arXiv:2302.05528v1 [cs.AI] 10 Feb 2023

对抗性多智能体系统中的协作行为学习*

Ni Wang[1][0000−0002−5867−0176]、Gautham P. Das[2][0000−0001−5351−9533]和AlanG. Millard[3][0000−0002−4424−5953]

[1] 巴黎-萨克雷大学，法国 niwang.fr@gmail.com[2] 林肯大学林肯农业机器人研究中心，英国 gdas@lincoln.ac.uk[3] 约克大学计算机科学系，英国 alan.millard@york.ac.uk

摘要。 本研究扩展了一个名为RoboSumo的现有虚拟多智能体平台，创建了TripleSumo——一个用于研究连续动作空间中、具有物理接触的对抗性环境下多智能体协作行为的平台。本文研究了这样一种情境：两个智能体，即“Bug”和“Ant”，必须组队将另一个智能体“Spider”推出竞技场。为实现这一目标，新加入的智能体“Bug”会在“Ant”和“Spider”之间正在进行的比赛中接受训练。“Bug”必须形成对其他智能体动作的感知能力，推断双方的策略，并最终学习一种动作策略以实现协作。研究采用深度确定性策略梯度（DDPG）强化学习算法，并使用结合密集奖励与稀疏奖励的混合奖励结构。通过平均获胜概率和获胜所需的平均步数，对协作行为进行定量评估。

关键词： 多智能体协作 · 强化学习。

1 引言

随着 AlphaGo 的成功，游戏中的人工智能日益受到世界各地研究人员的关注 [1]。在物理世界中，智能体会在多个时空尺度上展现智能行为 [2]。我们设计了 [3]，多智能体平台 RoboSumo[4] ，通过元学习研究智能体在非平稳、竞争性多智能体环境中的持续适应潜力。在该平台上，两个同质或异质的智能体相互对抗，学习玩一场“相扑”游戏。成功将对手推出方形竞技场的智能体赢得比赛。

*本研究得益于阿卜杜拉国王科技大学（KAUST）助理教授 Shinkyu Park 的建议。[4] https://github.com/openai/robosumo

RoboSumo 支持在对抗场景中研究多智能体物理交互，但该平台不支持探索智能体之间的合作行为。本文对 RoboSumo 进行了扩展，加入一个新的智能体，使其能够与现有智能体之一组队。这个新智能体必须学会一项策略，与预先指定的队友合作，并与对手对战。我们使用深度确定性策略梯度（DDPG）训练该系统。该强化学习算法利用离策略数据和贝尔曼方程学习 Q函数，同时利用 Q函数学习策略 [4]。我们通过定性观察智能体在模拟中的行为，以及“平均胜率”和“获胜所需步数”这两个定量参数，对训练结果进行评估。用于训练、测试和评估的代码已公开供公众访问[5]。本研究的主要贡献包括：建立一个虚拟平台，以便在具有丰富物理接触的场景中探索合作与竞争交互；并报告使用 DDPG 训练该系统后，在两项评估指标上的基线结果。

本文下一节回顾基于虚拟平台的多智能体游戏相关工作；第 3 节介绍我们对 RoboSumo 的扩展，并建立 TripleSumo；第 4 节详述我们使用 DDPG 算法训练智能体的方法，并评估训练结果。最后一节总结研究发现，并概述未来工作的计划。

2 相关工作

游戏提供了具有挑战性的环境，可让研究人员以安全且可复现的方式快速测试强化学习（RL）的新算法和新理念 [5]。因此，近年来涌现出一系列新颖的虚拟游戏平台，推动了强化学习研究的发展。有些研究聚焦于由单个智能体完成任务，而许多研究则基于对抗模型，让智能体彼此竞争，以改善系统的整体表现。

2020 年，‘Google Research Football’[5] 被设计为一个开源平台，智能体在基于物理模拟的三维场景中接受训练，以进行足球比赛。该平台提供了三个难度各异的场景。作者实现了三种强化学习算法，即重要性加权演员-学习者架构（IM-PALA）、近端策略优化算法（PPO）和 Ape-X DQN，并报告了基准结果。该热门平台被证明有助于开发人工智能方法，例如 ‘TiKick’[6]。然而，‘Google Research Football’假设多智能体设置中的动作同步执行，这限制了其用途。为此，研究人员开发了‘Fever Basketball’ [7] ，这是一个用于多智能体强化学习的异步体育游戏环境。

