arXiv:2302.09277v1 [cs.LG] 18 Feb 2023

通过互助促进多智能体强化学习中的合作

Yunbo Qiu Yue Jin LebinYu JianWang Xudong Zhang

清华大学电子工程系，中国北京

摘要

近年来，多智能体强化学习（MARL）在协作任务中取得了长足进展。然而，在局部奖励机制中，每个智能体只能获得局部奖励，而无法获得由所有智能体共享的全局奖励，传统多智能体强化学习算法对智能体之间相互影响的考虑并不充分。在协作任务中，智能体之间的相互影响尤为重要，因为智能体需要进行协调，以获得更好的性能。本文提出了一种新颖的互助型多智能体强化学习（MH-MARL）算法，指导智能体相互帮助，从而促进合作。MH-MARL利用期望动作模块，为每个特定智能体生成其他智能体的期望动作。随后，在训练过程中，将这些期望动作传递给其他智能体，以供其进行选择性模仿。实验结果表明，MH-MARL在成功率和累积奖励两方面均提升了多智能体强化学习的性能。

索引词— 多智能体强化学习、协作、互助

1. 引言

近年来，多智能体强化学习（MARL）在协作任务中取得了巨大成功，例如电子游戏[1], 电力分配[2], 和协作导航[3]。奖励机制的设计对于多智能体强化学习至关重要，它能够指导智能体完成相应目标。根据具体应用场景，可以设计局部奖励机制[4, 5]或全局奖励机制[6, 7], 分别为单个智能体提供奖励，或为所有智能体提供共享奖励。全局奖励机制涉及信用分配问题 [6]。采用局部奖励机制的传统算法会优化智能体的局部奖励，而这些奖励无法明确反映智能体之间的相互影响，因此会导致智能体之间的协调不足[8]。智能体之间的相互影响在多智能体强化学习中至关重要，尤其是在协作任务中；在此类任务中，智能体需要进行协调，从整个多智能体系统的角度实现更好的性能。

主要有三种方法将智能体之间明确的相互影响纳入传统的多智能体强化学习：通信方法、师生方法以及基于局部奖励与全局奖励的方法。通信方法通过智能体之间的通信协议来提升智能体的合作。 [9] 在进行通信的同时训练智能体的策略。 [10] 进一步研究何时进行通信以及如何使用通信消息。 [11] 形成每个智能体的意图，然后将其传达给其他智能体。然而，通信方法要求具备明确的通信信道，以便在智能体之间交换消息，而这在许多应用中可能并不存在。

师生方法指导智能体相互学习。 [12]引导智能体向已经具备经验的智能体学习。 [13, 14] 考虑了智能体在共享环境中接受训练并相互教学的框架。 [15, 16] 深入研究了如何重复利用从教师智能体那里接收到的指导信息。在 [17],中，智能体通过共享经验隐式地相互教学。然而，师生方法要求智能体彼此相同，才能直接利用其他智能体关于动作的指导信息。

局部-全局奖励方法同时利用局部奖励和全局奖励，其中全局奖励可以由局部奖励构建。 [18]分别通过局部奖励和全局奖励学习两种策略，并限制这两种策略之间的差异。 [8] 维护一个集中式评论家，从全局奖励中学习，同时维护一个去中心化评论家，从局部奖励中学习，并同时优化这两个评论家。然而，局部-全局奖励方法要求根据局部奖励机制设计全局奖励。此外，从奖励机制的角度来看，这类方法只考虑单个智能体与整个多智能体系统之间的关系，而没有考虑智能体之间的详细关系。

本文提出了一种新颖的基于互助的多智能体强化学习算法（MH-MARL），用于指导智能体在传统多智能体强化学习算法训练的同时相互帮助，其中在传统多智能体强化学习算法中加入了期望动作模块。在采用局部奖励方案的传统多智能体强化学习中，智能体只优化自身策略，而忽略其他智能体的性能。相比之下，在 MH-MARL 中，每个智能体在优化自身回报的同时，学习生成能够帮助提高其他智能体回报的动作。具体而言，每个智能体在接受传统多智能体强化学习算法训练时，会选择性地模仿其他智能体所期望的动作

