DOI:10.11918/202510031

CM-MADDPG:面向多无人机博弈对抗的协同决策算法

高志云[1,2],唐艺宸[2],靳舒葳[2],梁宇玲[3]

(1.天津市城市空中交通系统技术与装备重点实验室(中国民航大学),天津 300300;2.中国民航大学 交通科学与工程学院,天津 300300;3.沈阳工业大学 人工智能学院,沈阳 110870)

摘 要:为解决多无人机博弈对抗中现有决策方法在复杂场景下奖励设计烦琐,提升算法训练稳定性与样本效率。将多智能体强化学习与元强化学习结合,提出评论家与元策略联合增强的多智能体深度确定性策略梯度(Critic and Meta enhanced multi-agent deep deterministic policy gradient,CM-MADDPG)算法。首先,通过在集中式评论家侧引入适度的结构与学习增强,提升跨机交互的价值估计精度并抑制过估计,从而使训练更稳定,稳态性能更高。其次,在训练期间采用元策略进行自适应调度(如动态调整损失权重与探索强度),提升样本效率与收敛速度,且不增加推理期开销。最后,在相同实验条件下,红方无人机分别采用多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient,MADDPG)、评论家增强的多智能体深度确定性策略梯度(Critic multi-agent deep deterministic policy gradient,Critic-MADDPG)、元策略增强的多智能体深度确定性策略梯度(Meta enhanced multi-agent deep deterministic policy gradient,Meta-MADDPG),以及CM-MADDPG与随机机动的蓝方无人机进行博弈对抗。仿真结果表明,CM-MADDPG在最终胜率、各无人机奖励,以及收敛速度等指标上均取得显著提升。证明CM-MADDPG在不增加推理计算开销的前提下成功缓解稀疏奖励与训练不稳定问题,提高收敛速度与样本效率。研究成果可为多无人机博弈理论提供一种新思路和途径。

关键词:多无人机博弈对抗;多智能体深度确定性策略梯度;元策略增强;集中式评论家;多智能体强化学习;样本效率

中图分类号:TP181;TP273 **文献标志码**:A **文章编号**:0367-6234(2026)08-0173-10

CM-MADDPG:a cooperative decision-making algorithm for multi-unmanned aerial vehicle game confrontation

Gao Zhiyun[1,2], Tang Yichen[2], Jin Shuwei[2], Liang Yuling[3]

(1. Key Laboratory of Technology and Equipment of Tianjin Urban Air Transportation System (Civil Aviation University of China), Tianjin 300300, China; 2. College of Transportation Science and Engineering, Civil Aviation University of China, Tianjin 300300, China; 3. School of Artificial Intelligence, Shenyang University of Technology, Shenyang 110870, China)

Abstract: To address the cumbersome reward design of existing decision-making methods in complex scenarios of multi-unmanned aerial vehicle game confrontation and to improve corresponding training stability and sample efficiency, this paper proposed a Critic and Meta enhanced multi-agent deep deterministic policy gradient (CM-MADDPG) algorithm by combining multi-agent reinforcement learning with Meta-reinforcement learning. First, this paper introduced appropriate structural and learning enhancements to the centralized Critic side to improve the value estimation accuracy of cross-unmanned aerial vehicle interaction and suppress overestimation, thereby making training more stable and achieving higher steady-state performance. Second, a Meta-strategy was adopted during the training period for adaptive scheduling (such as dynamically adjusting loss weights and exploration intensity) to improve sample efficiency and convergence speed without increasing inference-time overhead. Finally, under identical experimental conditions, red unmanned aerial vehicles respectively adopting multi-agent deep deterministic policy gradient (MADDPG), Critic multi-agent deep deterministic policy gradient (Critic-MADDPG), Meta enhanced multi-agent deep deterministic policy gradient (Meta-MADDPG), and CM-MADDPG conducted a game confrontation against blue unmanned aerial vehicles with random maneuvering. Simulation results show that CM-MADDPG achieves significant improvements in final win rate, rewards of each unmanned aerial vehicle, and convergence speed. This proves that CM-MADDPG successfully alleviates sparse reward and training instability issues without increasing inference computational overhead and improves convergence speed and sample efficiency.

收稿日期:2025-10-14;**录用日期**:2025-11-25;**网络首发日期**:2026-05-13

网络首发地址:https://link.cnki.net/urlid/23.1235.T.20260513.1529.003

基金项目:国家自然科学基金(62403471);中央高校基本科研业务费(3122023QD18);天津市城市空中交通系统技术与装备重点实验室(TJKL-UAM-202402)

作者简介:高志云(1992—),女,博士,讲师

通信作者:高志云,zygao@cauc.edu.cn

The research results of this paper provide a new idea and approach for theories of multi-unmanned aerial vehicle game confrontation.

Keywords: multi-unmanned aerial vehicle game confrontation; multi-agent deep deterministic policy gradient; Meta-strategy enhancement; centralized Critic; multi-agent reinforcement learning; sample efficiency

无人机因具有便携性、隐蔽性、高机动性等一系列优点,已被广泛地应用于空中通信[1-2]、对地打击[3-5]、监视侦察[6-8]等多类任务。由于通信半径有限、恶劣天气频发,以及电磁环境复杂,传统的人为远程操控难以胜任高机动、高对抗的任务需求。因此,发展面向态势理解与自主管理的智能化无人机,使其实现自治决策,已成为重要研究趋势。

作为智能无人机应用中颇具代表性的方向,无人机对抗因其多主体协同—博弈特性而受到学术界与工程界的高度重视。在多机对抗中,无人机需依据观测实时选择机动、规避与打击指令,以获取态势与几何优势、把握交战时机并降低被击中风险。由此,自主机动决策成为核心能力,即利用数学优化、机器学习,以及其他智能技术,在对抗过程中自动产生可靠、可执行的控制输出以获得作战主动权[9]。在过去几十年里,国内外研究人员围绕该问题进行了持续研究,提出了多种模型与算法框架,并在仿真与实验层面不断验证与完善。

自主决策空战算法的研究大体可归纳为 3 条技术路线:1)数学求解范式。此类方法可追溯至20 世纪 60 年代,通过解析建模与严密推导获得策略解。然而,早期研究面向任务简单,已难以适应当前的任务环境。2)机器搜索思路[10]。代表性方法包括蒙特卡洛树搜索[11]与决策树等。该类方法通常先进行态势评估与威胁排序[12],再依据评估结果与排序输出动作方案[13]。但其核心依赖专家知识与先验规则,导致泛化能力不足,面对快速变化的战场条件时适应性与鲁棒性有限。3)强化学习方法。通过交互试错与回报驱动的策略迭代逐步提升机动决策水平,显示出在非线性高维任务中的潜在优势。在上述 3 类技术路线中,以强化学习为核心的技术路线已成为主流研究与应用方向。相较于传统数学求解范式对严格模型假设及复杂解析推导的高度依赖,以及机器搜索思路受限于预设专家知识与先验规则的固有局限,强化学习展现出显著的技术优势:其无需依赖上述前提条件,仅需在动态空战环境中持续开展训练,并以回报信号为核心导向进行策略的迭代优化,即可有效适配高维、强耦合的复杂战场场景。同时,该路线在应对战场态势动态变化时,能够表现出更为优异的泛化能力与鲁棒性,为空战决策的智能化升级提供了关键技术支撑[14]。