同样基于虚拟足球游戏， [2] 通过开发一种结合模仿学习、单智能体与多智能体强化学习以及基于种群的训练的方法，研究了物理具身多智能体系统中的多尺度综合决策，

[5] https://github.com/niart/triplesumo

并利用了行为的可迁移表征。该研究采用多种分析技术评估智能体行为，包括现实世界体育分析中的统计数据。

[8] 引入了一个“捉迷藏”游戏，用于研究智能体学习工具使用的过程。通过在这一环境中进行训练，智能体构建出一系列六种不同的策略和反制策略。这项工作展现了多智能体协同适应的光明前景，这种协同适应可能产生复杂且智能的行为。

关于基于强化学习的多智能体协作方法， [9] 解决了Q学习在非平稳环境中的局限性，即随着智能体数量增加，策略梯度会产生方差。他们提出了一种行动者-评论家方法的改进方案，该方案考虑其他智能体的行动策略，并为每个智能体采用策略集成的训练机制，从而实现更加稳健的多智能体策略。

CollaQ [10] 将每个智能体的Q函数分解为“自身项”和“交互项”，并采用多智能体奖励归因（MARA）损失对训练进行正则化。该方法在“星际争霸多智能体挑战”中得到验证，并证明其性能优于现有的最先进技术。

在多智能体追逃问题中， [11]使用共享经验，通过课程学习为给定数量的追捕者训练策略，并在运行时由每个智能体独立执行该策略。他们设计了一种结合个体与群体奖励的奖励结构，以鼓励形成良好队形。

同样，在追逃问题中， [12] 提出了一种新的几何方法，用于学习2名追捕者对抗单个逃逸者场景下的合作行为，以缩短逃逸者的捕获时间。事实证明，该方法借助对冗余追捕者的分类与移除，具备良好的可扩展性。

这些虚拟游戏为对抗性环境中的多智能体团队合作提供了有用的实验平台和学习方法。然而，连续域中交互智能体之间的物理接触却很少得到研究。我们的工作通过开发一个平台，促进对物理接触丰富环境中的多智能体合作研究，满足了这一需求。本文介绍了扩展RoboSumo平台的TripleSumo，并展示了在对抗性多智能体场景中学习合作行为的强化学习方法。

3 TripleSumo

基于采用软件工具包 OpenAI/Gym[13] 和物理引擎 MuJoCo[14] 构建的 RoboSumo 框架，TripleSumo 为系统增加了一个智能体（见图1）。在这一场景中，‘Ant’（红色）与‘Bug’（蓝色）组成团队，在方形竞技场（“榻榻米”）上对抗对手‘Spider’（绿色）（见图2）。智能体的行为在指定的连续动作空间中进行训练和观察。为简化接口，这项初步工作将所有智能体设为四足且大小相同。不过，这三个智能体的接触力各不相同（见表3）。智能体的形态与物理特征

