# 基于多智能体强化学习的大规模无人机集群对抗

王泊涵[1]，吴婷钰[2]，李文浩[2]，黄达[2]，金博[2,3]，杨峰[4]，周爱民[2]，王祥丰[2,3]
(1. 国防科技大学 系统工程学院，湖南 长沙 410073；2. 华东师范大学 计算机科学与技术学院，上海 200062；
3. 上海自主智能无人系统科学中心 可信人工智能研究所，上海 200062；4. 中国人民解放军 军事科学院，北京 100039)

**摘要：**攻击成本低、体系生存率高且具备细粒度灵活作战能力的无人机集群作战未来将成为重要的战争形态。高效而适应性地对无人机集群进行细粒度任务规划，对提高集群作战效能具有重要意义。多智能体强化学习在解决群体序列决策任务时存在维度灾难及组合爆炸，多适用于小规模场景。将对抗环境中的无人机集群任务规划问题建模为马尔可夫博弈问题，基于平均场理论将大规模无人机间复杂交互过程建模为单机与集群平均影响间的交互，建立大规模集群对抗仿真环境，提出基于平均场的多智能体强化学习算法用于求解集群任务规划问题，其高效性和灵活性得到充分验证。

**关键词：**无人机集群；细粒度任务规划；多智能体强化学习；马尔可夫博弈；动态对抗；平均场
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X (2021) 08-1739-15
DOI: 10.16182/j.issn1004731x.joss.21-0476

## Large-scale UAVs Confrontation Based on Multi-agent Reinforcement Learning

*Wang Bohan*[1], *Wu Tingyu*[2], *Li Wenhao*[2], *Huang Da*[2], *Jin Bo*[2,3], *Yang Feng*[4], *Zhou Aimin*[2], *Wang Xiangfeng*[2,3]
(1. School of System Engineering, National University of Defense Technology, Changsha 410073, China;
2. School of Computer Science and Technology, East China Normal University, Shanghai 200062, China;
3. Institute of Trusted Artificial Intelligence, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 200062, China;
4. Academy of Military Sciences PLA China, Beijing 100039, China)

**Abstract:** UAV swarms operation with low attack cost, high survival rate, and fine-grained flexible capabilities will be the important form in future war. Efficient and adaptive fine-grained task planning of UAV is very important in enhancing the drone swarm operation effectiveness. Due to dimensional disaster and explosive combination, multi-agent reinforcement learning can only be applicable to the sequential decision-making tasks in small-scale scenarios. The swarm mission planning in an adversarial environment is modeled as a Markov game, and the mean field theory is adopted to simplify the complex interaction between large-scale UAVs, as the interaction model between a single UAV and average impact of the multiple nearby UAVs. The multi-agent reinforcement learning algorithm based on the mean field theory is used to solve the adaptive task planning problem of the drone swarm in the large-scale confrontation simulation environment, and its high efficiency and flexibility are verified.

**Keywords:** UAV swarm; fine-grained task planning; multi-agent reinforcement learning; Markov game; dynamic confrontation; mean field

## 引言

经过数十年发展，无人机在替代人类执飞任务中展现了巨大优势，尤其是 4D (Dull，Dirty，Dangerous and Deep)任务的复杂、多变、难以预测使得以小体积却速度快的低成本无人机脱颖而出，

收稿日期：2021-05-25 修回日期：2021-07-12
基金项目：国家自然科学基金(12071145)；科技部新一代人工智能重大专项(2020AAA0107400)；之江实验室开放课题(2021KE0AB03)
第一作者：王泊涵(1987-)，男，硕士生，高工，研究方向为指挥控制系统。E-mail: cnpeking@qq.com
通讯作者：王祥丰(1987-)，男，博士生，副教授，研究方向为多智能体强化学习、分布式优化等。E-mail: xfwang@cs.ecnu.edu.cn

从而能够有效避免人员伤亡[1-2]。而现代战争由于其复杂多变的战场环境，仅凭单架无人机已无法全面应对敌方的攻击，而以多架无人机组成的无人机编队更能够在进行协同互补的基础上适应作战环境，从而实现战斗能力的全面提升。具体来说，相对于单一作战、全面防御的单无人机作战模式，无人机集群作战模式不仅融合了单无人机的强大功能，并且更加注重无人机集群协同作战、共同抗击，以集结单一的作战能力优势以及集群协作能力优势，在复杂战斗环境中展现了以下 3 个优势[3-4]：①攻击成本低。无人机集群一般由大量低成本、生产效率高且功能单一的无人机组成，单架无人机的研发、生产以及战损成本远低于功能完备的单机作战平台。美国海军研究局研发的“鱼叉”导弹成本高达 120 万美元，而以低成本优势研发的“郊狼”无人机成本仅为 1.5 万美元。因此采用由大量廉价无人机组成的集群作战平台取代全功能大型作战平台将成为现代战争的趋势。②体系生存率高。无人机集群规模庞大、分布零散、单目标特征小，这使得敌方防御系统很难对其进行识别以及锁定；例如美国国防战略能力办公室在 2014 年 9 月通过 F-16 战机投放的“山鹑”无人机机身尺寸不足一部智能手机，且重量仅为 0.45 kg。此外，在无人机集群协同作战过程中，即使部分无人机受到攻击或被摧毁，集群整体依然能够不受影响并完成作战任务，这得益于无人机集群所具备的去中心化、自主化特点。③能够实现细粒度灵活作战。无人机集群中的每架无人机都可分别根据局部信息分别执行各自的作战任务，包括巡航、侦察、出击等，从而实现远超单一作战平台的集群作战体系。例如，美国诺斯洛普公司发现将 15 架 X-47B 无人机与有人机组成小规模混合集群，集群中的每架无人机配合有人机协同执行巡逻、侦察、监视以及攻击任务，实现作战能力的巨大提升。

因此，无人机的作战样式正在从单机作战向集群作战方向转变[5-7]，且备受各国重视。2016 年上半年，美国空军发布了近 20 年来的小型无人机飞行规划，该规划中对实现无人机集群协同作战方式进行详细研究。此外，俄罗斯在同年披露将在 10 年后发布下一代战斗机的作战方式，预期利用战斗机协同控制无人机进行集群作战。

现有方法一般将无人机集群任务规划拆解为协同搜索、饱和攻击以及动态对抗 3 个方面，分别设计算法进行解决[4]。具体来说，无人机集群系统的每架无人机在作战过程中对未知区域进行搜索，定位、监视敌方无人机，从而实现后续任务的有效推进。现有协同搜索算法一般基于全局搜索图[8-9]，采用中心化[10-11]或分布式方法[12-13]，对集群中所有无人机的航线进行规划。然而，航线规划过程中还要考虑无人机实时避障、防撞等。现有方法通过在原始图搜索问题中加入约束条件来解决这些问题[14-16]，但都复杂度较高，无法扩展到大规模场景，且很难适应外部环境的动态变化。饱和攻击即在无人机集群发现敌方目标后，通过恰当的作战单位分配来实现对敌方高价值目标的精确打击[17]。现有算法将作战单位分配建模为整数规划问题[18-20]，并采用中心化[21-23]或去中心化[24-26]的传统优化算法进行求解。接下来无人机一般基于现有算法设计的人工规则来对目标实施饱和攻击[27-29]。上述现有算法的设计思想使得该框架同样很难扩展到大规模场景，且同样也不具有适应动态环境的能力。也有一些现有算法专门解决无人机集群的动态对抗问题，通过出击相同规模的无人机集群来拦截敌方无人机集群的饱和攻击[30]。现有算法主要基于人工规则对动态对抗场景中的无人机集群任务规划问题进行求解[31-32]。

多智能体深度强化学习是多智能体系统与强化学习、深度学习的结合领域，致力于通过在一个公共环境中让智能体不断与环境进行交互试错，基于深度学习来解决多个智能体的序列决策问题[33]。而对抗环境中的无人机集群细粒度的任务规划就是一个典型的多智能体系统序列决策问题。其中，

