arXiv:2603.02783v1 [cs.RO] 3 Mar 2026

面向机器人群体的生成对抗模仿学习：从人工示范和训练策略中学习

Mattes Kraus[1] 和 Jonas Kuckling[1,2,3]

摘要—**在模仿学习中，机器人应从期望行为的示范中学习。群体机器人学领域的大多数模仿学习研究，都将现有策略的轨迹展开作为示范。本文提出了一个基于生成对抗模仿学习的框架，旨在从人工示范中学习群体行为。我们在六种不同任务中评估了该框架，分别从手动演示和经 PPO 训练的策略生成的示范中进行学习。结果表明，模仿学习过程能够学会在定性上有意义的行为，其表现与所提供的示范相近。此外，我们在真实机器人实验中，将学习到的策略部署到由 TurtleBot 4 机器人组成的群体上。实验中呈现的行为保留了视觉上可辨识的特征，其性能与仿真中的表现相当。**

I. 引言

机器人群是一种去中心化的同质多机器人系统，其中每个机器人仅依据本地可获取的信息自主行动 [1], [2]。这些特性使其适用于构建具备可扩展性、灵活性和鲁棒性等理想属性的多机器人系统 [2]。尽管机器人群已展现出执行广泛任务的能力 [1], [3], [4], ，为特定任务设计控制软件仍是群体机器人学面临的一项关键挑战 [5], [6]。这是因为群体需要在个体层面进行编程，而期望的群体行为则是由机器人之间大量且往往难以预测的交互涌现而成。因此，手动设计方法通常依赖试错，也就取决于设计者的技能 [7]。

自动化方法会将设计问题转化为优化问题 [7]。优化算法不再为特定任务寻找合适的控制软件，而是搜索能使某种（特定于任务的）性能度量最大化的控制软件实例。通常假定，使特定于任务的性能度量最大化的控制软件实例也能以令人满意的方式完成任务。然而，设计这些性能度量本身就是一项颇具挑战的问题。如果它们只衡量期望结果，那么在设计过程的早期阶段可能无法提供足够的信息 [8], [9], [10]。相反，如果设计者还将行为应当“如何”实现编码其中，性能度量就容易受到奖励劫持 [11]。因此，得分很高的行为可能并不符合期望。

模仿学习方法为指定期望行为提供了另一种途径。设计者无需提供明确的性能度量，而是提供期望行为的示范。随后，模仿学习算法将推导出一种策略，使其呈现出与示范相同的特征 分布。

在本研究中，我们提出了一种面向机器人群体的模仿学习框架，该框架采用生成对抗模仿学习（GAIL） [12]。我们关注群体级特征，以捕捉涌现集体动力学，而非个体智能体的行为。示范由人类操作员提供，他们可以通过定制开发的示范工具控制机器人群体。为了评估人工提供的示范所产生的影响，我们还研究了从采用近端策略优化（PPO）训练的策略中获取示范进行模仿学习的方法 [13]。此外，我们通过真实机器人实验验证了该框架，将学习策略部署到由 TurtleBot 4 机器人组成的真实群体中。

II. 相关工作

在群体机器人学中探讨模仿学习的研究通常采用行为克隆、显式特征匹配、逆强化学习或生成对抗模仿学习等方法。在行为克隆中，利用示范收集一组状态-动作（或观测-动作）对。随后将问题转化为监督学习问题，即训练策略，根据输入的状态预测相应的动作 [14], [15], [16]。在显式特征匹配中，利用示范计算一组特征（智能体级或群体级），以描述期望的行为。之后，模仿学习过程旨在显式地最小化示范中的特征与生成行为的特征之间的差异 [17], [18]。在逆强化学习方法中，特征分布的差异惩罚并未被显式给出。相反，利用这些特征来学习奖励函数，该函数可能会为某些特征赋予不同的重要性 [19], [20], [21]。

[1] 德国康斯坦茨，康斯坦茨大学计算机与信息科学系[2] 康斯坦茨大学卓越集群“集体行为高级研究中心”[3] 康斯坦茨大学 Zukunftskolleg。JK 感谢卡尔·蔡司基金会和 Zukunftskolleg 提供的资助。本研究得到德国卓越战略项目下卓越集群“集体行为高级研究中心”的支持——EXC 2117——422037984。实验由 MK 和 JK 设计，MK 执行。真实机器人实验由 MK 和 JK 执行。论文由 MK 和 JK 撰写并编辑，所有作者均阅读并批准了最终版本。研究由 JK 指导。jonas.kuckling@uni-konstanz.de

在生成对抗模仿学习方法中，不存在显式的目标函数，而是并行训练判别器和生成器（策略） [12]。策略生成状态-动作对，判别器则尝试判断任意状态-动作对是来自示范，还是由模型生成。判别器经过训练以降低分类误差；如果策略生成的状态-动作对被判别器错误地判定为来自示范，策略就会获得奖励。多机器人系统中最早采用生成对抗模仿学习的工作之一是图灵学习 [22]。Li等人采用协同进化，为每个模仿智能体分别进化一个判别器网络和一个生成器模型。判别器只能获取智能体当前的线速度和角速度。Wu等人使用状态-动作对，通过 OpenAI-ES 训练判别器；该方法使用每个智能体的状态-动作对 [23]。作为策略，他们采用模块化的深度注意力网络。Agunloye等人采用多智能体生成对抗模仿学习，为一群无人机学习群体行为 [24]。为生成同质策略，作者采用了带参数共享的多智能体近端策略优化。群体行为通过智能体级特征建模，这些特征用于计算智能体与环境中所有相关元素之间的距离。