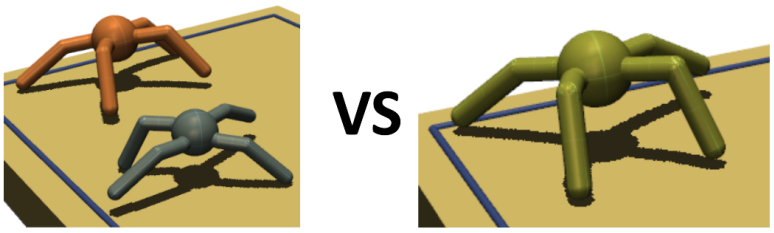

图 1. 智能体“Spider”（绿色，右侧）与“Bug”（蓝色，中间）和“Ant”（红色，左侧）组成的一对之间的竞争关系

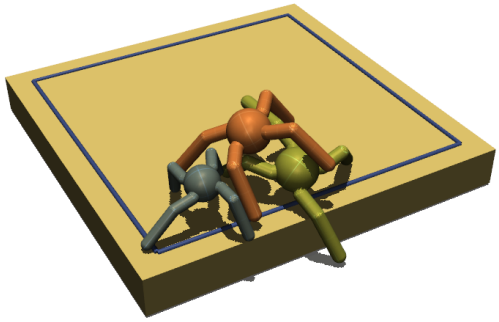

图 2. 模拟虚拟平台“TripleSumo”：两名智能体（红色和蓝色）在榻榻米上与绿色智能体对抗。

可根据未来研究需求自由选择。游戏开始后，三个智能体通过物理接触进行互动，比赛持续到其中任一智能体的质心越出竞技场边缘为止。‘Ant’和‘Spider’这两个智能体已通过 DDPG 预先训练，以开展持续进行的游戏（见补充视频1）。如果‘Spider’成功将‘Ant’或‘Bug’推出竞技场，它就会赢得游戏。反之，如果‘Ant’和‘Bug’组成的团队成功将‘Spider’推出竞技场，则该团队将赢得游戏。

4 实验结果

本节采用常用的强化学习算法 DDPG 来训练新的智能体。为了推断其对手和搭

接触力的控制范围（单位：牛顿）	
“Spider”（绿色）	[-0.22, 0.22]
‘Ant’（红色）	[-0.20, 0.20]
‘Bug’（蓝色）	[-0.18, 0.18]

表 1. 异构智能体接触力的控制范围

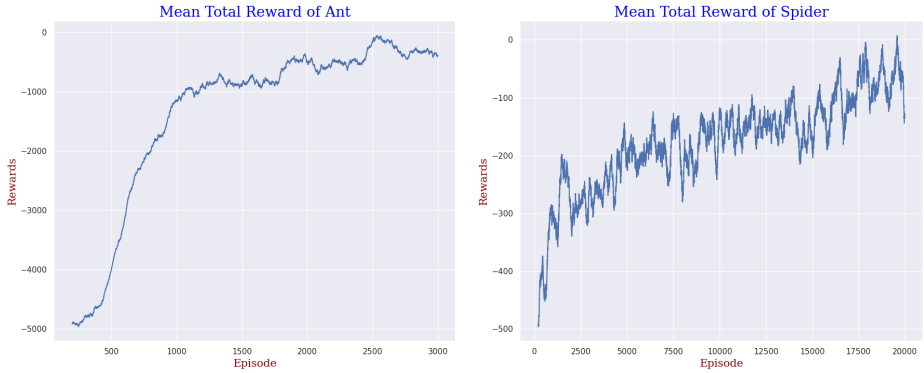

图3. ‘Ant’在 3,000 个训练轮次中获得的平均总奖励（左）；‘Spider’在 20,000 个训练轮次中获得的平均总奖励（右）；对每个图表应用 200 的滚动均值滤波器。

档的策略，智能体“Bug”在“Ant”和“Spider”对战的过程中接受训练，以学习动作策略[6]，。首先使用 DDPG 对“Ant”进行 3,000 个训练轮次的训练，随后对“Bug”进行 20,000 个训练轮次的训练，从而生成这场对战（图 3）。奖励结构与公式 (1)、(2) 类似，但方向相反。

4.1 奖励塑形

本实验所用的奖励函数由两部分组成：

$$Reward \;=\; Dense\ Reward \;+\; Sparse\ Reward \tag{1}$$

在“稠密”部分，智能体每一步移动都对应一个数值，该数值会累加到当前训练轮次的总奖励中。这种“稠密奖励”由四项组成——“对手速度奖励”、“伙伴速度奖励”、“自身速度奖励”以及一个固定惩罚项（公式 2）；其中，C_1、C_2 和 C_3 是分别插入稠密奖励函数各项中的常数系数。

$$\begin{aligned} Dense\ reward = {} & C_1 \;*\; opponent\ velocity\ reward \\ & + C_2 \;*\; self\ velocity\ reward \\ & + C_3 \;*\; partner\ velocity\ reward \\ & + still\ punishment \end{aligned} \tag{2}$$