从无人机的对抗场景与数目来看,当前的研究主要以一对一[15-17]的无人机空战对抗场景为主。但由于受制于任务模块单一、可挂载设备有限,以及探测范围不足,单架无人机难以在高强度对抗中取得优势。面向未来应用,由多种类型平台组成的无人机集群更具现实意义。通过在集群内部实现协同机动与队形调整,并在战术上实施相互支援与遮蔽,可显著提升体系防御能力与打击覆盖度,从而更好地应对复杂、多变的作战场景。然而与一对一场景相比,面向多对一及多对多无人机对抗的系统性研究成果较少。Wang 等[18]以二对一为例,通过改进多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)的经验筛选机制,实现了高效协同攻击。Yang 等[19]以 10 × 10 的红、蓝双方无人机集群攻防对抗场景为例,提出基于分解机制与优先经验回放的多智能体深度确定性策略梯度(decomposed and prioritized experience replay-based multi-agent deep deterministic policy gradient, DP-MADDPG)通过改进 MADDPG 的经验筛选与策略评价机制,在非对称多无人机协同攻击中展现出显著优势。Wang 等[20]将研究焦点聚焦在大规模编队对抗,提出加权平均场式强化学习,以平均场将全体交互近似为“中心机体—虚拟代理”作用显著降低复杂度,并通过注意力筛选高价值观测以提升学习的稳定性与效率,实验表明该方法适用于 50 架以上平台的对抗任务。结合现有成熟算法与本文在工程项目中的实践,发现多无人机对抗场景下现有方法在落地应用仍存在规模适应性不足、稀疏回报与训练不稳,以及样本效率与时效性受限不足等问题。

本文研究 3 × 3 多无人机空中对抗问题,将多智能体对抗过程建模为离散时间二维空间上的马尔可夫博弈,并采用集中式训练与分散式执行(centralized training and decentralized execution, CTDE)范式。针对奖励稀疏、训练不稳与样本效率低等痛点,提出评论家与元策略联合增强的多智能体深度确定性策略梯度(Critic and Meta enhanced multi-agent deep deterministic policy gradient, CM-MADDPG),实验在自建 UAV 3 × 3 基准上开展,红方为待训练策略,蓝方采用随机机动且可开火的统一基线以保证对比公平。仿真结果表明,与 MADDPG、评论家增强的多智能体深度确定性策略梯度(Critic multi-agent deep

deterministic policy gradient, Critic-MADDPG)、元策略增强的多智能体深度确定性策略梯度(Meta enhanced multi-agent deep deterministic policy gradient, Meta-MADDPG)3 类算法互相对照,CM-MADDPG 在最终胜率、奖励与收敛速度等指标上取得稳定的提升,并在扰动条件下保持更好的鲁棒性,同时维持与基线相当的推理时延与计算复杂度,更适合工程化部署。

1　系统模型与对抗环境

1.1　任务与场景描述

本文构建红、蓝双方的多无人机对抗基准。红方与蓝方各由 3 架固定翼无人机组成(集合分别记为 R,B;全体为 $N=R\cup B$)。在二维方域 $\Omega=[-S,S]^2$ 内机动并进行对抗交互。红方采用训练中的智能体策略,蓝方在训练阶段采取随机策略以形成统一的难度基准;推理期双方均可执行射击、规避等动作,计分规则对称,强调对称攻防与团队协同。对抗目标是在有限步长内,通过机动接敌、对准与开火获取更高的队伍累积回报;本文同时记录胜负指标用于统计分析。

为保证统计可比性,本文建立了固定时间步长与回合长度的仿真协议:离散步长 $\Delta t=0.1$,单回合最大步数 $T_{\max}=100$。此外,本文定义了基础计分、逼近与对准塑形、边界与编队惩罚,以及命中、落空与冷却期开火结算,形成可解释评价的“任务—评分”闭环。

1.2　作战空间、时间与记号

二维作战域 $\Omega=[-S,S]^2$,时间离散为 $t=0,1,\cdots,(T_{\max}-1)$,步长 $\Delta t=0.1$;单回合最大决策步数 $T_{\max}=100$。红、蓝双方各由 3 架无人机组成:红方编号 $R=\{\mathrm{adv}_1,\mathrm{adv}_2,\mathrm{adv}_3\}$,蓝方编号 $B=\{\mathrm{agt}_1,\mathrm{agt}_2,\mathrm{agt}_3\}$。第 i 架无人机位置记为 $\boldsymbol{p}_i(t)=[x_i(t),y_i(t)]^{\mathrm{T}}$,航向角为 $\psi_i(t)$,标量速度为 $v_i(t)$。

其中无人机默认参数最小线速度 $v_{\min}=0$、最大线速度 $v_{\max}=1.0$、最大角速度 $\omega_{\max}=2.0$、最大加速度步长 $a_{\max}=0.2$、命中判据 $r_{\mathrm{hit}}=0.15$、对准阈值 $c_{\min}=0.75$、边界安全系数 $S_{\mathrm{safe}}=0.92$、队形最小距离 $d_{\mathrm{team}}=0.25$、开火冷却步数 $T_{\mathrm{cool}}=5$。

1.3　无人机动力学与动作映射

考虑任意无人机 $i\in N$,其状态 $\boldsymbol{s}_i(t)=[x_i(t),y_i(t),\psi_i(t),v_i(t)]^{\mathrm{T}}$。在二维方域 $\Omega=[-S,S]^2$ 上,以离散步长 Δt 按单轮运动学模型推进,并作硬边界裁剪为

$$\begin{cases} x_i(t+1)=\mathrm{clip}(x_i(t)+v_i(t)\cos\psi_i(t)\Delta t,-S,S)\\ y_i(t+1)=\mathrm{clip}(y_i(t)+v_i(t)\sin\psi_i(t)\Delta t,-S,S)\\ \psi_i(t+1)=\psi_i(t)+\dot{\psi}_i(t)\\ v_i(t+1)=\mathrm{clip}(v_i(t)+\Delta v_i(t),v_{\min},v_{\max}) \end{cases} \tag{1}$$

其中:

$$\Delta v_i(t)\in\{-a_{\max},0,a_{\max}\}$$
$$|\dot{\psi}_i(t)|\leqslant\omega_{\max}$$

每个个体的离散动作集合 $A=\{0,1,2,3,4,5\}$ 依次对应:保持、左转、右转、加速、减速、开火。给定 $a_i(t)\in A$,其与控制量的关系为

$$(\dot{\psi}_i(t),\Delta v_i(t))=\begin{cases}(0,0), & a_i=0\\ (+\omega_{\max},0), & a_i=1\\ (-\omega_{\max},0), & a_i=2\\ (0,+a_{\max}), & a_i=3\\ (0,-a_{\max}), & a_i=4\\ (0,0), & a_i=5(\text{触发设计逻辑})\end{cases} \tag{2}$$

1.4　“感知—对抗—奖励”一体化建模

为适配 CTDE 范式,本文在执行期为每架无人机提供本地可见的紧凑观测,在训练期的集中式评论家侧访问全体个体的联合观测与动作信息,以提升跨机交互的价值建模能力。

1)执行期采用 CTDE 观察,每个智能体的局部可观测为 14 维向量,按固定拼接顺序,红、蓝双方各 3 架(x,y)坐标与自身的速度与航向角(v_i,ψ_i),观测向量可表示为

$$\boldsymbol{o}_i(t)=[x_{\mathrm{adv}_1},y_{\mathrm{adv}_1},\cdots,x_{\mathrm{adv}_3},y_{\mathrm{adv}_3};x_{\mathrm{agt}_1},y_{\mathrm{agt}_1},\cdots,x_{\mathrm{agt}_3},y_{\mathrm{agt}_3},v_i,\psi]^{\mathrm{T}}\in\mathbf{R}^{14} \tag{3}$$

2)目标关联,令对手集合为

$$\begin{cases} O(i)=\begin{cases}B, & i\in R\\ R, & i\in B\end{cases}\\ d_i^{\min}(t)=\min\limits_{j\in O(i)}\|\boldsymbol{p}_j(t)-\boldsymbol{p}_i(t)\|_2 \end{cases} \tag{4}$$

$$\begin{cases} a_i(t)=\cos(\psi_i(t)-\angle(\boldsymbol{p}_{j^*}(t)-\boldsymbol{p}_i(t)))\\ j^*=\operatorname{argmin}\limits_{j\in O(i)}\|\boldsymbol{p}_j(t)-\boldsymbol{p}_i(t)\|_2,a_i\in[-1,1] \end{cases} \tag{5}$$

