基于老鹰捉小鸡的多智能体对抗系统：设计与实现报告

Author Name

2025 年 11 月 18 日

目录

摘要

本文围绕“老鹰—母鸡—小鸡”三方智能体的空中对抗场景，构建了一个基于强化学习的多智能体协同对抗仿真平台。系统采用 MAAC-R 算法，使各角色在共享环境中依据独立观测学习最优策略，并通过奖励塑形实现协作与博弈的平衡。本文提出了面向保护、拦阻与安全约束的细粒度奖励设计，并在环境层引入物理碰撞与击退机制，以提升策略的鲁棒性与仿真逼真度。实验结果表明，该平台能稳定支持训练与评估流程，并具有良好的扩展性与展示效果。本文同时给出了主要模块的设计原则与关键算法实现，以便读者理解系统方法及其实验结论。

1 引言

1.1 项目背景与动机

多智能体系统在动态环境中的协同与对抗行为学习是人工智能领域的重要研究方向，其核心挑战在于平衡竞争目标与协作需求，并在信息不完全的条件下实现鲁棒决策。本项目聚焦这一问题，设计了一个“老鹰（UAV）-母鸡（Protector）-小鸡（Target）”三方交互场景：老鹰以捕获小鸡为目标主动探索，小鸡通过规避老鹰并向母鸡靠拢寻求保护，母鸡则通过动态调整位置在老鹰与小鸡间形成有效阻挡，三者构成“对抗-协作”交织的复杂动态关系。

该场景的研究价值体现在：它是现实中多主体博弈场景（如无人机协同围捕、群体机器人防护）的抽象缩影，其中“部分可观测性”“目标冲突与协同的耦合”“动态环境适应性”等特性，为多智能体强化学习（MARL）算法的设计与验证提供了典型测试平台。为系统解决这一问题，本项目设定三个核心研究目标：

- 构建**可复现的二维空中对抗实验环境**，包含明确的物理约束（如空间边界、碰撞交互规则）和动态状态更新机制，确保实验结果的稳定性与可比性；
- 为三方角色设计**差异化的观测-动作-奖励闭环**：基于角色特性定义局

部观测空间（如老鹰感知未捕获目标位置、母鸡聚焦安全半径内个体）、离散动作集合（如转向角度、移动步长）及角色专属奖励函数，实现行为导向的精细化建模；

- 通过**奖励塑形技术**调和多目标冲突，将老鹰的“捕获收益”、母鸡的“护卫成效”与小鸡的“生存概率”统一到同一优化框架，使三方策略在动态博弈中实现协同进化。

从强化学习理论视角，该环境可形式化为部分可观测马尔可夫博弈（POMG）模型 $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$：其中 $\mathcal{S}$ 为全局状态空间（包含所有智能体位置、状态及环境参数）；$\mathcal{A} = \mathcal{A}_{\text{UAV}} \times \mathcal{A}_{\text{Prot}} \times \mathcal{A}_{\text{Tgt}}$ 为三方联合动作空间；$\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ 为状态转移概率函数，刻画动作对环境状态的动态影响；$\mathcal{R} = \{\mathcal{R}_{\text{UAV}}, \mathcal{R}_{\text{Prot}}, \mathcal{R}_{\text{Tgt}}\}$ 为角色化奖励函数集合，分别量化各方行为的收益；$\gamma \in (0, 1)$ 为折扣因子，平衡即时与长期收益。

各智能体仅能获取局部观测 $o_i \in \mathcal{O}_i$（$\mathcal{O}_i \subset \mathcal{S}$），无法知晓全局状态，因此需在信息不完全的条件下学习具有泛化能力的鲁棒策略，实现“对抗中存协作，协作中含博弈”的动态平衡。

环境可表述为 $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$，其中 $\mathcal{S}$ 为全局状态空间，$\mathcal{A} = \mathcal{A}_{UAV} \times \mathcal{A}_{Prot} \times \mathcal{A}_{Tgt}$ 为联合动作空间；$\mathcal{P}$ 为状态转移概率；$\mathcal{R}$ 为按角色定义的奖励函数；$\gamma \in (0, 1)$ 为折扣因子。各角色仅可获得局部观测 $o_i \in \mathcal{O}_i$，由此需在部分可观测条件下学习鲁棒策略。