外部环境，例如地形、气象以及敌方无人机集群、我方无人机集群共同构成一个多智能体系统。该系统中，每架我方无人机都表示一个智能体。整个多智能体系统要完成的任务就是通过所有智能体的协同来歼灭敌方无人机集群。将对抗环境中的无人机集群控制问题建模为一个多智能体强化学习问题后，相比于已有算法有如下优点：①完全通过目标驱动的多智能体强化学习算法可以将对抗环境中的无人机集群任务规划问题中的协同搜索、饱和攻击以及动态对抗 3 个子问题当成一个整体进行学习，从而大幅降低了问题的建模复杂度。②基于无模型的多智能体强化学习算法不需要事先知晓外部环境的物理模型，而是通过让智能体不断与环境交互试错来从中习得有效策略。这种端到端学习的特点也避免了繁琐的基于专家经验的人工规则的设计。③多智能体强化学习算法可以通过修改奖励函数来高效且灵活地处理除主要任务之外的诸多约束，例如避障、防撞等。然而，现有的多智能体强化学习算法与上述已有算法类似，仅能解决少数智能体场景[34-35]。

本文主要研究目标是应用多智能体强化学习算法高效处理大规模无人机集群的细粒度任务规划问题。虽然无人机集群由于规模巨大使得其中交互过程十分复杂，但无人机集群中的无人机交互过程存在一个显著特点，即每架无人机只与少量的其他无人机直接进行交互。通过直接交互的链式关系，无人机集群中任意两架无人机都可实现全局的间接交互[36]。这类大规模复杂交互系统的可扩展性可以通过平均场理论[37]来解决。具体来说，无人机集群群内的复杂交互过程可以通过单架无人机与整体(局部)无人机集群的平均效应相互作用来近似。这种学习过程是在两架无人机而不是多架无人机之间相互加强：单架无人机的最优任务规划策略的学习是基于无人机集群的联合策略的动态变化，而无人机集群的联合策略则根据单架无人机的任务规划策略进行更新。

基于上述思想，本文将对抗环境的无人机集群细粒度任务规划的协同搜索、饱和攻击以及动态对抗整体建模为 1 个马尔可夫博弈，并应用无模型的多智能体强化学习算法解决该博弈问题。该算法将敌我双方战损情况以及避障损失、撞击损失等约束融入奖励函数的设计中，通过目标驱动的端到端的训练过程让无人机集群自动地习得有机融合协同搜索、饱和攻击以及动态对抗的细粒度策略。

为了高效解决多智能体强化学习算法的可扩展性，本文将平均场理论引入到无模型多智能体强化学习算法的设计中，提出了平均场 Q 学习算法。在本文建立的大规模无人机集群对抗仿真环境中，基于平均场的多智能体强化学习算法在胜率上明显优于现有的解决多智能体序列决策问题的多智能体基准算法。

本文的核心贡献主要包括：①将对抗环境中无人机集群细粒度任务规划问题中协同搜索、饱和攻击及动态对抗整体建模为一个马尔可夫博弈，并引入多智能体强化学习算法解决该问题。②将敌我双方战损情况以及避障损失、撞击损失等约束融入多智能体强化学习算法的奖励函数设计中，通过目标驱动的端到端的训练过程让无人机集群自动地习得有机融合协同搜索、饱和攻击以及动态对抗的细粒度策略。③使用结合平均场理论的无模型多智能体强化学习算法用以解决算法在大规模无人机集群下的可扩展性问题，并在本文建立的大规模无人机集群对抗仿真环境中验证了基于平均场的多智能体强化学习算法的高效性以及灵活性。

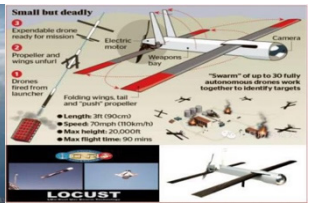
## 1 相关工作

### 1.1 无人机集群在研项目

美国空军在 20 世纪 90 年代末期提出无人机集群作战的概念，以美国国防部为领导的无人机集群作战项目在功能、体系上互为补充[5,17,38]。美国国防高级研究局在 2015 年 9 月发布的

“Gremlins”项目(如图 1(a)所示)，该项目中的运输机通过发射无人机集群来承担巡航、侦察、攻击以及返航无人机的接受任务。该项目旨在取代 F-35 等高成本装备来突破敌防御系统。美国海军在 2015 年 4 月公布的低成本无人机集群项目(如图 1(b)所示)提出基于自组网技术，增强集群之间的消息传递以及信息共享，并在 2016 年 4 月成功组建 30 架“郊狼”无人机的编队并成功完成无人机协同作战的飞行实验。而早在 2012 年美国已经开展“山鹑”微型无人机集群项目的研究(如图 1(c)所示)，并于 4 年后利用 F-16 战机进行了空中投放试验。2017 年 1 月，来自美国海军的 3 架 F/A-18F 战斗机成功投放了 103 架无人机并完成了编队决策任务，其中每架战斗机的投放速度高达 0.6 马赫。同样在 2017 年 1 月，美国防高级研究局提出能够由地面站轻松控制无人机集群的进攻性蜂群使能项目(如图 1(d)所示)，开发并测试专为城市作战的蜂群战术。

(a) “Gremlins”项目




(b) “Locust”项目


(c) “Gremlins”项目

(d) “Locust”项目

图 1 美军无人机集群典型在研项目

Fig. 1 Typical ongoing research project of U.S. military UAVs

我国无人机集群技术起步稍晚但发展迅速。2016 年 9 月，国内第一个无人机集群作战系统——“集群式箱式发射折叠翼无人机”系统研发成功；2016 年 11 月国内某公司完成接近 70 架的固定翼无人机的试飞试验，且在 2017 年 7 月和 2018 年 5 月再次分别实现了接近 200 架的固定翼无人机的编队起飞、空中集结等复杂策略。

## 1.2 无人机集群作战任务规划技术

复杂对抗环境下的无人机集群作战任务引起了国内外研究的热潮，无人机集群作战任务规划技术主要可以分为协同搜索方式、饱和攻击方式以及动态对抗方式。

### 1.2.1 无人机集群协同搜索

在无人机集群系统执行任务过程中，对未知区域进行搜索，从而对于敌方目标进行定位是执行后续任务的基础与前提。传统以搜索论[39]为基础的方法基于最大似然预先设计协同搜索航线，但由于外部环境的不确定性以及无人机集群中交互过程的动态性，使得预先设计无法实现。因而目前方法常基于搜索图[8-9]构造外部环境的二维离散地图，并根据无人机反馈对地图进行实时的搜索路径规划。基于搜索图的算法早期一般基于中心化架构，由中心节点对搜索图进行实时更新。文献[10]通过设计全局目标概率图实现了集群的协同搜索。文献[11]将集群协同搜索问题建模为模型预测控制问题并通过遗传算法进行模型求解。然而，当无人机集群规模较大时，中心化架构计算量及通信量都将呈指数级增长。文献[12]基于分布式模型预测控制算法将协同搜索任务转化为每架无人机的局部优化模型，结合粒子群优化算法寻求各局部优化问题的纳什最优解。文献[13]在此基础上针对时敏目标进行了优化，并基于时域滚动方法实现了协同搜索决策的输出。

考虑到无人机集群的实时避障，文献[14]提出了基于模拟退火的粒子群优化算法。文献[15]针对集群遭遇突发威胁，将搜索模式分为正常协同搜索以及紧急避障，基于 Dubins 曲线进行实时的突发避障路径规划。文献[16]在文献[15]的基础上研究了无人机回收约束，从而保证集群能够在航程内返回起点。

