团队竞争马尔可夫博弈中的对手感知的基于角色学习

Paramita Koley 印度理工学院卡拉格普尔分校 印度卡拉格普尔
paramita.koley@iitkgp.ac.in

Aurghya Maiti 哥伦比亚大学 美国纽约
a
m5887@columbia.edu

Niloy Ganguly 印度理工学院卡拉格普尔分校 印度卡拉格普尔
niloy@cse.iitkgp.ac.in

Sourangshu Bhattacharya 印度理工学院卡拉格普尔分校 印度卡拉格普尔
sourangshu@cse.iitkgp.ac.in

arXiv:2301.05873v1 [cs.MA] 14 Jan 2023

摘要

在多智能体马尔可夫博弈中，团队竞争是多智能体强化学习中日益重要的一种设定，因为它普遍适用于对许多现实情境进行建模。多智能体演员-评论家方法是团队竞争设定下学习最优策略的最合适方法，因为它们能够灵活地学习特定于智能体的评论家函数，并且这些函数还可以从其他智能体那里学习。在许多现实世界的团队竞争场景中，智能体的角色会自然涌现，以帮助团队成员之间进行协调与协作。然而，现有的涌现角色学习方法严重依赖Q学习设置，而该设置不允许学习特定于智能体的Q函数。本文提出RAC，这是一种新颖的方法，用于学习团队中智能体多样且动态的涌现角色。在所提出的方法中，智能体还可通过预测对手团队中智能体的角色获益。RAC采用带有角色编码器和对手角色预测器的演员-评论家框架来学习最优策略。在两款游戏上的实验表明，RAC学习到的策略比采用最先进的基线方法学习到的策略获得更高的奖励。此外，实验结果表明，团队中的智能体能够学到多样且对手感知的策略。

关键词

多智能体强化学习、团队竞争、涌现角色、演员—评论家

1 引言

许多现实场景都可以建模为**团队竞争马尔可夫博弈** [20] （也称为混合合作竞争博弈），在这类博弈中，每个智能体都会因自身动作获得个人奖励，并与同一团队的其他智能体共享共同结果带来的奖励。例如，在踢足球时，每位球员都可能因个人表现获得奖励，同时获得球队获胜带来的共享奖励。在一些社会互动中，例如公司争夺客户时，每位员工都可能因个人表现获得奖励，同时获得团队表现带来的共享奖励。在这类奖励情境中，**角色**会自然涌现，因为它们能让团队内的智能体有效合作。对于足球运动员，角色可以是“前锋”和“后卫”；对于公司员工，角色则可以是“生产”和“销售”。在这项工作中，我们研究在团队竞争马尔可夫博弈中，如何在强化学习（RL）设定下学习具有涌现角色的策略。

我们采用多智能体强化学习（MARL）来学习此设置下各个智能体的最优策略。尽管针对共享奖励设置的深度多智能体强化学习技术 [9, 23, 27, 29, 39], 有很多，但针对混合奖励设置的技术相对较少。遵循集中式训练与分散式执行理念的多智能体演员—评论家算法 [11, 20],是混合奖励设置中最流行的多智能体强化学习方法，可用于建模团队竞争场景。MADDPG [20] 是最早采用演员—评论家框架在混合奖励设置中学习确定性策略的技术之一。MAAC [11] 在评论家模块中引入注意力机制，使其能够有选择地关注从其他智能体收集的信息，从而在复杂的多智能体环境中实现更好的可扩展性。然而，据我们所知，演员—评论家框架中的角色导向策略学习，尤其是在竞争性博弈中的应用，尚未得到探索。

另一方面，许多多智能体强化学习系统 [2, 30]已探索为智能体预定义角色，以降低设计复杂度。针对特定领域问题中基于动态角色涌现的多智能体强化学习，相关研究并不多 [1, 13, 15, 28, 36, 43]。近期，Wang等人 [37] 在其名为 ROMA 的系统中，提出了面向合作任务的多智能体强化学习自动角色涌现方法。该方法基于 QMIX [27]平台，在智能体之间进行基于潜在角色的任务分解。类似地，Wang等人 [38] 提出了 RODE，这是一种双层基于角色的学习框架，通过依据动作的效果对动作进行聚类，将联合动作空间分解为受限的角色动作空间。最后，该框架使用角色选择器在规模较小的角色空间中搜索角色，同时角色策略在缩减后的动作观测空间中学习，从而显著提高学习效率。Liu等人 [19] 提出了 ROGC，这是一种基于面向角色的图卷积的多智能体强化学习框架：该框架通过对智能体进行分类来生成角色，利用图卷积模块实现高效的角色内通信，并最终通过将角色信息嵌入算法来生成有效的个体策略。然而，上述所有框架都基于 QMIX 构建，其全局效用函数由各智能体的局部效用函数组合而成，而这种组合依赖于所有智能体共享相同奖励这一假设。因此，该设置无法直接扩展到混合奖励设置，因而无法通过联合学习，在存在对手的情况下进行面向角色的策略学习。

无论何种形式。近期，研究者已在许多场景中探索对手感知的演员—评论家算法，包括学习时间动态对手模型的TDOM-AC [40] , 、最大熵正则化对手模型 ROMMEO [35], 、采用变分贝叶斯的对手模型 PRR [34], 等。这些研究验证了对手模型对于提升竞争性博弈表现的必要性。然而，这些方法主要局限于一对一竞争，且未考虑角色感知对手建模。但在团队竞争设置中，每个智能体都在团队内采用角色感知策略；在对手建模中考虑角色，有助于降低设计复杂度，从而使智能体受益。

在本研究中，我们提出了 RAC，这是一种在混合奖励设置中结合角色学习与角色感知对手建模的演员-评论家算法。RAC 使用基于 MAAC 的演员—评论家框架 [11], , 该框架允许评论家网络使用所有智能体的联合观测，从而将对手的经验纳入评论家网络。我们的主要贡献有两点。首先，我们在演员—评论家框架中加入了角色编码器，以学习每个团队动态且多样化的涌现角色。此外，在对手建模方面，我们为每个团队加入了一个对手角色预测网络，该网络在执行时可供策略使用。对手角色预测网络与角色编码器、评论家网络和策略网络联合训练。具体而言，在集中式训练期间，对手角色预测器经过训练，以模仿对手实际的角色编码分布。在执行期间，这有助于智能体选择应对方式。学习到的角色会根据智能体的轨迹动态变化，并且在团队内的不同智能体之间各不相同。使用我们提出的回合制棋盘游戏 Touch-Mark 和市场进行的实验表明，在直接竞争中，RAC 智能体的表现优于 MAAC 智能体。此外，实验还表明，RAC 学到的动态角色符合每款游戏中直观的角色特定行为，从而证实了涌现角色的有效性。

总而言之，我们的贡献如下：

- 据我们所知，我们提出了 RAC，这是首个面向混合奖励设置、基于演员—评论家框架的对手感知角色导向策略学习方法。
- RAC 鼓励团队内部形成多样化角色，并确保角色编码能够捕捉智能体的轨迹。
- 最重要的是，RAC 学习一个对手角色预测器模块，在集中式训练期间利用对手角色编码，并在执行期间将其用于策略。最终，对手角色预测器会引导智能体针对对手选择策略性响应。
- 通过在两款回合制棋盘游戏 Touch-Mark 和 Market 中提高奖励，我们从实证角度验证了 RAC 的策略学习效果有所提升；同时，智能体表现出的直观角色特定行为也证明了所学习角色的可识别性。