。期望动作由期望动作模块生成。选择性模仿遵循这样的原则：如果帮助其他智能体不会严重损害自身性能，那么智能体就会选择学习帮助其他智能体。需要注意的是，MH-MARL 考虑了智能体之间的相互帮助，而不要求在执行过程中进行通信。此外，智能体也不必彼此相同。我们在群体导航环境中评估了 MH-MARL[19] ，其中智能体之间的合作对于成功至关重要。实验结果表明，MH-MARL 在成功率和奖励方面都提升了性能。本文的主要贡献如下：

- 提出了一种新颖的算法 MH-MARL，通过利用期望动作模块来指导智能体有选择地相互帮助，同时优化各自的策略，从而促进多智能体强化学习的局部奖励机制中的合作。
- 通过实验验证了 MH-MARL 在合作任务中的有效性，以及其能够构建于不同基础多智能体强化学习算法之上的能力。

2. 背景

2.1. 具有局部奖励机制的马尔可夫博弈

具有局部奖励机制的马尔可夫博弈通常以类似于 [4, 5]的方式建模。在每个时间步 t，每个智能体i 根据其局部观测 o_i 和自身的策略 $\pi_i(o_i)$选择动作 a_i 与环境进行交互。每个智能体的观测 o_i 和动作 a_i 的组合分别称为联合观测 $\boldsymbol{o}$ 和联合动作 $\boldsymbol{a}$。智能体与环境交互后，每个智能体 i 都会接收到新的局部观测 o_i' 和局部奖励 r_i。每个智能体的策略 π_i 的目标是最大化其回报的期望，也称为累积奖励：$R_i = \sum_{t=0}^{\infty} \gamma^t r_{i,t}$，其中 γ 为折扣因子。

2.2. 多智能体强化学习框架

演员-评论家架构以及集中训练、分散执行（CTDE）是多智能体强化学习框架中常用的两个组成部分。

在演员-评论家架构中，每个智能体 i 都会学习一个“演员”和一个“评论家”。“演员”表示策略函数 π_i，可根据观测 o_i生成动作 a_i 。“评论家”表示动作价值函数，也称为 Q 函数，用于估计给定当前动作 $\boldsymbol{a}$和观测 $\boldsymbol{o}$时策略的回报。

在集中训练、分散执行[4], 中，训练期间，智能体可以访问其他智能体的观测和动作。这有助于智能体从更全面的视角更准确地估计预期回报。在执行期间，每个智能体只需根据自身的观测选择动作，从而使智能体能够以分布式方式执行。

MADDPG[4]是一种采用局部奖励机制的代表性多智能体强化学习算法。MADDPG中的Q函数通过最小化以下损失进行优化：

$$L_{i,critic} = \mathbb{E}_{\boldsymbol{o},\boldsymbol{a},r_i,\boldsymbol{o}'\sim\mathcal{D}}[(Q_i(\boldsymbol{o}, a_1, ..., a_n) - y_i)^2], \quad (1)$$

其中， $\mathcal{D}$ 是用于存储和重用智能体经验的经验回放缓冲区，y_i 是通过以下方式获得的目标值：

$$y_i = r_i + \gamma Q_i'(\boldsymbol{o}', a_1', ..., a_n')|_{a_j'=\pi_j'(o_j')}, \quad (2)$$

其中， Q_i' 和 π_j' 是作为备份以稳定训练的目标函数。

MADDPG 中的策略函数通过最小化以下损失进行优化：

$$L_{i,actor} = \mathbb{E}_{\boldsymbol{o},\boldsymbol{a}\sim\mathcal{D}}[-Q_i(\boldsymbol{o}, a_1, ..., a_i, ..., a_n)|_{a_i=\pi_i(o_i)}]. \quad (3)$$

3. 算法

传统多智能体强化学习算法并未充分考虑智能体的动作对彼此产生的主动且明确的相互影响，尽管在集中训练、分散执行中，Q函数会将其他智能体的动作和观测纳入考量，但主要是为了解决非平稳性问题。智能体之间的相互影响至关重要，尤其是在协作任务中，智能体应通过团队协作实现更好的性能。在局部奖励机制下，智能体只优化自身回报，因此对相互影响缺乏考量的问题更加明显。

因此，我们提出了一种新颖的基于互助的多智能体强化学习算法（MH-MARL），如图1所示，该算法通过在传统演员-评论家架构中加入期望动作模块，指导智能体相互帮助。MH-MARL旨在引导每个智能体帮助其他智能体，同时避免严重损害自身性能，从而在协作任务中实现“人人为我，我为人人”的效果。MH-MARL可以建立在局部奖励机制下任何采用演员-评论家架构和集中训练、分散执行的传统多智能体强化学习算法之上。本文选择在具有代表性的多智能体强化学习算法MADDPG[4]上构建MH-MARL。训练期间，借助期望动作模块的互助包含两个过程：为每个智能体生成期望动作，以及利用其他智能体生成的期望动作。