### 1.2.2 无人机集群饱和攻击

无人机集群发现敌方目标后，需要通过恰当的作战单位分配来实现对敌方高价值目标的精确打击，从而为后续武装力量进入保驾护航[17]。饱和攻击任务关键在于任务分配，即在一定约束下，合理地将多个攻击任务分配给多个无人机子集群。现有方法主要使用混合整数显示规划模型[18]、动态网络流优化模型[19]以及多维多选择背包模型[20]对任务分配问题进行建模，并常采用粒子群算法[21]、蚁群算法[22]以及遗传算法[23]等中心化算法对上述模型进行求解。同样考虑到中心化架构的局限性，现有工作提出了基于合同网协议的市场竞拍算法[24]、分布式马尔可夫决策方法[25]以及多智能体满意决策方法[26]等分布式算法。

在进行合理的任务分配后，饱和攻击需要进行攻击时机以及位置的确定。文献[27]基于排队论研究了三层防御系统的突防概率，提供了饱和攻击的理论基础。文献[28]基于一系列的简单行为规则实现了无人机集群的多点饱和攻击。文献[29]基于一致性算法提出了一种无人机集群构成指定队形的控制方法。

### 1.2.3 无人机集群动态对抗

针对上文所述的无人机集群饱和攻击，涌现了无人机集群动态对抗方式进行抗衡，具体来说，通过出动无人机集群进行拦截[30]。早期工作提出了一些用于无人机集群动态对抗的概念模型。文献[40]提出了基于多智能体理论的爱因斯坦对抗模型。文献[41]提出了信念、目标、战术意图等空战自主决策思维。文献[30]建立了上层为指挥层，下层为任务层的分层无人机集群对抗建模框架。现有的动态对抗算法大多基于分治法与人工规则相结合的思想。文献[31]基于多智能体理论建立了集群对抗模型，并设计相应的行为规则集合。文献[32]在文献[31]的基础上引入由巡航、接近、远离、攻击、支援等构成的规则集合。

## 1.3 多智能体强化学习

现有的处理对抗场景的多智能体强化学习算法主要包括基于神经网络的敌手建模[42]、策略函数共享[43]，以及基于中心化训练去中心化执行框架的通信学习[44]、协作学习[45-46]等。然而，上述方法只适用于至多数十个智能体的场景。当智能体数目增加时，不仅 $Q$ 值函数的输入空间大小呈指数级增长，更为严重的是由于其余智能体探索行为导致的累积噪声将使得 $Q$ 值函数的学习极为困难。本文使用的基于平均场的多智能体强化学习算法通过在联合动作空间上使用平均场估计[37]来处理上述问题。由于平均场估计将无人机集群中的复杂交互转化为两个实体之间的交互(单架无人机与邻近无人机分布之间的交互)使得 $Q$ 值函数的参数量与智能体的数目无关，从而该方法可以有效缓解由于其余智能体导致的探索噪声问题[47]，且使得每个智能体都能够高效决定对自身有益的局部决策。

基于平均场的多智能体强化学习算法与平均场博弈[48-50]具有相似之处。平均场博弈主要研究由于个体决策导致的群体行为。用数学语言来描述，群体行为动态由 2 个随机微分方程确定。其中后向方程确定个体值函数的后向动态，前向方程确定群体行为分布的前向动态。尽管后向方程等效地描述了贝尔曼方程在马尔可夫决策过程中表示的内容，但平均场博弈的主要目标而是基于模型的规划，并推断个体密度随时间的变化情况。平均场理论[37]在物理学中同样被广泛应用，然而基于平均场的多智能体强化学习算法的不同点在于其关注于无模型的最优策略学习，即系统动态以及奖励函数对于学习算法来说是未知的。最近，文献[51]同样将强化学习与平均场博弈相结合。然而文献[51]主要关注从已有专家数据中使用逆强化学习来同时学习奖励函数以及平均场博弈的前向方程，本文使用的方法则聚焦于在时序差分学习模型下学习一个可计算的 $Q$ 值函数。

## 2 背景知识

多智能体强化学习属于强化学习与博弈论的相交领域，两者结合形成了马尔可夫博弈的总体框架。

一个由 $N$ 个智能体组成的马尔可夫博弈 $\Gamma$ 可由元组 $\Gamma \triangleq \left(S, A^1, A^2, \cdots, A^N, r^1, r^2, \cdots, r^N, p, \gamma\right)$ 进行表示。其中 $S$ 表征状态空间，$A^j$ 表征智能体 $j \in \{1,2,\cdots,N\}$ 的动作空间。智能体 $j$ 的奖励函数定义为 $r^j : S \times A^1 \times A^2 \times \cdots \times A^N \to R$。转移概率 $p : S \times A^1 \times A^2 \times \cdots \times A^N \to \Omega(S)$ 表征了状态随时间的随机演化，$\Omega(S)$ 则是状态空间概率分布集合。常数 $\gamma \in [0,1)$ 表示奖励折扣因子。在时间步 $t$，所有智能体同时采取行动后分别收到由于采取了之前行动得到的即时奖励 $r_t^j$。

智能体根据策略采取行动。对于智能体 $j$，其对应的策略定义为 $\pi^j : S \to \Omega(A^j)$，其中 $\Omega(A^j)$ 表示智能体 $j$ 动作空间 $A^j$ 上的概率分布集合。$\pi \triangleq \left[\pi^1, \pi^2, \cdots, \pi^N\right]$ 被定义为所有智能体的联合策略。给定初始状态 $s$，智能体 $j$ 在联合策略 $\pi$ 下的值函数被写作期望未来累积折扣奖励：

$$v_\pi^j(s) = \sum_{t=0}^{\infty} \gamma^t E_{\pi,p}\left[r_t^j \mid s_0 = s, \pi\right] \tag{1}$$

在 $N$ 智能体博弈框架下，基于公式(1)以及贝尔曼方程，智能体 $j$ 在联合策略 $\pi$ 下的 $Q$ 函数(或动作值函数) $Q_\pi^j : S \times A^1 \times A^2 \times \cdots \times A^N \to R$ 可以形式化为：

$$Q_\pi^j(s,a) = r^j(s,a) + \gamma E_{s' \sim p}\left[v_\pi^j(s')\right] \tag{2}$$

式中：$s'$表征下一时间步的状态。值函数 $v_\pi^j$ 可以通过公式(2)中的 $Q$ 函数进行表征：

$$v_\pi^j(s) = E_{a \sim \pi}\left[Q_\pi^j(s,a)\right] \tag{3}$$

公式(2)中的 $Q$ 函数通过引入所有智能体做出的联合动作 $a \triangleq \left[a^1, a^2, \cdots, a^N\right]$，从单智能体博弈中扩展到了 $N$ 智能体博弈场景下。公式(3)是对联合动作求取期望值。

本文将对抗场景下无人机集群的细粒度任务规划问题建模为一个离散时间下的马尔可夫博弈。该博弈被假定是不完备的，但具有完美信息[52]，即每架无人机既不了解博弈状态转移概率也不知道其余无人机的奖励函数，但能够观察并响应其余无人机先前的动作以及即时奖励。

## 3 平均场多智能体强化学习

联合动作 $a$ 的维度随着智能体数目 $N$ 的增大而呈指数级增长。由于所有智能体同时行动并根据联合动作计算各自的值函数，因而当无人机集群规模过大时 $Q$ 函数 $Q^j(s,a)$ 的计算将变得极为困难。为了解决这个问题，平均场多智能体强化学习算法通过仅使用局部双边交互将 $Q$ 函数进行分解：

$$Q^j(s,a) = \frac{1}{N^j} \sum_{k \in N(j)} Q^j\left(s, a^j, a^k\right) \tag{4}$$

式中：$N(j)$ 为智能体 $j$ 邻近智能体的序号集合，大小为 $N^j = \left|N(j)\right|$。将 $Q$ 函数通过该智能体与邻居的双边估计进行分解后，不仅大幅降低了无人机集群中的交互复杂度，还能够保留集群中任意两架无人机之间的间接交互[36]。