1.2 研究目标与主要贡献

针对多智能体混合博弈场景下的策略学习问题，结合课程项目的实践要求，本研究围绕“环境构建-策略学习-评估验证”全流程设定核心目标，具体包括：

- 实现基于多智能体强化学习（MARL）的三方策略框架，使老鹰、母鸡、小鸡能根据局部观测自主决策，形成动态对抗与协作行为；
- 构建标准化评估体系与可视化工具，支持动态过程直观展示，并形成固定随机种子与参数配置下的可复现实验流程。

基于上述目标，本项目的主要贡献体现在：

- 提出并实现了一个 ** 三角色协同对抗学习平台 **，通过分离观测空间、策略网络与经验回放机制，解决了混合博弈中“对抗-协作”目标的策略干扰问题；

- 设计了 ** 细粒度角色化奖励机制 **，将追踪效率（老鹰）、阻挡有效性（母鸡）、护卫半径保持（母鸡）、安全距离（小鸡）与边界约束等行为指标转化为可量化的奖励分量，实现策略与目标行为的精准对齐；

- 开发了 ** 自动化评估与可视化模块 **，支持训练过程中最新模型的自动加载、性能指标实时记录（如每步捕获数、存活小鸡比例）及交互过程动态渲染，提升了实验效率与结果可解释性。

1.3 研究范围与约束条件

本项目的研究范围聚焦于二维封闭空间内的多智能体交互行为，具体设定如下：

- **空间范围**：实验环境为二维矩形区域，坐标范围定义为 $x \in [0, 2000]$ 且 $y \in [0, 2000]$（单位：虚拟长度单位），所有智能体的位置坐标会被实时裁剪至该边界内，确保物理交互的空间一致性；

- **动作空间**：采用 12 维离散动作集合，通过量化方位角（每 30° 为一个离散转向选项）实现智能体的方向控制，简化连续动作空间的决策复杂度；

- **角色能力边界**：为体现角色分工差异，为三方智能体设定差异化参数：- 观测半径：老鹰侧重全局目标感知（半径较大），母鸡聚焦局部护卫范围（半径中等），小鸡仅感知近距离威胁与保护者（半径较小）；- 运动约束：老鹰速度与角速度上限高于母鸡，母鸡略高于小鸡，模拟“捕食者-护卫者-目标”的动力学特性差异；- 交互半径：包含老鹰的捕获半径（触发捕获判定）、母鸡的护卫半径（定义安全区域）及碰撞交互半径（触发击退机制）。

研究过程中的约束条件主要包括：

- **计算资源约束**：受限于课程项目的训练环境（如单 GPU 算力），超参数选择需平衡模型性能与训练效率；

- **可视化开销约束**：评估阶段的实时渲染（如动态位置更新、轨迹绘制）会增加计算负载，因此支持通过配置项关闭在线渲染，仅保存离线动画用于后续分析；

- **环境简化约束**：为聚焦策略学习核心问题，本版本暂未引入真实世界的通信带宽限制与观测噪声，而是通过几何约束和奖励机制间接模拟信息不完全性；

- **扩展性边界**：当前框架支持智能体数量的小幅调整（如老鹰 1-3 只、小鸡 5-20 只），但大规模群体场景（如 50+ 智能体）可能受限于状态空间维度与训练稳定性，需后续优化。

2 理论与方法

2.1 理论基础

2.1.1 强化学习基础

在马尔可夫决策过程（MDP）框架下，智能体通过与环境交互最大化长期回报。策略梯度方法直接优化参数化策略，价值函数近似用于降低方差、稳定训练过程。Actor-Critic 结构结合二者优势，使策略更新与价值评估相互促进。

设策略 $\pi_\theta(a \mid s)$ 与价值函数 $V_\phi(s)$，目标为最大化 $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_t \gamma^t r_t]$。典型 A2C 更新为：

$$\nabla_\theta J \approx \mathbb{E}[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, \delta_t], \quad \delta_t = \big(r_t + \gamma V_\phi(s_{t+1})\big) - V_\phi(s_t), \qquad (1)$$

评论器以时序差分目标最小化均方误差 $\|\delta_t\|^2$，从而提升策略更新的稳定性。

2.1.2 多智能体学习基础

多智能体环境具有非平稳性、局部可观测性与策略耦合等挑战。独立学习可提升并行性与模块化，但可能引入不稳定；集中式学习可改善全局协调，但成本较高。项目采用“共享环境、独立网络”的折中架构，并通过奖励塑形与通信网络提升协同效率。