首先，我们关注如何生成期望动作。在集中训练、分散执行中，每个智能体的Q函数都将所有智能体的动作作为输入的一部分，这表明其他智能体的动作会影响该智能体的回报。因此，如果每个智能体都希望最大化自身性能，那么除了调整自身动作之外，还可以依赖其他智能体动作的调整。为了明确其他智能体的哪些动作能够最大程度地帮助某个智能体，每个智能体 i 维护一个期望策略函数 μ_i，根据联合观测生成其他智能体的期望动作

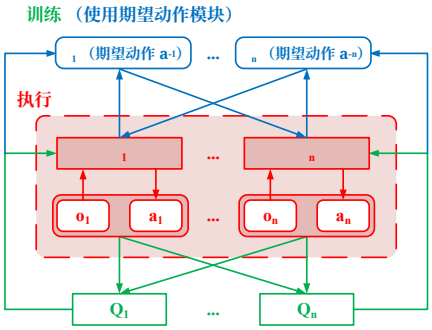

图 1。MH-MARL算法概览。传统多智能体强化学习中的Actor模块和评论家模块分别以红色和绿色表示。蓝色表示MH-MARL中的期望动作模块，该模块为其他智能体生成期望动作，以供其有选择地模仿。

，其方法是最小化以下损失：

$$L_{i,expect} = \mathbb{E}_{\boldsymbol{o}\sim\mathcal{D}}[-Q_i(\boldsymbol{o}, a_i, \boldsymbol{a_{-i}})|_{a_i=\pi_i(o_i),\boldsymbol{a_{-i}}=\mu_i(\boldsymbol{o})}], \tag{4}$$

其中，经验从经验回放缓冲区 $\mathcal{D}$ 中随机提取，作为（3）； a_i 由智能体当前的策略π_i生成； $\boldsymbol{a_{-i}}$ 表示智能体 i所期望的其他智能体的动作。随着训练的进行，通过最小化（1）进行训练后，Q函数将更加准确，随后 μ_i 能够更准确地生成其他智能体的期望动作。

其次，我们考虑当每个智能体优化其自身策略函数 π_i 时，如何利用其他智能体的期望动作。对于每个智能体 i，如果我们考虑优化 π_i仅仅是为了帮助另一个智能体 j，那么由π_i 生成的动作应当模仿 $\mu_j(\boldsymbol{o}, i)$（智能体 i在 $\mu_j(\boldsymbol{o})$中的期望动作），这通过最小化 $\mu_j(\boldsymbol{o}, i)$ 与智能体自身动作 $\pi_i(o_i)$ 之间的差异来实现。然而，$\mu_j(\boldsymbol{o}, i)$ 在某些情况下可能会严重损害智能体 i 的性能，因此完全采用期望动作会妨碍智能体 i 的优化。因此，该智能体i 仅考虑模仿不会严重降低其自身性能的期望动作。这种选择性模仿通过最小化以下损失来实现：

$$\begin{aligned} L_{i,help} =& \mathbb{E}_{\boldsymbol{o},\boldsymbol{a}\sim\mathcal{D}}[\frac{1}{n-1}\sum_{j\neq i}[||\pi_i(o_i) - \mu_j(\boldsymbol{o}, i)||_2 \\ & \cdot \epsilon(Q_i(\boldsymbol{o}, a_1, ..., a_i, ..., a_n)|_{a_i=\mu_j(\boldsymbol{o},i)} + \eta \\ & - Q_i(\boldsymbol{o}, a_1, ..., a_i, ..., a_n)|_{a_i=\pi_i(o_i)})]], \end{aligned} \tag{5}$$

其中， $\epsilon(x) = max(0, \frac{x}{|x|})$ 是用于比较的阶跃函数， η 是一个正常数。由于每个智能体 i 都可以从其他所有智能体那里接收期望动作，因此，选择性模仿会在所有这些期望动作上取平均。

除了帮助其他智能体之外，每个智能体还通过最小化（3）来优化其策略函数，以最大化自身回报，这与传统的多智能体强化学习相同。帮助其他智能体的选择性模仿与优化自身回报的传统多智能体强化学习