群体机器人学中的模仿学习仍面临一个主要挑战：如何提供示范。上述大多数研究都假设已有能够生成所需行为的策略。实际而言，这会导致自举悖论：如果策略已经存在，就没有必要模仿它；而如果无法轻易生成该策略，模仿学习过程就无法获得专家示范。只有少数研究直接探讨了由人工操作者提供示范的问题。值得注意的是，Gharbi等人 [20] 和Szpirer等人 [21] 使用的示范仅包含环境中所有机器人的最终位置。人工操作者可以在模仿学习过程开始前轻松提供这些示范。Mattson和Brown通过设计轮转系统 [14]，解决了如何向整个群体提供示范的问题。人工操作者控制一台机器人，其余群体机器人则由随机策略控制。随后，受人工操作者控制的机器人会通过行为克隆，得出一个模仿该示范的策略。在接下来的步骤中，人工操作者依次切换到其他机器人，为每台机器人提供示范，其余机器人则使用各自学习到的策略。

III. 方法

要将模仿学习技术应用于群体机器人学，需要考虑若干因素。特征既可以定义在智能体级，也可以定义在集体级。虽然智能体级特征与策略的联系更为紧密，但无法捕捉群体的所有特性。因此，我们聚焦于群体级特征（见第 III-B 节）。这使我们能够将该问题建模为单智能体模仿学习问题（见第 III-D 节）。

表 I：TurtleBot 4 机器人的简化模型。

传感器	范围	Dim.
速度 $\lvert v\rvert$	$[0, 0.31m]$ m/s	1
激光雷达	$[0, 200]$ 厘米	5
颜色	$\{0;\ 1\}$	3
保险杠	$\{0;\ 1\}$	1

执行器	范围	Dim.
线速度 v	$[0, 0.31]$ 米/秒	1
角速度 ω	$[-1.9, 1.9]$ 弧度/秒	1

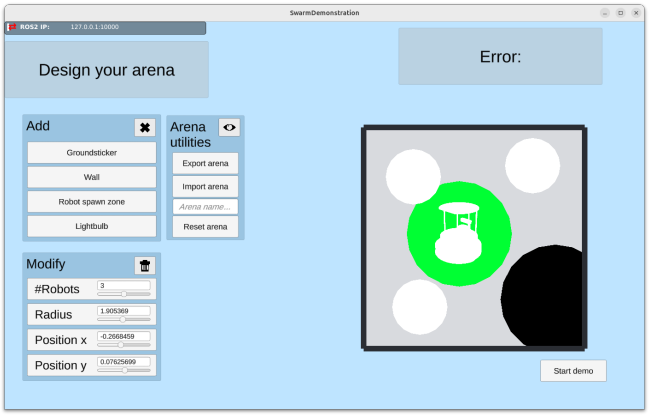
A. 机器人模型

我们为由三台 TurtleBot 4 机器人[1]组成的群体提供示范并设计策略。这些机器人常用于机器人学研究，配备了多种传感器，包括红外接近传感器、二维激光雷达和用于碰撞检测的保险杠。为降低传感器数据的维度，我们采用简化机器人模型（见表 I）。该机器人被建模为线速度为 v 、角速度为 ω的差速驱动智能体。在感知方面，机器人可以获取自身当前的绝对速度 $|v|$。它还可以通过激光雷达感知邻近机器人。激光雷达读数被汇总为五个大小相等的扇区（右前、右侧、左前、左侧、后方），每个扇区覆盖 $72°$ 。每个扇区只考虑距离最近的机器人。如果某个扇区内没有其他机器人，则读数上限设为 200 厘米。为使机器人能够对环境中的不同区域作出反应，我们假设机器人配备了地面传感器。地面传感器可以检测三种不同的灰度颜色：黑色、白色和灰色。地面颜色以一个独热编码的三维向量表示。最后，机器人可以使用保险杠检测与其他机器人或环境的碰撞。

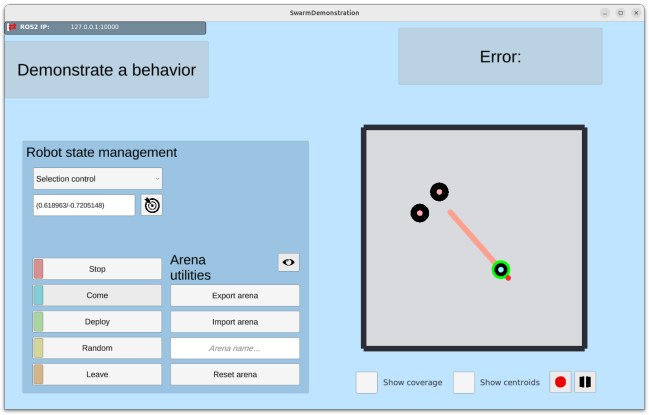
B. 群体行为的特征

我们选取五项特征来描述群体行为：平均速度、聚集、覆盖、颜色切换时间和颜色访问频率。这些特征在任务执行期间的每个时间步计算。平均速度 定义为 $\frac{1}{N}\sum_{i=1}^{N}|v_i|$，其中 N 为群体规模，v_i 为机器人 i的线速度。聚集特征描述机器人到群体中心的平均距离。其计算方式为 $\frac{1}{N}\sum_{i=1}^{N}|x_i - x|$ ，其中 N 为群体规模， x_i 为机器人 i 的当前位置， $\overline{x}$ 为群体的质心位置。为了计算覆盖，我们将场地划分为一个 4×4 网格。每个网格单元记录自上次被机器人访问以来经过的时间。聚集和覆盖率指标取自 [25]。我们还纳入两项指标，用于描述机器人相对于环境中彩色区域的行为。对于颜色访问频率，我们记录每个机器人访问过的地面颜色序列。由于灰色是环境的默认颜色，我们只关注白色和黑色地面色块。每当机器人进入一个彩色色块时，我们就根据当前颜色和上一次访问的地面颜色增加相应计数器。由此得到四个计数器，

