arXiv:2304.14102v2 [cs.RO] 7 Jul 2023

# SocNavGym：用于社会导航的强化学习 Gym

Aditya Kapoor*[1]、Sushant Swamy*[2]、Pilar Bachiller[3] 和 Luis J. Manso[4]

摘要——**自主机器人在有人类活动的环境中导航时，遵循社会规范至关重要。近年来，机器学习，尤其是深度强化学习，在社会导航领域受到广泛关注。这在一定程度上是因为由此产生的策略不受人类在代码复杂度或所处理变量数量方面的局限。不幸的是，深度强化学习算法缺乏安全保证，且数据需求庞大，因此无法在真实世界中进行学习。为弥合这一差距，人们经常使用仿真环境。我们提出了 SocNavGym，这是一个先进的社会导航仿真环境，能够生成丰富多样的社会导航场景，并助力智能社交智能体的开发。SocNavGym 轻量、快速、易于使用，而且可以轻松配置，以生成不同类型的社会导航场景。它还可配置为支持不同的手工设计和数据驱动的社会奖励信号，并生成多种评估指标，用于衡量智能体的表现。此外，我们还提供了一个案例研究，展示如何使用 SocNavGym 训练 Dueling-DQN 智能体学习社会导航策略。结果表明，SocNavGym 可用于从零开始训练智能体，使其能够在简单及复杂的社会场景中导航。我们的实验还表明，与使用基于启发式方法的奖励函数训练的智能体相比，使用数据驱动奖励函数训练的智能体表现出更先进的社会规范遵从性。**

## I. 引言

社会规范遵从性是部署机器人的关键，不仅关系到行人的舒适度，也有助于提升效率。该领域的大量研究充分证明了这一点 [1]。直到最近，绝大多数社会导航（SN）算法仍由人类手工设计并编写代码。这种方法的局限之一是可扩展性较差，尤其是在所考虑的变量数量方面；部分原因是相关方程变得复杂，难以理解、实现，或许最重要的是难以调试。其他令纳入更多变量变得困难的因素包括缺乏关于行人舒适度的系统理论，以及世界状态只能被部分观测这一事实（例如，行人的隐藏目标、意图、情绪）。由于这些局限，业界正明显转向能够考虑更多变量的数据驱动方法。最常见的采用学习方法的社会导航方案，是将手工设计算法与基于学习的人体轨迹预测器结合起来。这类工作的例子包括[2], [3], [4], [5], [6], [7], ，其中通过搜索树和其他基于启发式的算法来规划动作，并依赖对人体状态的估计 [8]。其他基于学习的工作还包括社会合规性估计器，例如社会导航图神经网络（SNGNN） [9]。第三类颇具前景且近来日益受到关注的社会导航算法，则采用深度强化学习（DRL）和示范学习（LfD） [10], [11], [12], [13], [14], [15]。

DRL系统包含三个尤为关键的组成部分：1）DRL 算法；2）神经网络架构；3）提供观测和奖励的环境。虽然DRL研究主要聚焦于前两个方面，但现有文献的一项局限在于，现有Gym环境中使用的奖励函数，无一旨在反映用户对智能体情境社会合规性的整体意见。现有奖励函数更适合解决多智能体避碰问题，而非社会导航；后者是一个不同且相对简单的问题。据我们所知，基于DRL的社会导航文献所使用的奖励函数全都是分段定义的函数，其中用于考量智能体社会规范遵从性的部分只是一个简单的线性方程，或类似的轮廓函数，且仅考虑行人的位置。我们强烈认为，要通过DRL学习符合社会规范的策略，就应精心设计奖励函数，将用户意见纳入考量。

逆强化学习（IRL）在社会导航方面展现出了良好的前景 [16], [17], [18]。如果使用充足的数据进行训练，学习得到的奖励函数有助于解决这些问题，因为可以训练 DRL 架构对其进行优化，从而实现社会合规策略学习。理论上，由此产生的行为可以像这样学习得到的奖励函数一样有原则且有效。基于这一思路，我们的假设是，将学习得到的舒适度模型用作奖励函数，与其他简单或手工设计的奖励相比，能够提高社会规范遵从性。因此，我们的训练环境 SocNavGym 将 SNGNN-v2 [19] 集成到其奖励函数中；据我们所知，这是文献中唯一能够提供逐步社会合规性评分的模型。

SocNavGym 设计所遵循的要求如下：

1) 轻量、快速、易于使用和配置，并尽可能减少软件依赖项；2) 模拟丰富多样且逼真的社会场景，让智能体能够应用于尽可能多的现实场景

[1]Aditya Kapoor 就职于印度 Tata Consultancy Services 研究与创新部门。aditya.kapoor1@tcs.com[2]Sushant Swamy 就读于印度果阿的 Birla Institute of Technology and Science。f20191031@goa.bits-pilani.ac.in[3]Pilar Bachiller 就职于西班牙埃斯特雷马杜拉大学计算机与通信技术系。pilarb@unex.es[4]Luis J. Manso 就职于英国阿斯顿大学计算机科学系自主机器人与感知实验室。l.manso@aston.ac.uk

* 为共同贡献。

中；3) 易于修改和扩展；4) 提供高质量的社会奖励，确保近乎最优的策略符合真实行人的预期；5) 提供成熟的评估指标，以衡量智能体的表现。

据我们所知，现有的导航健身环境中没有一个能够满足上述所有要求（详见第二节）。

SocNavGym 是通过扩展 OpenAI 的多智能体粒子环境（MPE） [20] 构建的二维仿真环境。我们借鉴了现有文献中的相关工作，并将其中一些最相关的特征融入 SocNavGym。这些附加特征包括静态障碍物类实体、静态和动态智能体队形，以及不同实体之间的交互（例如，人看着笔记本电脑、人与人交谈）。我们还集成了在用户意见数据集 [21], [22] 上训练的数据驱动奖励函数 SNGNN-v2 [19], a。此外，我们训练了一个 DuelingDQN [23] 模型，以展示我们的仿真环境可用于在各种丰富的社交情境中训练深度强化学习算法。