式中:$d_i^{\min}(t)$为最邻近敌距,用于度量逼近或者远离的趋势;朝向对准度 $a_i(t)$为无人机航向与最近对手连线方向的夹角余弦,1 表示完全正对,-1 表示完全背对,是射击几何约束与对准奖励的统一依据;$\boldsymbol{p}_i(t)=[x_i(t),y_i(t)]^{\mathrm{T}}$ 为第 i 架无人机在时刻 t 的二维坐标;$\boldsymbol{p}_j(t)=[x_j(t),y_j(t)]^{\mathrm{T}}$ 为第j架(通常是对手)无人机在时刻 t 的二维坐标。

3)射击与冷却(红、蓝双方对称),一次有效射击需同时满足:

$$\begin{cases} a_i(t)=5 \\ \mathrm{cd}_i(t)=0 \\ \exists j\in O(i):\|\boldsymbol{p}_j(t)-\boldsymbol{p}_i(t)\|_2\leqslant r_{\mathrm{hit}}\wedge a_i(t)\geqslant c_{\min} \end{cases} \tag{6}$$

式中：r_{hit}为射击命中的距离阈值，射击有效需要满足 3 个条件：执行开火指令 $a_i(t)=5$，冷却结束 $\mathrm{cd}_i(t)=0$，存在 $j\in O(i)$ 使得 $\|\boldsymbol{p}_j(t)-\boldsymbol{p}_i(t)\|_2\leqslant r_{\mathrm{hit}}$ 且 $a_i(t)\geqslant c_{\min}$。

$$\mathrm{cd}_i(t+1)=\begin{cases} T_{\mathrm{cool}} \\ \max\{\mathrm{cd}_i(t)-1,0\} \end{cases} \tag{7}$$

命中后将对最近满足条件的目标 j^* 进行奖励、惩罚结算，并将开火方的冷却计数置为 T_{cool}；冷却期间强行开火会触发轻惩罚，用以抑制无效尝试。红、蓝双方完全对称，蓝方同样可以攻击并按同一规则计分。

4）奖励函数以稀疏为主并进行轻量塑形，即

$$\begin{aligned} r_i(t)=&r_i^{\mathrm{base}}+r_i^{\mathrm{chase}}(t)+r_i^{\mathrm{align}}(t)+\\ &r_i^{\mathrm{bdry}}(t)+r_i^{\mathrm{space}}(t)+r_i^{\mathrm{fire}}(t) \end{aligned} \tag{8}$$

式中：基础步惩罚 $r_i^{\mathrm{base}}=-0.003$；逼近奖励 $r_i^{\mathrm{chase}}=0.015[d_i^{\min}(t-1)-d_i^{\min}(t)]$；对准奖励 $r_i^{\mathrm{align}}(t)=0.008a_i(t)$；贴边惩罚中，若 $\max\{|x_i|,|y_i|\}>S_{\mathrm{safe}}$，则 $r_i^{\mathrm{bdry}}=-0.012$，否则为 0；编队惩罚中，若 $\exists k\neq i$ 且 $\|\boldsymbol{p}_k-\boldsymbol{p}_i(t)\|_2\leqslant d_{\mathrm{team}}$，则 $r_i^{\mathrm{space}}=-0.012$，否则为 0；开火结算中，命中 $r_i^{\mathrm{fire}}=+10.0$，$r_{j^*}^{\mathrm{fire}}=-10.0$；落空 $r_i^{\mathrm{fire}}=-0.12$；冷却期强开 $r_i^{\mathrm{fire}}=0.1\times(-0.12)$。上述规则红、蓝双方对称：蓝方亦可攻击并按同一准则计分。

1.5 回合取消与胜负判定

单回合在步数达到 $t=T_{\max}$ 时回合结束。分别累加红、蓝两队在整回合内的个体奖励得到如下团队回报，即

$$\begin{cases} R_{\mathrm{red}}=\sum\limits_{i=R}\sum\limits_{t=0}^{T_{\max}-1}r_i(t) \\ R_{\mathrm{blue}}=\sum\limits_{i=B}\sum\limits_{t=0}^{T_{\max}-1}r_i(t) \\ \mathrm{win}_{\mathrm{red}}=\prod\{R_{\mathrm{red}}>R_{\mathrm{blue}}\} \end{cases} \tag{9}$$

以 $R_{\mathrm{red}}>R_{\mathrm{blue}}$ 判为红方胜。除单回合胜负外，实验部分还统计 200 回合滑动胜率曲线与团队/个体回报曲线，用于刻画不同算法的收敛速度、稳态性能与方差水平。

2 MADDPG 算法与 CM-MADDPG 算法

2.1 MADDPG 算法

多智能体强化学习通常在马尔可夫决策过程（Markov decision process，MDP）框架下刻画，可定义为高维元组 $(S,A_1,L,A_N,R_1,L,R_N,P,\gamma)$。其中 S 为状态空间；N 为智能体数量；A_i 为第 i 个智能体的动作集合；R_i 为该智能体接收到的奖励函数；P 为环境的状态转移概率；γ 为折扣因子。

MADDPG 是 OpenAI 团队[15]于 2017 年提出的多智能体强化学习算法，旨在应对协作、竞争，以及两者并存的复杂交互场景。算法基于确定性策略梯度，将多主体间的信息交换与主体—环境的持续交互结合，使个体能够在高维、动态且非平稳的条件下形成协调的行为策略，属于分布式决策方法在多智能体领域的代表。其训练/执行机制采用 CTDE 思想：训练时引入集中式评论家，可访问其他主体的观测以提升价值估计与稳定性；执行时则保持去中心化，各主体仅凭本地观测独立作出动作选择。

作为单智能体深度确定性策略梯度（deep deterministic policy gradient，DDPG）在多主体领域的推广，MADDPG 令系统内每个智能体各自遵循 DDPG 结构：一方面通过演员网络学习确定性策略 $\boldsymbol{\pi}_{\mathrm{actor}}$，另一方面借助评论家网络评估动作价值 Q_{critic}，同时为二者设置目标网络（target networks）以减小估计方差并提升收敛稳定性。算法采用 off-policy 范式，结合经验回放与 Q-learning 风格的目标构造进行参数更新。Q 计算公式为

$$Q=Q(s_t,a_1,a_2,\cdots,a_n,\theta) \tag{10}$$

式中 s_t 下标 t 表示第 t 个离散时间步的环境状态。在每个智能体内均配置 Actor-Critic 结构：训练阶段，Critic 可接收额外的全局信息（如其他智能体的观测与动作），用于指导 Actor 学习；而在测试阶段移除 Critic，由 Actor 依据自身观测独立输出下一步动作。该种 CTDE 策略绕开了将单智能体 Q-learning 或策略梯度（policy gradient）直接迁移到多智能体环境时因仅能使用本体状态/动作而产生的非平稳性与回放失效等问题。借此，MADDPG 使每个智能体所面对的有效环境更接近平稳过程，其系统动力学可形式化描述为：

$$\begin{aligned} &P(s'|s,a_1,a_2,\cdots,a_n,\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_n)=\\ &P(s'|s,a_1,\cdots,a_n)=\\ &P(s'|s,a_1,a_2,\cdots,a_n,\boldsymbol{\pi}_1',\boldsymbol{\pi}_2',\cdots,\boldsymbol{\pi}_n') \end{aligned} \tag{11}$$

