强化学习代理
用于二维射击游戏

Thomas Ackermann[1] , Moritz Spang[2] , Hamza A. A. Gardi[2,3]
{[1] Faculty for Mathematics, [2] Department of Electrical Engineering and Information Technology, [3] IIIT at ETIT},
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

摘要—强化学习代理在复杂游戏环境中常常遭受稀疏奖励、训练不稳定和样本效率低下的问题。本文提出了一种结合离线模仿学习与在线强化学习的混合训练方法，适用于二维射击游戏代理。我们实现了一个多头神经网络，其中行为克隆和 Q 学习有独立的输出，并通过带有注意力机制的共享特征提取层统一。

初步使用纯深度 Q 网络的实验显示了显著的不稳定性，代理经常倒退回不良策略，尽管偶尔表现出色。为了解决这个问题，我们开发了一种混合方法论，该方法首先对基于规则的代理演示数据进行行为克隆，然后过渡到强化学习。

我们的混合方法在对抗基于规则的对手时，胜率持续高于 70%，显著优于纯强化学习方法，后者表现出高方差和频繁性能退化。多头架构能够在保持训练稳定性的同时有效进行知识转移。结果表明，在复杂多智能体环境中，结合基于演示的初始化与强化学习优化为开发游戏 AI 代理提供了一种稳健的解决方案，证明了单纯探索是不够的。

***Index Terms*—强化学习，游戏代理，视频游戏，射击游戏。**

I. 介绍

强化学习（RL）作为人工智能中的基础范式，通过与动态环境的交互训练智能体进行顺序决策。通过优化行为以奖励反馈而非显式监督的方式，RL 使开发能够自主学习复杂任务的系统成为可能，这一特性对于机器人、自主系统、个性化医疗、金融等领域至关重要 [1], [2]。

强化学习的一个最显著优势是其在现实世界问题中的广泛适用性，这些问题的最优策略难以预先制定。例如，在机器人技术中，强化学习有助于学习控制策略以完成诸如操纵和在非结构化环境中移动等任务 [3]。在医疗保健领域，强化学习算法通过随着时间适应个体患者的反应来优化慢性疾病的治疗策略，显示出前景 [4]。

尽管有这些令人鼓舞的方向，但在现实世界中部署强化学习常常受到样本效率低、缺乏可解释性和训练过程中失败成本高昂等挑战的阻碍。因此，视频游戏已成为推进强化学习研究的重要工具。游戏提供了一个丰富且交互式的环境，它们是安全的、可扩展的，并且非常适合用于衡量算法进展 [5], [6]。值得注意的突破展示了强化学习在复杂游戏环境中应用的巨大潜力。DeepMind 的 AlphaGo 在古老的棋盘游戏中实现了超人表现，通过结合深度神经网络与蒙特卡洛树搜索（MCTS）和强化学习击败了世界冠军 [7]。在此基础上，AlphaZero 进一步证明了仅使用强化学习就能掌握多种棋类游戏而无需人类数据 [8]。

这些成就突显了游戏作为控制实验平台的价值，在这种平台上，强化学习代理可以学会在高维度、部分可观察和随机的环境中操作。从这些领域获得的见解现在正被转移到机器人技术、自动驾驶和其他高风险领域中，在这些领域，泛化能力、鲁棒性和实时适应性是至关重要的。

在这篇论文中，我们探讨了通过强化学习为一款用 Python 开发的 2D 射击游戏实现代理的各种策略。借助最先进的策略，我们的目标是找到最优策略，使代理在游戏中获胜率最大化。当代理作为玩家能够击败地图上的敌人（总共击中敌人 3 次，同时自身被击中少于 3 次）时，游戏即告胜利。

玩家和敌方实体拥有相同的一组可用动作。训练完成后，RL 代理将在射击游戏中作为敌人使用。因此，独自对抗先进的敌方人工智能变得有趣且具有挑战性。

为了找到实现 RL 代理的最佳策略，我们的方法如下：

1) 分析当前文献中在射击游戏中实现 RL 代理的最先进策略。
2) 选择最适合且高效的策略来教授 RL 代理如何玩游戏
3) 在射击游戏背景下实施这些策略
4) 评估关于最大化胜率的策略结果
5) 根据评估，选择一个适合二维射击游戏中强化学习代理的策略