## II. 相关工作

许多高保真模拟器已被开发出来，用于收集数据并训练机器人应用例如所需的模型，如 CARLA [24], iGibson [25]。Nomad [26] 和 Menge [27] 等人群仿真平台旨在模拟个体和人群。相比之下，SEAN [28]和 SocialGym [29] 旨在训练和评估社交智能体，但它们依赖 Unity [30]和机器人操作系统（ROS），由于所需的额外复杂性和计算资源，对许多 DRL 研究人员而言可能并不理想。尽管这些模拟器能提供良好的物理仿真，但它们需要大量计算能力，而对于这一特定应用而言，这种投入似乎并不合理。此外，为了专注于社会导航中的决策，可以省去感知任务，这也进一步削弱了对高保真模拟器的需求。SocNavGym 和其他最先进的社会导航健身环境（例如，SARL [10], DSRNN [11]）采用低保真模拟器，但与 SocNavGym 不同，其他环境未能重现真实而复杂的社交场景。在 SARL [10],中，场景由围成一圈的人类组成，他们使用 ORCA策略朝着直径的相反方向移动，周围则是一片空旷区域。在 DSRNN [11], 中也有静态和动态群体，但它们仍与典型的有人类活动的环境大相径庭。SocialGym [29] 更进一步，引入了类似房间的场景，其中包含走廊和更真实的行人移动。然而，它依赖 ROS [31] ，增加了额外的开销，而且其设置和安装过程并不如预期那样简便。此外，要对环境进行任何更改，用户都必须了解 ROS 框架，这并不理想。另外，这些环境都未提供一种有原则的社交奖励函数来考虑人类意见，因此使用它们训练的智能体会将人类视为动态障碍物，而非社会性个体。

仿真环境中发现的另一个缺陷是缺乏对常见交互的支持，例如人与人之间以及人与其他无生命实体之间的交互。在为真实场景训练智能体时，这类交互非常重要，因为智能体所造成的干扰程度取决于这些交互。例如，即使智能体没有进入相关人员的个人空间，它也可能对正在交谈的两个人造成相当大的干扰（人际交互）。无论这些交互是否能被智能体直接观察到，或是否需要智能体推断，都应将其纳入考虑（SocNavGym 同时支持这两种选项）。

除了 SocialGym [29]和 SEAN [28],之外，所有经过评估的社会导航健身环境都未提供足够的评估指标（例如，到达目标所用时间、与人的最小距离、不适程度），无法用于衡量社会性表现。尽管这些健身环境提供了部分评估指标，但没有任何指标或奖励函数能够兼顾效率和人类舒适度。

## III. 健身环境的设计决策与特性

SocNavGym 遵循标准的 OpenAI Gym API 规范。安装和设置也很简单，既可以从源代码安装，也可以通过 pip 安装。所有可调整参数均在项目文档中说明，无需访问代码库，即可通过配置文件进行修改。项目还提供了若干示例配置文件，以方便使用所提出的工具。此外，项目提供了 DuelingDQN [23] 基线作为参考，相关说明见第四节。我们提出了一款新的开源[1] 健身环境，满足第二节所述的设计要求。具体而言，SocNavGym 具有以下特性：

1) 轻量级随机场景生成与模拟，2) 支持静态物体、走廊和墙体，以围出不同形状和大小的房间（图 1.i-j）），3) 创建和分散群体队形并模拟互动（例如，人看着笔记本电脑、人群）（图 1 (a & e)、(b & f)、(c & g)、(d & h)），4) 支持可通过离散或连续动作空间控制的全向和非完整约束机器人，5) 集成结合社会与功能因素的学习得到的奖励函数（SNGNN-v2 [19]）、其他研究中使用的简单基于启发式的奖励函数 [11]，，以及用于创建自定义奖励函数的应用程序接口，6) 提供多种额外评估指标，用于评测智能体的表现（尤其是指南论文 [32]中建议的指标），7) 将注视作为一种可添加到观测空间中的人类属性，并支持模拟有限视野（FoV）（图 1 (l)），

[1]https://github.com/gnns4hri/SocNavGy m

8) 允许限制智能体的视野和传感器范围（图 1 (k)），9) 支持人类采用 ORCA [33]和社会力模型 [34] 策略，并可选择考虑或忽略机器人，10) 可选且可配置的智能体感知输入噪声，11) 手动控制智能体以收集示教学习数据集。

### *A. Gym* 特性描述

SocNavGym 不依赖高保真物理引擎或任何软件框架，从而避免运行仿真时产生额外的开销和延迟（特性 1）。以所提出的仿真器为基础的 SARL [10] 和 DSRNN [11]等先前工作只能处理环境中的人类实体。在我们的 Gym 环境中，不同回合运行期间生成的社会场景可以具有不同的形状和大小，其中既包括植物、桌子、笔记本电脑等无生命实体，也包括不同规模的人群等有生命实体（特性 2）。在现实的社交场景中，人群和人-物交互可以在任意时刻形成和解散，而我们的 Gym 环境支持这种动态变化（特性 3）。这种交互形成与消散的动态特性对于学习有意义的社交情境是必要的，因为当交互发生过渡，即形成或消散时，干扰程度会发生变化。

关于机器人智能体的转向系统，先前工作仅支持全向智能体，其学习到的策略无法直接迁移到非全向智能体。因此，我们允许用户选择希望学习策略的智能体（特性 4）。

以往研究中使用的奖励函数可以说是为碰撞规避而非不适避免而设计的。为了学习真正考虑社会规范的策略，我们在奖励函数中加入了不适度估计器，根据环境情境（特征 5）对造成干扰的智能体施加惩罚。我们还展示了一个应用案例：使用所提出的奖励函数训练一个 DuelingDQN 智能体，并使用指南论文 [32], 提出的评估指标，对其行为进行基准测试，参见表 II。人类的注视可以为导航决策提供有用的信息。因此，对于每个被观测到的人，SocNavGym 都会提供一个标志，指明机器人智能体是否处于其视场内（特征 7）。该环境还允许用户模拟视场或探测范围有限的机器人所受的实体硬件约束（特征 8）。环境中的人类可由 ORCA 或 SFM 策略控制（特征 9），用户还可以选择在计算人类策略时是否将智能体纳入考量。