因此，当 $\boldsymbol{\pi}_i\neq\boldsymbol{\pi}'$ 时，整个环境仍然是稳定的。

以 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_n)$ 统称各智能体策略的参数，并用 $\boldsymbol{\pi}=(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_n)$ 表示所有主体采用的策略族[21]。面向第 i 个智能体，定义其目标函数为期望累计回报为

$$J(\theta) = E[R_i] = E_{s\sim p^{\pi}, a_i \sim \pi_{\theta_i}}\left[\sum_{t=0}^{\pi}\gamma^t r_i, t\right] \tag{12}$$

式中:r_i 为第 i 个智能体的奖励,γ 为折扣率。

针对随机策略梯度,求解策略梯度的公式为

$$\nabla_{\theta_i} = E_{s-P\pi, a_i-\pi\theta_i}[\nabla_{\theta_i}\ln\pi_i(a_i|o_i)Q_i^{\mu}(x, a_1, a_2, \cdots, a_n)] \tag{13}$$

式中:o_i 为第 i 个智能体的观测值;$\boldsymbol{s} = [o_1, o_2, \cdots, o_n]$ 为所观测的向量,也就是状态 $Q_i^{\pi}(s, a_1, a_2, \cdots, a_n)$ 表征第 i 个智能体在当前情境下、面对全体联合动作时的集中式状态-动作价值评估。考虑到多智能体系统中每个主体均依据自身回报函数 R_i 独立更新,其学习目标可以自然适配不同交互关系,因此不仅能够完成合作类任务,也能够胜任竞争性或合作—竞争并存的任务设置。

针对每个智能体建立价值函数,解决了传统强化学习算法在多智能体领域的不足[22]。设经验回放缓冲区为 D,其中每条记录为 $(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1})$。其中:$\boldsymbol{s}_t = (o_t^1, \cdots, o_t^n)$ 为时刻 t 全体智能体的观测集合向量,$\boldsymbol{a}_t = (a_t^1, \cdots, a_t^n)$ 为联合动作向量,$\boldsymbol{r}_t = (r_t^1, \cdots, r_t^n)$ 为在 t 时刻执行动作后获得的奖励向量,$\boldsymbol{s}_{t+1} = (o_{t+1}^1, \cdots, o_{t+1}^n)$ 为下一时刻的联合观测向量。集中式 Critic 的参数更新与深度 Q 网络(deep Q-network, DQN)的思想一致,采用基于时序差分(temporal difference, TD)的做法。

系统的损失函数定义为

$$L(\theta_i) = E_{s,a,r,s'}[(Q_i^{\mu}(s, a_1, \cdots, a_n) - y)^2] \tag{14}$$

式中:$y = r_i + \gamma Q_i^{\mu'}(s', a_j', \cdots, a_n')\big|_{a_j=\mu_j(o_j)}$,$Q_i^{\mu}$ 为目标网络,$Q_i^{\mu}\mu' = [\mu_1', \mu_2', \cdots, \mu_n']$ 为目标策略具有滞后更新的参数 θ_i'。

对于 Actor 网络,参数的更新公式分别为:

$$\theta^{\mu} = \theta^{\mu} + \alpha_{\mu}\nabla_{\theta^{\pi}}J \tag{15}$$

$$\nabla_{\theta x}J = \frac{1}{k}\sum_{i=1}^{K}\nabla_{\theta^{\pi}}\pi(o, \theta^{\pi})\nabla Q(s, a_1, a_2, \cdots, a_n, \theta^{Q}) \tag{16}$$

式中:θ^{μ} 为 Actor 网络的参数值,α_{μ} 为 Actor 网络的学习率。

需要注意的是,上述 MADDPG 在推导上基于集中式评论家、去中心化执行、单一 Q 评估结合的标准框架,但在本文构建的 3 × 3 多无人机博弈场景中,还存在以下 3 点不利因素:1)联合观测-动作维度将随平台数线性增长,而某一时刻真正影响个体收益的往往只是少量关键目标,使用全连接多层感知机(multilayer perceptron, MLP)对整条联合向量直接回归容易引入与决策无关的噪声;2)奖励主要由命中事件触发,呈现出明显的稀疏脉冲特性,单一 Q 的 TD 目标在此情况下更容易产生过估计和训练振荡;3)蓝方随机机动导致环境对红方策略而言呈现非平稳状态,固定学习率和探索强度会使早期收敛阶段变长。基于这些局限,下文将在保持 MADDPG 原有 CTDE 框架的基础上,对评论家结构与学习节奏进行有针对性的增强。

2.2　CM-MADDPG 算法

与上文所提的标准 MADDPG 相比,CM-MADDPG 保留了 CTDE 训练范式、经验回放,以及目标网络软更新 3 条主线,同时在两处核心位置做了改进:1)在集中式评论家侧,将原始的全连接值函数替换为带注意力的双 Q 结构,使评论家能够在联合输入中自适应选择对当前被评估智能体更相关的敌、友平台,并通过最小化双 Q 目标抑制过估计;2)在训练阶段为 Actor 叠加低频的 Reptile 型元策略更新,使策略参数能够根据近期对抗态势自动调整学习幅度,从而在不改变执行期开销的前提下提升样本效率与收敛速度。

面向 3 × 3 多无人机博弈,尽管系统的联合状态-动作 (o, a) 随平台数量线性增长,但其交互结构呈现出高度的非均匀性,局部时刻仅有少数“几架关键智能体”真正影响第 i 架无人机的即时回报与未来价值。此外,以命中事件为主的稀疏脉冲奖励机制,叠加蓝方随机机动导致的环境非平稳性,进一步增加了决策难度。在经典 MADDPG 中,集中式评论家 $Q_i(o, a)$ 用 MLP 对全量联合输入回归,容易将“无关智能体”的噪声也卷入估计;再叠加单 Q 的过估计偏置与稀疏回报下早期探索-利用失衡,训练常表现为收敛慢与波动大。为此,本文在保持 CTDE 与 6 离散动作不变的前提下,首先,提出 CM-MADDPG:在评论家侧引入“注意力 + Twin-Q”以提升跨机交互的价值建模并抑制过估计;在学习节奏侧使用“Reptile 元外圈 + 自适应调度”在不增加推理期开销的前提下加速收敛并提高样本效率。红方第 i 个智能体的策略 $\pi_{\theta_i}(a_i|o_i)$ 采用 Gumbel-Softmax 硬样本实现离散动作的可导近似为

$$\boldsymbol{a}_i = \mathrm{one}_{\mathrm{hot}}\left(\underset{k}{\mathrm{argmax}}[(\log\pi_{\theta_i}(a_i = k|o_i) + g_k)/\tau_{\exp}]\right),\quad g_k \sim \mathrm{Gumbel}(0,1) \tag{17}$$

式中:$\tau_{\exp}$ 为探索温度,保证 Actor 更新可反传与执行时仍为 one-hot 离散指令;$\pi_{\theta_i}(a_i|o_i)$ 为第 i 个智能体的策略;θ_i 为对应目标网络参数;$g_k \sim \mathrm{Gumbel}(0, 1)$ 为 Gumbel 噪声。

其次,先进行价值建模增强,令系统有 n 个智能体,先对每个智能体的观测–动作做线性嵌入,即

$$\boldsymbol{z}_j = \sigma(W_o o_j + W_a a_j) \in R^d \tag{18}$$

式(18)为第 j 个智能体的观测–动作嵌入。其中 d 为嵌入的维度。嵌入后以被评估智能体 i 为查询做点积注意力获取上下文,即

$$\begin{aligned} &\boldsymbol{q}_i = W_q \boldsymbol{z}_i, \boldsymbol{k}_j = W_k \boldsymbol{z}_j, \boldsymbol{v}_j = W_v \boldsymbol{z}_j, \\ &\alpha_{ij} = \frac{\mathrm{e}^{q_i^{\mathrm{T}} k_j/\sqrt{d}}}{\sum_{l}^{n} \mathrm{e}^{q_i^{\mathrm{T}} k_l/\sqrt{d}}}, \boldsymbol{c}_i = \sum_{j=1}^{n} \alpha_{ij} \boldsymbol{v}_j \end{aligned} \tag{19}$$

