避免碰撞。团队和目标以随机的位置、顺序和朝向生成在通道的两侧。

在图 8b 中，我们报告了不同观测噪声值下的训练成功率。得益于针对物理异质智能体的推断行为类型判定，我们发现，当不添加噪声时，两个模型都能以最优方式完成任务。随着噪声增加，异质模型能够保持显著更好的性能。例如，当观测噪声为 0.2 时，HetGPPO 的成功率仍超过 80%，而 GPPO 则低于 40%。

现实世界部署。为展示异质策略的韧性，我们将场景 B（第 6.3 节）部署到现实环境中。任务设置如图 9b 所示，与仿真中的设置相同。我们使用两台全向 RoboMaster S1 地面机器人 [16]（图 9a），每台机器人都搭载定制的基于模型的控制器 [51]。我们分别在仿真环境（图 9c）和现实世界中（图 9d），对训练好的 HetGPPO 和 GPPO 模型各进行 10 次运行。如第 6.3 节所述，异质模型和同质模型都能在仿真中解决该场景，其中同质模型利用了针对物理同质智能体的推断行为类型判定。如图 9c 所示，两个模型的所有运行均在 15 秒内达到 100% 的任务完成率。另一方面，如图 9d 所示，同质模型的性能在现实世界中受到严重影响。这是因为在这个对称场景中，同质模型无法仅根据位置对智能体进行类型识别，而必须依赖速度观测来构建行为类型。然而在实际应用中，现实世界中的估计速度可能会受到诸如以下因素的影响而产生噪声

机器人测量和运动模型中的控制/过程延迟与可变性。因此，依赖这些观测会使齐次（无记忆）模型容易动态地错误切换行为类型（即无法区分机器人当前是朝向中心还是背离中心移动）。这导致图中所示的滚动采样中，机器人在通道附近交替地相互让行。在 10 次运行中，



(a) 机器人　　(b) 场景 B（现实世界）

(c) 仿真　　(d) 现实世界

图9：场景 B 的现实世界部署（图 5）。我们针对每个模型分别在仿真和现实世界中运行了 10 次。我们绘制了随时间变化的任务完成度（每个机器人到其目标位置的负距离之和经缩放后的值）。尽管在仿真中两个模型都能够完成任务，但现实世界中的不完善因素使同质模型在学习到的行为类型之间动态切换，导致机器人在中心区域附近多次交换位置。这造成了（d）中的之字形行为，其中某些滚动采样失败，或耗时超过分配的最长时间 60 秒。

只有 5 次能在 60 秒内完成。另一方面，异质模型不依赖行为类型识别，也不受部署噪声影响，其表现与仿真中的表现一样好。

8 结论

本文提出了一种在多智能体强化学习中学习异构策略的新范式。我们通过对同质模型可用于模拟异质行为的技术进行分类来阐述这一范式，并通过实证展示了这些技术的局限性。最后，我们展示了在仿真和现实世界中的多机器人任务中，策略异构性对性能和韧性的益处。虽然我们没有采用任何方法来控制智能体策略的异质程度，但我们观察到，训练本身已是一种良好的异质性正则化项。换言之，如果系统具有异构需求，HetGPPO 就能够学习这些需求；而如果系统受益于同质策略，HetGPPO 就会学到与 GPPO 相同的策略（但样本效率会有所降低）。未来，我们有兴趣开发机制，以衡量并主动调整团队中的策略异构程度，从而控制同质策略的样本效率与异构策略的韧性之间的权衡。

致谢

本研究得到 ARL DCIST CRA W911NF-17-2-0181、欧洲研究理事会（ERC）项目 949940（gAIa）的支持，并部分得到了 Arm 的捐赠资助。