每一步中，智能体都会收到“对手速度奖励”，该奖励分解为 X 分量和 Y 分量。X 分量的绝对值与对手沿 X 方向的速度成正比；如果对手正背离智能体移动，则该分量为正，

[6] 补充视频 1：https://www.youtube.com/watch?v=nxzi7Pha2GU

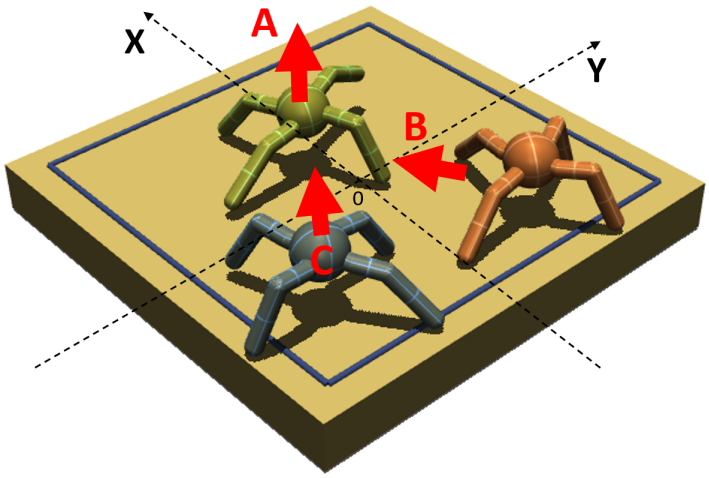

图4. 分配正向稠密奖励的方向：A. 对手速度奖励；B. 伙伴速度奖励；C. 自身速度奖励。

如果对手正朝智能体移动，则该分量为负。“对手速度奖励”的 Y 分量在 Y 方向上的定义方式类似。“自身速度奖励”和“伙伴速度奖励”的绝对值分别与智能体和其伙伴的速度成正比；如果智能体或其伙伴正朝对手移动，则奖励为正；如果它们远离对手移动，则奖励为负。也就是说，智能体或其伙伴朝对手移动，或者对手背离智能体移动时，智能体会获得奖励；反之则会受到惩罚。“静止惩罚”是一个负常数，以确保智能体在静止不动时会受到惩罚。图 4 展示了向智能体分配正向稠密奖励的方向。另一方面，奖励函数中的“稀疏”部分与游戏结果相关。如果团队赢得游戏，则分配 +500 分；如果输掉游戏，则分配 -400 分（见算法1）。如果训练轮次超过最大步数且没有一方获胜，则默认稀疏奖励为 0。

算法1 奖励函数的稀疏部分

```
当 训练轮次正在进行时执行
    如果 对手从榻榻米上掉下则
        奖励 += 500
    否则如果“Bug”或“Ant”从榻榻米上掉落则
        奖励 -= 400
    结束 if
结束 while
```

超参数	值
优化器	Adam
演员学习率	1e-4
评论家学习率	1e-3
折扣因子 (γ)	0.99
目标更新因子 (τ)	0.01
每训练轮次的步数	500
批量大小	64
最大训练轮数	20,000
经验回放缓冲区大小	1,000,000
更新步长	30

表 2. DDPG 实现中的超参数

每轮结果	局部胜率
团队获胜	+1
团队失败	-1
时间耗尽且无结果	0

表3. 特定训练轮次下局部胜率的定义

重要的超参数总结见表4.1。在为期20,000个训练轮次的训练过程中，如图5所示，智能体的密集奖励和稀疏奖励平均值均不断提高。

4.2 性能评估

通过此次训练得到的“Bug”动作策略，在包含 20,000 个训练轮次的测试中展现出智能体的协同行为，以及稠密奖励和稀疏奖励的稳定值（见图6）。此外，为了评估智能体协作的效率，我们使用了两个指标：MWR 和“获胜所需步数”，来分析训练结果。