为收集离线数据集，SocNavGym 提供了一个示例脚本，允许用户在房间中手动操控智能体，以收集（状态、动作、下一状态、奖励）元组（功能 11）。

观测以字典形式呈现，每种实体类型对应一个键，相应的值是该类型实体组成的列表，具体包括“目标”“人类”“物体”“墙体”和“关系”。所有实体都具有一些基本属性，包括标识符、二维位置坐标、方向、包围球半径，以及线性和角速度等相对速度。环境返回的观测可以包含环境中所有实体的信息（完全可观测），也可以仅包含部分可观测实体的信息。用户可根据需要调整机器人的量程和视场宽度来配置此项（功能 8）。同样，SocNavGym 也可以配置为使观测相对于智能体或全局参考坐标系。此外，还可以设置分布参数，为特定实体的感知信息添加高斯噪声（功能 10）。这些观测变化可通过 Gym 包装器进行配置。如果将 SocNavGym 配置为提供相关信息，观测空间中的“关系”部分将包含关系信息（例如，人类交谈、某人注视物体以及人类注视，功能 7）。

### *B.* 评估指标

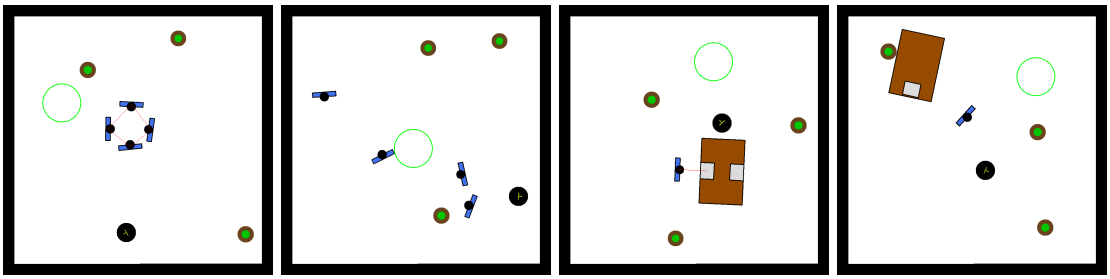
SocNavGym 通过每个回合的步骤所返回的 info 变量，提供事件以及 [32] 中建议的评估指标（见表 I）。这些指标旨在帮助更好地了解智能体在训练过程中的表现（特性 6）。除了这些事件和评估指标外，还可以通过 info 变量提供的数据获取对应于多个回合执行情况的全局指标。它还会返回一个邻接矩阵，用于识别仿真过程中每个时刻环境中的交互实体。可以通过辅助函数获取这些全局指标，从而评估社交智能体在多次运行中的表现。

### *C.* 人类轨迹生成

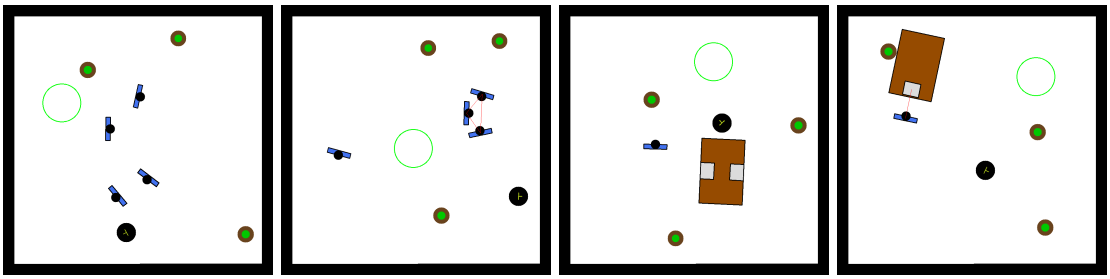
可以使用 ORCA 或社会力模型生成仿真的行人轨迹。为了让每个人类的行为各不相同，SocNavGym 可以配置为使用其中一种模型，也可以为每个生成的行人随机选择模型。用户还可以选择在计算仿真人类的动作时忽略机器人。生成时，会为动态行人和人群指定一个目标位置；到达目标后，如果回合尚未结束，则会再次采样一个目标位置。SocNavGym 还可以配置为创建和解散群组，以及实现人-人交互和人-物交互。对于人群疏散，每个人类或子群体都会获得一个不同的目标，必须前往该目标或留在原地（静止）。此外，用户可以更改行人的视野，使其策略仅考虑视野内的实体。

### *D.* 奖励函数

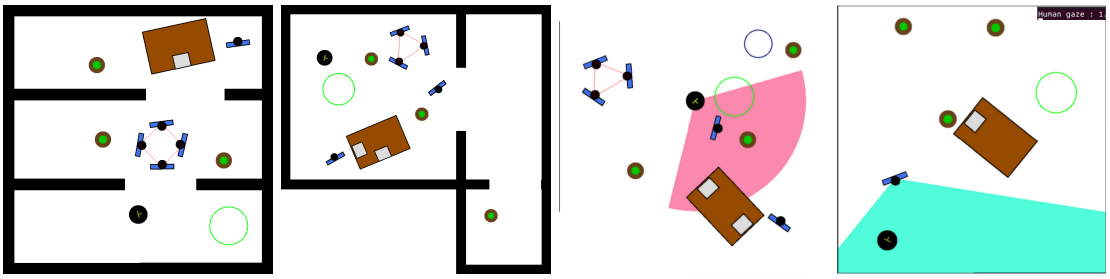
SocNavGym 可使用流行的基于深度强化学习的社会导航方法所采用的奖励函数（例如， [35], [10]），该分段函数：a）在机器人到达目标时给予正奖励；b）在机器人与



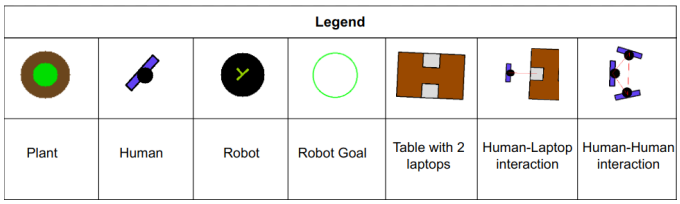
（a）人群疏散前　（b）人群聚集前形成　（c）人机交互疏散前　（d）人机交互交互形成