2 相关文献

多智能体行动者-评论家方法主要遵循集中式训练与分散式执行这一理念，是深度多智能体强化学习（MARL）方法中最主流、最受欢迎的技术之一 [7–9, 20, 23, 27, 33]。在这些工作中，MADDPG [20], 是对DDPG [16], 的多智能体改编，它考虑了混合奖励设置下的连续动作空间和确定性策略，并成功训练智能体在合作及竞争场景中采用复杂的协作策略。MAAC [11], 是 SAC [10], 的扩展，它在评论家模块中引入注意力机制，有选择地关注从其他智能体收集的信息，从而在复杂的多智能体环境中实现了大规模扩展。此外，还有一系列基于价值函数分解的方法，它们以集中式端到端方式训练去中心化策略，并采用联合动作价值网络 [27]。许多后续的基于演员-评论家的多智能体强化学习算法针对各种问题进行了改进，例如 SAC [26] 旨在提升可扩展性，SEAC [4]旨在共享经验，LICA [44] 针对信用分配问题，TESSER-ACT [21] 旨在将评论家网络张量化，双层演员-评论家 [42] 针对智能体能力不均等的多智能体协调问题，DAC-TD [6]旨在隐私感知框架中训练智能体，VDAC [31] 旨在将价值分解框架与演员-评论家方法相结合，Scalable Actor Critic [17] 旨在随机智能体网络中实现可扩展学习，等等。然而，这些工作均未在其模型中考虑角色涌现范式。

另一方面，在人类社会中，角色分工对于完成任何复杂的集体任务而言都十分常见且高效 [3]。因此，许多多智能体系统也会将任务分解为预定义角色，从而降低设计复杂度并加快学习速度 [2, 5, 25, 30, 41]。近年来，Majumdar等人 [22]提出了一系列工作，在基于种群的训练范式中，通过解耦并自动调整个体目标与全局目标的权重，使MADDPG智能体学习不同的策略。Liu等人 [18]研究了动态组成下的协调问题，其中，拥有全局视角的教练智能体将个体策略分配给动态变化的玩家智能体。Jiang等人 [12] 通过生成动态内在奖励，促使个体性涌现，从而实现子任务专门化。Su等人 [32] 提出了另一种演员-评论家算法，使智能体能够在合作团队博弈中学习多样化策略。此外，针对机器人足球 [1, 15, 24, 36],、足球环境 [28], 、群体系统 [43], 和图像特征提取任务 [13] 等专门领域，也有一系列面向角色的多智能体强化学习研究，展示了如何通过将复杂策略分解为更简单的子策略来学习这些复杂策略。

相比之下，Wang等人提出了ROMA [37] 和RODE [38],这两种动态、自适应的面向角色的多智能体强化学习框架，它们基于QMIX [27] 构建，适用于相当一般的设置，并展示了团队中动态角色的涌现现象，验证任务为SMAC基准任务。最近，Liu等人提出了ROGC [19], , 这是一种基于图卷积的面向角色的多智能体强化学习框架。然而，它们都受到共享奖励设置的固有限制，即仅适用于合作任务。

我们的工作与 [37], 十分相似，主要区别在于，我们将对手建模与基于团队的角色学习融入演员-评论家框架，以利用在对手存在的情况下进行训练的优势。近期有少数研究探索了演员-评论家框架中的对手建模，例如，TDOM-AC [34] 用于学习时变对手模型，ROMMEO [35]用于学习最大熵正则化对手模型，PRR [40]则采用变分贝叶斯学习对手模型，等等。这些研究主要考虑双人竞争博弈的情形，其中每个智能体都会维护一个对手策略原型。然而，它们并未

符号	含义
$\mathcal{S}, \mathcal{A}$	状态和动作空间
$\mathcal{P}, \mathcal{R}$	转移函数和奖励函数
N, K	智能体和团队的总数
$\mathrm{N}(k)$	团队中的智能体集合 k
$\bar{o} = (o_1, \ldots, o_N) \in O$	观测向量
$\bar{a} = (a_1, \ldots, a_N) \in \mathcal{A}$	动作向量
$\pi = (\pi_1, \ldots, \pi_N)$	智能体策略
$\tau^t = (\tau_1^t, \ldots, \tau_N^t)$	截至时间 t的轨迹
$Q = (Q_1, \ldots, Q_N)$	价值-动作函数
$h_S = (h_S^1, \ldots, h_S^K)$	自身角色编码器
$h_O = (h_O^1, \ldots, h_O^K)$	对手角色编码器
$\rho = (\rho_1, \ldots, \rho_N)$	自我角色
$\hat{\rho} = (\hat{\rho}_1^o, \ldots, \hat{\rho}_N^o)$	预测对手角色
λ	衰减因子
$\mathcal{L}_Q$	评论家损失
$\mathcal{L}_{MI}$	互信息损失
$\mathcal{L}_D$	散度损失
$\mathcal{L}_{Opp}$	对手损失

表1：重要符号列表。

考虑团队竞争情形。将这些框架扩展到我们这种具备角色感知能力的竞争性团队设定，并非易事。

3 对手感知的基于角色的多智能体强化学习

本节将在团队竞争马尔可夫博弈场景下，提出用于具有涌现角色的多智能体策略学习的算法 RAC [20]。

3.1 预备知识

马尔可夫博弈由元组 S, A, P, R 表征，其中 S 表示状态集合，A 表示每个 N 智能体的动作集合。因此，联合动作空间为 A^N。状态转移函数将每个状态与联合动作组合映射为未来状态上的概率分布。$\mathrm{P}: \mathrm{S} \times \mathrm{A}^N \to \mathbb{P}(\mathrm{S})$。对于团队竞争博弈（也称为混合合作—竞争博弈 [20]），令 K 表示 K团队集合，$\mathrm{N}(k)$ 表示团队 k,$k \in \mathrm{K}$ 中智能体的集合。奖励函数 $\mathrm{R} = (r_1, \ldots, r_N): \mathrm{S} \times \mathrm{A}^N \to \mathbb{R}^N$ 为每个智能体 i 指定奖励方案（r_i）。逐智能体奖励函数可以对团队竞争博弈 [20],进行建模，而这正是本文关注的对象。此外，在分散式执行设置中，令 $o_i \in \mathrm{O}$ 表示智能体 i的观测。每个智能体 i学习一个策略 $\pi_i: \mathrm{O} \to \mathbb{P}(\mathrm{A})$，将其自身观测映射为其动作集合 A 上的概率分布。我们还将每个智能体 i 截至时间 t 的轨迹 $\tau_i(t) \in T = (\mathrm{O} \times \mathrm{A})^*$ 定义为一系列观测-动作对 $\{(o_i^j, a_i^j)\ j = 1, \ldots, t$。我们还将联合动作向量和观测向量分别表示为：$\bar{a} = (a_1, \ldots, a_N)$ 和 $\bar{o} = (o_1, \ldots, o_N)$。智能体旨在通过优化特定于智能体的累积折扣奖励函数 J,来学习其策略 π_i：

$$\mathcal{J}_i(\pi_i) = E_{\bar{a} \sim \pi, s \sim \mathcal{P}} \left[\sum_{u=0}^{\infty} \gamma^u r_i(s^u, \bar{a}^u) \right] \tag{1}$$

3.2 多智能体演员-评论家

演员-评论家方法及其扩展因具有灵活性高、奖励估计方差低等优点而被广泛用于多智能体强化学习 [10, 11]，这些优点有助于更快地学习策略参数。一种常见的扩展是软演员-评论家 [10](SAC)，其策略梯度的奖励（公式 1）加入了每个状态下策略分布的期望熵。因此，其奖励函数变为：

$$\mathcal{J}_i(\pi_i) = E_{\bar{a} \sim \pi, s \sim \mathcal{P}} \left[\sum_{u=0}^{\infty} \gamma^u (r_i(s^u, \bar{a}^u) + \alpha \mathcal{H}(\pi_i(s^u)) \right] \tag{2}$$