指标“胜率”（WR）定义为在训练或测试的某个阶段赢得游戏的概率。在训练期间，某个训练轮次的局部胜率若团队获胜则记为“+1”；若团队失败则记为“-1”；若该轮次达到最大步数仍未分出胜负，则记为“0”。局部胜率的定义见表3。“平均胜率”是指训练或测试达到某个训练轮次时，局部胜率的滚动平均值。直观而言，正的胜率表示团队强于对手，负的胜率表示团队弱于对手，而零胜率表示双方的能力水平相近。采用滚动平均值

[7] 补充视频 2：https://www.youtube.com/watch?v=C4xGuIyeY5A

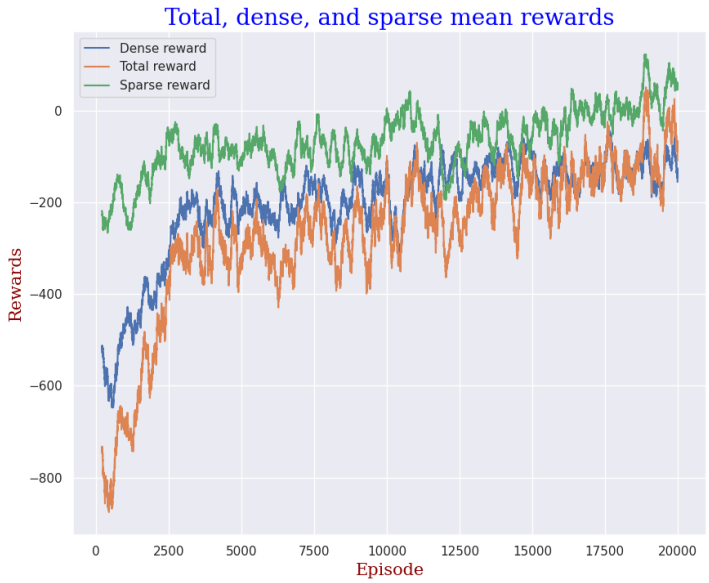

图5. 训练期间，‘Bug’在每个训练轮次中获得的总奖励（橙色）、稠密奖励（蓝色）和稀疏奖励（绿色）的平均值。应用了窗口大小为200的滚动均值滤波器。由于奖励函数中包含恒定的‘静止惩罚’项，最终的稠密奖励和总奖励仍低于零。

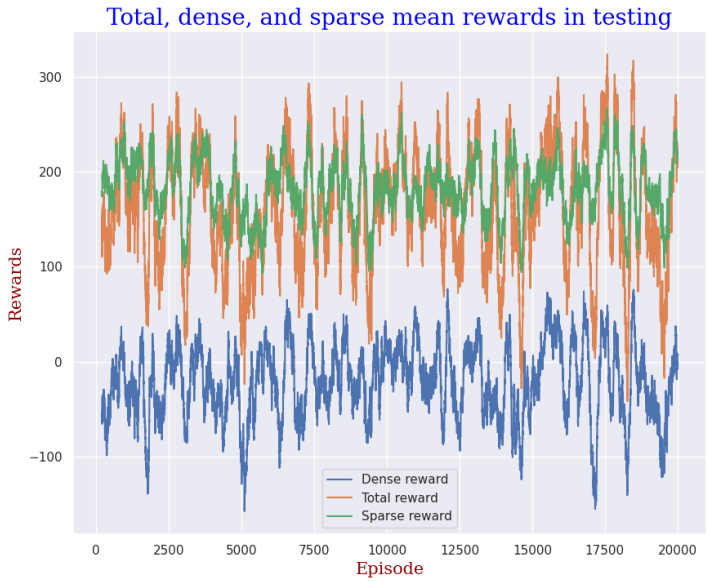

图6. 在包含20,000个训练轮次的测试中，‘Bug’获得的总奖励（橙色）、稠密奖励（蓝色）和稀疏奖励（绿色）的平均值。应用了窗口大小为200的滚动均值滤波器。