(e) 人群疏散后 (f)人群聚集后　(g) 人机交互交互疏散　(h) 人机交互交互形成



(i) 方形的走廊 area　(j) L形区域　(k) 机器人有限视场和传感器范围视野　(l) 人类注视




(m) 实体标签

图1：环境中人群动态编队与疏散以及人机交互场景的前期（a-d）和后期（e-h）。图 i 和图 j 分别为正方形区域和L形区域。图 k 展示机器人的传感器范围，图 l 展示人的注视方向。

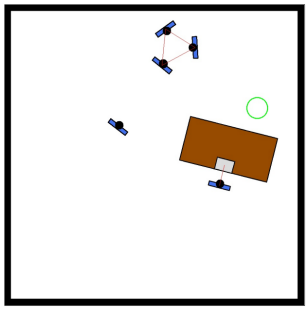
人距离过近时给予线性惩罚；c）奖励缩短与目标之间的距离。用数学符号表示：

$$r(s_t, a_t) = \begin{cases} -1, & \textbf{if } d_{min}^t \leq r_{agent} + r_{entity} \\ (d_{min}^t - \delta_{disc}) \cdot \alpha \cdot dt, & \textbf{if } 0 < d_{min}^t < \delta_{disc} \\ 1.0, & \textbf{if } d_{goal}^t \leq \rho_{agent} \\ (d_{goal}^{t-1} - d_{goal}^t) \cdot \beta, & \textbf{otherwise} \end{cases} \tag{1}$$

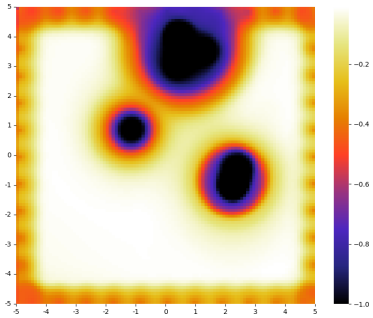
其中，$r(s_t, a_t)$ 是智能体在时间 $t$ 的状态 $s_t$ 下执行动作 $a_t$ 时获得的奖励，$d_{min}^t$ 是时间 $t$ 时机器人与最近的人类之间的最小可接受距离，$r_{agent}$ 和 $r_{entity}$ 分别是智能体和环境中任意实体的半径，$\delta_{disc}$（默认值为 0.6 米）是智能体与人之间的不适阈值距离，$\alpha$ 是不适缩放因子，$dt$ 是仿真时间步长，$\rho_{agent}$ 是目标半径，$d_{goal}^t$ 是时间 $t$ 时到目标位置的 L2 距离，而 $\beta$ 是一个

| 指标名称 | 描述 |
| --- | --- |
| 超出地图 | 如果智能体超出地图，则为 True |
| 与人类碰撞 & 与物体碰撞 & 与墙壁碰撞 | 如果智能体与人、物体或墙壁发生碰撞，则为真 |
| 碰撞 | 如果智能体与任何实体发生碰撞，则为真 |
| 成功 | 如果智能体已到达目标，则为真 |
| 超时 | 如果回合因达到最大回合长度而终止，则为 True |
| 未能取得进展 | 机器人未能缩短到目标的距离的时间步数 |
| 停滞时间 | 机器人速度为 0 的时间步数 |
| 到达目标所需时间 | 机器人到达其目标所需的时间步数 |
| STL | 按时间长度加权的成功率 |
| SPL | 按路径长度加权的成功率 |
| 路径长度 | 机器人走过的总路径长度 |
| 最小速度 & 最小加速度 & 最小加加速度 | 机器人达到的最小速度或加速度或加加速度 |
| 平均速度和平均加速度和平均加加速度 | 机器人的平均速度或加速度或加加速度 |
| 最大速度和最大加速度和最大加加速度 | 机器人达到的最大速度或加速度或加加速度 |
| 碰撞时间 | 轨迹中任意时刻与人发生碰撞的最短碰撞时间。 |
| 与人的最短距离 | 与任意人的最短距离。 |
| 个人空间遵从度 | 机器人不处于任意人的个人空间（0.45米）内的步数百分比。 |
| 与障碍物的最短距离 | 与任何物体的最小距离。 |
| 平均障碍物距离 | 与任何物体的平均距离。 |
| 不适感 DSRNN | 根据 DSRNN 奖励函数计算的不适感奖励 |
| 距离奖励 | 前后两个时间步中智能体到目标的距离之差 |
| SNGNN 奖励 | SNGNN 奖励为 $SNGNN\ value - 1$。（ $DISCOMFORT\ SNGNN - 1$） |
| 存活奖励 | 智能体获得的存活奖励（旨在用作惩罚） |
| 最近人类距离 | 与人的最近距离 |
| 最近障碍物距离 | 与障碍物的最近距离 |