其中，α 是用户定义的正则化系数。请注意，他们并未明确讨论团队竞争场景下的多智能体强化学习。随后，多注意力演员-评论家（MAAC） [11]扩展了 SAC，以支持对具有个体奖励的智能体策略进行集中式训练。MAAC 中的学习算法包含两个主要组成部分：(1) 对 i^{th}智能体的策略 π_i 进行 SAC 风格的更新；(2) 对 i^{th} 智能体的评论家函数 Q_i进行注意力学习。策略更新使用以下策略梯度进行：

$$\begin{aligned} \nabla_{\pi_i} \mathcal{J}_i(\pi_i) = E_{\bar{o} \sim D, \bar{a} \sim \pi} [\nabla_{\pi_i} \log(\pi_i(a_i|o_i)(Q_i(\bar{o}, \bar{a}) \\ - b(\bar{o}, \bar{a}_{-i}) - \alpha \log(\pi_i(a_i|o_i)))] \end{aligned} \tag{3}$$

α是熵项的正则化常数。D是回放缓冲区，其中包含截至当前强化学习训练时刻 t，由策略生成的各回合每一步对应的元组（$\bar{o}$，$\bar{a}$，$\bar{r}$，$\bar{o}'$）。在观测 $\bar{o}$下执行动作 $\bar{a}$会产生所有智能体的奖励 $\bar{r}$和下一观测 $\bar{o}'$。b（o，a_{-i}）是按如下方式计算的多智能体基线。

$$b(\bar{o}, \bar{a}_{-i}) = E_{a_i \sim \pi_i(o_i)} [Q_i(\bar{o}, (a_i, \bar{a}_{-i}))] \tag{4}$$

$$= \sum_{a_i \in \mathcal{A}} \pi_i(a_i|o_i) Q_i(\bar{o}, (a_i, \bar{a}_{-i})) \tag{5}$$

该基线用于计算优势函数，这有助于解决多智能体信用分配问题 [11]。

智能体 i 的评论家函数 Q_i 取决于智能体 i的嵌入g_i（o_i，a_i）以及其他智能体的注意力加权平均奖励 x_i：

$$Q_i(\bar{o}, \bar{a}) = f_i(g_i(o_i, a_i), x_i) \tag{6}$$

$$x_i = \sum_{j \neq i} \alpha_{ij} \mathcal{V}(g_j(o_j, a_j)) \tag{7}$$

$$\alpha_{ij} \propto \exp(v_i^T W_i^T W_j v_j) \tag{8}$$

其中，f_i 和 V 是前馈神经网络。α_{ij} 是由智能体 i指向智能体 j 的动态注意力函数，取决于嵌入 $v_i = g_i(o_i, a_i)$ 和 $v_j = g_j(o_j, a_j)$。W_i和 W_j是针对各智能体的注意力函数“键”变换矩阵。评论家函数通过时序差分损失进行学习，该损失根据使用当前和过去策略收集的轨迹计算得出，这些轨迹存储在回放缓冲区 D 中。损失函数为

更新智能体 i 的评论家模块的形式如下：

$$\mathcal{L}_Q = \sum_{i=1}^{N} E_{(\bar{o},\bar{a},r,\bar{o}')\sim D}\left[(Q_i(\bar{o},\bar{a}) - y_i)^2\right] \text{, where}$$
$$y_i = r_i + \gamma E_{\bar{a}'\sim\bar{\pi}(\bar{o}')}\left[\bar{Q}_i(\bar{o}',\bar{a}') - \alpha\log(\bar{\pi}_i(a_i|o_i))\right] \quad (9)$$

其中，$\bar{Q}_i$和 $\bar{\pi}_i$ 分别是智能体 i的目标评论家和目标策略。在这项工作中，我们基于上述多智能体演员-评论家框架，将每个智能体中涌现角色的学习纳入其中。接下来，我们将介绍合作场景中一种常用的角色学习框架。

3.3 角色感知策略学习

角色已用于许多多智能体强化学习应用，以促进智能体之间的合作与协调 [1, 28, 36, 43]。最近，Wang等人提出了 ROMA, [37], 这是一种用于在合作博弈场景中学习涌现角色的多智能体强化学习框架。在 ROMA 中，角色的设计旨在体现两个关键属性：（1）**可辨识性：**角色应能通过其行为加以辨识，并具有足够的多样性以完成任务；（2）**专门化：** 具有相似角色的智能体应专门承担相似的职责。为确保可辨识性，该方法最大化角色分布 $p(\rho_i|o_i)$ 与轨迹条件角色分布 $p(\rho_i|\tau_i,\ o_i)$ 之间的互信息 $I(\rho_i;\ \tau_i)$。为确保专门化与多样性，该方法维护一个差异度模型 d_ϕ，用于衡量两个智能体的轨迹之间的距离，并寻求最大化 $I(\rho_i;\tau_j)\ + d_\phi(\tau_i,\ \tau_j)$。借助这些正则化项，ROMA 可确保智能体在每条轨迹中于团队内采用角色感知策略。

ROMA 建立在 QMIX [27],之上，QMIX 面向智能体根据联合观测和动作获得共享奖励的场景。ROMA 也采用 Q学习模式，无需额外的策略网络。它为每个智能体学习局部效用网络，其 Q_i 输出会传递至混合网络，以计算全局时序差分损失。在执行过程中，每个智能体都根据其局部效用函数导出的最优策略采取动作。在 ROMA 中，局部效用函数 $Q_i(o_i,\ a_i,\ \rho_i)$ 还将智能体的角色 ρ_i作为输入。角色 ρ_i被定义为一个带有分布的潜在（涌现）变量，该分布以智能体的轨迹 τ_i为条件。然而，ROMA 使用的 Q学习框架无法简单地扩展到竞争场景，因为将局部效用函数 Q_i组合成一个全局效用函数 Q，很大程度上依赖于所有智能体共享共同奖励这一假设；而在奖励混合的竞争场景中，这一假设并不成立。此外，局部效用函数 Q_i 仅将第 i 个智能体的观测和动作（o_i 和 a_i）作为输入，因此智能体无法从对手（奖励相互独立的其他智能体）的经验中学习。

在下一节中，我们将在演员-评论家框架下提出一种新颖的表述方法，用于在竞争场景中学习涌现角色，该方法更适合纳入混合奖励。在演员-评论家框架中，评论家函数 $Q_i(\bar{o},\ \bar{a})$ 将所有智能体的观察和动作作为输入，因此能够促进奖励解耦的智能体之间共享信息；此外，智能体还能利用训练期间可获得的共享对手信息，学习对手感知策略。

3.4 对手感知角色学习

算法1： RAC

初始化回放缓冲区， D
$T_{update} \leftarrow 0$
对于 $eps = 1\cdots$个回合，**执行**
  $\tau_i^0 \leftarrow 0$
  $D_i \leftarrow \emptyset, \forall i$
  重置环境，并获取每个智能体的初始 o_i^1 i
  对于 $t = 1\cdots$回合长度 **执行**
    使用 (11) 为所有智能体采样角色 ρ^t，$\hat{}\rho^t$
    从各自的策略中采样动作
    $a_i^t \sim \pi_i(o_i^t, \rho_i^t, \hat{\rho}^{o}{}_{,i}^{\ t})$ 对每个智能体 i
    每个智能体从环境接收 o_i^{t+1}，r_i^t
    $\tau_i^t \leftarrow \tau_i^{t-1} + (o_i^t,\ a_i^t)$
    $D_i \leftarrow D_i + (o_i^t, \tau_i^{t-1}, a_i^t, r_i^t, o_i^{t+1})$
  end
  $T_{update} = T_{update} + 1$
  如果 $T_{update} \geq$ 每次更新的最小步数 **那么**
    对于 $j = 1\cdots$更新次数 **执行**
      $H_B \leftarrow B$ 从 D采样的回合
      更新(H_B)
    end
    $T_{update} \leftarrow 0$
  end
end