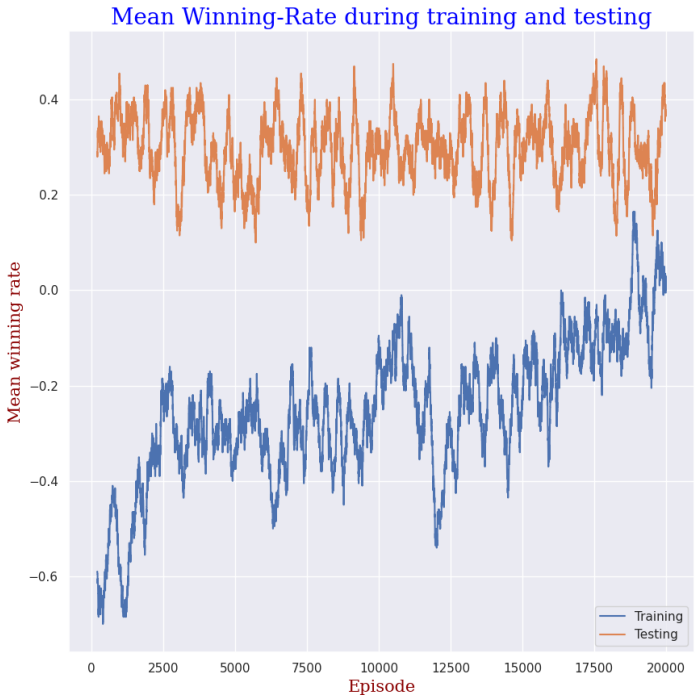

图 7. 随着训练的进行，团队的 MWR 逐渐提高（蓝色）。MWR 从负值上升至正值，表明团队最终的能力强于对手。在 20,000 个训练轮次的测试中，团队的 MWR 始终为正（介于 0.15 和 0.47 之间）（橙色）。对每个图表均应用了长度为 200 的滚动均值滤波器。



图 8. 随着训练的进行，获胜所需步数的平均值逐渐减少。应用了长度为 200 的滚动均值滤波器。仅统计团队获胜的游戏。

λ 的滤波器，在训练轮次 α （记作 E_α）时的平均胜率（记作 MWR）为：

$$MWR_\alpha = \frac{sum\ of\ all\ WRs\ in\ recent\ \lambda\ epochs\ until\ E_\alpha}{\lambda} \tag{3}$$

图 7 绘出了训练和测试期间 MWR 的变化情况，表明随着训练的推进，团队的能力不断提升，并在最终策略中保持稳定。换言之，智能体“Bug”已经学会与“Ant”合作，并为团队合作作出贡献。

另一个指标“赢得一场比赛所需的步数”用于反映成功团队合作所耗费的时间成本。其滚动平均值（以类似公式 3 的方式计算）反映了训练过程中团队合作效率的变化。图 8 显示，随着训练的进行，团队的合作效率不断提高，从而减少了赢得比赛所需的步数或时间。

5 讨论与未来工作

这项初步工作将现有的虚拟多智能体平台 Robo-Sumo 扩展为 TripleSumo，其中包含三名选手进行相扑游戏。预先设定两名智能体组成团队，与另一名智能体对战。作为基线研究，我们在一场正在进行的比赛中采用混合奖励结构，并使用强化学习算法 DDPG 训练新加入的智能体，以考察合作行为。结果表明，随着训练的推进，平均奖励中的稀疏部分和密集部分均有所增加，并最终趋于收敛。团队合作的能力和效率均有所提升，这体现为平均胜率上升、赢得一场比赛所需的平均步数下降，表明智能体在这一对抗性环境中实现了成功合作。尽管本研究仅关注 DDPG，我们未来的研究将考察其他强化学习算法在多智能体系统中学习合作策略的表现。

本文提出的场景类似于在双追捕者单逃逸者场景中学习合作行为，以缩短捕获逃逸者所需的时间，并可直接应用于群体机器人系统。一个自然的扩展方向是提高环境的复杂性，加入更多追捕智能体，使其成为多追捕者单逃逸者场景，并纳入有关猎物位置的概率观测。这样，追捕智能体便能够演化出不同类别的行为，例如“拦截者”“护卫者”和“冗余者”。

这项研究的下一步是考察 TripleSumo 中一种更复杂的情景：合作伙伴并非预先配对。三名或更多智能体彼此对战时，游戏将持续进行，直到竞技场上只剩下一名智能体。所有智能体都将学会自由选择与非预先指定的伙伴组队，以便尽可能长时间地留在游戏中。