[1]https://clearpathrobotics.com/turtlebot-4/



(a) 竞技场构建阶段



(b) 示范构建阶段

图 1：示范工具的两个屏幕截图。用户可以构建实验竞技场，并提供群体行为的示范。

分别对应白-白、白-黑、黑-白和黑-黑四种组合。颜色切换时间 定义为机器人在任意颜色的地面色块之间切换所需时间的平均值。该特征描述四种可能组合的平均时间。这样，每个时间步总共得到 23 项群体级特征。

C. 示范工具/模拟器

为了提供示范，我们开发了一款示范工具（见图 1）。该工具基于 Unity 游戏引擎[2] ，可供下载[3]。示范工具允许用户为群体设计不同的环境（见图 1a）。用户可以通过放置墙壁和黑色或白色地面色块，来修改 4m × 4 m 的场地。此外，用户还可以为机器人指定生成区域。在实验开始时，机器人会随机出现在这些区域内。配置好的场地可以导出和导入，因此能够轻松重复实验。

在为群体设计环境之后，用户可以提供示范（见图 1b）。使用示范工具，用户可以向群体或其中的单个机器人下达高层级命令。先前的研究表明，与向用户提供低层级控制相比，高层级交互命令能产生更好的示范 [26]。可用的交互命令改编自 Kolling 等人 [27]。Kolling 等人定义了两种控制模式：选择控制 和信标控制。在选择控制中，用户可以通过点击来选择哪些机器人应改变其行为。信标控制是一种环境控制方法，用户并不直接改变机器人的行为。相反，他们可以放置信标，即与特定行为相关联的触发区域。每当机器人进入信标时，它会立即切换到该信标所关联的行为。

Kolling 等人定义了每个机器人都可以执行的七种行为：停止、随机、靠近、离开、航向、会合和部署。然而，操作者使用各种行为的频率并不相同。因此，我们没有实现会合 和航向。其余行为的实现方式如下：执行停止 行为时，机器人保持静止。执行随机 行为时，机器人以弹道运动的形式进行随机游走。机器人以最高速度直线行进，直到遇到障碍物。随后，它会原地转动随机数量的时间步，然后继续直线行进。对于靠近和离开行为，需要在场地内选择一个目标位置。执行靠近 行为时，机器人以最高速度驶向目标。接近目标时，它会逐渐减速。当目标进入半径为 20 cm 的范围内时，机器人停止。执行离开 行为时，目标会对机器人产生排斥作用。发生碰撞时，机器人会切换到随机行为。执行部署 行为时，机器人以每个机器人的位置为生成点计算沃罗诺伊图。机器人以最高速度驶向分配给自己的区域的质心 。与其他机器人发生碰撞时，它会略微向右改变方向，以避免进一步碰撞。

为了收集用户演示，示范工具能够提供机器人行为的轻量级仿真。为了尽量减小示范与设计之间仿真上下文的差异，我们还将示范工具用作模仿学习过程的仿真器。GAIL 库与示范工具之间的通信通过 ROS 2 和 Unity Robotics Hub[4]实现。

D. SwarmGAIL

我们将机器人群的模仿学习问题简化为单智能体生成对抗模仿学习问题。对于我们的模仿学习算法，我们采用了略作调整的 GAIL [12] 版本（见图 2）。该算法学习一个单一策略（生成器），以轮询方式控制群体中的每个机器人。该策略以机器人模型生成的观测（见表 I）为输入，并输出焦点机器人的执行器指令。由于该策略是去中心化的，且仅依赖于

[2]https://unity.com/products/unity-engine[3]https://github.com/CPS-Konstanz/SwarmGAIL

[4]https://github.com/Unity-Technologies/Unity-Robotics-Hub

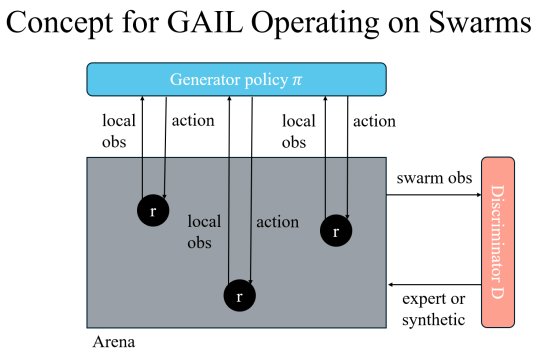
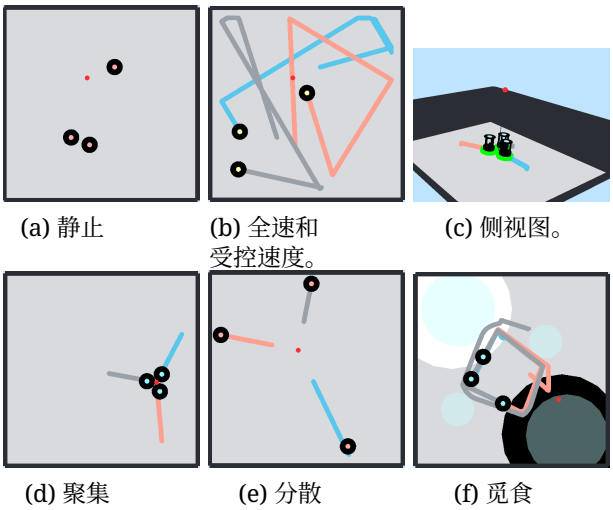