### 3.1 平均场估计

公式(4)中的双边交互 $Q^j\left(s, a^j, a^k\right)$ 可以使用平均场理论[37]进行估计。在对现实无人机集群作战需求与算法复杂度进行权衡后，本文考虑离散的动作空间。每个智能体 $j$ 的动作 $a^j$ 是一个通过独热编码表示的离散类别变量，独热编码中的每一位表示 $D$ 个可能动作 $a^j \triangleq \left[a_1^j, a_2^j, \cdots, a_D^j\right]$ 中的一个。如图 2 所示。

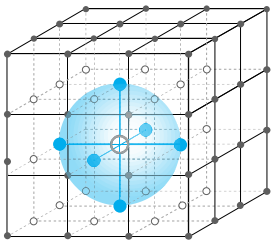


图 2 平均场估计
Fig. 2 Mean field approximation

平均场多智能体强化学习算法根据智能体 $j$ 的邻近智能体集合 $N(j)$ 计算平均动作 $\bar{a}^j$，并将每个邻居智能体 $k$ 的独热动作 $a^k$ 表示为平均动作 $\bar{a}^j$ 以及一个微小扰动 $\delta a^{j,k}$ 之和：

$$a^k=\bar{a}^j+\delta a^{j,k},\text{其中}\bar{a}^j=\frac{1}{N^j}\sum_k a^k$$

式中：$\bar{a}^j\triangleq\left[\bar{a}_1^j,\bar{a}_2^j,\cdots,\bar{a}_D^j\right]$，可以解释为智能体 $j$ 的邻近智能体采取的动作的经验分布。

根据泰勒定理，双边 $Q$ 函数 $Q^j\left(s,a^j,a^k\right)$ 如果关于智能体 $k$ 采取的动作 $a^k$ 二阶可微，则有

$$\begin{aligned}&Q^j(s,a)=\frac{1}{N^j}\sum_k Q^j\left(s,a^j,a^k\right)=\\&\frac{1}{N^j}\sum_k\left[Q^j\left(s,a^j,\bar{a}^j\right)+\nabla_{a^j}Q^j\left(s,a^j,\bar{a}^j\right)\cdot\delta a^{j,k}\right]+\\&\frac{1}{N^j}\sum_k\left[\frac{1}{2}\delta a^{j,k}\cdot\nabla^2_{\tilde{a}^{j,k}}Q^j\left(s,a^j,\tilde{a}^{j,k}\right)\cdot\delta a^{j,k}\right]=\\&Q^j\left(s,a^j,\bar{a}^j\right)+\nabla_{a^j}Q^j\left(s,a^j,\bar{a}^j\right)\cdot\left[\frac{1}{N^j}\sum_k\delta a^{j,k}\right]+\\&\frac{1}{2N^j}\sum_k\left[\delta a^{j,k}\cdot\nabla^2_{\tilde{a}^{j,k}}Q^j\left(s,a^j,\tilde{a}^{j,k}\right)\cdot\delta a^{j,k}\right]=\\&Q^j\left(s,a^j,\bar{a}^j\right)+\frac{1}{2N^j}\sum_k R^j_{s,a^j}\left(a^k\right)\approx Q^j\left(s,a^j,\bar{a}^j\right)\end{aligned}\quad(5)$$

式中：$R^j_{s,a^j}\left(a^k\right)\triangleq\delta a^{j,k}\cdot\nabla^2_{\tilde{a}^{j,k}}Q^j\left(s,a^j,\tilde{a}^{j,k}\right)\delta a^{j,k}$ 表示泰勒多项式余项，$\tilde{a}^{j,k}=\bar{a}^{j,k}+\epsilon^{j,k}\delta a^{j,k}$，$\epsilon^{j,k}\in[0,1]$。

从图 2 可以看出，基于平均场估计，智能体 $j$ 与每个邻近智能体 $k$ 之间所有的双边交互 $Q^j(s,a^j,a^k)$ 被简化为中心智能体 $j$ 与虚拟平均智能体之间的交互(图 2 中蓝色区域即为邻居节点的平均作用)。该虚拟平均智能体通过智能体 $j$ 所有邻近智能体的平均作用抽象而出。上述交互过程就可以被简化并由公式(5)中的 $Q^j\left(s,a^j,\bar{a}^j\right)$ 表示。

在学习过程中，给定 1 条样本数据 $e=\left(s,\left\{a^k\right\},\left\{r^j\right\},s'\right)$，平均场 $Q$ 函数通过以下方式循环更新：

$$\begin{aligned}Q^j_{t+1}\left(s,a^j,\bar{a}^j\right)=&(1-\alpha)Q^j_t\left(s,a^j,\bar{a}^j\right)+\\&\alpha\left[r^j+\gamma v^j_t\left(s'\right)\right]\end{aligned}\quad(6)$$

$$v^j_t\left(s'\right)=\sum_{a^j}\pi^j_t\left(a^j s',\bar{a}^j\right)\mathbb{E}_{\bar{a}^j\left(a^{-j}\right)\sim\pi^{-j}_t}\left[Q^j_t\left(s',a^j,\bar{a}^j\right)\right]\quad(7)$$

从公式(6)及公式(7)中可以看出，基于平均场估计，原始的马尔可夫博弈问题被转化为求解中心智能体 $j$ 关于平均动作 $\bar{a}^j$ 的最优相应 $\pi^j_t$。

平均场多智能体强化学习算法使用一种迭代方法计算每个智能体 $j$ 的最优响应 $\pi^j_t$。具体来说，智能体 $j$ 所有邻近智能体的平均动作 $\bar{a}^j$ 首先通过智能体 $j$ 的 $N^j$ 个邻居采取的动作 $a^k$ 的平均值进行计算，其中 $a^k$ 由策略 $\pi^k_t$ 决定，受到上一时刻平均动作 $\bar{a}^k_-$ 的影响：

$$\bar{a}^j=\frac{1}{N^j}\sum_k a^k,a^k\sim\pi^k_t\left(s,\bar{a}^k_-\right)\quad(8)$$

在计算出平均动作后，智能体 $j$ 的策略 $\pi^j_t$ 由于对当前平均动作的依赖，从而也对应发生改变。

智能体 $j$ 的新的玻尔兹曼策略为：

$$\pi^j_t\left(a^j s,\bar{a}^j\right)=\frac{\exp\left(\beta Q^j_t\left(s,a^j,\bar{a}^j\right)\right)}{\sum_{a^{j'}\in\mathcal{A}^j}\exp\left(\beta Q^j_t\left(s,a^{j'},\bar{a}^j\right)\right)}\quad(9)$$

通过迭代计算公式(8)以及公式(9)，所有智能体的策略都进行了更新。

## 3.2 算法实现

平均场多智能体强化学习算法使用神经网络来建模公式(5)中的平均场 $Q$ 函数，$Q$ 函数被神经网络权重 $\phi$ 参数化。公式(6)中的更新规则从而可以转化为参数更新。平均场多智能体强化学习算法使用标准离线 Q 学习[53]来解决离散动作空间问题，本文简写为 MF-Q。在 MF-Q 算法中，智能体 $j$ 通过最小化下述损失函数进行训练：

$$L\left(\phi^j\right)=\left(y^j-Q_{\phi^j}\left(s,a^j,\bar{a}^j\right)\right)^2,$$

$$\nabla_{\phi^j}L\left(\phi^j\right)=\left(y^j-Q_{\phi^j}\left(s,a^j,\bar{a}^j\right)\right)\nabla_{\phi^j}Q_{\phi^j}\left(s,a^j,\bar{a}^j\right)$$

有了上述梯度，MF-Q 就可以基于随机梯度下降进行参数更新。除了像 MF-Q 中使用 $Q$ 函数来计算玻尔兹曼策略之外，平均场多智能体强化学习算法还可以直接使用由 $\theta$ 参数化的神经网络来建