算法1 MH-MADDPG（基于 MADDPG）

初始化 每个智能体的参数 θ_i i
对于 回合 = 1 至 E **执行**
  初始化一个用于动作探索的随机过程 $\mathcal{N}$
  接收初始观测 $\boldsymbol{o}$
  对于 t = 1 至 $T_{episode}$ **执行**
    对于每个智能体 i，根据当前策略和探索执行动作 $\boldsymbol{a}$ = $(a_1,...,a_n)$ 并获得奖励 $\boldsymbol{r}$ 和新的观测 $\boldsymbol{o'}$
    将 $(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{o'})$ 存入经验回放缓冲区 $\mathcal{D}$
    $\boldsymbol{o} \leftarrow \boldsymbol{o'}$
    对于智能体 $i = 1$ **执行**
      随机采样一个小批量 M，从 $\mathcal{D}$
      为智能体 i 生成期望动作，每个 $\mu_{j\neq i}(\boldsymbol{o}, i)$
      通过使用期望动作最小化（6）来更新演员
      通过最小化（1）来更新评论家
      通过最小化（4）来更新期望动作策略
    结束循环
    为每个智能体软更新目标网络 i:
    $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$
  结束循环
结束循环

通过以下损失函数同时实现：

$$L_{i,actor}^{MH} = L_{i,actor} + \alpha_i \cdot L_{i,help}, \tag{6}$$

其中， α_i 是两个损失项的混合比率。为自动平衡这两个损失项，每个小批次都会计算 α_i ，将两个损失项转换为相同的尺度：

$$\alpha_i = \frac{\beta}{M}\sum_{\boldsymbol{o},\boldsymbol{a}\sim\mathcal{D}}|Q_i(\boldsymbol{o}, a_1, ..., a_i, ..., a_n)|_{a_i=\pi_i(o_i)}|, \tag{7}$$

其中， M 是小批次的大小， β 是一个超参数。基于 MADDPG 构建的 MH-MARL 完整算法如算法 1 所示。

4. 实验

4.1. 主要实验

为验证MH-MARL的有效性，我们在群体导航环境[19], 中开展实验，其中智能体需要导航至目标区域，同时保持群体队形并避免碰撞。由于智能体需要精确控制彼此之间的距离，以避免发生碰撞和破坏群体队形，这是一项高度协作的任务。我们基于具有代表性的多智能体强化学习算法MADDPG[4] 构建MH-MARL，形成MH-MADDPG，并将其与四种基线算法进行对比：MADDPG、MADDPG-GR、DE-MADDPG和SEAC-MADDPG。MADDPG-GR使用全局奖励，该奖励是局部奖励之和，并由所有智能体共享。DE-MADDPG[8]同时学习用于优化局部奖励的分散式Q函数，以及用于优化全局奖励的集中式Q函数（全局奖励与MADDPG-GR中一样，是局部奖励之和）。SEAC-MADDPG

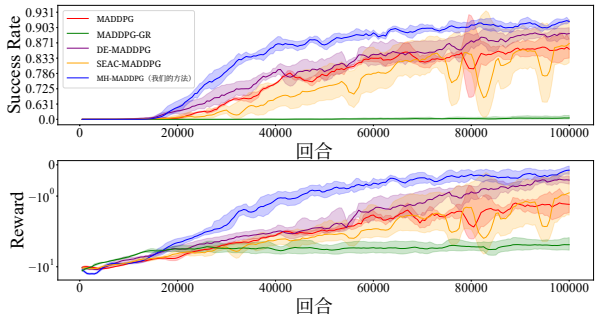

图 2。主要实验中成功率和奖励的收敛曲线，分别采用五次根尺度和对称对数尺度表示。

基于 MADDPG 设计，直接在智能体之间共享经验[17]。$\eta = 0.05$ 和 $\beta = 2$作为超参数。所有算法均使用 5 个随机种子运行。

成功率和奖励的收敛曲线如图 2 所示。结果表明，与四种基线算法相比，我们的算法 MH-MADDPG 在训练过程中学习速度最快，并在训练结束后取得了最佳的成功率和奖励。将局部奖励机制直接转换为全局奖励机制（MADDPG-GR）会导致较差的性能。DE-MADDPG 和 SEAC-MADDPG 在成功率和奖励方面均超过了其基础多智能体强化学习算法 MADDPG，但仍不及 MH-MADDPG。此外，由于利用了其他智能体的经验，SEAC-MADDPG 的稳定性最差。