图 2：我们的 SwarmGAIL 实现中的信息流。判别器接收群体级特征作为观测。策略以轮询方式使用，根据各机器人的局部观测为所有机器人提供动作。



(a) 静止　(b) 全速和受控速度。　(c) 侧视图。

(d) 聚集　(e) 分散　(f) 觅食

图 3：所有所考虑任务的示范行为表示。对于全速和受控速度，机器人所走的轨迹在视觉上看起来相似，但机器人的速度有所不同。

单个机器人的观测，因此我们之后可以在每个机器人上部署该策略的一个副本。

判别器旨在区分示范行为与生成行为。作为输入，我们提供特征—动作对（群体级特征的定义见第 III-B 节）。值得注意的是，这些特征不包含当前焦点机器人的局部观测。这确保判别器仅根据群体行为进行分类，而不会受到示范与生成行为之间观测差异的无意影响。

我们使用 imitation 库 [28]在 Python 3 中实现了这一调整后的 GAIL 版本。判别器是一个多层感知机，包含两个隐藏层，每层大小为 32。策略是一个多层感知机，包含两个隐藏层，每层大小为 64，并使用近端策略优化 [13]进行训练。由于技术原因，策略和判别器需要相同的观测空间。因此，我们考虑一个结合局部观测与群体特征的联合观测向量。在将该向量提供给判别器或策略之前，我们将观测向量中不必要的部分覆写为 0。

IV. 实验设置

A. 任务

为了测试我们的模仿学习框架，我们考虑了六项任务：静止、全速、受控速度、聚集、分散和觅食（见图 3）。对于每项任务，我们都定义了一种通用策略，用于提供示范；此外，还定义了一个奖励函数，并假设我们的示范使该函数达到最大值。

静止、全速、受控速度、聚集和分散这些任务都在空旷场地中进行。机器人初始位置大约在环境中央。在静止任务中，机器人不应向前移动（允许原地旋转）。示范通过选中所有机器人并为其指定停止 行为来完成。奖励函数定义为 $R_s(t) = -\bar{v}(t)$，其中 $\bar{v}(t)$ 是群体在时间步 t 的平均线速度。在全速任务中，机器人应尽可能快速地移动。因此，该任务的奖励函数定义为 $R_f(t) = \bar{v}(t)$。进行示范时，选中所有机器人，并为它们应用随机 行为。对于受控速度任务，机器人应以恒定速度 $v_{target} = 0.1\text{m/s}$ 移动。由于我们的示范工具中的行为都无法明确设置线速度，我们降低了示范时允许的最大速度。然后，我们选中所有机器人，并为它们应用随机 行为。该任务的奖励函数定义为 $R_c(t) = |v_{target} - \bar{v}(t)|$。

在聚集任务中，机器人应在场地内聚集，但聚集地点没有特定要求。奖励函数定义为 $R_a(t) = 1/\bar{g}$，其中 $\bar{g}$ 是群组特征，即机器人到群体质心的平均距离。示范通过选中所有机器人并为其指定靠近行为来提供，目标点选在群体质心附近。类似地，在分散任务中，机器人应尽可能分散在场地内。奖励函数定义为 $R_d(t) = -1/\bar{g}$。示范通过选中所有机器人并为其指定离开 行为来提供。觅食是我们考虑的最复杂行为。环境包含一个白色区域（巢穴）和一个黑色区域（源点），机器人则在场地中央生成。机器人需要从巢穴收集（虚拟）物品，并将其放置到源点。机器人进入相应区域后，物品便会自动取回或放置。每成功取回或放置一件物品，即可获得 10 分。示范通过信标控制提供。在灰色区域放置了两个信标。进入这些信标时，机器人会切换到靠近 行为，目标点分别位于巢穴和源点内。巢穴和源点内也设有信标，进入这些信标后，机器人将会

切换到靠近 行为，目标点为通往相应下一个目标的路径点。开始时，选中机器人并为其指定靠近 行为，使其前往最近的感兴趣区域（路径点、巢穴或源点）。

我们知道，这些都是基础任务，所需的群体行为也可以通过模仿以外的方法实现。事实上，文献中已有现成的此类行为实现方法 [29]。我们选择聚焦于这些基础任务，以降低示范失败的风险。这使我们能够将性能视为所提出框架的结果，而非所获取示范质量的结果。

B. 真实环境

除了模拟任务外，我们还考虑在真实环境中开展实验。机器人群在一个 2 m × 4m 的场地内运行，场地四周由墙壁围住。每个机器人都执行各自的分散式策略。为防止机器人和环境受到损害，我们实现了一个硬件保护层。一旦检测到（潜在）碰撞，硬件保护层就会禁止策略输出，并使机器人原地旋转，直至障碍物被清除。此外，我们通过 Vicon动作捕捉系统追踪机器人。这样便能追踪场地内机器人的位置和朝向，并计算这些实验中获得的奖励。