算法 2： 更新(H_B)

选择 $t_1, \cdots t_B$，每个 B 的起始时间步
回合
$\mathrm{L}_{MI} \leftarrow 0, \mathrm{L}_D \leftarrow 0, \mathrm{L}_{Opp} \leftarrow 0$
对于 $j = 1\cdots$回合长度 **执行**
  $S \leftarrow \emptyset$
  对于 $k = 1\cdots B$**执行**
    $S_k \leftarrow (\bar{o}^{t_k+j,\ k},\ \bar{\tau}^{t_k-1+j,\ k},\ \bar{a}^{t_k+j,\ k},\ \bar{r}^{t_k+j,\ k},\ \bar{o}^{t_k+1+j,\ k})$
    $S \leftarrow S + S_k$
    $\rho^{j,k}, \hat{\rho}^{j,k}, \mathcal{L}_{MI}^{j\ k}, \mathcal{L}_D^{j\ k}, \mathcal{L}_{Opp}^{j\ k} \leftarrow Role(S_k)$
    $\mathrm{L}_{MI} \leftarrow \mathrm{L}_{MI} + \mathrm{L}_{MI}^{j\ k}$,
    $\mathrm{L}_D \leftarrow \mathrm{L}_D + \mathrm{L}_D^{j\ k}$,
    $\mathrm{L}^{Opp} \leftarrow \mathrm{L}^{Opp} + \mathrm{L}_{Opp}^{j\ k}$,
  end
  使用以下内容更新评论家网络、策略和目标 $S, \rho^{j,:}, \hat{\rho}^{j,:}$
end
$\psi_{Role} \leftarrow \mathrm{argmin}\psi(\mathrm{L}_{MI} + \mathrm{L}_D + \mathrm{L}^{Opp})$

我们提出了 RAC，即一种基于角色的演员—评论家算法，旨在用于团队竞争性博弈中的对手感知学习。在这一设定中，团队内的智能体既获得个人奖励，也获得共享奖励；而不同团队的智能体不会获得任何共享奖励。本节为简化起见，介绍该算法在两个团队的设定下的情况，每个团队包含

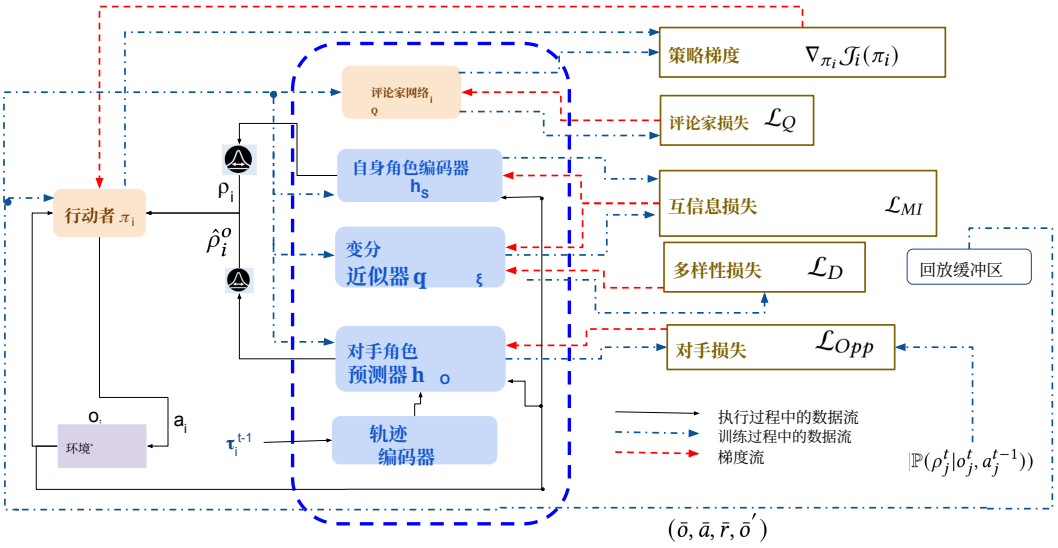


图1：我们方法的示意图。自身角色编码器 h_S 为第 i 个智能体生成角色 ρ_i，对手角色预测器 h_O 则预测对手的角色 $\hat{\rho}_i^o$。策略 π_i 利用自身观测 o_i 以及生成的角色编码（$\rho_i, \hat{\rho}_i^o$）生成动作 a_i。评论家网络 Q_i 利用 $((o_i, \rho_i, \hat{\rho}_i^o)\ a_i)$ 计算状态-动作价值。评论家损失用于训练评论家模块，而互信息损失、多样性损失和对手损失用于训练角色编码器模块。该框架可以端到端方式进行训练。

算法3： 角色(S)

根据公式(11)获得角色 $\rho^t, \hat{\rho}^t$

使用公式(15)、(16)和(17)计算 L_{MI}、L_D 和 L^{Opp}

返回 ρ^t、$\hat{\rho}^t$、$\mathcal{L}_{MI}$、$\mathcal{L}_D$、$\mathcal{L}^{Opp}$

任意数量的合作智能体，它们依据单个游戏的规则（见第 4 节）相互竞争。在这一设定中，一个值得探讨的问题是：学习对手感知的基于角色的策略是否有利于各团队。

RAC 基于上述 MAAC 框架构建，并使用智能体专属的策略网络 π_i 和评论家网络 Q_i。此外，RAC 还采用了另外两个网络：

(1) h_S：自我角色编码网络 (2) h_O：对手角色预测网络

另一个循环神经网络（门控循环单元）用于维护截至时间 t 的轨迹编码 τ_i^t。为简化起见，在本文余下部分，我们将符号 τ_i^t 重载，用来表示截至时间 t 的轨迹嵌入。直观而言，智能体 i 的策略取决于其观测 o_i、潜在角色 ρ_i 以及对对手智能体角色的预测 $\hat{\rho}_i^o = [\hat{\rho}_j]_{j \in Opp(i)}$，其中 $Opp(i)$ 表示智能体 i 的对手智能体集合。因此，$\pi_i(o_i, \rho_i, \hat{\rho}_i^o)$ 计算动作 a_i 的概率。自我角色编码网络 h_S 和对手角色预测网络 h_O 以当前观测 o_i^t 和智能体 h_S 的过去动作 a_i^{t-1}，或智能体 h_O 的轨迹编码 τ_i^{t-1} 为输入，并输出相应的高斯分布

roles:

$$(\mu_{\rho_i^t}, \sigma_{\rho_i^t}) = h_S(o_i^t, a_i^{t-1})\ ;\ \mathbb{P}(\rho_i^t | o_i^t, a_i^{t-1}) = \mathcal{N}(\mu_{\rho_i^t}, \sigma_{\rho_i^t}) \quad (10)$$

$$\hat{\mu}_{\rho_j^t}, \hat{\sigma}_{\rho_j^t} = h_O(o_i^t, \tau_i^{t-1}) \forall j \neq i\ ;\ \hat{\mathbb{P}}(\hat{\rho}_j^t | o_i^t, \tau_i^{t-1}) = \mathcal{N}(\hat{\mu}_{\rho_i^t}, \hat{\sigma}_{\rho_i^t}) \quad (11)$$