并以$[\boldsymbol{z}_i;\boldsymbol{c}_i]$回归 Q 这使评论家自适应聚焦对 i 的价值最相关的友机与敌机,天然削弱无关智能体的扰动。式中 $\boldsymbol{q}_i$、$\boldsymbol{k}_j$、$\boldsymbol{v}_j$ 分别为查询、键、值向量。为抑制 Q-learning 的过估计,本文采用 Twin-Q + 最小目标与目标网络软更新:设双路评论家 $Q_{i,\phi^{(1)}}$, $Q_{i,\phi^{(2)}}$, TD 目标和评论家损失分别为:

$$\begin{aligned} &y_i = r_i + \gamma(1-d) \min_{m\in\{1,2\}} Q_{i,\bar{\phi}^{(m)}}(\boldsymbol{o}', \boldsymbol{a}'_\pi), \\ &\boldsymbol{a}'_\pi = (\mathrm{GS}(\pi_{\bar{\theta}_1}(\boldsymbol{o}'_1)), \cdots, \mathrm{GS}(\pi_{\bar{\theta}_n}(\boldsymbol{o}'_n))) \end{aligned} \tag{20}$$

$$\begin{aligned} &L_{\mathrm{critic}} = \frac{1}{|B|} \sum_{(\cdot)\in B} \sum_{i\in R} \sum_{m=1}^{2} \left(Q_{i,\phi^{(m)}}(\boldsymbol{o},\boldsymbol{a}) - y_i\right)^2, \\ &\bar{\phi}^{(m)} \leftarrow \tau\phi^{(m)} + (1-\tau)\bar{\phi}^{(m)} \end{aligned} \tag{21}$$

式中:$Q_{i,\phi}(\boldsymbol{o},\boldsymbol{a})$为第 i 个智能体的集中式评论家(其中 $\bar{\phi}$ 为其目标网络参数);Twin-Q 中 $Q_{i,\phi^{(1)}}$、$Q_{i,\phi^{(2)}}$为双路评论家,用于取 min 抑制过估计;$\gamma \in (0,1)$为折扣因子;$d \in (0,1)$为终止标志;$(\boldsymbol{o}',\boldsymbol{a}')$为下一时刻变量;$B$ 为回放小批量;τ 为软更新系数。

Actor 在集中式价值评估下以最大化 Q 的等价目标更新,并加入轻量熵正则稳定探索,即

$$\begin{aligned} &\tilde{J}_{\mathrm{actor}}(\theta_i) = -E_B[Q_{i,\phi^{(1)}}(\boldsymbol{o},\boldsymbol{a}_\pi^{(i)})] - \\ &\lambda_H E_B[H(\pi_{\theta_i}(o_i))], \bar{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i \end{aligned} \tag{22}$$

式中:$H(\pi_{\theta_i})$为策略熵,λ_H 为熵正则系数。需要注意的是:注意力机制把影响者显式参数化,Twin-Q 则给出保守的目标,两者叠加显著降低价值估计的方差与偏置,表现为收敛曲线更平滑、胜率平台更高。

最后,进行学习节奏增强,由于稀疏奖励下早期爬坡慢,本文以低频元外圈提供方向性拉拢:每隔 K_{meta} 回合,复制当前 θ_i 得到 θ_i^{fast},在若干批次上做 K 步学习率内圈更新,然后按 Reptile 规则靠拢,即

$$\theta_i \leftarrow \theta_i + \beta(\theta_i^{\mathrm{fast}} - \theta_i) \tag{23}$$

式中:β 为外圈步长,θ_i^{fast} 为内圈更新后的快参数。该外圈仅发生在训练期,推理期参数与计算图不变。同时引入自适应调度缓解非平稳构造态势统计为

$$z_t = [\overline{\mathrm{win}}_{200}(t), \mathrm{Var}_B(y_i), \bar{\alpha}(t), \overline{d^{\mathrm{min}}}(t)] \tag{24}$$

式中:$\overline{\mathrm{win}}_{200}(t)$为 200 回合的滑动胜率,$\mathrm{Var}_B(y_i)$为批内 TD 目标方差,同时探索温度按照自适应升降温,即

$$\tau_{\exp}(t) = \tau_0/(1+\rho t)\cdot\exp\left(\eta\left(\frac{1}{2} - \min(t)\right)\right) \tag{25}$$

其可作用于当近期胜率低、目标方差高时鼓励更多探索;当胜率稳定提升时逐步降温,减少无谓随机性。该调度只依赖批内统计量,计算与存储几乎零开销,符合不增加推理期成本的工程约束。

凭借评论家侧结构化建模与学习节奏自适应的联合增强,CM-MADDPG 在不改变 CTDE 与执行期复杂度的前提下,显著提升了训练稳定性与样本效率。注意力集中式评论家配合 Twin-Q 目标,使价值估计更聚焦于与当前个体强相关的友机、敌方智能体交互,抑制过估计与无关噪声,训练曲线更平滑、方差更小;Reptile 元外圈与温度与权重自适应调度则在稀疏奖励与非平稳对手下加速早期爬坡,避免反复振荡。综合效果体现在:收敛速度更快,最终胜率更高、个体与团队回报更优,且对超参数与初始条件的鲁棒性更强;同时,因推理期仅保留标准 Actor,运行时延与资源占用与基线一致,便于工程部署。相较于“仅评论家增强”或“仅元策略”的单路改进,CM-MADDPG 在高维多机对抗中兼具可解释的价值建模与高效的学习节奏控制,从而在性能与稳定性上取得统合优势。

综上所述,CM-MADDPG 在不破坏原 MADDPG 的 CTDE 主体前提下,从价值函数的结构化建模和策略更新的自适应调度两个角度补足了其在高维联合输入、稀疏奖励、非平稳对手场景中的薄弱环节,因此能够在本文的 3 × 3 无人机对抗基准上获得更快的收敛和更高的稳态性能。

3 仿真实验

3.1 仿真环境介绍

为了验证由 CM-MADDPG 算法控制的无人机集群在战场对抗的有效性,本文使用 Python 语言搭建了异构无人机集群对抗仿真环境,基于 CUDA12.8 和 Torch2.8.0 搭建算法的训练框架。

本文在自建多无人机智能对抗平台上进行 3 × 3 仿真实验。仿真实验以固定空域为作战背景,红方为待训练的协同编队,承担搜索、逼近、对准与火力压制等协同任务;蓝方采用随机机动并具有对称开火能力,作为统一难度的基准对手。各无人机基

于简化动力学与离散动作进行机动与交战。对抗流程为:红、蓝双方从空域左右半区随机入场,红方在集中式训练—分散式执行框架下开展协同决策,与蓝方持续对抗;回合结束后按队伍累计回报判胜,并统计滑动胜率评估策略的协同与稳定性。为保证对比公平,除被消融的模块外,各方法共享同一环境与训练设定。具体对抗示意图如图 1 所示。