C. 实验方案

对于每项任务，我们进行五次相互独立的实验运行。在每次运行中，我们提供一次人工操作示范，用于生成五条不同的轨迹展开，从而得到五条由特征-动作对组成的轨迹。我们按照上述一般策略提供示范，并不断改进具体的示范内容，直到对群体性能感到满意为止。每次示范耗时 51 秒。在我们的实验中，这在等待目标行为出现所需的时间与评估所有策略所需的总时间之间实现了良好的权衡。随后，将这五条示范轨迹展开作为专家示范提供给模仿学习框架。为了减少评估策略所需的时间，我们将仿真中机器人的控制频率降低至 1 Hz。为跟踪模仿学习过程的发展，我们定期记录与判别器相关的信息，以及当前策略的副本（每次运行总共记录 100 个策略）。为了评估模仿策略的质量，随后在仿真中对其进行五次评估。此外，我们将其迁移至真实机器人，并在真实环境中评估一次。真实机器人的控制频率设为 10 Hz。

此外，我们还考虑从训练策略中进行模仿学习。为此，我们使用近端策略优化（PPO）和第 III-D 节所述的策略架构，针对每项任务执行标准的强化学习过程。随后，在仿真中对生成的策略进行五次评估，以生成五条策略展开轨迹（类似于五条人工操作

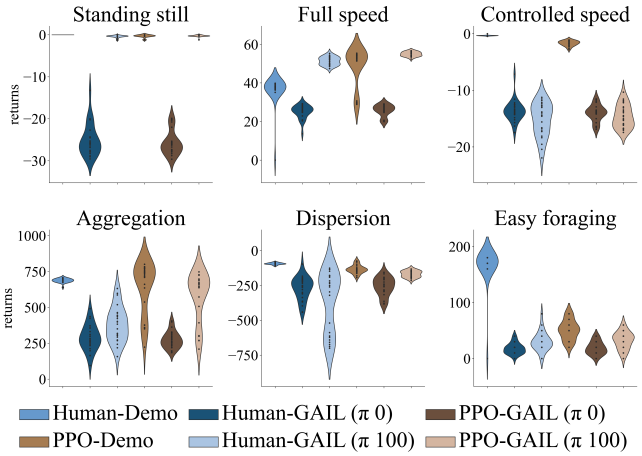


图4：所有所考虑任务中全部评估的回报（累积奖励）小提琴图。颜色表示示范的来源（蓝色表示由人工操作产生，Brown 表示由近端策略优化训练产生）。深色表示每项模仿学习实验的初始策略，浅色表示最终策略。黑色圆点表示对某个策略/示范进行的一次评估。回报越高，性能越好。

示范）。接着，在与人工操作示范相同的实验方案下，将这些轨迹作为模仿学习过程的输入。具体而言，我们还针对每项任务重复这一过程五次，从而得到五个不同的学习策略，每次重复对应一个。

所有代码、示范、策略、性能数据和视频均可在线获取[5]。

V. 结果

A. 手动演示与训练策略

图4展示了所有示范的轨迹展开性能。总体而言，采用手动控制或训练策略都能产生表现相近且良好的行为。值得注意的是，人工示范并非明确以最大化奖励函数为目标。表现质量上的差异通常源于任务和示范各自的特点。例如，在“静止”任务中，手动示范从一开始就在所有机器人上使用 停止 行为，因此很容易达到恰好为 0 的分数。然而，经 PPO 训练的策略难以始终将线速度设为恰好 0。0 ms^{-1}。因此，有些机器人会不时移动，从而累积少量惩罚。

在“全速”任务中，情况恰恰相反：经 PPO 训练的策略通常会发现人工演示者无法采用的策略。我们的示范采用随机行走 行为，使机器人执行弹道运动行为。机器人以最大速度直线移动，直到遇到墙壁，此时它们会原地转向。原地转向会导致一段时间内

[5]https://github.com/CPS-Konstanz/SwarmGAIL

机器人的线速度为 0。0 毫米/秒$^{-1}$，因此无法获得奖励。然而，在经 PPO 训练的策略中，机器人通常会沿圆周运动。这样可以使线速度（几乎）保持不变，同时将角速度设得很高，以避免碰撞。觅食任务也尤其值得关注。人工提供的示范通常能取得 150 分以上的成绩。然而，经 PPO 训练的策略在任何一次轨迹展开中都无法取得至少 100 分。这一观察结果支持了我们的看法：应当更加重视人工示范。尤其是在更复杂的任务中（觅食仍是这类任务的一个简单代表），人工示范很容易胜过其他方法生成的行为。

最后需要指出的是，与经 PPO 训练的策略相比，人工示范的性能方差较小。这有两个原因：首先，五个经 PPO 训练的策略中，有些可能无法充分完成任务；然而，手动示范从未出现过这个问题。其次，同一 PPO 策略在重复轨迹展开中的性能波动，比人工操作者在重复示范中的性能波动更大。只有在觅食任务中，人工操作示范的方差才高于由经 PPO 训练的策略生成的示范。

B. 生成对抗模仿学习

图4还展示了生成对抗模仿学习过程的性能。除了示范（人工示范和经 PPO 训练的示范）的性能外，我们还展示了初始策略（π 0）和最终生成的策略（π 100）的性能。由于初始策略是随机生成的，其性能仅作为基线，用于凸显性能提升的程度。