最后，智能体 i 的评论家网络 Q_i 以联合观测 $\bar{o}$、联合动作 $\bar{a}$、当前智能体 ρ_i 的角色嵌入以及其对手的角色预测 $\hat{\rho}_i^o$, $Q_i(\bar{o}, \rho_i, \hat{\rho}_i^o, \bar{a})$ 为输入，并输出后续价值的估计值。图 1 展示了 RAC 系统的整体架构。

在上述网络架构中，各智能体的策略 π_i 通过标准策略梯度步骤进行训练，而评论家网络 Q_i、自身角色编码器网络 h_S 和对手角色预测器网络 h_O 的参数，则通过评论家的集中式训练进行更新。为了更新智能体 i的策略网络，式 (3) 中的策略梯度进一步修改如下：

$$\nabla_{\pi_i} J_i(\pi_i) = E_{\bar{o} \sim D, \bar{a} \sim \pi, \rho \sim h_S, \hat{\rho} \sim h_O} \left[\nabla_{\pi_i} \log(\pi_i(a_i | o_i, \rho_i, \hat{\rho}_i^o) \right.$$
$$\left. (Q_i(\bar{o}, \rho_i, \hat{\rho}_i^o, \bar{a}) - b(\bar{o}, \rho_i, \hat{\rho}_i^o, \bar{a}_{-i}) - \alpha \log(\pi_i(a_i | \bar{o}_i, \rho_i, \hat{\rho}_i^o))) \right] \quad (12)$$

相应地，基线计算也修改如下：

$$b(\bar{o}, \rho_i, \hat{\rho}_i^o, \bar{a}_{-i}) = \mathbb{E}_{a_i \sim \pi_i(o_i, \rho_i, \hat{\rho}_i^o)} \left[Q_i(\bar{o}, \rho_i, \hat{\rho}_i^o, (a_i, \bar{a}_{-i})) \right]$$
$$= \sum_{a_i \in A_i} \pi_i(a_i | o_i, \rho_i, \hat{\rho}_i^o) Q_i(\bar{o}, \rho_i, \hat{\rho}_i^o, (a_i, \bar{a}_{-i})) \quad (13)$$

虽然 π_i 的参数是通过策略梯度更新进行训练的，但其他网络的参数则通过最小化

下文所述的聚合损失函数联合学习。首先，将式（9）中定义的评论家损失 L$_Q$修改如下：

$$\mathcal{L}_Q = \sum_{i=1}^{N} \mathbb{E}_{(\bar{o},\bar{a},\bar{r},\bar{o}')\sim D,\rho_i\sim h_S,\hat{\rho}_i^o\sim h_O}\left[(Q_i(\bar{o},\rho_i,\hat{\rho}_i^o,\bar{a})-y_i)^2\right], \text{where}$$

$$y_i = r_i + \gamma \mathbb{E}_{\bar{a}'\sim\bar{\pi},\rho_i'\sim h_S,\hat{\rho}_i^{o,\prime}\sim h_O}\Big[\bar{Q}_i(\bar{o}',\rho_i',\hat{\rho}_i^{o,\prime},\bar{a}') - \alpha\log(\bar{\pi}_i(a_i'|o_i',\rho_i',\hat{\rho}_i^{o,\prime}))\Big] \quad (14)$$

现在，我们介绍用于训练自身与对手角色编码器模块的损失函数。为了确保角色与轨迹之间的可辨识性，我们希望最小化条件熵 $H(\rho_i^t|o_i^t,\ \tau_i^{t-1})$，但这通常难以处理。因此，我们对以轨迹为条件的角色分布维持一个变分近似：$q_\xi(\rho_i^t|o_i^t,\ a_i^{t-1},\ \tau_i^{t-1})$，并最大化 $q_\xi(\rho_i^t|o_i^t,\ a_i^{t-1},\ \tau_i^{t-1})$ 与 $\mathbb{P}(\rho_i^t|o_i^t,\ a_i^{t-1})$ 之间的互信息。同时，我们最小化角色分布 $\mathbb{P}(\rho_i^t|o_i^t,\ a_i^{t-1})$ 的熵，以在角色编码器网络中编码历史信息。因此，相应的互信息损失 L$_{MI}$ 定义如下：

$$\mathcal{L}_{MI} = \sum_{i=1}^{N} \mathbb{E}_{\tau_i^{t-1},o_i^t\sim D}\left[D_{KL}(\mathbb{P}(\rho_i^t|o_i^t,a_i^{t-1})\| q_\xi(\rho_i^t|o_i^t,a_i^{t-1},\tau_i^{t-1}))\right] + H(\mathbb{P}(\rho_i^t|o_i^t,a_i^{t-1})) \quad (15)$$

此处，D_{KL} 表示 KL散度，$H(\cdot)$ 表示熵函数。为了促进团队内智能体行为的多样性，我们引入一种新的多样性损失 L$_D$，定义如下：

$$\mathcal{L}_D = \sum_{k\in\mathcal{K}} \sum_{(i,j)\in\mathcal{N}(k)} \mathbb{E}_{(\bar{\tau}^{t-1},\bar{o}^t)\sim D}\left[q_\xi(\rho_i^t|\tau_j^{t-1},o_j^t,a_i^{t-1})\right] \quad (16)$$

这里，K 是团队的集合，N (k) 是团队 k 中的智能体。该损失会惩罚将 ρ_i 识别为同一团队中另一智能体的轨迹τ_j 的概率。最后，为了训练对手角色预测网络 h_O，我们将对手损失 L^{Opp} 定义为：

$$\mathcal{L}_{Opp} = \sum_{\substack{i,j:team(i)\\ \neq team(j)}} \mathbb{E}_{(\bar{\tau}^{t-1},\bar{o}^t)\sim D}\left[D_{KL}(\hat{\mathbb{P}}(\rho_j^t|o_i^t,\tau_i^t)\|\mathbb{P}(\rho_j^t|o_j^t,a_j^{t-1}))\right] \quad (17)$$

该损失会惩罚从智能体 i 的角度对智能体 j 的预测对手角色分布 $\hat{\mathbb{P}}(\rho_j^t|o_i^t,\ \tau_i^t)$ 与智能体 j 的自身角色分布 $\mathbb{P}(\rho_j^t|o_j^t,\ a_j^{t-1})$ 之间的差异。我们将用于训练角色网络的复合损失函数定义为：

$$\mathcal{L}_{Role} = \mathcal{L}_D + \mathcal{L}_{MI} + \mathcal{L}_{Opp} \quad (18)$$

因此，L^Q+L$_{Role}$ 是用于训练RAC模型组合评论家模块的总损失。然而，我们观察到，在训练后期，角色网络的演变会导致策略学习不稳定，从而降低平均奖励。通过对角色损失L$_{Role}$使用指数衰减权重，可以缓解这一现象。总评论家损失变为：

$$\mathcal{L}_{tot} = \mathcal{L}_Q + \lambda^{\frac{u}{C}} \times \mathcal{L}_{Role} \quad (19)$$