参考文献

[1] Nora Ayanian。2019年。Dart：机器人团队中增强多样性的自主性。《国际机器人研究杂志》38，12-13（2019年），1329–1337。[2] Tucker Balch。2000年。层级社会熵：一种衡量机器人群体多样性的信息论度量。《自主机器人》8，3（2000年），209–238。[3] Tucker Balch 等。1997年。角色学习：机器人团队中的行为多样性。载于 AAAI 多智能体学习研讨会。[4] Spring Berman、Adám Halász、Vijay Kumar 和 Stephen Pratt。2007年。将机器人群部署到多个目的地的仿生群体行为。载于《2007年 IEEE 国际机器人与自动化会议论文集》。IEEE，2318–2323。[5] Daniel S Bernstein、Robert Givan、Neil Immerman 和 Shlomo Zilberstein。2002 年。马尔可夫决策过程去中心化控制的复杂性。《运筹学数学》27，4（2002年），819–840。[6] Matteo Bettini、Ryan Kortvelesy、Jan Blumenkamp 和 Amanda Prorok。2022年。VMAS：用于集体机器人学习的向量化多智能体模拟器。第16届国际分布式自主机器人系统研讨会（2022年）。[7] Jan Blumenkamp 和 Amanda Prorok。2021年。多智能体强化学习中对抗性通信的涌现。载于机器人学习会议。PMLR，1394–1414。[8] Elizabeth R Boroson 和 Nora Ayanian。2019年。异构多机器人同步定位与建图的三维关键点可重复性。载于2019年国际机器人与自动化会议（ICRA）。IEEE，6337–6343。[9] Olli Bräysy 和 Michel Gendreau。2005年。带时间窗的车辆路径问题，第二部分：元启发式算法。《交通科学》39，1（2005年），119–139。[10] Praneel Chand 和 Dale A Carnegie。2013年。使用能力受限机器人在层级异构多机器人系统中进行建图与探索。《机器人与自主系统》61，6（2013年），565–579。[11] Li Chenghao、Tonghan Wang、Chengjie Wu、Qianchuan Zhao、Jun Yang 和 Chongjie Zhang。2021年。在共享式多智能体强化学习中弘扬多样性。《神经信息处理系统进展》34（2021年）。[12] Filippos Christianos、Georgios Papoudakis、Muhammad A Rahman 和 Stefano V Albrecht。2021年。通过选择性参数共享扩展多智能体强化学习。载于国际机器学习会议。PMLR，1989–1998。[13] Christian Schroeder de Witt、Tarun Gupta、Denys Makoviichuk、Viktor Makoviychuk、Philip HS Torr、Mingfei Sun 和 Shimon Whiteson。2020年。在星际争霸多智能体挑战赛中，独立学习是否就足够了？arXiv 预印本 arXiv:2011.09533（2020年）。[14] Mark Debord、Wolfgang Hönig 和 Nora Ayanian。2018年。异构机器人团队的轨迹规划。载于2018年 IEEE/RSJ 国际智能机器人与系统会议（IROS）。IEEE，7924–7931。[15] Ankur Deka 和 Katia Sycara。2021年。人工智能竞争团队中异构策略的自然涌现。载于《群体智能进展：第12届国际会议》，13–25。[16] 大疆。访问日期：2023年1月17日。RoboMaster S1。https://www.dji.com/robomaster-s1。[17] Yousef Emam、Siddharth Mayya、Gennaro Notomista、Addison Bohannon 和 Magnus Egerstedt。2020年。针对能力不断演变且未知的异构多机器人团队的自适应任务分配。载于2020年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，7719–7725。[18] Yousef Emam、Gennaro Notomista、Paul Glotfelter 和 Magnus Egerstedt。2021年。利用鲁棒控制屏障函数为异构多机器人团队进行数据驱动的自适应任务分配。载于2021年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，9124–9130。[19] Jakob Foerster、Ioannis Alexandros Assael、Nando De Freitas 和 Shimon Whiteson。2016年。通过深度多智能体强化学习学习通信。《神经信息处理系统进展》29（2016年）。[20] Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。2018年。反事实多智能体策略梯度。载于《AAAI 人工智能会议论文集》第32卷。[21] Fabian Fuchs、Daniel Worrall、Volker Fischer 和 Max Welling。2020年。SE(3)-变换器：三维旋转平移等变注意力网络。《神经信息处理系统进展》33（2020年），1970–1981。[22] Brian P Gerkey 和 Maja J Mataric。2002年。推送者—观察者：一种容错的紧耦合机器人协同方法。载于《2002年 IEEE 国际机器人与自动化会议论文集》第1卷。IEEE，464–469。[23] Dani Goldberg 和 Maja J Mataric。1997年。将干扰作为设计和评估多机器人控制器的工具。载于 AAAI/IAAI，637–642。[24] Jayesh K Gupta、Maxim Egorov 和 Mykel Kochenderfer。2017年。利用深度强化学习进行协作式多智能体控制。载于国际自主智能体与多智能体系统会议。施普林格，66–83。