这种方法也反映了本文的结构。最后，我们将提供我们方法和研究结果的结论，并展望未来可以改进或适应的内容。

II. 基础知识

A. 强化学习

强化学习（RL）是一种计算框架，在该框架中，智能体通过与环境交互来学习做出决策以最大化累积奖励。这个过程形式化为一个马尔可夫决策过程（MDP），定义为元组 $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$，其中：

- $\mathcal{S}$ 是状态的集合，
- $\mathcal{A}$ 是动作的集合，
- $P(s'|s,a)$ 是在动作 a 下从状态 s 到状态 s' 的转移概率，
- $R(s,a)$ 是奖励函数，
- $\gamma \in [0,1)$ 是折扣因子。

代理的目标是找到一个策略 $\pi(a|s)$，以最大化预期回报，定义为：

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

价值函数 $V^\pi(s)$ 估计从状态 s 开始并遵循策略 π 的期望回报：

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,|\, S_t = s \right] \tag{2}$$

类似地，动作值函数 $Q^\pi(s,a)$ 预估从状态 s 执行动作 a 的预期回报：

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,|\, S_t = s, A_t = a \right] \tag{3}$$

为了从后续状态计算预期的未来价值，以下方程递归定义了遵循特定策略 π 处于状态 s 的期望值。

$$V^\pi(s) = \sum_a \pi(a|s) \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^\pi(s') \right] \tag{4}$$

最优值函数 $V^*(s)$ 定义为：

$$V^*(s) = \max_\pi V^\pi(s) \tag{5}$$

并且因此

$$V^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right] \tag{6}$$

为了估计这些价值函数，采用了各种算法。一个例子是时序差分（TD）学习，在这种学习中，估算会逐渐更新：

$$V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right] \tag{7}$$

其中 α 是学习率。策略梯度方法然后可以通过调整 $\pi_\theta(a|s)$ 的参数 θ 来直接优化策略，以最大化预期回报。

B. 行为克隆

行为克隆（BC）是模仿学习中的一种基础方法，其中代理通过监督学习模仿专家的行为来学习策略。与传统强化学习（RL）中的探索环境并通过奖励信号接收反馈不同，BC 直接使用从专家策略 π_E 收集的演示数据集 $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^N$ 将观察到的状态映射为专家动作。

在 BC 中的学习目标是找到一个由 θ 参数化的策略 π_θ，以最小化数据集上的监督损失函数。

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(\pi_\theta(s_i), a_i\right) \tag{8}$$

其中 ℓ 通常是离散动作的交叉熵损失或连续控制的均方误差。

尽管 BC 易于实现且高效，特别是在基于视觉的控制 [9] 等高维领域中，它会受到协变量转移问题的影响。由于学习到的策略可能会遇到与专家数据集中的状态不同的情况，小预测误差会在时间上累积，导致智能体进入在状态空间 [10] 中未见或表示不佳的区域。这种现象会严重降低性能。

为了应对这一问题，已经提出了诸如数据集聚合（DAgger）之类的方法，该方法迭代地通过学习到的策略生成的新轨迹并由专家 [11] 重新标记来扩充数据集。

尽管存在局限性，BC 仍然是一个强大的基线，并且在安全探索至关重要的领域或专家演示丰富的场景中特别有用。

C. 代理竞技场

Agentarena 是由 Thomas Ackermann 开发的一款二维射击游戏。游戏玩家的目标是在代理竞技场中通过射击击败敌人。竞技场的大小为 900 乘以 1200 像素，如下图所示。

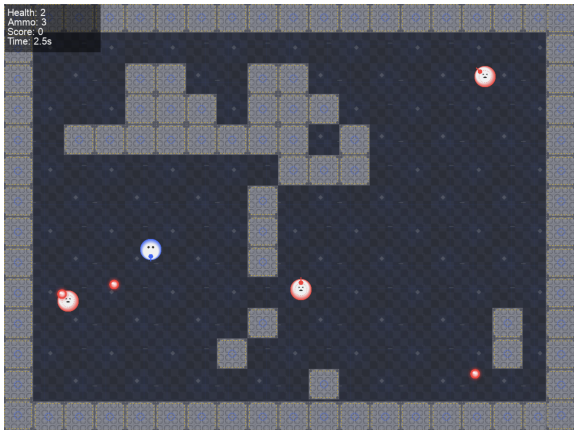

图 1. 代理竞技场游戏的 playing field，矩形竞技场被随机放置的墙壁所阻碍，这些墙壁作为玩家和敌人掩护的遮挡物。蓝色圆圈突出显示玩家实体，而红色圆圈突出显示敌人实体。