对于静止和全速这两个任务，我们观察到，初始策略正如预期一样相对较差。然而，最终训练得到的策略达到了示范的质量，在全速任务中甚至超过了手动示范。学习策略利用了能够同时控制线速度和角速度这一事实，而手动示范依赖于弹道运动。如前所述，碰撞后原地转向的弹道运动会产生少量惩罚。然而，在模仿策略中，机器人会设定角速度以避免碰撞。这一策略与使用近端策略优化训练的策略所采用的策略相似。不过，由人工操作示范训练得到的策略转向次数少于使用近端策略优化训练的策略，后者基本上会原地转向。这是因为弹道运动覆盖了整个场地，从而影响了聚集或覆盖等群体级特征。因此，模仿策略也会覆盖整个场地。相应地，模仿使用近端策略优化训练的示范得到的策略也学会了几乎原地转向，密切模仿其示范行为。这表明，模仿学习过程能够捕捉示范行为中的这些特异之处，并在相应的模仿策略中将其复现。然而，GAIL 未能正确模仿弹道运动的原地转向，这一点值得进一步研究，即使从理论上看，这种行为带来了更好的表现。

在受控速度任务中，模仿策略的性能未能达到示范的水平。事实上，与初始策略相比，模仿过程似乎对性能产生了负面影响。对初始策略和最终策略进行目视检查后发现，两者似乎都依靠随机停止和移动来控制线速度。这会导致较高的惩罚，因为平均速度大部分时间不是过低就是过高。然而，人工操作和由近端策略优化训练得到的示范始终以接近目标速度移动。目前尚不清楚为何示范行为没有在任一模仿学习过程中涌现。

在聚集任务中，使用由近端策略优化训练的策略所生成的轨迹展开作为示范的 GAIL 实验，成功学会了很好地模仿该行为。模仿人工示范的策略虽然比随机初始化的策略有所改进，但表现仍不如原始策略。对生成行为进行目视检查后发现，机器人基本处于静止状态，只是试图缓慢地相互靠近。因此，该行为的分数在很大程度上取决于机器人的初始位置。这可能是群体级特征与手动示范相互作用所致。人工操作者快速完成聚集后，机器人便保持静止。结果是，大多数特征的发展方式都与静止群体相似（早期平均速度有所提高的情况除外）。模仿学习过程似乎难以复现这种聚集行为，而主要关注示范中持续时间较长的静止阶段。然而，在近端策略优化训练的示范及其产生的模仿策略中，聚集过程是动态的，而非静态的（另有一对策略也学会了静态聚集）。这使群体级特征呈现出更具动态性的表征，显然更有利于模仿学习过程。

在分散任务中，模仿学习相较于随机初始策略仅有小幅改进。不过，这些策略的性能方差较小（其中一个策略似乎学会了聚集行为，除此之外皆是如此）。这可以解释为初始策略不包含非常结构化的运动。因此，它们的分数主要取决于初始位置。学习策略提供了更结构化的行为，使机器人在起始位置过于接近时能够彼此分开。然而，由于机器人的激光雷达读数范围仅为 2 米，它们只能分开到彼此无法感知对方为止。人工演示者还可以更进一步，将机器人尽可能分散地放置在场地中。

在最后考察的觅食任务中，人工示范明显优于模仿策略。我们最初认为，机器人或许会学会沿着与源点和巢穴重叠的圆形轨迹或螺旋轨迹移动（手动示范也从这一策略中汲取灵感，并设置了相应的路径点）。

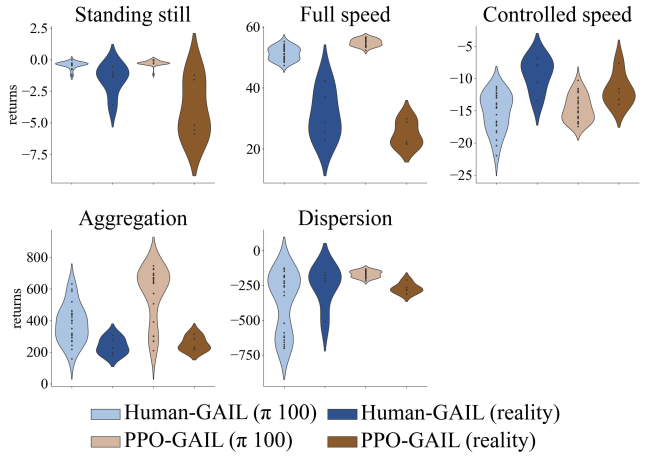


图 5：仿真（浅色）和现实（深色）中模仿策略的回报（累积奖励）小提琴图。颜色表示示范的源点（蓝色表示由人工操作产生的示范，Brown 表示由近端策略优化训练得到的示范）。黑色圆点对应对某个策略/示范的一次评估。回报越高，性能越好。

然而，初始策略通常会产生缓慢、随机移动的行为。最终训练出的策略（无论是基于人工操作示范还是 PPO 训练的示范）都表现出更快、但仍然随机移动的行为。移动速度的提升增加了机器人拾取和存放物品的可能性，但这些情况仍属偶然。由于 PPO 训练的示范也未能展现出任何明显有别于随机游走的行为，我们认为这可能表明环境缺乏可供机器人利用的导航结构。此外，与颜色相关的群体级特征可能进一步增加了模仿学习过程的难度。手动示范能够轻松地在地面色块之间导航，因此呈现出较高的访问频率和较低的颜色通行时间。学习策略的导航要困难得多，因此无法匹配这些特征。由此，判别器可以轻而易举地区分示范行为与学习到的行为。因此，策略改进的奖励景观是平坦的，因为任何可能的改进都无法战胜判别器。

C. 真实世界验证

图 5 展示了真实机器人实验的性能。我们出于两个原因，没有将觅食任务纳入考量。首先，真实的 TurtleBot 4 机器人没有执行该任务所需的地面传感器。虽然我们本可以提供虚拟地面传感器，但我们也调整了场地尺寸。这些策略在仿真中的表现并不特别好，而场地尺寸的变化很可能会进一步影响其性能。