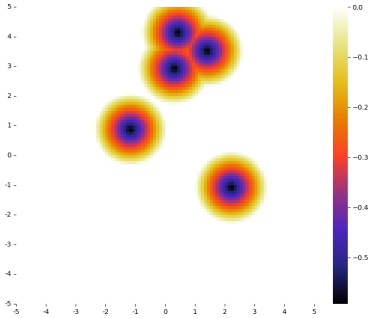
表 I：信息字典中的指标



（a）当前状态的环境快照

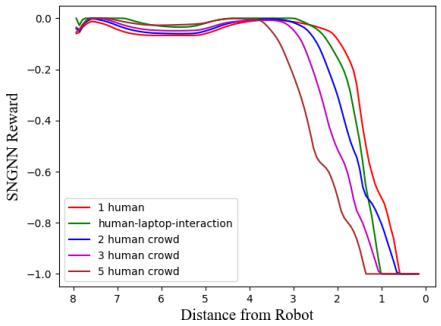


（b）当前状态下的环境的 SNGNNv2 不适度值环境的当前状态。

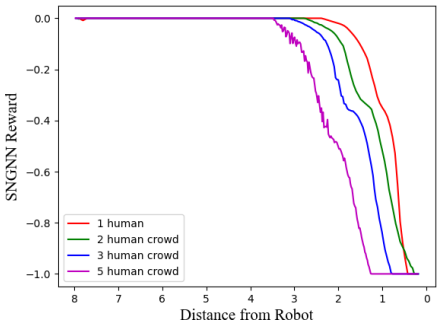


（c）当前状态下的环境的 DSRNN 不适度值环境的当前状态

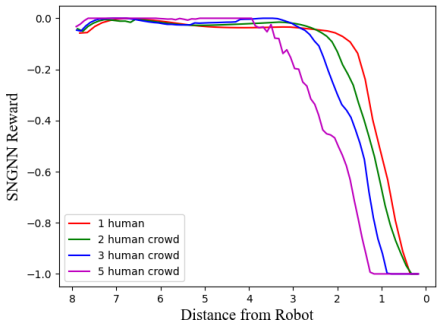
图 2：社交情境下 DSRNN 不适度和 SNGNNv2 不适度的热力图。尽管形状相似（呈圆形），但数据驱动奖励函数的分布明显呈非线性。为了便于可视化，我们将房间划分为 N×N 网格（N=100），并将机器人放置在每个网格单元中，以获取不适度值。我们仅针对人和交互生成这些值，以获得清晰的热力图。




（a）机器人向静止的行人。

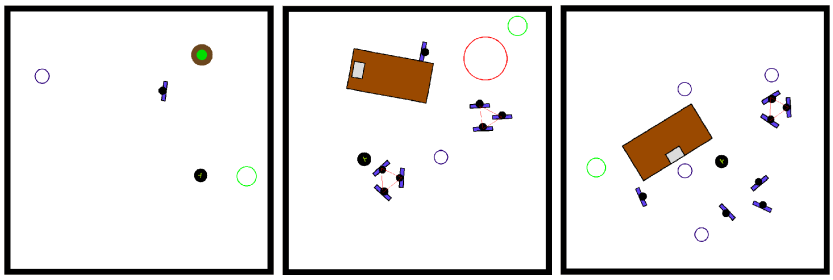



(b) 行人朝向一个静止的机器人移动。




(c) 行人和机器人都朝向彼此移动。

图3：当智能体 (a) 朝向静止实体移动、(b) 保持静止而实体朝向智能体移动，以及 (c) 实体彼此朝向移动时的 SNGNN-Discomfort。SNGNN-Discomfort 在评估场景时，同时考虑上下文以及不同关系的存在或缺失。



(a) 实验 1　　(b) 实验 2　　(c) 实验 3

图 4：实验中使用的仿真环境快照。

L2 距离的缩放因子。

奖励函数的第二部分是在房间内每个实体周围半径为 $\delta_{disc}$ 的范围内设置的势场；随着智能体与实体之间的距离缩短，智能体会受到惩罚。这部分是不适度惩罚，用作学习社会规范行为的反馈。

我们认为，该奖励函数无法准确反映人类不适，因为它没有考虑社交情境。例如，阈值$d^t_{min}$ 的选择是任意的，而且该函数没有考虑许多变量，例如速度、朝向或交互。另一个例子是，当智能体从人的后方接近和从人的前方接近时，会受到相似的惩罚；然而，正如研究所指出的 [36]，，这两种情况下造成的干扰有所不同。因此，我们提出使用一种基于整体社交情境、能够考虑用户意见的奖励函数。如前所述，鉴于使用手工设计的方程对复杂的社交偏好进行建模存在困难，我们提出了公式 2 中的奖-励函数，该函数以基于 SNGNN-v2 [19] 的函数替代上述不适度因素。利用从人类反馈中学习得到的不适度量，或许能够更好地处理社交情境：

$$r(s_t, a_t) = \begin{cases} -1, & \textbf{if } d^t_{min} \leq r_{agent} + r_{entity} \\ 1, & \textbf{if } d^t_{goal} \leq \rho_{agent} \\ (d^{t-1}_{goal} - d^t_{goal}) \cdot \beta + P(s_t), & \textbf{otherwise} \end{cases} \tag{2}$$

其中， $P(s_t)$ 是 SNGNN 不适度惩罚，定义如下：

$$P(s_t) = (SNGNN(s_t) - 1) \cdot \delta \tag{3}$$

$SNGNN\ (\cdot)$ 是 SNGNN 模型实现的函数， $\delta$ 是一个缩放因子。SNGNN 分数范围为 1（无干扰）到 0（极端干扰）。因此，为了将其用作惩罚项，需要从其输出中减去 1。

SNGNN-v2 [19] 经过训练，能够考虑社会现象。图 3所示的一个例子是：对于给定的机器人与人的距离，它如何根据不同情境下估计的不适度差异作出处理。同样，图2展示了两个奖励函数中使用的不适度值热力图。无论人的方向如何，也无论是否存在人群或交互，DSRNN 的不适度值显然都大致相同。这表明尽管 DSRNN 的奖励函数有助于研究无碰撞导航，但可以说，它并非解决社会导航问题的最佳选择。

## IV. 使用案例

我们将 Dueling-DQN 智能体与环境集成，用于比较基于 SNGNN 的奖励函数与文献中广泛使用的奖励函数的表现，后者称为 DSRNN 奖励函数。该智能体的神经网络由 5 个线性层组成，隐藏单元数分别为 512、256、128 和 64，并使用 LeakyReLU 作为激活函数。我们采用 Adam 优化器，固定学习率为 0.001，并使用上述两个奖励函数分别训练智能体。训练持续了 50K 个回合，每个回合的最大长度为 200 个离散时间步。

### *A.* 社会场景

智能体的观测空间包括参考坐标系中其他实体的位置、方向和速度。在每次实验中，所有实体的位置和方向均进行随机初始化，并从高斯分布中采样人使用 ORCA 或 SFM 运动时的超参数，以产生多样化的行为。