参考文献

1. Q. Yin、J. Yang、W. Ni、B. Liang 和 K. Huang，《游戏中的人工智能：技术、挑战与机遇》，*ArXiv*，第 abs/2111.07631 卷，2021。2. S. Liu、G. Lever、Z. Wang、J. Merel、S. M. A. Eslami、D. Hennes、W. M. Czarnecki、Y. Tassa、S. Omidshafiei、A. Abdolmaleki、N. Siegel、L. Hasenclever、L. Marris、S. Tunyasuvunakool、H. F. Song、M. Wulfmeier、P. Muller、T. Haarnoja、B. D. Tracey、K. Tuyls、T. Graepel 和 N. M. O. Heess，《从运动控制到模拟人形足球中的团队协作》，*ArXiv*，第 abs/2105.12196 卷，2021。3. M. Al-Shedivat、T. Bansal、Y. Burda、I. Sutskever、I. Mordatch 和 P. Abbeel，《通过元学习在非平稳竞争环境中持续适应》，*ArXiv*，第 abs/1710.03641 卷，2018。4. T. P. Lillicrap、J. J. Hunt、A. Pritzel、N. M. O. Heess、T. Erez、Y. Tassa、D. Silver 和 D. Wierstra，《使用深度强化学习进行连续控制》，*CoRR*，第 abs/1509.02971 卷，2016。5. K. Kurach、A. Raichuk、P. Stanczyk、M. Zajac、O. Bachem、L. Espeholt、C. Riquelme、D. Vincent、M. Michalski、O. Bousquet 和 S. Gelly，《Google Research Football：一种新型强化学习环境》，载于 *AAAI*，2020。6. S. Huang、W. Chen、L. Zhang、Z. Li、F. Zhu、D. Ye、T. Chen 和 J. Zhu，《TiKick：基于单智能体演示，迈向完整的多智能体足球比赛》，*ArXiv*，第 abs/2110.04507 卷，2021。7. H. Jia、Y. Hu、Y. Chen、C. Ren、T. Lv、C. Fan 和 C. Zhang，《Fever Basketball：一种用于多智能体强化学习的复杂、灵活且异步的体育游戏环境》，*ArXiv*，第 abs/2012.03204 卷，2020。8. B. Baker、I. Kanitscheider、T. Markov、Y. Wu、G. Powell、B. McGrew 和 I. Mordatch，《多智能体自课程机制中涌现的工具使用》，*ArXiv*，第 abs/1909.07528 卷，2020。9. R. Lowe、Y. Wu、A. Tamar、J. Harb、P. Abbeel 和 I. Mordatch，《用于混合合作竞争环境的多智能体演员-评论家方法》，*ArXiv*，第 abs/1706.02275 卷，2017。10. T. Zhang、H. Xu、X. Wang、Y. Wu、K. Keutzer、J. Gonzalez 和 Y. Tian，《通过奖励归因分解实现多智能体协作》，*ArXiv*，第 abs/2010.08531 卷，2020。11. C. de Souza、R. Newbury、A. Cosgun、P. Castillo、B. Vidolov 和 D. Kuli´c，《使用深度强化学习进行去中心化多智能体追逃》，*IEEE Robotics and Automation Letters* 6，4552–4559，2021，第 pp. – 卷。12. A. Von Moll、D. W. Casbeer、E. Garcia 和 D. Milutinovi´c，《多名追捕者追捕一名逃逸者》，载于 *2018 International Conference on Unmanned Aircraft Systems (ICUAS)* 2018，133–142，第 – 页。13. G. Brockman、V. Cheung、L. Pettersson、J. Schneider、J. Schulman、J. Tang 和 W. Zaremba，《OpenAI Gym》，*ArXiv*，第 abs/1606.01540 卷，2016。14. E. Todorov、T. Erez 和 Y. Tassa，《MuJoCo：用于基于模型的控制的物理引擎》，载于 *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* 2012，5026–5033，第 – 页。