[25] Natasha Jaques、Angeliki Lazaridou、Edward Hughes、Caglar Gulcehre、Pedro Ortega、DJ Strouse、Joel Z Leibo 和 Nando De Freitas。2019年。将社会影响作为多智能体深度强化学习的内在动机。载于《国际机器学习会议》。PMLR，3040–3049。[26] Chanyoung Ju 和 Hyoung Il Son。2019年。基于 Ramadge–Wonham 理论的异构农业野外机器人的建模与控制。《IEEE机器人与自动化快报》5，1（2019年），48–55。[27] Leslie Pack Kaelbling、Michael L Littman 和 Anthony R Cassandra。1998年。部分可观测随机领域中的规划与行动。《人工智能》101，1-2（1998年），99–134。[28] Soobum Kim、María Santos、Luis Guerrero-Bonilla、Anthony Yezzi 和 Magnus Egerstedt。2022年。面向时间敏感型应用的不同最大速度移动机器人的覆盖控制。《IEEE机器人与自动化快报》7，2（2022年），3001–3007。[29] Ryan Kortvelesy 和 Amanda Prorok。2022年。QGNN：利用图神经网络进行价值函数分解。arXiv 预印本 arXiv:2205.13005（2022年）。[30] Karol Kurach、Anton Raichuk、Piotr Stańczyk、Michał Zając、Olivier Bachem、Lasse Espeholt、Carlos Riquelme、Damien Vincent、Marcin Michalski、Olivier Bousquet 等。2020年。Google研究足球：一种新型强化学习环境。载于《AAAI 人工智能会议论文集》，第34卷。4501–4510。[31] Ling Li、Alcherio Martinoli 和 Yaser S Abu-Mostafa。2004年。在协作群体系统中学习并测量专业化。《适应性行为》12，3-4（2004年），199–212。[32] Eric Liang、Richard Liaw、Robert Nishihara、Philipp Moritz、Roy Fox、Ken Goldberg、Joseph Gonzalez、Michael Jordan 和 Ion Stoica。2018年。RLlib：分布式强化学习的抽象层。载于《国际机器学习会议》。PMLR，3053–3062。[33] Ryan Lowe、Yi I Wu、Aviv Tamar、Jean Harb、OpenAI、Pieter Abbeel 和 Igor Mordatch。2017年。面向混合合作竞争环境的多智能体演员-评论家。《神经信息处理系统进展》30（2017年）。[34] Matthew Malencia、Sandeep Manjanna、M Ani Hsieh、George Pappas 和 Vijay Kumar。2022年。利用异构机器人团队自适应采样潜在现象（ASLaP-HR）。arXiv 预印本 arXiv:2208.06053（2022年）。[35] Sandeep Manjanna、Alberto Quattrini Li、Ryan N Smith、Ioannis Rekleitis 和 Gregory Dudek。2018年。用于探索和战略性水体采样的异构多机器人系统。载于《2018年 IEEE 国际机器人与自动化会议（ICRA）》。IEEE，4873–4880。[36] Siddharth Mayya、Diego S D’antonio、David Saldaña 和 Vijay Kumar。2021年。异构多机器人系统中的弹性任务分配。《IEEE机器人与自动化快报》6，2（2021年），1327–1334。[37] Nathan Michael、Shaojie Shen、Kartik Mohta、Vijay Kumar、Keiji Nagatani、Yoshito Okada、Seiga Kiribayashi、Kazuki Otake、Kazuya Yoshida、Kazunori Ohno 等。2014年。地面与空中机器人协作绘制地震受损建筑的地图。载于《野外与服务机器人》。施普林格，33–47。[38] John F Nash Jr。1950年。多人博弈中的均衡点。《美国国家科学院院刊》36，1（1950年），48–49。[39] Gennaro Notomista、Siddharth Mayya、Yousef Emam、Christopher Kroninger、Addison Bohannon、Seth Hutchinson 和 Magnus Egerstedt。2021年。面向异构多机器人系统的弹性、能量感知任务分配框架。《IEEE机器人学汇刊》38，1（2021年），159–179。[40] Gennaro Notomista、Siddharth Mayya、Seth Hutchinson 和 Magnus Egerstedt。2019年。异构多机器人系统的最优任务分配策略。载于《2019年第18届欧洲控制会议（ECC）》。IEEE，2071–2076。[41] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。2019年。PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》32（2019年）。[42] Luciano CA Pimenta、Vijay Kumar、Renato C Mesquita 和 Guilherme AS Pereira。2008年。异构机器人网络的感知与覆盖。载于《2008年第47届 IEEE 决策与控制会议》。IEEE，3947–3952。[43] Amanda Prorok、M Ani Hsieh 和 Vijay Kumar。2017年。多样性对异构机器人群最优控制策略的影响。《IEEE机器人学汇刊》33，2（2017年），346–358。[44] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。2018年。QMIX：用于深度多智能体强化学习的单调价值函数分解。载于《国际机器学习会议》。PMLR，4295–4304。[45] Mikayel Samvelyan、Tabish Rashid、Christian Schroeder de Witt、Gregory Farquhar、Nantas Nardelli、Tim GJ Rudner、Chia-Man Hung、Philip HS Torr、Jakob Foerster 和 Shimon Whiteson。2019年。《星际争霸多智能体挑战赛》。载于《第18届国际自主智能体与多智能体系统会议论文集》。2186–2188。[46] María Santos、Yancy Diaz-Mercado 和 Magnus Egerstedt。2018年。具有异构感知能力的多机器人团队覆盖控制。《IEEE机器人与自动化快报》3，2（2018年），919–925。[47] Miguel Schneider-Fontan 和 Maja J Mataric。1998年。领地式多机器人任务划分。《IEEE机器人学与自动化汇刊》14，5（1998年），815–822。