我们在三个难度逐渐增加的社会场景中训练双重深度Q网络智能体。所有实验均使用非完整型智能体进行。训练 DRL 智能体时使用了以下环境设置：

- 实验 *1*：在本实验中，我们考虑一个 10 米 × 10 米的方形房间，其中有 1 株植物和 1 名动态行人（图 4.(a)）。
- 实验 *2*：一个 10 米 × 10 米的方形房间，里面有 1 张桌子，桌上随机放置 1 台笔记本电脑，还有 3 名静态人群成员、3 名动态人群成员，以及 1 名能够在回合过程中与桌上的笔记本电脑动态建立或中断现有交互的人（图 4.(b)）。
- 实验 *3*：一个 10 米 × 10 米的方形房间，内有 1 张桌子、随机放置在桌子上的 1 台笔记本电脑、由 3 个人组成的静态人群、由 3 个人组成的动态人群，以及一个可以与笔记本电脑进行交互的动态行人。与笔记本电脑交互的人以及动态

| | 实验1 DSRNN | 实验1 SNGNN | 实验2 DSRNN | 实验2 SNGNN | 实验3 DSRNN | 实验3 SNGNN |
|---|---|---|---|---|---|---|
| **不适感 SNGNN** | -5.495 | **-2.377** | -17.786 | **-4.639** | -22.376 | **-5.429** |
| **不适感 DSRNN** | -0.2 | **-0.143** | -0.481 | **-0.403** | -1 | **-0.341** |
| **个人空间遵从度** | 0.979 | **0.996** | **0.997** | 0.973 | **0.998** | 0.989 |
| **最近人距离** | **2.556** | 2.528 | 1.456 | **1.92** | 1.376 | **1.736** |
| **人与人碰撞率** | 0.02 | 0.02 | **0** | 0.06 | 0.01 | 0.01 |
| **墙壁碰撞率** | 0.02 | **0.01** | **0** | 0.01 | 0.04 | **0** |
| **物体碰撞率** | 0.07 | **0.04** | **0.02** | 0.06 | 0.12 | **0** |
| **碰撞率** | 0.09 | **0.06** | **0.02** | 0.12 | 0.13 | **0.01** |
| **最短碰撞时间** | 186.511 | **191.195** | 172.24 | **177.492** | 174.56 | **184.601** |
| **最近障碍物距离** | 1.039 | **1.298** | 0.806 | **0.845** | 0.751 | **0.884** |
| **平均障碍物距离** | 4.651 | **4.686** | **4.449** | 4.394 | **4.458** | 4.426 |
| **超时** | **0.1** | 0.26 | 0.5 | **0.08** | 0.53 | **0.04** |
| **停滞时间** | **18.61** | 23.28 | 32.25 | **13.72** | 81.19 | **0.37** |
| **到达目标所需时间** | **90.55** | 97.86 | 139.15 | **79.77** | 162.47 | **67.14** |
| **未能取得进展** | **5.42** | 12.27 | 29.06 | **2.05** | **6.55** | 7.84 |
| **路径长度** | **5.558** | 6.451 | 10.44 | **4.575** | **6.288** | 6.483 |
| **STL** | **0.644** | 0.607 | 0.388 | **0.746** | 0.204 | **0.888** |
| **SPL** | **0.678** | 0.624 | 0.393 | **0.75** | 0.27 | **0.871** |
| **最小速度** | 0.03 | 0.042 | 0.03 | 0.053 | 0.006 | 0.082 |
| **平均速度** | 0.08 | 0.084 | 0.082 | 0.09 | 0.052 | 0.098 |
| **最大速度** | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| **最小加速度** | 0.002 | 0.0 | 0 | 0.002 | 0 | 0 |
| **平均加速度** | 0.023 | 0.031 | 0.064 | 0.011 | 0.018 | 0.012 |
| **最大加速度** | 0.139 | 0.139 | 0.17 | 0.115 | 0.126 | 0.135 |
| **最小加加速度** | 0.002 | 0.001 | 0 | 0.004 | 0.001 | 0 |
| **平均加加速度** | 0.041 | 0.062 | 0.126 | 0.019 | 0.033 | 0.023 |
| **最大加加速度** | 0.214 | 0.218 | 0.303 | 0.141 | 0.224 | 0.183 |

表 II：使用两种奖励函数训练时 DuelingDQN 的表现。以上所有指标均为 100 个测试回合的平均值。

人群可以在回合中形成并解除其物体交互或人群交互（图 4.(c)）

通过修改配置文件中的变量，可以生成更复杂的场景。

### *B.* 结果

Dueling-DQN 智能体在上述社会场景中使用公式 1 和公式 2 的两种奖励函数进行训练。表 II 展示了训练后的智能体在不同评估指标下的表现。具体而言，研究考虑了两类不同的指标：社会规范遵从性指标和导航指标。社会规范遵从性涵盖 SNGNN-Discomfort（公式 2）、DSRNN 不适度（公式 1）、个人空间遵从度（PSC）、与人的最近距离以及人与人碰撞率等指标。其余指标构成导航指标。最小、平均和最大速度、加速度及加加速度为智能体的运动基元提供了额外信息。纳入第二组指标是为了帮助用户了解智能体在穿越目标位置时的表现。如表 II 所示，根据社会指标，与使用 DSRNN奖励的智能体相比，使用 SNGNN-v2奖励函数训练的智能体实现了更高的社会规范遵从性。此外，在实验 2 和实验 3 中，社会情境更为复杂，SNGNN 智能体在导航指标方面的表现也显著优于 DSRNN 智能体。这表明，借助这种新的不适度惩罚，智能体能够有效学习如何将社交情境与碰撞规避结合起来，从而为使用我们的工具创建新的人类感知机器人导航算法带来额外益处使用我们的工具创建新的人类感知机器人导航算法。

## V. 结论与未来工作

在本工作中，我们为人机交互社区提供了一个易于使用的社会导航环境，可用于训练社会导航智能体。该环境支持用户指定不同的环境配置，这些配置的复杂程度各不相同，并配有一个社会奖励函数。实验表明，考虑社会奖励能够提升社会性表现，同时对按时间长度加权的成功率（STL）几乎没有影响，或完全没有影响。尽管目前的工作已取得令人满意的成果，但仍存在一些局限性，我们正致力于在近期解决这些问题。这些问题包括：针对全向系统对 SNGNN-v2 进行显式再训练，以及为人类采用更复杂的导航策略。


## 致谢

大多数实验在由EPSRC核心设备基金资助的Aston EPS机器学习服务器上运行，资助编号为EP/V036106/1。


## 参考文献


[1] C. Mavrogiannis、F. Baldini、A. Wang、D. Zhao、P. Trautman、A. Steinfeld 和 J. Oh，“社会机器人导航的核心挑战：一项综述”，*ACM*人机交互汇刊，第 abs/2103.05668 卷，第 3583741 页，2023 年 2 月。 [在线]。可访问：https://dl.acm.org/doi/ 10 1145 3583741 。/ S.