该游戏允许识别状态空间参数，例如玩家和敌方实体的 x 和 y 坐标，以及这些实体是否开火或被击中。当一个实体被击中三次时，它将从游戏中移除。当所有敌人被消灭时，这算作胜利。如果玩家死亡，则算作失败。

III. 相关工作

近期在射击环境中的强化学习研究越来越多地转向结合行为克隆（BC）与强化学习（RL）的混合方法，尤其是为了应对复杂游戏环境中观察到的不稳定性和稀疏奖励信号问题。这些方法可以直接应用于我们的 2D 射击游戏场景中，在这种场景下仅靠基于规则的战略是不够的，纯粹探索会导致性能崩溃。

Goecks 等人 [12] 引入了学习循环（CoL）框架，该框架结合 BC 与离线 Q-learning 以促进模仿和探索之间的平稳过渡。他们的多阶段训练流程从专家数据的离线学习开始，并逐步纳入在线 RL。作者展示了在密集奖励环境和稀疏奖励环境中均提高了学习效率和稳定性，验证了混合损失设计对于复杂领域的重要性。

Reddy 等人 [13] 提出软 Q 模仿学习 (SQIL)，这是一种将专家演示注入到 RL 框架中的方法，通过为专家行为分配固定的正奖励，而为所有其他行为分配零奖励。这简化了集成，消除了对单独的 BC 目标的需要，但仍然通过强化学习引导策略改进。虽然不是显式地多头，但通过演示进行奖励塑造的原理与交替训练阶段的概念重叠。

类似地，Spick 等人 [14] 采用了基于 BC 的方法来初始化他们的 FPS 代理的行为，旨在快速准确地模仿人类游戏玩法，这与仅依赖 RL 的策略形成对比。在他们的工作中，BC 随后通过增加 RL 来进行扩展，以便在长时间训练期间进行探索并获取新技能。这种混合迭代框架带来了适当平衡 BC 和 RL 影响的挑战。微调这一平衡对于缓解诸如代理不稳定（例如卡住）等问题至关重要，同时仍能促进初始 BC 模型未能捕捉到的新行为出现。

近期，Torabi 等人 [15] 引入了一种通过行为克隆（BC）在专家行动访问受限的环境中改进策略的方法。他们表明，即使在这种受限的情况下，BC 也能生成一个强大的初始策略，当通过强化学习（RL）进行优化时，其性能超过了单独使用任意一种方法的效果。

这些研究共同证实了结合模仿和强化学习策略的优势。因此，研究问题是如何将这些策略应用于二维射击游戏中以最大化代理的胜率。我们的工作基于此领域的研究成果，实现了一种多头神经架构，该架构分离了模仿和 Q 学习输出，允许隔离梯度和模块化训练计划。与先前的工作不同，我们将这些技术应用到了一个自定义构建的具有离散动作空间的二维射击环境中，证明了混合 BC+RL 策略不仅适用于大规模的三维模拟环境，在轻量级但策略丰富的游戏设置中也能够普遍适用。

IV. 方法

开发适用于二维射击游戏的有效强化学习代理需要一种系统方法，结合多种技术。本节详细介绍了所使用的实现策略，从最初的纯强化学习尝试到最后将离线学习与在线强化学习相结合的训练方法。

A. 游戏环境：代理竞技场

AgentArena 担任我们强化学习实验的测试环境。游戏具有俯视视角的二维画面，玩家代理必须在避免被消灭的同时消除敌方代理。该环境提供了对 RL 训练至关重要的几个关键特征：