模策略本身，从而得到在线演员-评论家算法[54]，本文将这类算法简写为 MF-AC。MF-AC 中的策略网络 $\pi_{\theta^j}$，即演员通过采样得到的策略梯度进行训练：

$$\nabla_{\theta^j} L\left(\theta^j\right) \approx \nabla_{\theta^j} \log \pi_{\theta^j}\left(s\right) Q_{\phi^j}\left(s, a^j, \bar{a}^j\right)\Big|_{a=\pi_{\theta^j}(s)}$$

MF-AC 中的评论家与 MF-Q 算法中的平均值 $Q$ 函数相同。在 MF-AC 训练过程中，需要交替更新演员以及评论家的参数直到收敛。MF-Q 以及 MF-Q 算法的伪代码如下，流程图见图 3~4。

**算法 1** 平均场 Q 学习算法(MF-Q)

输入：对所有 $j \in \{1,2,\cdots,N\}$ 初始化 $Q_{\phi^j}, Q_{\phi_-^j}$，以及 $\bar{a}^j$；

While 训练未完成 do

for $m = 1,2,\cdots,M$ do

对每个智能体 $j$，根据公式(9)采样动作 $a^j$；

对每个智能体 $j$，根据公式(8)计算新的平均动作 $\bar{a}^j$；

end

采取联合动作 $a = \left[a^1, a^2, \cdots, a^N\right]$，得到联合奖励 $r = \left[r^1, r^2, \cdots, r^N\right]$ 并转移到下一个状态 $s'$；

将 $\left\langle s, a, r, s', \bar{a}\right\rangle$ 存储到回放缓存 $D$ 中，$\bar{a} = \left[\bar{a}^1, \bar{a}^2, \cdots, \bar{a}^N\right]$；

for $j = 1,2,\cdots,N$ do

从 $D$ 中采样包含 $K$ 条经验 $\left\langle s, a, r, s', \bar{a}\right\rangle$ 的小批量训练数据；

令 $\bar{a}_-^J \leftarrow \bar{a}^j$，并从 $Q_{\phi^j-}$ 采样动作 $a_-^j$；

根据公式(8)，$y^j = r^j + \gamma v_{\phi_-^j}^{\mathrm{MF}}\left(s'\right)$；

通过最小化以下损失函数更新 $Q$ 网络：

$$\mathcal{L}\left(\phi^j\right) = \frac{1}{K}\sum\left(y^j - Q_{\phi^j}\left(s^j, a^j, \bar{a}^j\right)\right)^2$$

end

根据学习率 $\tau$ 更新每个智能体 $j$ 的目标网络参数：$\phi_-^j \leftarrow \tau\phi^j + \left(1-\tau\right)\phi_-^j$

end

**算法 2** 平均场演员—评论家算法(MF-AC)

输入：对所有 $j \in \{1,2,\cdots,N\}$ 初始化 $Q_{\phi^j}, Q_{\phi_-^j}, \pi_{\theta^j}, \pi_{\theta^j\_}$ 以及 $\bar{a}^j$；

While 训练未完成 do

对每个智能体 $j$ 采样动作 $a^j = \pi_{\theta^j}(s)$；

对每个智能体 $j$ 计算新的平均动作 $\bar{a} = \left[\bar{a}^1, \bar{a}^2, \cdots, \bar{a}^N\right]$；

采取联合动作 $a = \left[a^1, a^2, \cdots, a^N\right]$，得到联合奖励 $r = \left[r^1, r^2, \ldots, r^N\right]$ 并转移到下一个状态 $s'$；

将 $\left\langle s, a, r, s', \bar{a}\right\rangle$ 存储到回放缓存 $D$ 中；

for j $= 1,2,\cdots,N$ do

从 $D$ 中采样包含 $K$ 条经验 $\left\langle s, a, r, s', \bar{a}\right\rangle$ 的小批量训练数据；

根据公式(8)，$y^j = r^j + \gamma v_{\phi_-^j}^{\mathrm{MF}}\left(s'\right)$；

通过最小化以下损失函数更新评论家：

$$L\left(\phi^j\right) = \frac{1}{K}\sum\left(y^j - Q_{\phi^j}\left(s^j, a^j, \bar{a}^j\right)\right)^2;$$

通过使用以下采样策略梯度更新演员：

$$\nabla_{\theta^j} J\left(\theta^j\right) \approx \frac{1}{K}\sum \nabla_{\theta^j} \log \pi_{\theta^j}\left(s'\right) + Q_{\phi_-^j}\left(s', a_-^j, \bar{a}_-^j\right)\Big|_{a_-^j = \pi_{\theta_-^j}(s')}$$

end

根据学习率 $\tau_\phi$ 以及 $\tau_\theta$ 更新每个智能体 $j$ 的目标网络参数：

$$\phi_-^j \leftarrow \tau_\phi\phi^j + (1-\tau_\phi)\phi_-^j$$
$$\theta_-^j \leftarrow \tau_\theta\theta^j + (1-\tau_\theta)\theta_-^j$$

end

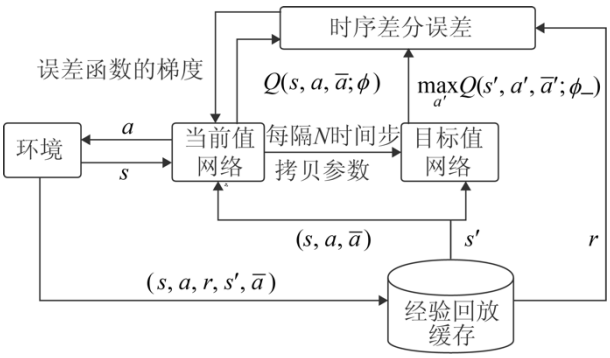



图 3 MF-Q 算法流程图

Fig. 3 Flow chart of MF-Q algorithm

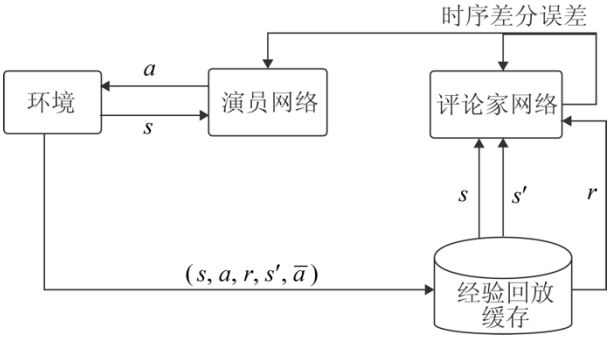



图 4 MF-AC 算法流程图
Fig. 4 Flow chart of MF-AC algorithm

# 4 仿真实验

## 4.1 仿真环境介绍

为了验证平均场多智能体强化学习算法在大规模无人机集群细粒度任务规划中的有效性，本文在自建的大规模无人机集群对抗仿真环境中进行实验。该仿真环境模拟真实的空域作战背景，选取 4 000 km×4 000 km 的空域范围作为作战区域，以 100 km 为单位距离进行网格划分，将整体空域范围划分为 40×40 的网格世界便于计算无人机的具体位置、模拟无人机的各种行为。在该作战区域内敌我双方进行战斗，双方分别拥有由 64 架无人机组成的无人机集群。本文自建的仿真环境对对抗环境下的无人机集群任务规划进行了以下简化。首先，集群内的所有无人机在离散时刻同时进行决策；其次，集群中的无人机之间需保持最小距离；最后，无人机攻击行为具有不可躲避性及瞬时性。具体来说，每架无人机的属性包含以下信息：①自身所在空域：由于仿真环境已经划分为网格世界，所以每架无人机的所在区域即为其中的某一个网格，大小为 100 km×100 km，且在多架无人机决定同时飞到某一特定区域时，随机选取一架无人机占领该空域，而其他无人机则取消飞往该区域的动作。②飞行速度：每架无人机最大飞行速度为 600 km/时间步，即以最大速度飞行可以在一个时间步内飞越两个网格的空域范围。③探测范围：每架无人机可侦查范围为以 600 km 为半径的圆形区域，主要观测其中敌我无人机数量、生命值以及自身是否处于合法空域作战范围内。④防御及攻击：每架无人机的初始生命值为 10，其每一次攻击的伤害值大小为 2，但每个时间步恢复 0.1 的生命值；当其生命值为 0 时即被消灭，并且消失在网格世界中。⑤攻击范围：每架无人机的攻击范围为以 150 km 为半径的圆形区域，可对在该攻击范围内的敌方无人机进行攻击(假定攻击可以瞬间生效)。