其中，$\lambda \in [0,1]$ 是衰减因子，u 是训练回合数，C 是影响跨回合衰减速率的常数。算法1包含我们算法的伪代码。表1列出了所有变量。

4 实验结果

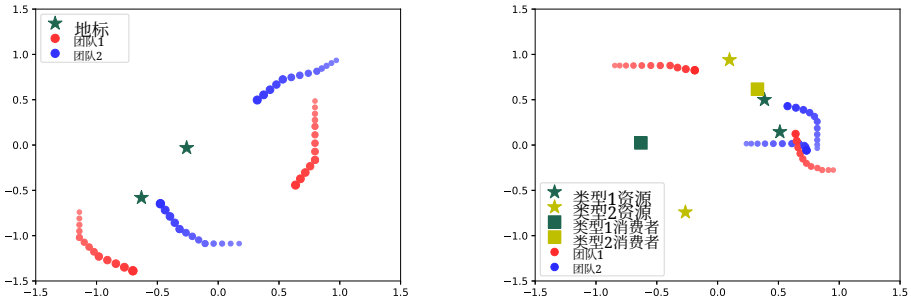
在本节中，我们研究两个主要目标。首先，我们考察采用角色网络如何影响算法与基线算法直接竞争时的性能。其次，我们考察角色网络在团队内部动态形成新兴角色的有效性。为此，我们采用了两个游戏，即 Touch-Mark 和 Market，它们是在 MPE [20, 23] 中设计的。

Touch-Mark [14]: 这是一款回合制棋盘游戏，由两个团队组成，每个团队包含多个智能体（本例中为 2 个）和两个静态地标。每回合开始时，所有智能体和地标的位置都会随机确定，每个团队都力求比对方团队的任何成员更早抵达至少一个地标。智能体初始速度为零，并逐渐加速，直至达到允许的最大速度。当某个智能体抵达地标时，该回合结束。获胜智能体所属的团队（即团队中的两名成员）将获得大量奖励 r_l，对立团队则受到 $-r_l$ 惩罚。此外，每个智能体在每个时间步都会受到少量惩罚，惩罚幅度与其到最近目标的距离成正比，以引导它向最近目标移动。此外，智能体可以与对手团队的智能体发生碰撞，使其偏离原定路线。碰撞会使两个智能体暂时停止移动。引入这一机制，是为了让智能体能够阻止对手团队的智能体抵达地标，从而帮助队友抢先抵达地标。

市场： 这是一款新颖的回合制棋盘游戏，分为两个团队，每个团队由多名玩家组成（本例中为 2 名）。每回合开始时，棋盘上会随机放置四个资源，每种资源各两个，同时每种资源各有一名消费者。智能体需要获取资源并将其送到对应的消费者处。获取资源后，智能体会持续受到惩罚，惩罚程度与其到相应消费者的距离成正比，以引导其前往对应的消费者处。获取和送出资源都会为相应团队带来高额奖励。回合的终止条件是达到 50 步的时间限制。消费者只有在收到一个对应类型的资源后才会继续存在。因此，要取得最高分，一个团队必须在对手之前获取两种类型的资源，并将它们分别送到对应的消费者处。

实验设置 我们使用单层前馈神经网络，后接一个门控循环单元，以获得轨迹的初始嵌入。对于每个团队，分别为 $\mathbb{P}(\rho_i^t|o_i^t,a_i^{t-1})$、$\hat{\mathbb{P}}(\rho_j^t|o_i^t,a_i^{t-1})$ 和 $q_\xi(\rho_i^t|\tau_i^{t-1},o_i^t)$ 使用双层多层感知机。评论家网络和策略网络的架构遵循 MAAC [11]，但对每个智能体而言，评论家模块和策略模块的输入 i 都经过修改，以包含该智能体的角色元组，其中包括自我角色 ρ_i^t 和预测对手角色$\hat{\rho}^o{}_i{}^t$。

基线方法 我们的主要基线方法是 MAAC [11]。我们还将本方法与 RAC$_{Team}$ 进行比较。RAC 与本方法相同，但在训练阶段不在团队之间共享信息。这模拟了竞争团队分别采用某种



(a) Touch-Mark (b) 市场

图2：Touch-Mark（左）是一种二维棋盘游戏，每个团队的成员都试图在对手团队的任何智能体到达任一地标之前，先到达其中一个地标。在市场（右）中，团队拾取散布在棋盘上的资源，并在对手完成相同操作之前将资源送至相应的消费者。

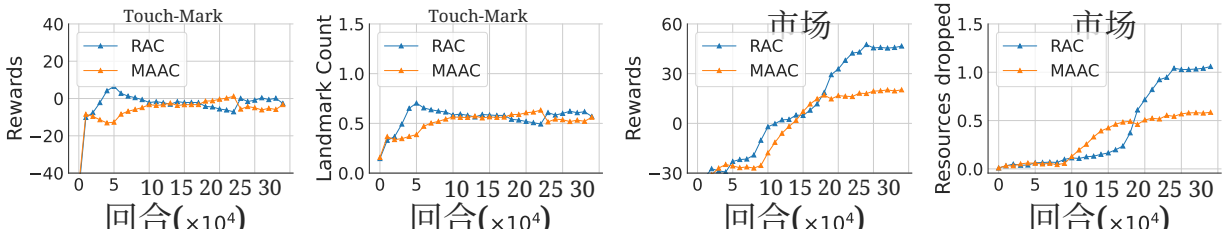


图3：性能比较。我们让 RAC 智能体与 MAAC 智能体进行对战；这些智能体分别通过自博弈训练了300000 训练步数。每经过1000 步，我们就让它们相互对战一次，并报告性能；结果取4组锦标赛对局的平均值。对于 Touch-Mark，我们报告平均团队奖励，以及每个团队到达地标的比例。对于市场，我们报告平均团队奖励，以及每个团队从资源来源处取走资源后，将其投放给各自消费者的比例。

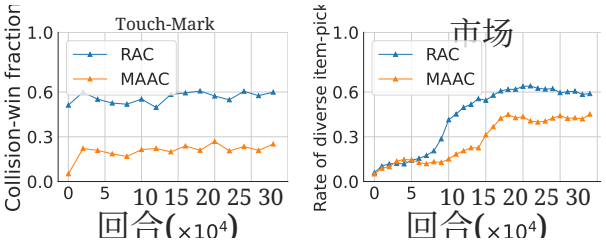


图4：角色涌现。对于 Touch-Mark，我们观察到，RAC 智能体造成的碰撞多于 MAAC 智能体。在市场中，我们观察到，RAC 智能体选择拾取的物品种类比 MAAC 智能体更多。

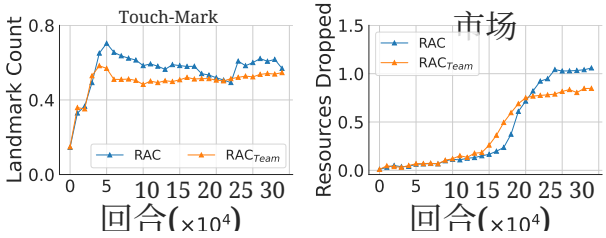


图5：我们将 RAC 与 RAC_{Team}进行比较；后者是移除了对手建模的 ROMA 扩展版本。然而，我们观察到，就性能而言，RAC 仍优于 RAC_{Team}。

协作式角色导向算法（如 ROMA），并且仅放弃在对手存在的情况下进行训练所带来的优势的情形。

指标 我们报告每回合的累积团队奖励，以衡量算法的性能。作为定性指标，我们报告 Touch-Mark 中每个团队每回合触及地标的比率，以及平均投放资源数量

Market 中每个团队的回合数。由于 Market 中有 2 个资源，每回合减少的平均资源量介于 0 到 2 之间。所有指标均为对所用的全部随机种子取平均后的结果。