状态空间：观察空间由玩家位置和生命值、敌人位置和生命值、子弹位置和方向以及墙壁位置组成。这导致了一个高维状态向量，能够捕捉所有相关游戏信息。

动作空间：代理可以选择来自 18 离散动作，结合 9 移动方向（包括静止）与 2 个射击选项（射击或不射击），结果共有 18 种可能的动作。

奖励结构：环境支持多种复杂程度不同的奖励函数，从基本的命中/未命中奖励到包括位置、弹药管理和生存策略在内的高级战术考量。

B. 神经网络架构

我们的方法利用了一种多头神经网络架构，旨在支持单一模型内的离线模仿学习和在线强化学习。该架构包含几个关键组件：

特征嵌入层：分离的嵌入网络处理不同实体类型（玩家、敌人、子弹、墙壁），并使用层规范化和 `LeakyReLU` 激活函数。这使得模型能够为每种游戏实体学习专门的表示。

注意力机制：多头注意力层从不同数量的实体中聚合信息，使用玩家状态作为查询，其他实体作为键和值。这使模型能够在处理不同数量的敌人和子弹时专注于相关实体。

双输出头：网络具有两个独立的输出头 —— 一个用于强化学习的 Q 学习头部，该头部输出用于动作选择的 Q 值；另一个是用于行为克隆的模仿学习头部，该头部输出监督学习的动作概率。

完整的网络架构通过嵌入层处理输入特征，应用注意力机制进行实体聚合，并根据训练模式通过特定任务的头部产生输出。

C. 训练方法演变

我们的训练方法根据实证结果和稳定性考量经过了多次迭代。

*1) 初始纯强化学习：*初始方法采用了标准的深度 Q-网络（DQN）训练，包括经验回放和 ϵ 贪婪探索。然而，该方法遇到了显著的稳定性问题：

稀疏奖励：代理很少收到奖励，导致学习速度慢且样本效率低。早期的奖励函数仅对成功的命中或清除提供反馈，从而导致长时间没有学习信号。

政策不稳定：即便代理实现了良好的表现，它也会频繁地回归到较差的策略，表明学习动态不稳定。这表现为获胜率不一致以及在改进时期后突然的表现下降。

可扩展性问题：复杂的多敌环境使得纯粹的探索效率低下，特别是在同时面对多个对手时。

*2) 奖励函数开发：*为了处理稀疏奖励问题，我们开发了越来越复杂的奖励函数：

基础奖励：简单的命中/未命中奖励，消灭敌人有额外奖励，受到伤害则有惩罚。

高级奖励：包含了诸如弹药管理、相对于敌人的位置、墙壁碰撞惩罚和子弹躲避奖励等战术考虑。高级奖励函数使用双曲正切归一化来限定奖励值：

$$R_{advanced} = \tanh(R_{events} + R_{tactical} + R_{positional}) \quad (9)$$

其中 R_{events} 表示基于事件的奖励，$R_{tactical}$ 捕获战术决策，而 $R_{positional}$ 鼓励战略定位。归一化函数被选用来限制神经网络输出范围。

*3) 离线学习的介绍：*由于纯强化学习的不稳定性，我们引入了使用行为克隆的离线学习：

数据收集：我们通过记录基于规则的代理与各种对手进行游戏的方式收集了演示数据。这提供了一大数据集，其中包含代表熟练游戏玩法的状态-动作对。最初我们使用的是由人类玩家收集的游戏数据。结果发现该数据集在选择的动作方面存在严重的偏见。

行为克隆：模仿学习头部使用监督学习在演示数据上进行了训练，学会了根据游戏状态预测专家动作。交叉熵被选为模仿学习的损失函数：

$$\mathcal{L}_{BC} = \mathbb{E}_{(s,a)\sim D}[\text{CrossEntropy}(\pi_\theta(s), a)] \quad (10)$$

其中 D 表示演示数据集，而 $\pi_\theta(s)$ 表示来自模仿头的动作概率。

动作平衡：为了处理演示数据中的偏差，我们实现了加权采样以平衡训练过程中动作的分布，防止诸如移动而不射击等常见动作的过度表示。

D. 混合训练方法

最终的方法结合了离线和在线学习的动态计划：

*1) 训练计划设计：*混合方法根据衰减计划在离线和在线训练阶段之间交替进行：

$$r_{\text{offline}}(t) = r_{\text{initial}} + \frac{t}{T} \cdot (r_{final} - r_{\text{initial}}) \quad (11)$$