据此建模的每架无人机的状态空间、动作空间以及奖励函数如下：①状态空间：分为局部空间信息、全局空间信息以及非空间信息。局部空间信息包含每架无人机探测范围内的局部信息；全局空间信息包含一个缩略地图查看整体态势，该缩略地图的观测精度压缩为大地图的百分之一，空域距离范围为 400 km×400 km，缩略地图的每个网格此时等价于原始环境的 100 个网格，网格中的值为该缩略网格内所能观测到的无人机数量占总体无人机数量的比例。非空间信息包含每架无人机的个体信息，包括自身 ID 的独热编码、上一次采取的动作、上一次获得奖励值以及归一化后的相对位置。②动作空间：每架无人机的动作分为移动、攻击两种。其中移动范围和攻击范围仍然以网格为单位进行设计，覆盖一定半径的圆形区域，与其探测范围及攻击范围相对应，共有 21 个动作，包括 13 个移动动作和 8 个攻击动作，如图 5 所示。③奖励函数：奖励函数的设计主要考虑两方面的因素。其一，鼓励我方无人机集群在尽可能较少的移动及较快的时间内对敌方无人机集群进行歼灭，每跨域一个网格空域范围则施加–0.005 的奖励；其二，由于我方最终的战斗目标为歼灭尽可能多的敌方无人机，因此每消灭一架敌方无人机则获得 5 的奖励。此外，还需要每架无人机尽可能地避免死亡或被攻击。因此每当被攻击或者死亡时，施加–0.1 的惩罚；同时，每架无人机还应当减少无效移动或攻击的次数，每当攻击一个空的网格区域时，施加–0.1 的惩罚。

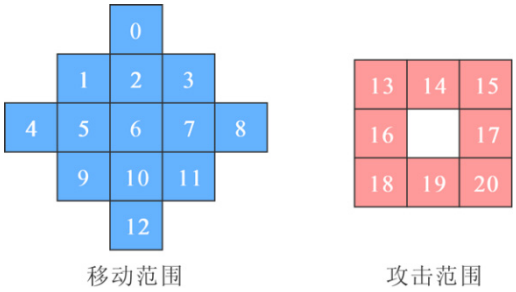



图 5 动作空间设计

Fig. 5 Design of action space

## 4.2 仿真实验设置

### 4.2.1 数据收集

在本文自建的大规模无人机集群仿真环境中，平均场 Q 学习算法 MF-Q、平均场演员—评论家算法 MF-AC 以及若干基准算法均采用相同的数据收集技术。为了提升训练效率，所有无人机均视为同质的，且共享神经网络参数。在本文仿真环境中，MF-Q 算法、MF-AC 算法以及基准算法独立 Q 学习 DQN、演员评论家算法 AC 均采取自博弈的训练方式。所有算法环境周期均采用截断采样，无论胜负与否，周期长度设置为固定的 400 个时间步，共训练 2 000 个环境周期。对于属于离线策略方法 MF-Q 及 DQN，其经验回放缓存的大小设置为 8 000；对于属于在线策略方法的 MF-AC 及 AC，每轮训练中采样得到的所有数据即为经验回放缓存的全部信息。

### 4.2.2 评价指标

本文将集群内所有无人机在每轮训练中的累积奖励和作为评价指标。具体来说，敌我每方的累积奖励和为所有 N=64 架无人机在 M=200 个环境周期中的环境奖励值之和的平均值。此外，由于在训练阶段各算法采用自博弈的方式进行训练，因此在评估模型效果时采取交叉对战方式，MF-Q, MF-AC, DQN, AC 4 种算法分别两两进行对战，每轮训练时间步仍然设置为固定的 400 个时间步，每一次共评估 50 个环境周期，根据所有的交叉战斗结果统计得到 4 种算法的两两胜负率以及平均累积奖励和。

## 4.3 仿真结果及分析

### 4.3.1 自博弈训练仿真结果

第一阶段分别采用 4 种算法 MF-Q, MF-AC, DQN, AC 进行自博弈对抗训练。在自博弈过程中，2 个采用相同算法以及采用领导者—跟随者参数更新模式的无人机集群进行相互对抗。在给定领导者—跟随者参数更新规则以及胜负判定方法的情况下，双方随机初始化策略网络参数。一方无人机集群(领导者)策略始终基于另一方(跟随者)进行学习；另一方(跟随者)则根据博弈胜负即当学习的一方(领导者)获胜时，依据领导者—跟随者参数更新规则，基于领导者参数进行软更新来调整其策略网络的相应参数，使其向领导者参数方向进行逼近的同时保持部分己身参数。这样在保证自博弈算法有效性的同时，可以加强自博弈过程的随机性，从而提升最终策略的鲁棒程度。接着，领导者以及跟随者通过反复迭代上述过程直至收敛，最后取领导者策略作为最终策略。在自博弈训练过程中，以每个环境周期的累积奖励和作为评估标准。

由于在自博弈的训练过程中，仅当领导者总体回报值高于跟随者时，算法才对该环境周期领导者策略对应的总体回报值进行记录，因此 4 种算法在基于自博弈重复训练 50 个环境周期的过程中，每个环境周期中每条记录的时间点无法通过同步记录对齐，因此计算总体回报的均值及标准差将不具有物理意义，无法反映出算法真实的性能。为了展示各算法对于不同随机种子的鲁棒性，本文随机选取了其中 4 次实验结果作为展示，如图 6 所示。

综合多次自博弈的训练曲线图可以看出，MF-Q 算法和 DQN 算法的收敛速度远快于 MF-AC 算法和 AC 算法，前两者的收敛环境周期数不超过 150，而后者的收敛环境周期轮次在 1 000~2 000 之间浮动。一方面说明基于 Q 学习的相关算法得益于其离线学习的优势，从大量的任意非当前策略经验中更快地学习到了协同策略，从而趋于更稳定的学习过程；另一方面，由于 Q 学习往往估计动作价值

函数的最大值并通过不断迭代得到最大期望动作价值函数，从而在一定程度上引入了正向的偏差估计，即使智能体所处的环境本身仍在动态变化，但是对动作价值函数的一定高估可能对每个智能体有效处理其他智能体带来的动态变化有所帮助。

此外，横向对比引入平均场理论在 Q 学习算法和演员评论家算法上的效果发现，对于 Q 学习算法来说，将智能体与其他智能体之间的动态交互转化为整体动态影响增强了离线 Q 学习算法的收敛速度，MF-Q 算法总体回报值的方差较 DQN 更小，最终收敛到的最优策略更为稳定。而对于演员评论家算法来说，MF-AC 算法在 2 000 个环境周期内并不一定收敛，平均场理论的引入可能对在线学习的方式引入了更大的不确定性。