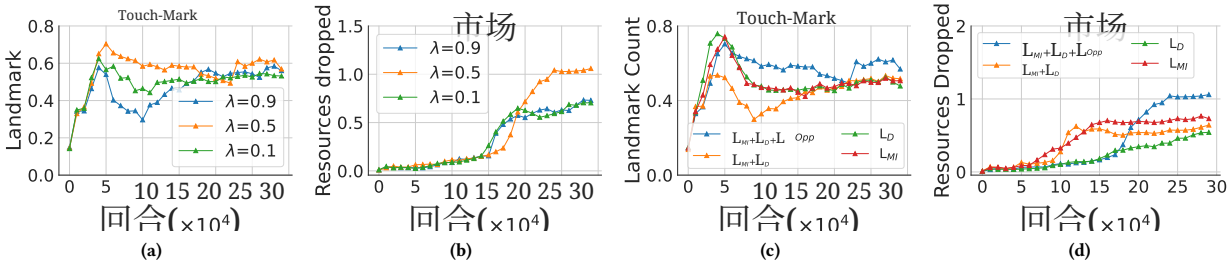


图6：（左）超参数调优。此处在 $[0,1]$ 中改变衰减参数 λ，并报告 RAC 团队与一组随机抽取的 MAAC 模型对战时的表现。（右）消融实验。报告了 RAC 的各种消融版本的表现，结果表明，在这两个环境中，三个不同的目标函数对于取得最佳表现都是必不可少的。

4.1 结果

直接对抗赛： 在此，我们通过让 RAC 和 MAAC 直接相互对战来比较二者。我们使用从随机种子开始、经过 3×10^5 个训练回合自博弈训练的 RAC 和 MAAC 团队。在训练期间，我们每隔固定间隔让它们相互对战 1000 个回合。我们对 4 对 RAC 和 MAAC 团队重复这一实验，并在图 3 中报告 Touch-Mark（左）和 Market（右）的平均指标。我们观察到，随着训练的推进，RAC 始终优于 MAAC。在 Market 中，性能差距相当显著，而在 Touch-Mark 中，提升则微乎其微。我们对此作如下解释：Market 的设定相对复杂，如果没有明确的子任务专门化辅助，智能体很难采取策略性行动。RAC 的角色驱动策略学习框架似乎在此发挥了作用：它缩小了需要探索的观测-动作空间，并通过制定对手感知型策略行为，迅速引导智能体实现更好的策略学习。Touch-Mark 的设定较为简单，即使没有外部辅助来进行隐式角色划分或子任务专门化，智能体也能较快学会有效策略，因此 RAC 相较于 MAAC 仅有微小提升。此外，我们观察到，随着 Touch-Mark 训练接近尾声，RAC 智能体的表现有所下降。我们对此现象作如下解释：RAC 加速了 Touch-Mark 中策略行为的学习，具体而言，它将碰撞和追逐地标的职责分配给团队成员。然而，随着 MAAC 智能体在下一阶段的训练中迅速掌握该策略，RAC 的优势几乎消失。因此，随着 MAAC 团队专业能力的提升，RAC 团队的表现有所下降，最终仅略胜 MAAC 智能体。

角色学习： 此外，我们探究 RAC 是否确实在帮助队友选择多样化角色。我们还研究 RAC 智能体的策略行为。在 Touch-Mark 中，正如我们之前提到的，虽然碰撞会带来较小的惩罚，但可以策略性地利用与对手的碰撞，延缓对手智能体到达地标。同样，在 Market 中，如果队友选择获取不同的资源，他们就能在未来获得更高的奖励。为了检验算法能否训练智能体采取此类策略性动作，我们汇总了两款游戏中的角色区分统计数据，如图 4 所示。对于 Touch-Mark，我们报告获胜回合中，除触碰地标的智能体之外的智能体与对手发生碰撞的比例。同样，对于 Market，我们报告团队成员采集不同类型资源的回合比例。在这两种情况下，我们都发现，与 MAAC 智能体相比，RAC 智能体更倾向于采取策略性动作，这或许可以解释 RAC 在直接竞争中表现更好的原因。这种多样化行为背后的关键因素是 L_D 和 L^{Opp} 的运用，它们分别促使角色网络为队友分配多样化且能考虑对手的角色分布。

与 RAC 的比较$_{Team}$ 扩展 [37]： 图5比较了 RAC 与 RAC$_{Team}$。后者是将 ROMA [37] 简单扩展到演员-评论家框架的结果：我们在 [37]中使用了用于角色学习的两个正则化项，但移除了对手建模。相反，两支团队分别使用简化角色模块增强型 MAAC 训练智能体，而每支团队采用共享奖励设置。最后，分别让 RAC 和 RAC$_{Team}$训练出的智能体与从一组共同的 MAAC 团队中随机选出的 MAAC 团队对战，并在图5中绘制平均比较结果。RAC 的优越性验证了我们的对手感知型角色导向框架，相较于将 ROMA [37] 直接扩展到演员-评论家，为竞争性团队博弈中的智能体提供了额外的指导层，从而进一步彰显了我们方法的动机。

超参数（λ）调优： 图 [6a-6b]展示了针对随机选取的 MAAC 团队时，RAC 团队在不同衰减参数 λ下的表现。在 Touch-Mark 和市场中，表现都会因所选的 λ而显著不同，其中 $\lambda = 0.5$ 的效果最好。$\lambda = 0.9$ 会导致角色损失衰减得较慢，从而使策略学习更加不稳定；而 $\lambda = 0.1$ 则会导致衰减过快。这项实验验证了我们基于现实经验提出的直觉假设：角色演化持续超过一定程度会损害团队表现。

与消融实验的比较： 图 [6c-6d] 比较了 RAC 与 RAC 的三种消融版本，后者分别仅最小化 L_D、L_{MI} 或 $L_D + L_{MI}$，而 RAC 最小化 $L_D + L_{MI} + L^{Opp}$。在这两种博弈中，RAC 的表现均优于其他方法。在市场博弈中，我们还观察到，与另外两种基线方法相比，L_{MI} 能够学习到与轨迹关联性更强、更易辨识的角色，从而更好地训练投放技能。

5 结论与未来工作

我们提出了一种面向团队竞赛、用于演员-评论家框架的对手感知的基于角色学习算法。该算法在演员-评论家框架中结合了自身角色编码器和对手角色预测器，以学习最优策略。我们在两种场景中分析了该方法，并展示了它如何提升学习质量。未来，我们计划将该算法扩展到更一般的设置，使多个团队能够相互竞争。