可以看出，由人工示范和 PPO 训练的示范生成的策略，在真实机器人上评估时表现相近。这并不令人意外，因为这些策略在仿真中的表现本来就相对接近。值得注意的是，通过目视检查机器人群，可以发现每项任务都会产生明显不同且易于识别的行为（唯一区别是，分散策略导致了类似聚集的行为）。这进一步证明，我们的框架能够成功模仿示范行为，而且生成的行为在定性上能够稳健地应对现实差距。

现实差距对策略的影响各不相同。硬件保护层对行为的影响最大。在聚集和全速等任务中，硬件保护层会阻止碰撞，而在仿真中，碰撞被视为行为中可接受的一部分。因此，机器人会减速并转向避开彼此，或者停在比仿真中更远的位置。有趣的是，这种影响似乎有利于受控速度任务：机器人在现实中的表现优于仿真中的表现。静止和分散任务的性能未受到硬件保护层的显著影响，因为在实验运行期间，机器人很少彼此相遇。这使我们相信，如果在仿真中对硬件保护层进行适当建模，我们的方法或许能够很好地迁移到真实机器人上。

VI. 结论

在这项工作中，我们提出了一种用于机器人群体的模仿学习框架。我们将模仿学习问题建模为单智能体问题，并使用 GAIL 求解。判别器根据五种群体级特征对样本进行分类，而策略则根据机器人各自的局部观测控制机器人。我们还提供了一种工具，以便进行人工操作示范。结果表明，除受控速度和觅食任务外，对于所有任务，模仿学习过程都能够学到在定性上有意义的行为，其表现与提供的示范相近。此外，我们还在真实机器人上测试了这些策略。策略迁移效果总体较好，在真实机器人上仍能直观地识别出这些行为。不过，仿真中没有对真实机器人的硬件保护层进行建模，因此影响了迁移策略的性能。

然而，我们的工作也揭示了该方法的一些局限性。尽管考虑的任务相对简单，但在某些情况下，示范行为并不总能被准确模仿。我们认为，这在一定程度上是由于临时选定的群体级特征影响了模仿学习过程；不过，还需要开展更多研究来确定限制因素。另一个潜在因素是，在觅食等更复杂的任务中，人类示范和近端策略优化训练的示范都表现出更大的性能方差。我们认为，还需要进一步研究，为这些复杂任务提供可靠的示范

任务。此外，未来的工作应致力于更好地理解特征如何影响模仿学习过程（例如，通过开展消融实验），以及研究人员如何系统地选择这些特征，以确保模仿行为具有良好的性能。我们将相同的特征同时用作判别器和奖励函数的输入，这一做法也值得进一步考量。这可能导致信息泄露，即判别器隐式地学会哪些特征与示范所假定要最大化的奖励函数相关。未来的工作应考虑采用一种更加解耦的特征空间与奖励函数方法。这一点在转向更复杂的任务时尤为重要。

尽管存在这些局限性，我们的方法也为研究开辟了新的方向。我们认为，将这一方法与文献中的其他方法进行比较，从具备不同群体机器人学专业水平的专家那里获取示范，并研究训练策略在不同群体规模下的扩展能力，都是值得关注的。

致谢

我们感谢 Mahabaleshwar Ammu 在搭建真实机器人实验方面提供的帮助。

参考文献

[1] M. Brambilla、E. Ferrante、M. Birattari 和 M. Dorigo，《群体机器人学：从群体工程视角出发的综述》，群体智能，第7卷，第1期，第1–41页，2013年。[2] M. Dorigo、M. Birattari 和 M. Brambilla，《群体机器人学》，*Schol- arpedia*，第9卷，第1期，第1463页，2014年。[3] M. Schranz、M. Umlauft、M. Sende 和 W. Elmenreich，《群体机器人行为及当前应用》，*Front. Robot. AI*，第7卷，第36页，2020年。[4] M. Dorigo、G. Theraulaz 和 V. Trianni，《群体机器人学：过去、现在与未来的观点》*IEEE*会议录，第109卷，第7期，第1152–1165页，2021年。[5] G. Francesca 和 M. Birattari，《机器人群体的自动化设计：成果与挑战》，*Front. Robot. AI*，第3卷，第29期，第1–9页，2016年。[6] J. Kuckling，《群体机器人学中机器人学习与进化的最新趋势》，*Front. Robot. AI*，第10卷，2023年。[7] M. Birattari、A. Ligot、D. Bozhinoski、M. Brambilla、G. Francesca、L. Garattoni、D. Garz´on Ramos、K. Hasselmann、M. Kegeleirs、J. Kuckling、F. Pagnozzi、A. Roli、M. Salman 和 T. St¨utzle，《机器人群体的自动离线设计：宣言》，*Front. Robot. AI*，第6卷，第59页，2019年。[8] J. Lehman 和 K. O. Stanley，《放弃目标：仅通过寻找新颖性实现进化》，进化计算，第19卷，第2期，第189–223页，2011年。[9] S. Doncieux、N. Bredeche、J.-B. Mouret 和 A. Eiben，《进化机器人学：是什么、为何以及走向何方》，*Front. Robot. AI*，第2卷，第4页，2015年。[10] M. Divband Soorati 和 H. Hamann，《适应度函数设计对进化机器人学性能的影响：先验知识的作用》，载于*GECCO 2015*，S. Silva 编。美国纽约州纽约：ACM，2015年，第153–160页。[11] A. Y. Ng、D. Harada 和 S. J. Russell，《奖励变换下的策略不变性：奖励塑造的理论与应用》，载于第十六届机器学习国际会议论文集，ICML '99 系列。美国加利福尼亚州旧金山：Morgan Kaufmann Publishers Inc.，1999年，第278–287页。[12] J. Ho 和 S. Ermon，《生成对抗模仿学习》，神经信息处理系统进展，第29卷，2016年。[13] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，《近端策略优化算法》，*arXiv*预印本 *arXiv:1707.06347*，2017年。