[2] Liu、P. Chang、Z. Huang、N. Chakraborty、K. Hong、W. Liang、D. L. McPherson、J. Geng 和 K. Driggs-Campbell，“基于注意力的交互图实现意图感知机器人群体导航”，2022 年。

[3] C. Chen、S. Hu、P. Nikdel、G. Mori 和 M. Savva，“用于人群导航的关系图学习”，载于 IEEE/RSJ智能机器人与系统国际会议（IROS）。美国内华达州拉斯维加斯：IEEE，2020 年 10 月，第 10 007–10 013 页。[4] A. Alahi、K. Goel、V. Ramanathan、A. Robicquet、L. Fei-Fei 和 S. Savarese，“Social LSTM：拥挤空间中的人类轨迹预测”，载于 IEEE计算机视觉与模式识别会议（CVPR）。美国内华达州拉斯维加斯：IEEE，2016 年 6 月，第 961–971 页。[5] A. Gupta、J. Johnson、L. Fei-Fei、S. Savarese 和 A. Alahi，“Social GAN：利用生成对抗网络生成符合社会规范的轨迹”，载于 IEEE/CVF计算机视觉与模式识别会议。美国犹他州盐湖城：IEEE，2018 年 6 月，第 2255–2264 页。[6] Y. Huang、H. Bi、Z. Li、T. Mao 和 Z. Wang，“STGAT：用于人类轨迹预测的时空交互建模”，载于 IEEE/CVF计算机视觉国际会议（ICCV）。韩国首尔：IEEE，2019 年 10 月，第 6271–6280 页。[7] S. Eiffert、H. Kong、N. Pirmarzdashti 和 S. Sukkarieh，“使用生成式循环神经网络和蒙特卡洛树搜索进行动态环境中的路径规划”，CoRR，第 abs/2001.11597 卷，2020 年。 [在线]。可访问：https://arxiv.org/abs/ 2001 11597 。M. L. Puterman，《马尔可夫[8] 决策过程：离散随机动态规划》，Wiley 概率与统计系列。美国新泽西州霍博肯：Wiley-Interscience，2005 年。[9] P. Bachiller、D. Rodriguez-Criado、R. R. Jorvekar、P. Bustos、D. R. Faria 和 L. J. Manso，“用于建模人类感知机器人导航中扰动的图神经网络”，Multimedia Tools and Applications，第 81 卷第 3 期，第 3277–3295 页，2022 年。[10] C. Chen、Y. Liu、S. Kreiss 和 A. Alahi，“人群-机器人交互：基于注意力的深度强化学习实现人群感知机器人导航”，载于机器人与自动化国际会议（ICRA）。加拿大魁北克省蒙特利尔：IEEE，2019 年 5 月，第 6015–6022 页。 [在线]。可访问：https://ieeexplore. ieee.org/document/8794134/ [11] Y. F. Chen、M. Everett、M. Liu 和 J. P. How，“利用深度强化学习进行社会感知运动规划”，载于 IEEE/RSJ智能机器人与系统国际会议（IROS）。加拿大不列颠哥伦比亚省温哥华：IEEE，2017 年 9 月，第 1343–1350 页。 [在线]。可访问：http://ieeexplore.ieee.org/document/8202312/ [12] Z. Zhou、P. Zhu、Z. Zeng、J. Xiao、H. Lu 和 Z. Zhou，“融合深度强化学习与在线规划的人群中的机器人导航”，Applied Intelligence，第 52 卷第 13 期，第 15 600–15 616 页，2022 年 10 月。 [在线]。可访问：https: //link.springer.com/ 10 1007 10489 022 03191 2。/s - - - L. Qin、Z.[13] Huang、C. Zhang、H. Guo、M. Ang 和 D. Rus，“动态行人环境中的深度模仿学习自主导航”，载于 IEEE机器人与自动化国际会议（ICRA）。中国西安：IEEE，2021 年 5 月，第 4108–4115 页。[在线]。可访问：https://ieeexplore.ieee.org/document/9561220/ [14] J. P. de Vicente 和 A. Soto，“Deepsocnav：通过模仿人类行为实现社会导航”，CoRR，第 abs/2107.09170 卷，2021 年。 [ ] 2107 09170 在线。可访问：[15] https://arxiv.org/abs/ 。K. Narayanan、L. Posada、F. Hoffmann 和 T. Bertram，“从示范中获取基于视觉的移动机器人导航行为动力学”，IFAC Proceedings Volumes，第 46 卷第 5 期，第 37–44 页，2013 年。 [在线]。可访问：https://linkinghub.elsevier.com/ retrieve/pii/S 1474667015361942 [16] H. Kretzschmar、M. Spies、C. Sprunk 和 W. Burgard，“通过逆强化学习实现符合社会规范的移动机器人导航”，The International Journal of Robotics Research，第 35 卷第 11 期，第 1289–1307 页，2016 年。 [在线]。可访问：https://doi.org/10.1177/ 0278364915619772[17]M. Fahad、Z. Chen 和 Y. Guo，“学习行人如何导航：一种深度逆强化学习方法”，载于 IEEE/RSJ智能机器人与系统国际会议（IROS）。西班牙马德里：IEEE，2018 年 10 月，第 819–826 页。 [在线]。可访问：https://ieeexplore.ieee.org/document/8593438/ [18] O. Saha 和 P. Dasgupta，“使用逆强化学习改进高效机器人导航的奖励估计”，载于 NASA/ESA自适应硬件与系统会议（AHS）。美国加利福尼亚州帕萨迪纳：IEEE，2017 年 7 月，第 245–252 页。 [在线]。可访问：http://ieeexplore.ieee.org/document/8046385/ [19] D. Rodriguez-Criado、P. Bachiller 和 L. J. Manso，“使用图神经网络生成人类感知导航地图”，载于 Artificial Intelligence XXXVIII，M. Bramer 和 R. Ellis 编。瑞士沙姆：Springer International Publishing，2021 年，第 19–32 页。