其中，$r_{\text{offline}}(t)$ 是第 t 轮的离线训练比例，T 是总的训练轮次数量，r_{initial} 是初始离线比例（通常为 0.8），而 r_{final} 是最终离线比例（通常为 0.2）。

2) 多头训练协议: 在离线阶段，模仿学习头使用行为克隆对演示数据进行训练。在线阶段，Q-学习头使用带有经验回放的 DQN 进行训练。两个头共享相同的特征提取层，允许两种学习模式之间的知识迁移。

基于相位的训练: 集被组织成固定长度的阶段（通常为 50 集），离线比率决定了每个阶段中使用离线训练与在线训练的集的数量。

梯度隔离: 两种训练模式使用单独的优化器和损失函数，以防止模仿学习和强化学习目标之间的干扰。

E. 实验设计

1) 环境配置: 实验在标准化的环境设置下进行：单一敌人配置以减少复杂性，固定场地大小为 1200×900 像素，最大回合长度为 1000 步，并优化了子弹速度和玩家移动速度以实现平衡的游戏体验。

2) 训练参数: 标准超参数在实验中保持一致：学习率为 0.001 并使用步进衰减，折扣因子 γ 为 0.99，初始探索率 ϵ 为 0.8 衰减至 0.1，离线训练的批量大小为 512，经验回放缓冲区大小为 $20,000$ 次转换。

3) 评估指标: 性能使用多个指标进行衡量：对抗基于规则的对手的胜率、平均回合奖励、回合长度、准确性（每次射击的成功命中数）以及训练稳定性（随时间变化的表现波动）。

这一全面的方法论解决了在复杂游戏环境中训练强化学习代理的关键挑战，同时保持了训练的稳定性并实现了性能的一致性提升。

V. 结果

为了评估我们混合训练方法的有效性，我们进行了几项实验来比较不同的智能体架构和训练配置。我们的评估框架测试了训练过的智能体与三种不同类型对手的表现：一种是随机选择行动且每个行动概率相等的随机智能体，以及两种复杂程度不同的基于规则的智能体。

A. 实验设置

所有模型均通过行为克隆在大约 200 轮基于规则的游戏玩法中收集的演示数据上经过 300 个周期的预训练智能体进行初始化。在此期间，两个基于规则的智能体相互竞争，并收集了游戏观察结果和智能体各自的行动。预训练持续时间是根据验证性能监控选择的，如图 2 所示，在超过 300 个周期后，我们观察到验证错误增加，这表明可能存在过拟合。随后，每个模型经历了 1000 轮在线强化学习训练。这一轮次数量是根据计算限制和经验观察得出的，即经过延长的训练期后胜率趋于平稳，如图 4 所示。

我们评估了三种不同的模型配置以评估网络架构和探索参数的影响:

- 具有初始探索率的大神经网络 $\epsilon_0 = 0.4$
- 大型神经网络带有初始探索率 $\epsilon_0 = 0.8$
- 小型神经网络，初始探索率为 $\epsilon_0 = 0.8$

B. 训练阶段分析

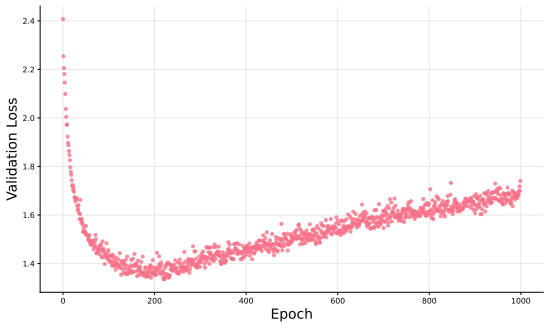

图 2. 验证损失在模仿学习预训练期间逐渐减少，直到大约 300 个周期后开始上升，这表明过拟合。这一趋势指导了我们对预训练持续时间的选择。

图 2 展示了行为克隆阶段验证损失的变化轨迹。模型显示了明显的进步，验证损失逐渐下降直到大约 300 个周期，之后损失开始增加，表明过拟合的开始。这一实证证据证明了我们在 300 个周期时终止预训练以保持泛化能力的决策是合理的。