图 1 给出了在所构建的 UAV 3 × 3 对抗环境中,一次典型对抗回合的空间轨迹示意。红方 3 架无人机均自作战区域左侧起始位置出发:其中红方 1 号自左侧中部向场景中心方向加速推进,在机动过程中始终保持对右侧蓝方目标的指向性,在接近到预设射程并满足朝向约束后完成了一次有效开火,是本回合的主要攻击机。红方 2 号则沿下侧安全边界附近向前推进,其轨迹相对平缓,主要承担保持编队间距、避免队内过近惩罚并向作战中心输送火力的平台角色。红方 3 号自左侧上部向中部收敛,其机动方向与红方 1 号形成一定夹角,可视作对中心交战区域的侧向支援与态势补位,从而提高红方整体的包围度与攻击机会。蓝方 3 架无人机均位于作战区域右半区起始:蓝方 1 号位于中部并首先与红方前出机发生遭遇,在进入红方射程后成为本次射击的受击目标;蓝方 2 号起始于上方,机动过程中向下和向内侧回避,体现出在随机机动策略与边界约束共同作用下的规避趋势;蓝方 3 号位于下方并沿右侧下缘移动,其轨迹与红方下路线形成一定距离,有助于保持蓝方编组的空间分散性。

从图 1 可以看出,直观反映了红方策略的主动逼近与集中打击特征,以及蓝方随机机动在有界二维空域内的分散、回撤行为,为后续对不同算法下对抗效果的量化比较提供了场景依据。

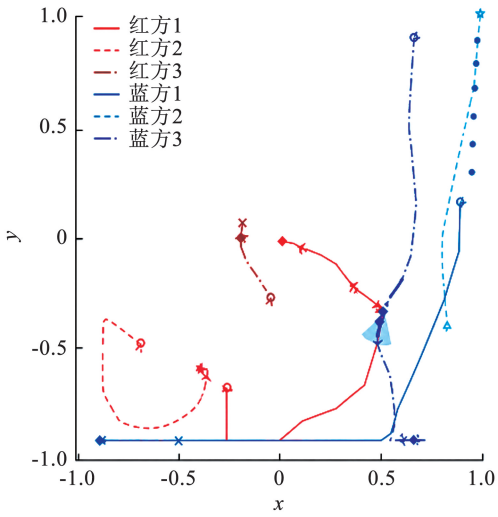


图 1　红、蓝双方对抗示意

Fig. 1　Schematic diagram of confrontation between red and blue sides

实验涉及的相关奖励参数详见表 1, 算法部分超参数设置见表 2。

表 1　奖励参数

Tab. 1　Reward parameters

名称	触发条件	数值
基础步惩罚 r_i^{base}	每步	-0.003
逼近奖励 r_i^{chase}	与最近敌机距离缩短	$0.015\cdot(d_i^{\min}(t-1)-d_i^{\min}(t))$
朝向对准 r_i^{aligm}	与最近敌机方向对齐	$0.008\cdot\alpha_i(t),\alpha_i\in[-1,1]$
贴边惩罚 r_i^{bdry}	$\max(\lvert x_i(t)\rvert,\lvert y_i(t)\rvert)>S_{\text{safe}}$	-0.012
队内过近惩罚 r_i^{space}	存在队友 $k\neq i$ 且 $\lVert p_k-p_i\rVert<0.25$	-0.012
命中奖励(射手) r_i^{fire}	开火且命中	$+10.0$
被命中惩罚(被击者)	对方命中我方	-10.0
未命中惩罚(射手)	开火未命中或未满足判据	-0.12
冷却期硬开火	冷却未到仍开火	$0.1\times(-0.12)=-0.012$
射程阈值	判定命中前提	$\lVert\boldsymbol{p}_j-\boldsymbol{p}_i\rVert\leqslant r_{\text{hit}}=0.15$
朝向阈值	判定命中前提	$\cos\angle\geqslant c_{\min}=0.75$
冷却步数	开火后进入冷却	$T_{\text{cool}}=5$

表 2　超参数表

Tab. 2　Hyperparameters

名称	数值	名称	数值
折扣因子 γ	0.95	Critic 学习率	1×10^{-3}
目标网络软更新系数 tau	0.02	注意力/嵌入维度	32
batch_size	1 024	元策略触发间隔(回合)	1 000
Actor 学习率	1×10^{-3}	元策略外圈步长	3×10^{-3}

仿真实验中单步总奖励为:$r_i(t)=r_i^{\text{base}}+r_i^{\text{chase}}(t)+r_i^{\text{align}}(t)+r_i^{\text{bdry}}(t)+r_i^{\text{space}}(t)+r_i^{\text{fire}}(t)$。命中后不移除实体,回合结束($T_{\max}=100$)按队伍累计回报判胜。

任务目标为在给定回合内实现更高的队伍累计

回报与更高的回合胜率:命中奖励与被命中惩罚对称计分,未命中与冷却期硬开火施加轻惩;回合结束不移除实体,通过红、蓝双方累计回报比较判胜,并统计滑动胜率以衡量收敛与稳定性。仿真实验中红方采用不同算法:MADDPG、Critic-MADDPG、Meta-MADDPG,以及本文 CM-MADDPG 进行训练与对照,蓝方保持随机策略不变,以公平评估所提方法在协同机动、对准与射击环节的综合效果。

3.2 结果与分析

在自建的 UAV 3 × 3 多无人机对抗平台中,采用"红方学习、蓝方随机"的统一基准,分别以 MADDPG、Critic-MADDPG、Meta-MADDPG,以及 CM-MADDPG 作为红方策略,按表 2 所列相同超参数与训练日程开展 30 000 回合的训练与评测。图 2 ~ 5 分别给出了 4 种红方策略(MADDPG、Critic-MADDPG、Meta-MADDPG,以及 CM-MADDPG)在 30 000 回合训练过程中的胜率演化情况。图中:浅蓝色为逐回合胜负结果,橙色为 200 回合训练数滑动平均,反映整体趋势。

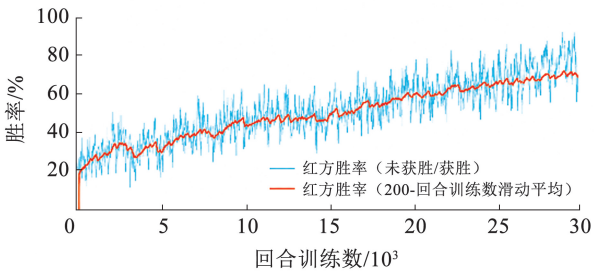


图 2 MADDPG 与随机策略对比

Fig. 2 Comparison between MADDPG and random policy

从图 2 可以看出,使用原始 MADDPG 时,红方胜率在很长一段时间内呈低斜率上升,前 2 000 回合训练数胜率仅由约 0.20 缓慢提升到 0.35 ~ 0.40,之后随训练推进才逐步爬升到 0.70 ~ 0.75 左右,同时橙色曲线始终伴随较明显的起伏,说明在稀疏命中奖励和 3 机协同这种耦合场景下,基线方法对经验的利用效率和训练稳定性都受到限制。

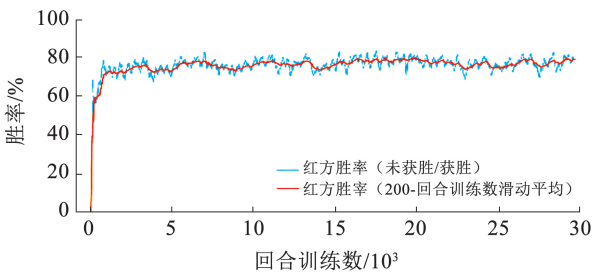


图 3 Critic-MADDPG 与随机策略对比

Fig. 3 Comparison between Critic-MADDPG and random policy

从图 3 可以看出,Critic-MADDPG 中,红方胜率在前期表现更好,大约到 1 000 回合训练数胜率即可突破 0.70,并在较长区间内保持在 0.75 ~ 0.85 之间,表明对联合观测-动作的信息做结构化建模以后,价值估计更准确、策略提升更快,但曲线仍能看到一定周期性的波动,说明价值侧的改进并不能完全消除样本分布变化带来的不稳定。