[14] C. Mattson 和 D. S. Brown，《R2BC：从单智能体示范中进行多智能体模仿学习》，2025年。 [在线]。可访问：https://hitl-robot-learning.github.io/pdfs/r2bc.pdf [15] I. M. Rahman、S. White、K. Crockett、Y. Gu、D. A. A. Dutra 和 G. A. S. Pereira，《通过学习智能体级控制器模仿群体行为》，载于2023年美国控制会议（ACC）。美国加利福尼亚州圣迭戈：IEEE，2023年5月，第7–13页。 [在线]。可访问：https://ieeexplore.ieee.org/document/10156561/ [16] S. Zhou、M. Phielipp、J. A. Sefair、S. I. Walker 和 H. B. Amor，《克隆群体：通过模仿学习预测和控制多机器人系统》，载于2019年IEEE/RSJ智能机器人与系统国际会议（IROS），2019年11月，第4092–4099页，arXiv:1912.02811 [cs]。 [在线]。可访问：http://arxiv.org/abs/1912.02811 。K. Alharthi、Z. S. Abdallah 和 S. Hauert，《从群体行为的视频观测中自动提取易于理解的控制器》，载于群体智能国际会议。Springer，2022年，第41–53页。[18] ——，《幽灵群体：仅从环境变化中学习群体规则》，载于欧洲遗传编程会议（EvoStar的一部分）。Springer，2025年，第1–17页。[19] A. ˇSoˇsi´c、W. R. KhudaBukhsh、A. M. Zoubir 和 H. Koeppl，《群体系统中的逆强化学习》，载于第16届自主智能体与多智能体系统国际会议（AAMAS 2017）论文集，2017年。 [在线]。可访问：https://ifaamas.org/Proceedings/aamas2017/pdfs/p1413.pdf [20] I. Gharbi、J. Kuckling、D. Garz´on Ramos 和 M. Birattari，《告诉我你的需求：通过示范进行逆强化学习，以自动设计机器人群体》，载于2023年IEEE机器人与自动化国际会议（ICRA）。美国新泽西州皮斯卡塔韦：IEEE，2023年，第5064–5070页。[21] J. Szpirer、D. G. Ramos 和 M. Birattari，《自动设计可执行复合任务的机器人群体：一种基于逆强化学习的方法》，载于2024年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2024年，第5791–5798页。[22] W. Li、M. Gauci 和 R. Groß，《图灵学习：一种推断行为的无度量方法及其在群体中的应用》，群体智能，第10卷，第3期，第211–243页，2016年9月。 [在线]。可访问：http://link.springer.com/10 1007 11721 016 0126 1 。/s - - - [23] Y. Wu、T. Wang、T. Liu、Z. Zheng、D. Xu 和 X. Peng，《结合深度注意力网络的群体系统对抗模仿学习》，复杂与智能系统，第11卷，第1期，第26页，2025年1月。 [在线]。可访问：https://link.springer.com/10.1007/s40747-024-01662-2 [24] A. O. Agunloye、S. D. Ramchurn 和 M. D. Soorati，《学习模仿多机器人系统中的空间组织》，载于2024年IEEE/RSJ智能机器人与系统国际会议（IROS）。阿拉伯联合酋长国阿布扎比：IEEE，2024年10月，第3463–3469页。 [在线]。可访问：https://ieeexplore.ieee.org/document/ 10801662/ [25] L. G. Jeston-Fenton、S. Abpeikar 和 K. Kasmarik，《在手持设备上可视化群体指标以实现人群交互》，载于2022年第26届国际信息可视化会议（IV）。奥地利维也纳：IEEE，2022年7月，第149–154页。 [在线]。可访问：https://ieeexplore.ieee.org/document/10017712/ [26] A. Hussein、H. Nguyen 和 H. A. Abbass，《不同自主程度下群体元宇宙中的机器教学》，英国皇家学会哲学汇刊A辑：数学、物理与工程科学，第383卷，第2289期，第20240149页，2025年。[27] A. Kolling、S. Nunnally 和 M. Lewis，《迈向人类对机器人群体的控制》，载于2012年第7届ACM/IEEE人机交互国际会议（HRI），2012年3月，第89–96页，ISSN：2167-2148。 [在线]。可访问：https://ieeexplore.ieee.org/document/ 6249619 [28] A. Gleave、M. Taufeeque、J. Rocamonde、E. Jenner、S. H. Wang、S. Toyer、M. Ernestus、N. Belrose、S. Emmons 和 S. Russell，《imitation：整洁的模仿学习实现》，arXiv:2211.11972v1 [cs.LG]，2022。 [在线]。可访问：https: //arxiv.org/abs/2211 11972 。T. K. Kaiser、M. J. Begemann、T. Plattenteich、L. Schilling、G. Schildbach 和 H. Hamann，《ROS2SWARM——用于群体机器人行为的ROS 2软件包》，载于ICRA 2022。IEEE，2022年，第6875–6881页。 [29]