对于局部可观测与干扰噪声，本文在观测构建与奖励分量中引入几何约束（如半径与距离的规范化），并通过裁剪/归一化提升数值稳定性；对于策略耦合与竞争关系，采用分角色的奖励与损失函数，避免互相干扰导致的学习退化。

2.2 算法框架与实现

2.2.1 MAAC-R 框架

本项目基于多智能体 Actor-Critic 框架，提出 MAAC-R（Multi-Agent Actor-Critic with Role-specific Rewards）方案，核心特征是为非对称角色设计独立决策网络与差异化奖励机制。其实现细节如下：

- **基础架构**：MAAC-R 中，老鹰（UAV）、母鸡（Protector）与小鸡（Target）采用角色分离的学习模式——默认配置下，老鹰使用启发式策略（优先追踪最近未捕获小鸡）以降低初期训练复杂度；母鸡与小鸡则各自配备独立的 Actor-Critic 网络及优化器，网络参数与经验数据完全隔离。动作空间统一为 12 维离散方位控制（每 30° 一个方向选项），对应智能体的转向决策。

- **与标准 MAAC 的差异**：原始 MAAC 通常采用统一奖励函数与共享价值评估机制，而 MAAC-R 的核心改进在于：- 角色化奖励设计：母鸡的奖励包含阻挡成功率（挡在老鹰与小鸡连线之间的比例）、护卫半径保持率（安全范围内小鸡数量占比）及边界惩罚；小鸡的奖励包含靠近母鸡的距离奖励、远离老鹰的躲避奖励及存活激励（每步未被捕获的正向奖励）；- 奖励归一方式：各奖励分量通过区间裁剪与线性缩放归一化至 [-1, 1] 区间，而非 Z-score 归一，以平衡不同分量的梯度贡献。

- **训练流程**：主循环遵循强化学习时序更新逻辑：1. 环境重置：初始化智能体位置、存活状态及固定随机种子，确保实验可复现；

2. 动作采样：母鸡与小鸡的 Actor 网络基于当前局部观测 o_i，采用指数衰减的 ϵ-贪心策略（$\epsilon = \epsilon_{\text{start}} \times (\epsilon_{\text{decay}}^{\text{episode}})$）输出离散动作；老鹰则通过启发式规则生成动作（如追踪最近目标）；

3. 环境交互：执行联合动作后，环境返回下一状态、角色化奖励及终止信号（如小鸡全部被捕或达到最大步数）；

4. 经验存储：按角色将 $(o_i, a_i, r_i, o_i', \text{done})$ 存入专属经验回放池，默认采用优先级经验回放（PER）机制提升样本效率；

5. 网络更新：当经验量达到阈值时，随机采样批次数据，通过时序差分（TD）目标（$y = r + \gamma \cdot Q'(s', a') \cdot (1 - \text{done})$）更新 Critic 网络（损失函数为均方误差），并通过策略梯度定理（$\nabla J \propto \mathbb{E}[\log \pi(a|s) \cdot \delta]$，其中 δ 为

TD 误差）更新 Actor 网络；

6. 模型保存与日志：每 100 局训练保存一次检查点（含网络参数与优化器状态），评估阶段自动检索并加载最新权重，简化复现流程。

2.2.2 PMI 协作评估网络

PMI（Pairwise Mutual Information，成对互信息）网络用于量化“母鸡-小鸡”对的协作紧密性，为奖励加权提供依据。其实现细节如下：

- **输入与结构**：网络输入包含三元组特征：母鸡与最近小鸡的相对位置$(\Delta x, \Delta y)$、母鸡视野内的小鸡数量、老鹰与目标小鸡的相对距离。通过两层全连接网络（隐藏层维度 256，ReLU 激活）输出 [0,1] 区间的标量评分，反映当前协作需求强度。

- **作用机制**：当前版本中，PMI 网络参数固定（不参与训练），其输出作为母鸡护卫奖励的加权系数——当评分接近 1 时（如小鸡远离母鸡且老鹰逼近），放大护卫奖励以强化保护行为；当评分接近 0 时（如小鸡已在安全范围），缩小奖励以避免冗余行为。需说明的是，PMI 仅用于奖励塑形的软加权，不直接参与动作选择或决策阈值设定。