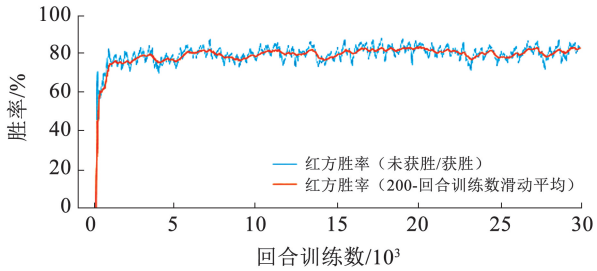


图 4 Meta-MADDPG 与随机策略对比

Fig. 4 Comparison between Meta-MADDPG and random policy

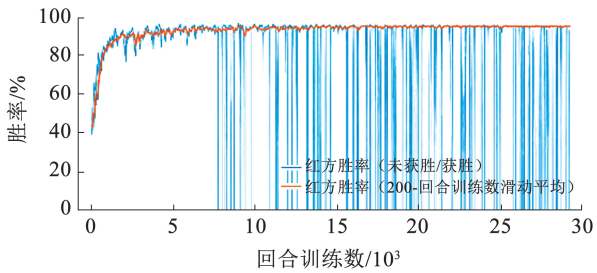


图 5 CM-MADDPG 与随机策略对比

Fig. 5 Comparison between CM-MADDPG and random policy

从图 4 可以看出,Meta-MADDPG 中,红方胜率在早期收敛上更为突出,约 500 回合训练数滑动胜率就能升到 0.85 左右,随后长期维持在 0.85 ~ 0.90区间,整体曲线比前两种更平顺,说明低频的 Reptile 式参数对齐确实提高了样本效率、弱化了梯度振荡。最终的 CM-MADDPG 同时叠加了评论家侧结构增强和训练期元策略调度,在图 5 中展示约 1 000 回合训练数后便进入高位平台,滑动胜率基本保持在 0.95 以上并逼近 1.00,虽然逐回合的未获胜/获胜曲线仍会偶发个别失败,但并未影响整体平均水平,体现出在该对抗基准上最快的收敛速度、最高的稳态胜率,以及最小的方差。由此可以得出:评论家增强主要改善中后段的价值建模能力,元策略主要加速早段学习和稳定更新,而二者结合时可以在整个训练阶段都获得优于单一改进的综合性能。

图 6 ~ 9 分别给出了 MADDPG、Critic-MADDPG、Meta-MADDPG,以及 CM-MADDPG 为红方策略时红、蓝双方 6 架无人机各自的奖励情况。

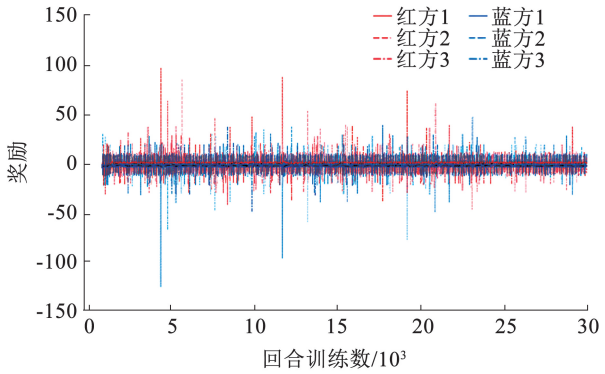


图 6　红方 MADDPG 时各无人机奖励情况

Fig. 6　Reward status of each unmanned aerial vehicle when red side adopts MADDPG

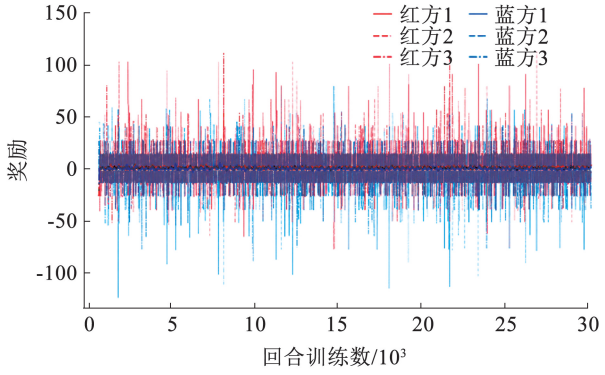


图 7　红方 Critic-MADDPG 时各无人机奖励情况

Fig. 7　Reward status of each unmanned aerial vehicle when red side adopts Critic-MADDPG

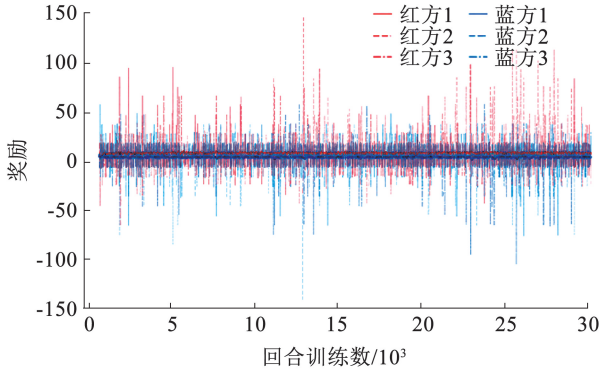


图 8　红方 Meta-MADDPG 时各无人机奖励情况

Fig. 8　Reward status of each unmanned aerial vehicle when red side adopts Meta-MADDPG

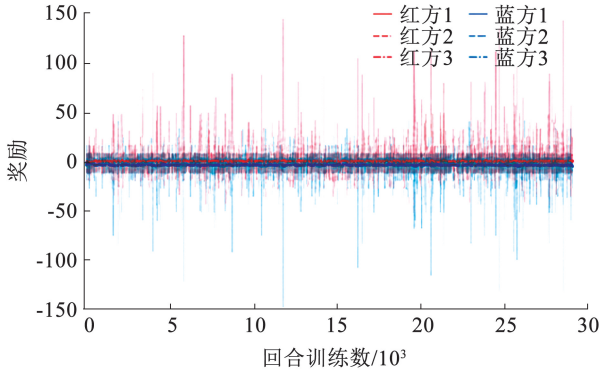


图 9　红方 CM-MADDPG 时各无人机奖励情况

Fig. 9　Reward status of each unmanned aerial vehicle when red side adopts CM-MADDPG

从 6 架无人机的回报演化曲线可以看出，不同算法在回报幅值、红、蓝双方分离度，以及波动性上呈现出由弱到强的层级关系。从图 6 可以看出，MADDPG 基线图中红、蓝双方的曲线整体都围绕 0 附近的负区间窄幅摆动，红方 3 机与蓝方 3 机的平均回报训练数几乎重合，只是在少量回合训练中出现由命中触发的尖锐正脉冲，但这些脉冲既不频繁也难以持续，说明在稀疏奖励、6 机同时博弈的条件下，原始 MADDPG 很难稳定地从经验池中分辨出有效攻击与合理机动的样本。从图 7 可以看出，在加入集中式评论家增强后的 Critic-MADDPG 曲线中，红方 3 条线的正向脉冲明显增多，同时峰值也更高，说明价值网络对接近并命中对手这类联合状态–动作的收益刻画更准确，红、蓝双方的回合训练数平均水平开始拉开，但仍可以看到回报训练数上下浮动较大，说明效果仍然不够稳定。从图 8 可以看出，在引入元策略外圈后的 Meta-MADDPG 中，在前 5 000 回合训练数左右就能把红方个体的滑动平均值推到接近 0 甚至略高的位置，而蓝方仍然围绕负值波动，这表明元策略帮助 Actor 更快适应当前经验分布、减少了早期的无效探索，训练后段的抖动也有所收敛。从图 9 可以看出，最密集、幅度最高的正向脉冲几乎都来自使用了 CM-MADDPG 的红方 3 机，而且它们的滑动平均值长期压在蓝方之上，红方曲线的基线也更平直，说明评论家侧的注意力建模已经把各自需要关注的目标平均筛选出来，而元策略又把这种能力更快地固化到策略参数里，于是命中奖励能够抵消贴边、扎堆、落空等负项，形成稳定的正收益区间。与胜率曲线对应，图 6 ~9回报图进一步说明：单独的评论家增强主要针对价值估计精确性的问题，单独的元策略主要针对学习效率提升的问题；二者联合后，既提高了红方个体获得高回报的频度，又降低了训练过程中回报的整体方差，从而呈现出性能最优且收敛最稳定的对抗效果。