[48] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017年。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*（2017年）。[49] Esmaeil Seraj、Zheyuan Wang、Rohan Paleja、Daniel Martin、Matthew Sklar、Anirudh Patel 和 Matthew Gombolay。2022年。通过学习实现合作异构团队协作中的高效多样化通信。载于《第*21*届国际自主智能体与多智能体系统会议论文集》。1173–1182。[50] Beining Shang、Richard Crowder 和 Klaus-Peter Zauner。2014年。基于机器人硬件差异的群体行为分类。载于《*2014*年第*4*届仿真与建模方法、技术及应用国际会议（*SIMULTECH*）》。IEEE，631–636。[51] Ajay Shankar、Sebastian Elbaum 和 Carrick Detweiler。2021年。Freyja：一套用于敏捷、精准室外飞行的完整多旋翼系统。载于《*2021*年 *IEEE* 国际机器人与自动化会议（*ICRA*）》。IEEE，217–223。[52] Riccardo Spica、Eric Cristofalo、Zijian Wang、Eduardo Montijano 和 Mac Schwager。2020年。用于自主双人无人机竞速的实时博弈论规划器。《*IEEE*机器人学汇刊》36，5（2020年），1389–1403。[53] Sainbayar Sukhbaatar、Rob Fergus 等。2016年。通过反向传播学习多智能体通信。《神经信息处理系统进展》29（2016年）。[54] Justin K Terry、Nathaniel Grammel、Ananth Hari、Luis Santos 和 Benjamin Black。2020年。重新审视多智能体深度强化学习中的参数共享。*arXiv* 预印本 *arXiv:2005.13625*（2020年）。[55] Ceyer Wakilpoor、Patrick J Martin、Carrie Rebhuhn 和 Amanda Vu。2020年。用于未知环境建图的异构多智能体强化学习。*arXiv* 预印本 *arXiv:2010.02663*（2020年）。[56] Mingyu Wang、Zijian Wang、John Talbot、J Christian Gerdes 和 Mac Schwager。2021年。多车辆竞争场景中自动驾驶汽车的博弈论规划。《*IEEE*机器人学汇刊》37，4（2021年），1313–1325。[57] Tonghan Wang、Heng Dong、Victor Lesser 和 Chongjie Zhang。2020年。ROMA：具有涌现角色的多智能体强化学习。载于《国际机器学习会议》。PMLR，9876–9886。[58] T Wang、T Gupta、B Peng、A Mahajan、S Whiteson 和 C Zhang。2021年。RODE：学习角色以分解多智能体任务。载于《国际学习表征会议论文集》。[59] Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。2018年。图神经网络的能力有多强？载于《国际学习表征会议》。[60] Javier Yu、Joseph A Vincent 和 Mac Schwager。2022年。DiNNO：用于多机器人协作学习的分布式神经网络优化。《*IEEE*机器人与自动化快报》7，2（2022年），1896–1903。[61] Kaiqing Zhang、Zhuoran Yang 和 Tamer Başar。2021年。多智能体强化学习：理论与算法精选综述。《强化学习与控制手册》（2021年），321–384。