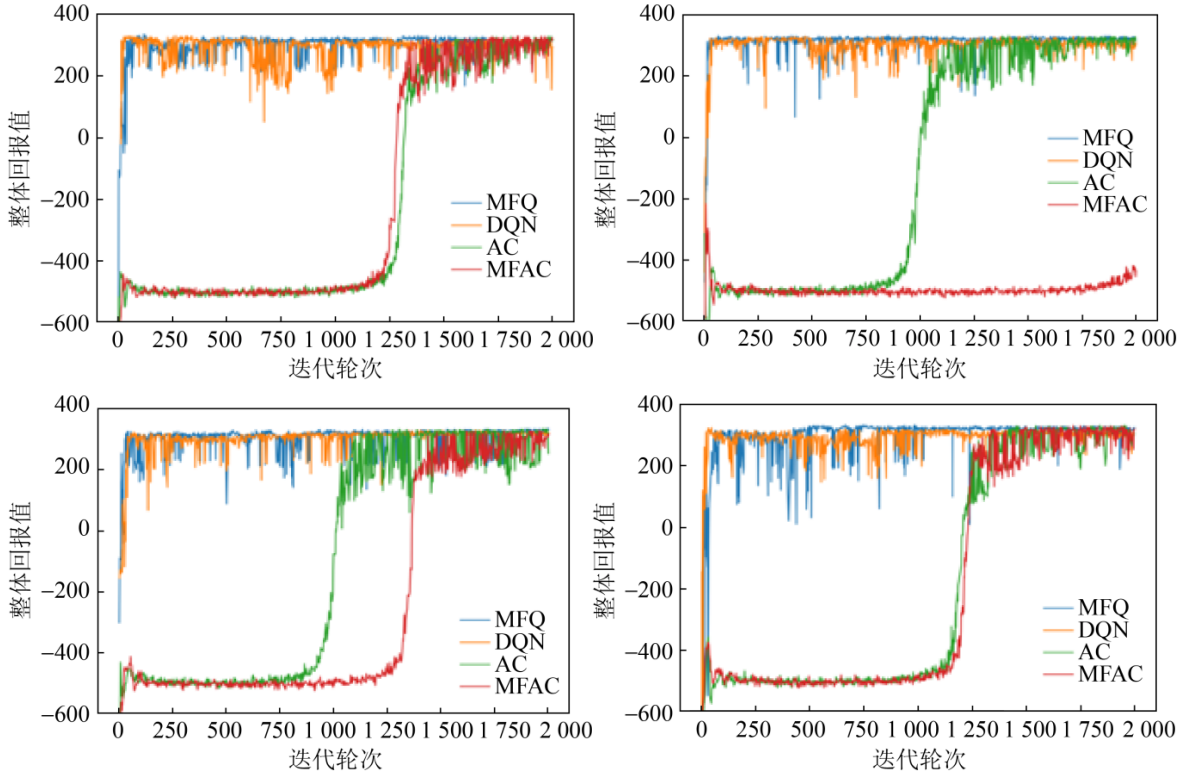



图 6　交叉战斗学习曲线图

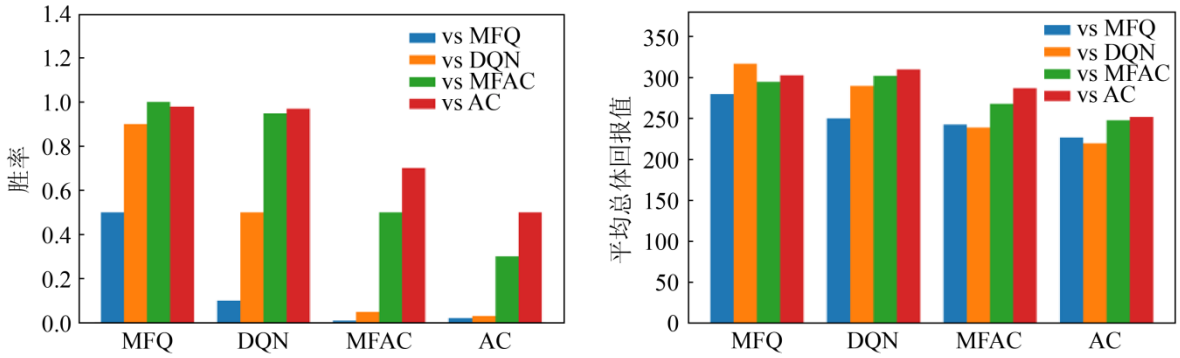
Fig. 6　Learning curves of cross-battle

### 4.3.2 交叉战斗仿真结果

将图 6 中自博弈训练所得的模型在自建的大规模无人机集群对抗仿真环境中进行交叉战斗模拟，图 7 展示了 4 种算法在两两对抗下的平均胜率(图 7(a))以及平均总体回报值(图 7(b))。在交叉战斗阶段，每种算法在自博弈阶段训练所得的策略以及奖励函数引入了不同的战斗方式，从平均胜率和平均总体回报值来看，MF-Q 和 DQN 两种离线学习算法仍然以较大优势战胜 MF-AC 和 AC 两种在线学习算法，并且 MF-Q 无论是在平均胜率还是平均总体回报值都要优于 DQN 算法。

选取上述交叉战斗过程中 MF-Q 算法与 DQN 算法的某一回合战斗进行仿真模拟可视化(图 8)。在战斗开始阶段，红方军队的 64 架无人机采用 MF-Q 算法训练所得策略，蓝方军队的 64 架无人机采用 DQN 算法训练所得策略，双方各自以方阵形式排开(图 8(a))。在战斗过程中，红方军队首先大规模铺开阵势进行协同搜索(图 8(b))，随后自发形成多个小范围对敌方军队进行包围(图 8(c))、动态转移以及对抗(图 8(d))，从而学会追击(图 8(e))和围攻(图 8(f))等智能化行为。最终，红方军队战胜蓝方军队(图 8(g))。上述红蓝双方的战斗过程展

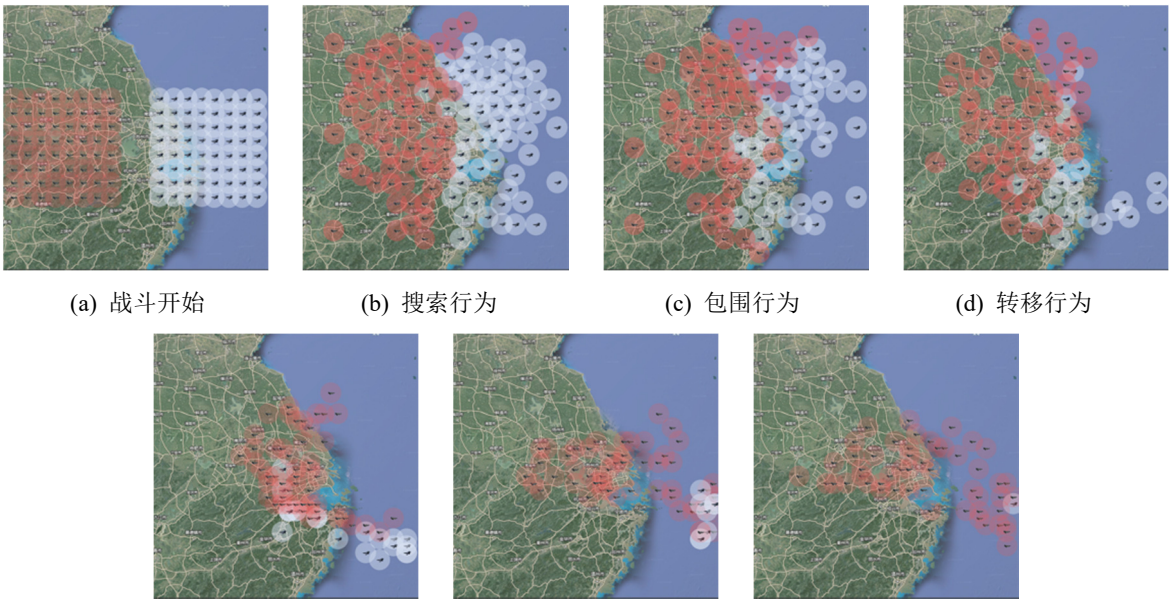
现了基于平均场多智能体强化学习算法在大规模无人机集群对抗场景下的有效性，从而使得无人机集群能够自发习得融合协同搜索、饱和攻击以及动态对抗的细粒度任务规划策略。