参考文献

[1] Samuel Barrett 和 Peter Stone。2015。与复杂领域中的未知队友合作：临时团队合作的机器人足球案例研究。载于*AAAI*人工智能会议论文集，第29卷。[2] N Bonjean、W Mefteh、MP Gleizes、C Maurel 和 F Migeon。2014。ADELFE 2.0。《面向智能体的设计流程手册》。M. Cossentino、V. Hilaire、A. Molesini 和 V. Seidita。[3] Eamonn Butler。2012。《浓缩的国富论》。独立研究中心。[4] Filippos Christianos、Lukas Schäfer 和 Stefano Albrecht。2020。用于多智能体强化学习的共享经验演员-评论家。神经信息处理系统进展33（2020），10707–10717。[5] Massimo Cossentino、Salvatore Gaglio、Luca Sabatucci 和 Valeria Seidita。2005。PASSI 和 Agile PASSI 多智能体系统元模型及其统一方案的比较。载于中东欧多智能体系统国际会议。施普林格，183–192。[6] Martin Figura、Yixuan Lin、Ji Liu 和 Vijay Gupta。2022。基于时序差分误差聚合的协作式演员-评论家。*arXiv*预印本 *arXiv:2207.12533*（2022）。[7] Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。2018。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。[8] Jakob Foerster、Nantas Nardelli、Gregory Farquhar、Triantafyllos Afouras、Philip HS Torr、Pushmeet Kohli 和 Shimon Whiteson。2017。稳定深度多智能体强化学习的经验回放。载于国际机器学习会议。PMLR，1146–1155。[9] Jayesh K Gupta、Maxim Egorov 和 Mykel Kochenderfer。2017。利用深度强化学习实现协作式多智能体控制。载于自主智能体与多智能体系统国际会议。施普林格，66–83。[10] Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。2018。软演员-评论家：采用随机演员的离策略最大熵深度强化学习。载于国际机器学习会议。PMLR，1861–1870。[11] Shariq Iqbal 和 Fei Sha。2019。用于多智能体强化学习的演员-注意力-评论家。载于国际机器学习会议。PMLR，2961–2970。[12] Jiechuan Jiang 和 Zongqing Lu。2021。个体性的涌现。载于国际机器学习会议。PMLR，4992–5001。[13] Shihab Hamad Khaleefah、Salama A Mostafa、Aida Mustapha、Rozanawati Daraman、Mohammad Faidzul Nasrudin 和 Shahrinaz Ismail。2021。探索智能体与多智能体在图像特征提取中的作用。载于*2021*年第*4*届智能体、多智能体系统与机器人国际研讨会（*ISAMSR*）。IEEE，76–83。[14] Paramita Koley、Aurghya Maiti、Sourangshu Bhattacharya 和 Niloy Ganguly。2022。通过强化学习辅助激励机制弥补竞争不平等。*IEEE*计算社会系统汇刊（2022）。[15] David L Leottau、Javier Ruiz-del Solar、Patrick MacAlpine 和 Peter Stone。2015。应用于机器人足球个体行为的分层学习策略研究。载于机器人世界杯。施普林格，290–302。[16] Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。2016。利用深度强化学习实现连续控制。载于国际学习表征会议（海报）。[17] Yiheng Lin、Guannan Qu、Longbo Huang 和 Adam Wierman。2021。随机网络化系统中的多智能体强化学习。神经信息处理系统进展34（2021），7825–7837。[18] Bo Liu、Qiang Liu、Peter Stone、Animesh Garg、Yuke Zhu 和 Animashree Anandkumar。2021。用于动态团队组成的教练-球员多智能体强化学习。*arXiv*预印本 *arXiv:2105.08692*（2021）。[19] Yuntao Liu、Yuan Li、Xinhai Xu、Donghong Liu 和 Yong Dou。2022。ROGC：基于面向角色的图卷积的多智能体强化学习。载于*2022*年*IEEE*多媒体与展览国际会议（*ICME*）。IEEE，1–6。[20] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、OpenAI 的 Pieter Abbeel 和 Igor Mordatch。2017。面向混合协作竞争环境的多智能体行动者-评论家。载于神经信息处理系统进展。6379–6390。[21] Anuj Mahajan、Mikayel Samvelyan、Lei Mao、Viktor Makoviychuk、Animesh Garg、Jean Kossaifi、Shimon Whiteson、Yuke Zhu 和 Animashree Anandkumar。2021。TESSERACT：用于多智能体强化学习的张量化行动者。载于国际机器学习会议。PMLR，7301–7312。[22] Somdeb Majumdar、Shauharda Khadka、Santiago Miret、Stephen Mcaleer 和 Kagan Tumer。2020。面向样本高效多智能体协调的进化强化学习。载于国际机器学习会议。PMLR，6651–6660。[23] Igor Mordatch 和 Pieter Abbeel。2018。多智能体群体中扎根式组合语言的涌现。载于AAAI人工智能会议论文集，第32卷。[24] Ori Ossmy、Justine E Hoch、Patrick MacAlpine、Shohan Hasan、Peter Stone 和 Karen E Adolph。2018。多样性制胜：踢足球机器人与婴儿行走。《神经机器人学前沿》12（2018），19。[25] Lin Padgham 和 Michael Winikoff。2002。Prometheus：一种开发智能体的方法。载于智能体导向软件工程国际研讨会。施普林格，174–185。[26] Guannan Qu、Yiheng Lin、Adam Wierman 和 Na Li。2020。具有平均奖励的网络化系统可扩展多智能体强化学习。《神经信息处理系统进展》33（2020），2074–2086。[27] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。2018。QMIX：用于深度多智能体强化学习的单调价值函数因子分解。载于国际机器学习会议。4295–4304。[28] Julien Roy、Paul Barde、Félix Harvey、Derek Nowrouzezahrai 和 Chris Pal。2020。通过策略正则化促进多智能体深度强化学习中的协调。《神经信息处理系统进展》33（2020），15774–15785。[29] Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero 和 Yung Yi。2019。Qtran：学习通过变换进行因子分解，以实现合作式多智能体强化学习。载于国际机器学习会议。PMLR，5887–5896。[30] Nikolaos Spanoudakis 和 Pavlos Moraitis。2010。使用ASEME方法进行模型驱动的智能体系统开发。载于智能体导向软件工程国际研讨会。施普林格，106–127。[31] Jianyu Su、Stephen Adams 和 Peter Beling。2021。价值分解多智能体行动者-评论家。载于AAAI人工智能会议论文集，第35卷。11352–11360。[32] Kefan Su 和 Zongqing Lu。2022。散度正则化多智能体行动者-评论家。载于国际机器学习会议。PMLR，20580–20603。[33] Sainbayar Sukhbaatar、Rob Fergus 等。2016。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）。[34] Yuan Tian、Klaus-Rudolf Kladny、Qin Wang、Zhiwu Huang 和 Olga Fink。2022。具有时变对手模型的多智能体行动者-评论家。arXiv预印本 arXiv:2204.05576（2022）。[35] Zheng Tian、Ying Wen、Zhichen Gong、Faiz Punakkath、Shihao Zou 和 Jun Wang。2019。具有最大熵目标的正则化对手模型。arXiv预印本 arXiv:1905.08087（2019）。[36] Daniel Urieli、Patrick MacAlpine、Shivaram Kalyanakrishnan、Yinon Bentor 和 Peter Stone。2011。优化相互依赖的技能：模拟三维人形机器人足球案例研究。载于AAMAS，第11卷。769。[37] Tonghan Wang、Heng Dong、Victor Lesser 和 Chongjie Zhang。2020。ROMA：具有涌现角色的多智能体强化学习。载于第37届国际机器学习会议论文集。[38] Tonghan Wang、Tarun Gupta、Anuj Mahajan、Bei Peng、Shimon Whiteson 和 Chongjie Zhang。2020。RODE：学习角色以分解多智能体任务。载于国际学习表征会议。[39] Tonghan Wang、Jianhao Wang、Chongyi Zheng 和 Chongjie Zhang。2019。通过通信最小化学习近似可分解价值函数。载于国际学习表征会议。[40] Ying Wen、Yaodong Yang、Rui Luo、Jun Wang 和 Wei Pan。2019。用于多智能体强化学习的概率递归推理。arXiv预印本 arXiv:1901.09207（2019）。[41] Michael Wooldridge、Nicholas R Jennings 和 David Kinny。2000。用于面向智能体的分析与设计的 Gaia 方法。《自主智能体与多智能体系统》3，3（2000），285–312。[42] Haifeng Zhang、Weizhe Chen、Zeren Huang、Minne Li、Yaodong Yang、Weinan Zhang 和 Jun Wang。2020。用于多智能体协调的双层行动者-评论家。载于AAAI人工智能会议论文集，第34卷。7325–7332。[43] Tianle Zhang、Zhen Liu、Zhiqiang Pu、Tenghai Qiu 和 Jianqiang Yi。2021。具有角色涌现的分层协作式群体策略学习。载于2021年IEEE计算智能研讨会系列（SSCI）。IEEE，1–8。[44] Meng Zhou、Ziyu Liu、Pengwei Sui、Yixuan Li 和 Yuk Ying Chung。2020。为合作式多智能体强化学习学习隐式信用分配。《神经信息处理系统进展》33（2020），11853–11864。