A 实验设置

A.1 仿真

我们附上了用于仿真和训练的所有代码。仿真在 VMAS [6] 仿真器中进行。除场景 B 外，所有环境均为自定义创建；场景 B 则基于仿真器中已有的一个场景进行改编。训练通过 RLlib [32] 使用 PyTorch [41] 和用于多智能体训练的 PPO 算法实现。表 2 展示了所用的一般训练参数。针对不同环境会对这些参数做少量调整，具体可见随附的训练脚本。GPPO 和 HetGPPO 模型的实现及详细信息均可在代码中查阅。训练使用 NVIDIA GeForce RTX 2080 Ti 图形处理器。每个工作进程使用 Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz 从仿真器收集经验。

A.2 真实世界

真实世界实验使用配备 12 台摄像机的 Optitrack[7] 动作捕捉系统，为机器人提供位置信息。所使用的机器人是全向 RoboMaster S1 地面机器人[8]，正在机载运行定制的基于模型的控制器 [51]。

表 2：PPO 训练参数。

训练		PPO	
批量大小	60000	ϵ	0.2
小批量大小	4096	γ	0.99
SDG 迭代次数	40	λ	0.9
# 工作进程	5	熵系数	0
# 每个工作进程的环境数	50	KL 系数	0.01
学习率	5e-5	KL 目标值	0.01

B 场景 B 奖励结构

用于训练场景 B 的奖励由两个部分组成：位置奖励和碰撞奖励。

位置奖励与智能体到目标的相对距离变化量成正比。换言之，智能体朝目标移动时会获得正奖励，远离目标时则会获得负奖励。所有智能体共享的位置奖励，等于各自位置奖励之和。当两个智能体都到达目标后，它们还会持续获得额外的最终奖励。回合在经过 500 个时间步后结束。

碰撞奖励是在发生碰撞时给予每个智能体的固定惩罚。训练开始时，只有智能体之间的碰撞会受到惩罚。训练过程中会设置课程学习：当智能体的位置奖励足够高，表明它们已解决任务时，凹槽处发生的碰撞也会开始受到惩罚。这样做是为了让智能体能够先学会完成任务，然后通过消除碰撞来微调其性能。

C 仿真到真实迁移

为了将 VMAS 仿真器中训练得到的策略部署到真实世界，我们会反复调整一些仿真超参数，使其符合真实世界条件。这些参数仅取决于机器人及其与真实世界的交互。一旦完成调整，它们就可用于任何训练场景。

成功部署的关键参数是线性摩擦和阻力。由于我们使用的地面机器人运行速度相对较低，因此将阻力设为 0，并对线性摩擦进行调节。通过 3 次真实到仿真的二分搜索迭代，我们找到了适用于机器人-地面组合的正确摩擦值。除了摩擦之外，我们还在仿真中调节了最大加速度，使其与真实机器人的最大加速度相匹配。这些参数在简单的单机器人任务上进行了测试和验证，例如轨迹跟踪和移动到目标位置。

[7] https://optitrack.com/

[8] https://www.dji.com/uk/robomaster-s1