4　结　论

1）面向 3 × 3 红、蓝双方空中对抗任务，在 CTDE 范式下提出 CM-MADDPG：在集中式评论家侧引入轻量的结构化建模以增强跨机交互的价值估计，并在训练阶段叠加 Reptile 元策略外圈对 Actor 进行低频靠拢更新；其余过程与 MADDPG 保持一致，从而在不增加推理期开销的前提下改善学习过程。

2）基于自建 UAV 3 × 3 基准的对比与消融表明，CM-MADDPG 较 MADDPG 及其单独增强变体

(Critic-MADDPG、Meta-MADDPG)在收敛速度、稳态胜率及团队、个体回报等指标上均取得显著提升;训练曲线更平滑、方差更小,体现出更好的样本效率与训练稳定性。

3)所提方法保持推理期复杂度与基线一致,具备较好的工程可用性与鲁棒性;评论家增强侧重提升价值估计的准确性与稳定性,元策略增强侧重提升学习效率与早期收敛,二者形成互补并在本对抗基准上获得最优综合表现。

4)需要注意的是,当前实验仅在二维环境和随机蓝方的条件下开展仿真实验,因而对高维、强对抗场景的推广还需进一步验证。后续工作将重点在引入更强的蓝方策略、将算法的环境推广至三维空间环境中,以及在更大规模多机对抗上测试算法的泛化能力等 3 个方面展开相应研究。

参考文献

[1]吴志娟,林艳,张一晋,等. 基于多智能体协同的无人机簇群多域节能抗干扰通信[J]. 中国科学: 信息科学, 2023, 53(12): 2511
Wu Zhijuan, Lin Yan, Zhang Yijin, et al. Multi-agent collaboration based UAV clusters multi-domain energy-saving anti-jamming communication [J]. Scientia Sinica Informationis (Information Sciences), 2023, 53(12): 2511. DOI: 10.1360/SSI-2023-0015

[2]Jiang Xu, Chen Xinqi, Tang Jie, et al. Covert communication in UAV-assisted air-ground networks[J]. IEEE Wireless Communications, 2021, 28(4): 190. DOI: 10.1109/MWC.001.2000454

[3]Zhang Qirui, Wei Ruixuan. Ground attack strategy of cooperative UAVs for multitargets[J]. Complexity, 2019, 2019(1): 9428087. DOI: 10.1155/2019/9428087

[4]Chen Xia, Hu Naikuan. Effectiveness evaluation for UAV air-to-ground attack based on PSO-BP neural network [C]//2017 36th Chinese Control Conference (CCC). Dalian: IEEE, 2017: 3864. DOI: 10.23919/ChiCC.2017.8027961

[5]Ernest N, Cohen K, Kivelevitch E, et al. Genetic fuzzy trees and their application towards autonomous training and control of a squadron of unmanned combat aerial vehicles [J]. Unmanned Systems, 2015, 3(3): 185. DOI: 10.1142/s2301385015500120

[6]Saska M, Vonášek V, Chudoba J, et al. Swarm distribution and deployment for cooperative surveillance by micro-aerial vehicles[J]. Journal of Intelligent & Robotic Systems, 2016, 84(1): 469. DOI: 10.1007/s10846-016-0338-z

[7]Savkin A V, Huang Hailong. A method for optimized deployment of a network of surveillance aerial drones[J]. IEEE Systems Journal, 2019, 13(4): 4474. DOI: 10.1109/JSYST.2019.2910080

[8]Huang Hailong, Savkin A V. An algorithm of reactive collision free 3-D deployment of networked unmanned aerial vehicles for surveillance and monitoring[J]. IEEE Transactions on Industrial Informatics, 2020, 16(1): 132. DOI: 10.1109/TII.2019.2913683

[9]Zhang Jiandong, Yang Qiming, Shi Guoqing, et al. UAV cooperative air combat maneuver decision based on multi-agent reinforcement learning[J]. Journal of Systems Engineering and Electronics, 2021, 32(6): 1421. DOI: 10.1109/JSEE.2021.00015

[10]He Wenjian, Qi Xiaogang, Liu Lifang. A novel hybrid particle swarm optimization for multi-UAV cooperate path planning[J]. Applied Intelligence, 2021, 51(10): 7350. DOI: 10.1007/s10489-020-02082-8

[11]何旭, 景小宁, 冯超. 基于蒙特卡洛树搜索方法的空战机动决策[J]. 空军工程大学学报(自然科学版), 2017, 18(5): 36
He Xu, Jing Xiaoning, Feng Chao. Air combat maneuver decision based on MCTS method [J]. Journal of Air Force Engineering University (Natural Science Edition), 2017, 18(5): 36. DOI: 10.3969/j.issn.1009-3516.2017.05.007

[12]Rashki M. The soft Monte Carlo method[J]. Applied Mathematical Modelling, 2021, 94: 558. DOI: 10.1016/j.apm.2021.01.022

[13]嵇慧明,余敏建,乔新航,等. 改进 BAS-TIMS 算法在空战机动决策中的应用[J]. 国防科技大学学报,2020,42(4):123
Ji Huiming, Yu Minjian, Qiao Xinhang, et al. Application of the improved BAS-TIMS algorithm in air combat maneuver decision[J]. Journal of National University of Defense Technology, 2020, 42(4): 123. DOI: 10.11887/j.cn.202004019

[14]刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1
Liu Quan, Zhai Jianwei, Zhang Zongchang, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1. DOI: 10.11897/SP.J.1016.2018.00001

[15]Lowe R, Wu Yi, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: MIT Press, 2017: 6382. DOI: 10.5555/3294771.3294860

[16]Rashid T, Samvelyan M, de Witt C S, et al. Qmix: monotonic value function factorisation for deep multi-agent reinforcement learning[PP/OL]. V2. arXiv (2018-06-06) [2026-05-08]. https://doi.org/10.48550/arXiv.1803.11485

[17]Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward [C]//Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Stockholm: ACM, 2018: 2085. DOI: 10.5555/3237383.3238080

[18]Wang Fei, Zhu Xiaoping, Zhou Zhou. MADDPG-based multi-UAV autonomous collaborative attack in confrontation scenarios [C]// Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning (CACML2024). Singapore: ACM, 2024: 271. DOI: 10.1145/3654823.3654873

[19]Yang Jianfeng, Yang Xinwei, Yu Tianqi. Multi-unmanned aerial vehicle confrontation in intelligent air combat: a multi-agent deep reinforcement learning approach[J]. Drones, 2024, 8(8): 382. DOI: 10.3390/drones8080382

[20]Wang Baolai, Li Shengang, Gao Xianzhong, et al. Weighted mean field reinforcement learning for large-scale UAV swarm confrontation [J]. Applied Intelligence, 2023, 53(5): 5274. DOI: 10.1007/s10489-022-03840-6

[21]Sun Rui, Silver D, Tesauro G, et al. Introduction to the special issue on deep reinforcement learning: an editorial [J]. Neural Networks, 2018, 107: 1. DOI: 10.1016/j.neunet.2018.08.001

[22]Lü Lei, Chen Ziming, Lu Zhenyu. A novel neural-network gradient optimization algorithm based on reinforcement learning[C]//2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). Guangzhou: IEEE, 2020: 106. DOI: 10.1109/SPAC49953.2019.237884

(编辑 张 红)