4.2. 消融实验

为验证 MH-MARL 各组成部分的有效性，我们设计了 MH-MADDPG 的三种消融算法：“无帮助”指移除 MH-MARL 中期望动作模块的算法，即 MADDPG；“无选择性”移除了式（5）中的阶跃函数，以完全采用期望动作；“无多智能体强化学习”移除了式（6）中的传统多智能体强化学习损失项（3），从而仅帮助其他智能体。

消融实验的收敛曲线如图 3 所示。这表明，在整个训练过程中，MH-MADDPG 整体上明显优于三种消融算法，验证了 MH-MARL 各组成部分的有效性。具体而言，与“无帮助”（MADDPG）相比，“无多智能体强化学习”几乎无法通过训练提升性能，而“无选择性”则会阻碍训练早期的学习。这表明，相较于帮助其他智能体，强调智能体自身性能的重要性。

4.3. 使用另一种基础多智能体强化学习算法的实验

在之前的实验中，我们选择 MADDPG 作为实现 MH-MARL 的基础多智能体强化学习算法。然而，MH-MARL 的设计使其能够基于任何采用演员-评论家架构和集中训练、分散执行的多智能体强化学习算法构建

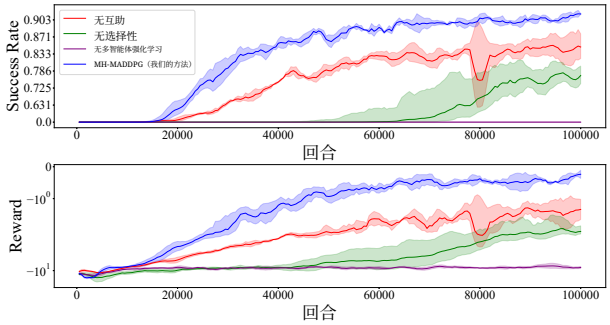

图 3。消融实验的成功率和奖励收敛曲线。

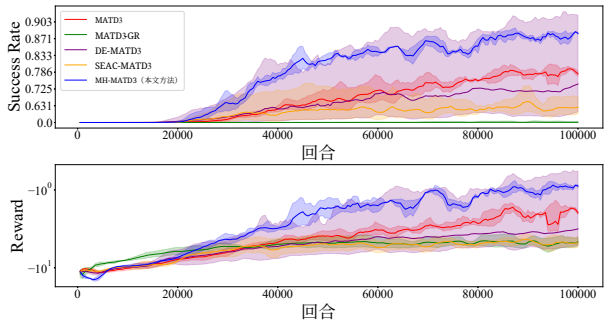

图 4。以 MATD3 作为基础多智能体强化学习算法的实验中，成功率和奖励的收敛曲线。

局部奖励机制。在本小节中，我们选择基于 MATD3[5]，而非 MADDPG 构建 MH-MARL。基线算法也进行了调整，采用 MATD3 作为其基础多智能体强化学习算法。

成功率和奖励的收敛曲线如图 4 所示。该图表明，MH-MATD3 在成功率和奖励方面仍大幅超越基线算法，验证了基础多智能体强化学习算法选择的灵活性。具体而言，DE-MATD3 和 SEAC-MATD3 的性能甚至不如其基础算法 MATD3，且与 MH-MATD3 不同，DE-MATD3 的稳定性较差。

5. 结论与未来工作

我们提出了一种新颖的算法 MH-MARL，该算法利用期望动作模块指导智能体相互帮助。智能体首先为其他智能体生成期望动作，然后其他智能体在接受传统多智能体强化学习算法训练的同时，有选择地模仿相应的期望动作。MH-MARL 强调将优化智能体自身性能作为首要目标，并将互助视为促进合作的次要目标。我们通过实验验证了 MH-MARL 在提升协作任务性能方面的优越性。实验结果还表明，MH-MARL 可以构建在不同的基础多智能体强化学习算法之上。在不久的将来，我们希望将 MH-MARL 应用于其他协作任务。

6. 参考文献

[1] Oriol Vinyals、Igor Babuschkin、Wojciech M Czarnecki、Micha¨el Mathieu、Andrew Dudzik、Junyoung Chung、David H Choi、Richard Powell、Timo Ewalds、Petko Georgiev 等，《使用多智能体强化学习达到《星际争霸II》宗师级水平》，《自然》，第 575 卷，第 7782 期，第 350–354 页，2019年。