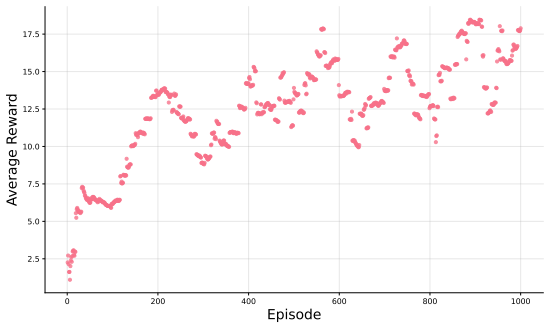

图 3. 在线强化学习训练期间的平均奖励进展，显示了稳步改进并最终收敛。

在线训练阶段，如图 3 和 4 所示，展示了我们混合方法的有效性。平均奖励在整个训练过程中表现出持续改进（图 3），而胜率则显示出快速的初始改进后保持稳定高水平表现（图 4）。值得注意的是，由于预训练，胜率从一个合理的基线开始，避免了与纯强化学习方法相关的通常较差的初期性能。

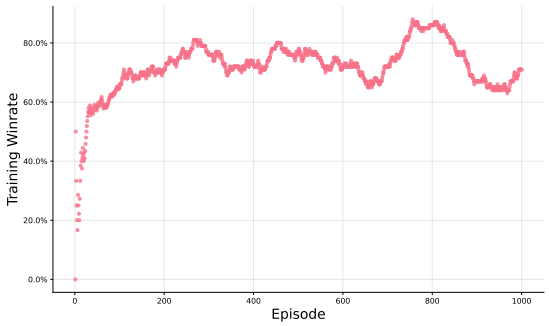

图 4. 胜率在强化学习训练过程中的演变。初始较低的胜率迅速提高并在高水平上稳定，这展示了预训练为 RL 优化提供稳定基础的有效性。

C. 面对不同对手的性能

敌人类别	胜率（%）	平均剧集长度（步骤）
Rule Based	76.0	117.8
Rule Based 2	86.0	180.5
Random	72.0	117.6

表 I
大型神经网络架构在初始探索率 $\epsilon_0 = 0.4$ 下的性能

敌人类别	胜率（%）	平均剧集长度（步骤）
Rule Based	94.0	73.3
Rule Based 2	76.0	264.2
Random	66.0	129.8

表 II
大型神经网络架构在初始探索率 $\epsilon_0 = 0.8$ 下的性能

敌人类别	胜率（%）	平均剧集长度（步骤）
Rule Based	96.0	105.3
Rule Based 2	92.0	97.3
Random	78.0	207.0

表 III
小神经网络架构在初始探索率 $\epsilon_0 = 0.8$ 下的性能

D. 性能分析

1) 优于训练来源的性能: 最值得注意的是，所有模型配置在与基于规则的代理对战时获得了 76-96%的胜率，该代理作为预训练演示数据的来源。这表明混合方法成功使代理超越了最初的教师，验证了离线学习与在线优化相结合的有效性。

2) 探索率影响: 比较大型网络配置揭示了探索策略的关键重要性。更高的初始探索（$\epsilon_0 = 0.8$）显著提高了对抗基于规则的代理的表现（胜率 94%对 76%），并且明显缩短了回合次数（73.3 步对 117.8 步）。这表明在早期训练阶段增加探索允许代理发现超出训练数据中展示的更高效策略，从而导致更高的成功率和更加决定性的胜利。

3) 网络架构影响: 与预期相反，小型神经网络架构在多个指标上优于大型架构。小型网络在对抗基于规则（96%）和基于规则 2（92%）对手时达到了最高的胜率，同时对随机对手也保持了有竞争力的表现（78%）。这一发现表明，在这个特定任务领域内，架构设计和训练方法比原始网络容量更为关键。另一个可能的原因是在 BC 阶段我们迅速遇到了过拟合的问题。很可能数据不足或数据过于相似，因为我们是针对同一代理进行训练的。

4) 针对特定对手的表现模式: 结果揭示了不同对手类型中的表现模式差异，这些差异为理解所学策略的本质提供了见解：

基于规则的对手： 代理在对抗基于规则的对手时表现出持续的高性能（获胜率 76-96%），小网络显示出特别强的结果。不同的回合长度（73.3-264.2 步）表明，代理适应了策略以高效利用特定的对手模式。