3 环境与建模

3.1 环境设计与约束

3.1.1 场景与边界设计

环境采用二维坐标系与硬边界裁剪,确保位置始终位于 $[0, x_{\max}]\times[0, y_{\max}]$。状态描述包含通信向量、局部观测与边界状态，动作空间为 12 维离散方位控制。碰撞—击退—锁向机制用于模拟保护者对 UAV 的物理阻挡，提高场景真实性。

停滞检测以“近 k 步位移总量”为判据，当位置变化低于阈值且持续 k 步以上时施加惩罚；转圈检测以“路径长度显著大于位移”为信号，连续触发将递增惩罚强度。该两类检测有效抑制无效运动与原地打转，促使策略产生更具目的性的行动。

3.1.2 智能体建模与行为

老鹰（UAV） 状态包含位置、朝向及对目标/保护者/队友的观测；动作为离散方位控制；奖励强调高效捕获、避免越界与重复围捕，并引入与保护者交互的惩罚项。

小鸡（Target） 状态体现威胁感知与护卫依赖；动作为规避与靠近策略；奖励鼓励保持在护卫范围并远离老鹰，被捕则施加强惩罚，以形成协同躲避行为。

母鸡（Protector） 状态以拦截几何与护卫半径为核心；动作侧重挡在老鹰与小鸡连线之间；奖励综合护卫得分、阻挡得分与失败惩罚，并加入逼近/远离的动态奖励与运动停滞惩罚。

此外，保护者在空间上形成“手臂阻挡”判定带，UAV 进入该带将受到母鸡朝向的方向锁定与空间击退效果；该物理机制通过几何最近点与法向方向计算实现，保证在不同速度与半径参数下依旧稳定工作。

3.2 奖励设计与评估指标

3.2.1 UAV 奖励设计

$$R_{UAV} = \alpha R_{track} + \beta R_{boundary} + \gamma R_{duplicate} + \omega R_{protector} \tag{2}$$

3.2.2 保护者奖励设计

$$R_{protector} = \alpha R_{protect} + \beta R_{block} + \gamma R_{failure} + R_{approach} + R_{retreat} + R_{movement} \tag{3}$$

3.2.3 目标奖励设计

$$R_{target} = \alpha R_{safety} + \beta R_{danger} + \gamma R_{capture} + R_{approach} + R_{escape} + R_{movement} \tag{4}$$

3.3 性能评估与分析

3.3.1 评估指标

评价涵盖目标捕获率、平均局长、奖励收敛速度与协同效率等核心指标，并辅以轨迹聚类与覆盖率统计用于行为模式分析。

3.3.2 仿真分析

通过不同超参数与算法变体的对比，观察训练稳定性与策略转移性能；针对奖励分量进行敏感性分析，评估其对协同行为的驱动效果。

4 实验结果与分析

4.1 实验设置与复现

说明硬件与软件环境、关键训练参数与评估方法学；给出复现实验的命令行与配置入口，并说明是否启用在线渲染与输出保存。

4.2 结果与性能评估

4.2.1 训练表现指标

展示各角色的学习曲线、奖励收敛特性与训练效率，并分析不同超参数对收敛速度与稳定性的影响。

4.2.2 评估表现指标

报告多场景下的成功率、行为分析与对比指标，结合轨迹统计与覆盖率曲线刻画策略质量与协同性能。

4.3 可视化与行为分析

通过轨迹图与热力图呈现空间行为模式，对奖励分量进行分解与统计检验，并附上学习行为的视频演示以支撑结论与展示效果。

5 结论与展望

5.1 主要发现

MAAC-R 在混合对抗—协作环境中表现出良好的学习稳定性与策略质量；奖励塑形在引导协同行为与抑制不良模式（如越界、重复围捕、原地转圈）方面效果显著；环境层的物理交互提升了策略鲁棒性与仿真逼真度。

5.2 经验总结

多智能体系统的设计需要在模型复杂度、训练稳定性与可视化展示之间取得平衡；超参数调优与可复现实验流程是确保结论可靠性的关键；验证与分析环节应同时关注数值指标与行为模式。

5.3 未来改进

后续可进一步探索三维环境、引入现实约束（传感噪声、通信延迟）、设计更高级的协同协议，并开展软硬件闭环实验以提升系统应用价值。

同时，可在奖励设计中引入基于任务层面的稀疏回报与层次化目标，结合局部密集回报以提高训练效率；在通信机制上探索注意力与图网络以提升多主体间的信息共享质量。

6 参考文献

7 附录