[2] JianhongWang、Wangkun Xu、Yunjie Gu、Wenbin Song 和 Tim C Green，“电力分配网络上的主动电压控制的多智能体强化学习”，*Advancesin Neural Information ProcessingSystems*，第 34 卷，第 3271–3284 页，2021年。

[3] YueJin、Yaodong Zhang、JianYuan 和 XudongZhang，“采用交错深度强化学习在未知环境中进行高效的多智能体协作导航”，载于*ICASSP 2019-2019 IEEE*国际声学、语音与信号处理会议*(ICASSP)*。IEEE，2019年，第 2897–2901 页。

[4] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和IgorMordatch，“面向混合协作竞争环境的多智能体演员-评论家方法”，载于神经信息处理系统国际会议第 *31* 届会议论文集，2017年，第 6382–6393 页。

[5] Johannes Ackermann、Volker Gabler、TakayukiOsa 和 Masashi Sugiyama，“使用双重集中式评论家减少多智能体领域中的高估偏差”，载于 *NeurIPS*深度强化学习研讨会，2019年。

[6]Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson，“反事实多智能体策略梯度”，载于人工智能 *AAAI* 会议论文集，2018年，第32卷。

[7]Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson，“Qmix：深度多智能体强化学习的单调价值函数分解”，载于*International Conference on Machine Learning*。PMLR，2018年，第4295–4304页。

[8] Hassam Ullah Sheikh 和 Ladislau B¨ol¨oni，“针对同时具有个体奖励和团队奖励的问题的多智能体强化学习”，载于 *2020*年国际联合神经网络会议（*IJCNN*）。IEEE，2020年，第1–8页。

[9] Sainbayar Sukhbaatar、Rob Fergus 等，《通过反向传播学习多智能体通信》，神经信息处理系统进展，第 29 卷，2016 年。

[10] Jiechuan Jiang 和 Zongqing Lu，“学习面向多智能体合作的注意力通信”，*Advances in Neural Information Processing Systems*，第 31 卷，2018年。[11]Woojun Kim、Jongeui Park 和 Youngchul Sung，“多智能体强化学习中的通信：意图共享”，载于 *InternationalConference onLearning Representations*，2020年。[12] Lisa Torrey 和 Matthew Taylor，“经济型教学：强化学习中智能体向智能体提供建议”，载于 *Proceedings of the2013 international conference onAutonomous agents andmulti-agent systems*，2013年，第 1053–1060 页。

[13]Felipe Leno Da Silva、Ruben Glatt 和 Anna Helena Reali Costa，《在多智能体强化学习中同时进行学习与指导》，载于第*16*届自主智能体与多智能体系统会议论文集，2017年，第1100–1108页。

[14] Shayegan Omidshafiei、Dong-Ki Kim、Miao Liu、Gerald Tesauro、Matthew Riemer、Christopher Amato、Murray Campbell 和 Jonathan P How，《在合作多智能体强化学习中学习教学》，载于*AAAI*人工智能会议论文集，2019年，第33卷，第6128–6136页。

[15] Changxi Zhu、Yi Cai、Ho-fung Leung 和 Shuyue Hu，《在教师-学生范式中通过重用先前建议进行学习》，载于第*19*届国际自主智能体与多智能体系统会议论文集，2020年，第1674–1682页。

[16] ErcumentIlhan、Jeremy Gow 和DiegoPerez Liebana，“通过建议模仿在深度强化学习中进行动作建议”，载于第 *20* 届自主智能体与多智能体系统国际会议论文集，2021年，第 629–637 页。

[17] Filippos Christianos、Lukas Sch¨afer 和 Stefano Al- brecht，“用于多智能体强化学习的共享经验演员-评论家算法”，神经信息处理系统进展，第 33 卷，第 10707–10717 页，2020年。

[18] Li Wang、Yupeng Zhang、Yujing Hu、Weixun Wang、ChongjieZhang、Yang Gao、Jianye Hao、TangjieLv 和 Changjie Fan，“个体奖励辅助的多智能体强化学习”，载于 *InternationalConferenceonMachine Learning*。PMLR，2022年，第 23417–23432 页。

[19] Yunbo Qiu、Yue Jin、Lebin Yu、Jian Wang、Yu Wang 和 Xudong Zhang，《利用非专家策略改进群集控制中多智能体强化学习的样本效率》，*IEEE Internet ofThings Journal*，第 1–14 页，2023年。