随机对手： 对抗随机代理的表现依然较低（胜率在 66%-78%之间），这可以归因于随机行为本质上难以预测的特性。随机代理表现出混乱无序的行为模式，这些模式很难被学习或利用，尽管它们缺乏策略规划，但仍是非常具有挑战性的对手。与随机对手对抗时较长的回合长度（129.8-207.0 步）表明了由于对手不可预测的模式而导致的更长决斗。

5) 情节长度分析: 各配置下剧集长度的变化为战术效率提供了额外的见解：

- **短剧集**（73.3-117.8 步）对抗基于规则的代理表明有效利用了可预测的模式
- **扩展情节**（180.5-264.2 步）与基于规则的 2 号代理人的对比表明存在更复杂的策略互动
- **变量长度** 对抗随机对手反映了随机行为的不可预测性

E. 训练稳定性与收敛性

混合训练方法在初步实验中表现出比纯强化学习方法更优越的稳定性。预训练阶段提供了一个稳定的基石，防止了在这个领域通常观察到的灾难性遗忘和策略崩溃。

如图 4 所示，胜率由于行为克隆初始化而开始处于合理的基线水平，然后迅速提升并在高水平上稳定下

来。这与典型的纯 RL 训练曲线形成了鲜明对比，后者在达到熟练程度之前常常会表现出长时间的低性能。

行为克隆初始化后紧接着强化学习优化的结合证明了在实现稳定性和性能提升方面是有效的。代理成功地学会了利用对手模式，同时发展出超出演示数据中存在的战术行为，这从它们能够在所有测试配置中超越其基于规则的教师的表现得到了证实。

VI. 结论

本研究提出了一种强大的预训练框架，该框架有效地将行为模仿（BC）与深度强化学习（RL）结合起来，以训练一个适用于 2D 射击游戏环境的有能力的智能体。所提出的方案通过用从专家演示中学习到的策略初始化智能体，并随后通过在线强化学习对其进行改进，解决了纯 RL 常见的限制问题，如训练不稳定、稀疏奖励和样本效率低。

经验评估表明，混合代理显著优于仅通过强化学习训练的代理，在与基于规则的对手对抗中胜率超过 90%。此外，多头神经网络架构与共享注意力基础特征提取的集成促进了模仿和强化学习模式之间的无缝知识转移。这有助于提高训练稳定性和收敛行为。

值得注意的是，混合训练计划——从离线学习逐渐过渡到在线学习——对于平衡探索与策略保留至关重要。实验还强调了早期阶段探索参数的重要性，这使得代理能够发现超越演示数据中的更有效策略。

总体而言，研究结果证实了利用演示驱动的策略初始化作为复杂多智能体环境中强化学习优化基础的有效性。未来的工作可能侧重于纳入课程学习、动态奖励塑造和多智能体协调，以进一步提高代理在日益复杂的场景中的适应性和性能。

参考文献

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[2] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[3] D. Kalashnikov *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[4] M. Komorowski *et al.*, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, pp. 1716–1720, 2018.

[5] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," in *Journal of Artificial Intelligence Research*, 2013.

[6] M. Kempka *et al.*, "Vizdoom: A doom-based ai research platform for visual reinforcement learning," in *IEEE Conference on Computational Intelligence and Games (CIG)*, 2016.

[7] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[8] ——, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[9] M. Bojarski *et al.*, "End to end learning for self-driving cars," 2016.

[10] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 627–635.

[11] S. Ross and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 627–635.

[12] V. Goecks, A. Lee, J. Hays, J. Valasek, and P. Stone, "Integrating behavior cloning and reinforcement learning for improved performance in dense and sparse reward environments," in *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

[13] S. Reddy, A. Dragan, and S. Levine, "Sqil: Imitation learning via regularized behavioral cloning," *arXiv preprint arXiv:1905.11108*, 2019.

[14] R. Spick, T. Bradley, A. Raina, P. V. Amadori, and G. Moss, "Behavioural cloning in vizdoom." [Online]. Available: http://arxiv.org/pdf/2401.03993

[15] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018.