[20] I. Mordatch 和 P. Abbeel，《多智能体群体中具身组合语言的涌现》，AAAI人工智能会议论文集，第 32 卷，第 1 期，2018 年 4 月。[在线]。可访问：https://ojs.aaai.org/index.php/AAAI/article/view/ 11492[21] L. J. Manso、P. Nu˜nez、L. V. Calderita、D. R. Faria 和 P. Bachiller，《SocNav1：用于评测和学习社交导航规范的数据集》，《数据》，第 5 卷，第 1 期，第 7 页，2020 年 1 月。 [在线]。可访问：https://www.mdpi.com/ 2306 5729 5 1 7 - / / / R. Baghel、A. Kapoor、P.[22] Bachiller、R. R. Jorvekar、D. Rodriguez- Criado 和 L. J. Manso，《用于生成社交导航数据集的工具包》，CoRR，第 abs/2009.05345 卷，2020 年。 [在线]。可访问：https://arxiv.org/abs/ 2009 05345 . Z.[23] Wang、N. de Freitas 和 M. Lanctot，《用于深度强化学习的对决网络架构》，CoRR，第 abs/1511.06581 卷，2015 年。 [在线]。可访问：http://arxiv.org/abs/1511. 06581[24] A. Dosovitskiy、G. Ros、F. Codevilla、A. M. L´opez 和 V. Koltun，《CARLA：一款开放式城市驾驶模拟器》，CoRR，第 abs/1711.03938 卷，2017 年。 [在线]。可访问：http://arxiv.org/abs/1711. 03938[25] O. Saha 和 P. Dasgupta，《利用逆强化学习改进奖励估计，以实现高效机器人导航》，载于 NASA/ESA自适应硬件与系统会议（AHS）。美国加利福尼亚州帕萨迪纳：IEEE，2017 年 7 月，第 245–252 页。 [在线]。可访问：http://ieeexplore.ieee.org/document/8046385/ [26] M. Campanella、S. Hoogendoorn 和 W. Daamen，《nomad模型：理论、发展与应用》，《交通运输研究进展》，第 2 卷，第 462–467 页，2014 年，2014 年行人和疏散动力学会议。[27] S. Curtis、A. Best 和 D. Manocha，《Menge：用于模拟人群运动的模块化框架》，《集体动力学》，第 1 卷，2016 年 3 月。[28] N. Tsoi、A. Xiang、P. Yu、S. S. Sohn、G. Schwartz、S. Ramesh、M. Hussein、A. W. Gupta、M. Kapadia 和 M. V´azquez，《Sean 2.0：将机器人导航的社交情境形式化并生成》，《IEEE机器人与自动化快报》，第 7 卷，第 4 期，第 11 047–11 054 页，2022 年。[29] J. Holtz 和 J. Biswas，《SocialGym：社交机器人导航基准测试框架》，载于 IEEE/RSJ智能机器人与系统国际会议（IROS）。日本京都：IEEE，2022 年 10 月，第 11 246–11 252 页。 [在线]。可访问：https://ieeexplore.ieee. org/document/9982021/ [30] J. K. Haas，《Unity游戏引擎的发展史》，伍斯特理工学院，技术报告，2014 年。[31] M. Quigley、K. Conley、B. Gerkey、J. Faust、T. Foote、J. Leibs、R. Wheeler 和 A. Ng，《ROS：一种开源机器人操作系统》，载于 IEEE机器人与自动化国际会议，第 3 卷，2009 年 1 月。[32] A. Francis、C. Perez-D’Arpino、C. Li、F. Xia、A. Alahi、R. Alami、A. Bera、A. Biswas、J. Biswas、R. Chandra、H.-T. L. Chiang、M. Ev- erett、S. Ha11、J. Hart、J. P. How、H. Karnan、T.-W. E. Lee、L. J. Manso、R. Mirksy、S. Pirk、P. T. Singamaneni、P. Stone、A. V. Taylor、P. Traut- man、N. Tsoi、M. Vazquez、X. Xiao、P. Xu、N. Yokoyama、A. Toshev 和 R. Mart´ın-Mart´ın，《社交机器人导航算法评估原则与指南》，arXiv 预印本 arXiv:2306.16740，2023 年。[33] J. van den Berg、M. Lin 和 D. Manocha，《用于实时多智能体导航的互惠速度障碍》，载于 IEEE机器人与自动化国际会议。美国加利福尼亚州帕萨迪纳：IEEE，2008 年，第 1928–1935 页。[34] D. Helbing 和 P. Moln´ar，《行人动力学的社会力模型》，《物理评论E》，第 51 卷，第 5 期，第 4282–4286 页，200 1995 [ ] 10 1103 51 4282 年 5 月。[35] 在线。可访问：https://doi.org/ . physreve. . S. Liu、P. Chang、W. Liang、N. Chakraborty 和 K. Driggs-Campbell，《用于机器人人群导航的分散式结构化RNN与深度强化学习》，载于 IEEE机器人与自动化国际会议（ICRA），2021 年，第 3517–3524 页。[36] E. T. Hall、R. L. Birdwhistell、B. Bock、P. Bohannan、A. R. Diebold Jr、M. Durbin、M. S. Edmonson、J. Fischer、D. Hymes、S. T. Kimball 等，《人际距离学 [及评论与回复]》，《当代人类学》，第 9 卷，第 2/3 期，第 83–108 页，1968 年。