(a) 交叉战斗平均胜率　　(b) 交叉战斗平均总体回报值

图 7　交叉战斗结果

Fig. 7　Results of cross-battle



(a) 战斗开始　(b) 搜索行为　(c) 包围行为　(d) 转移行为

(e) 追击行为　(f) 围攻行为　(g) 战斗结束

图 8　交叉战斗仿真结果

Fig. 8　Simulation results of cross-battle

## 5　结论

本文将对抗环境中无人机集群的细粒度任务规划的协同搜索、饱和攻击以及动态对抗 3 个方面有机地结合在一起，通过马尔可夫博弈进行建模，并引入多智能体强化学习算法进行问题求解。为了解决多智能体强化学习算法在大规模无人机集群场景下的可扩展性问题，本文使用了与平均场理论相结合的多智能体强化学习算法，将无人机集群间的复杂交互过程简化为智能体与邻近智能体平均作用之间的双边交互。结合传统离线 Q 学习算法以及在线演员评论家算法，本文将平均场 Q 学习算法 MF-Q 以及平均场演员-评论家算法 MF-AC 应用到自建的大规模无人机集群对抗仿真环境中。仿真实验结果表明，MF-Q 与 MF-AC 在胜率、算法收敛速度以及最终期望奖励上均要

明显优于基准多智能体强化学习算法，从而验证了基于平均场的多智能体强化学习算法在求解对抗环境中的无人机集群细粒度任务规划问题时的高效性以及灵活性。

## 参考文献：

[1] Danoy G, Brust M R, Bouvry P. Connectivity Stability in Autonomous Multi-level UAV Swarms for Wide Area Monitoring[C]// Association for Computing Machinery. New York: DIVANet, 2015.

[2] Shukla A, Karki H. Application of Robotics in Onshore Oil and Gas Industry—A review Part I[J]. Robotics and Autonomous Systems (S0921-8890), 2016, 75(part B): 490-507.

[3] 陈方舟, 黄靖皓, 赵阳辉. 美军无人“蜂群”作战技术发展分析[J]. 装备学院学报, 2016, 27(2): 34-37.
Chen Fangzhou, Huang Jinghao, Zhao Yanghui. Analysis on Unmanned Swarm Fighting System of U.S. Armed Forces[J]. Journal of Equipment Academy, 2016, 27(2): 34-37.

[4] 邢冬静. 无人机集群作战自主任务规划方法研究[D]. 南京: 南京航空航天大学, 2019.
Xing Dongjing. Autonomous Mission Planning Method for Unmanned Aerial Vehicle Swarm Operations[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2019.

[5] 段海滨, 李沛. 基于生物群集行为的无人机集群控制[J]. 科技导报, 2017, 35(7): 17-25.
Duan Haibin, Li Pei. Autonomous Control for Unmanned Aerial Vehicle Swarms Based on Biological Collective Behaviors[J]. Science & Technology Review, 2017, 35(7): 17-25.

[6] Robert P Otto. Small Unmanned Aircraft Systems (SUAS) Flight Plan: 2016-2036. Bridging the Gap Between Tactical and Strategic[R]// Air Force Deputy Chief of Staff Washington DC United States, 2016.

[7] Fan D D, Theodorou E, Reeder J. Model-based Stochastic Search for Large Scale Optimization of Multi-agent UAV Swarms[C]// IEEE Symposium Serieson Computational Intelligence. India: IEEE, 2018: 2216-2222.

[8] Huang Q W, Yao J, Li Q et al. Cooperative Searching Strategy for Multiple Unmanned Aerial Vehicles based on Modified Probability Map[M]. Theory, Methodology, Tools and Applications for Model in and Simulation of Complex Systems. Singapore: Springer, 2016: 279-287.

[9] Paradzik M, Ince G. Multi-agent Search Strategy based on Digital Pheromones for UAVs[C]// Signal Processing and Communication Application Conference. Turkey: IEEE, 2016: 233-236.

[10] Yang Y, Polycarpou M, Minai A. Multi-UAV Cooperative Search Using an Opportunistic Learning Method[J]. Journal of Dynamic Systems Measurement and Control-transactions of The ASME (S0022-0434), 2007, 129(5): 716-728.

[11] Tian J, Zheng Y, Zhu H Y, et al. A MPC and Genetic Algorithm Based Approach for Multiple UAVs Cooperative Search[C]// In CIS. Berlin: Springer, 2005.

[12] 彭辉, 沈林成, 朱华勇. 基于分布式模型预测控制的多UAV协同区域搜索[J]. 航空学报, 2010, 31(3): 593-601.
Peng Hui, Shen Lincheng, Zhu Huayong. Multiple UAV Cooperative Area Search Based on Distributed Model Predictive Control[J]. Acta Aeronauticaet Astronautica Sinica, 2010, 31(3): 593-601.

[13] 肖东. 异构多无人机自主任务规划方法研究[D]. 南京: 南京航空航天大学, 2017.
Xiao Dong. Autonomous Mission Planning Method for Heterogeneous Multiple Unmanned Aerial Vehicles[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2017.

[14] Jian J C, Zha W Z, Peng Z, et al. Cooperative Area Reconnaissance for Multi-UAV in Dynamic Environment[C]// 2013 9th Asian Control Conference (ASCC). Turkey: IEEE, 2013: 1-6.

[15] Gao C, Zhen Z, Gong H J. A Self-organized Search and Attack Algorithm for Multiple Unmanned Aerial Vehicles[J]. Aerospace Science and Technology (S1270-9638), 2016, 54: 229-240.

[16] Zhen Z, Xing D, Gao C. Cooperative Search-attack Mission Planning for Multi-UAV Based on Intelligent Self-organized Algorithm[J]. Aerospace Science and Technology (S1270-9638), 2018, 76: 402-411.

[17] 申超, 武坤琳, 宋怡然. 无人机蜂群作战发展重点动态[J]. 飞航导弹, 2016, 11: 28-33.
Shen Chao, Wu Kunlin, Song Yiran. Key Trends in the Development of Drone Swarm Operations[J]. Aerodynamic Missile Journal, 2016, 11: 28-33.

[18] Seil A, Kim H. Simultaneous Task Assignment and Path Planning Using Mixed-integer Linear Programming and Potential Field Method[C]// 13th International Conference on Control, Automation and Systems. (ICCAS 2013). Korea: IEEE, 2013: 1845-1848.

[19] Eun Y J, Bang H. Cooperative Control of Multiple UCAVs for Suppression of Enemy Air Defense [C]//

AIAA 3rd “Unmanned Unlimited” Technical Conference, Workshop and Exhibit. Chicago: AIAA, 2004: 6529.

[20] Alighanbari M, How J P. Decentralized Task Assignment for Unmanned Aerial Vehicles[C]// 44th IEEE Conference on Decision and Control. Spain IEEE, 2005: 5668-5673.

[21] Jiang X W, Zhou Q, Ye Y. Method of Task Assignment for UAV based on Particle Swarm Optimization in Logistics[C]// 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence. New York: ACM, 2017: 113-117.

[22] Parag C. Pendharkar. An Ant Colony Optimization Heuristic for Constrained Task Allocation Problem[J]. Journal of Computer Science (S1877-7503), 2015, 7: 37-47.

[23] Bello-Orgaz G, Ramirez-Atencia C, Fradera-Gil J, et al. GAMPP: Genetic Algorithm for UAV Mission Planning Problems[M]// Intelligent Distributed Computing IX. Switzerland: Springer, Cham, 2016: 167-176.

[24] Li J, Zhang K, Xia G. Multi-AUV Cooperative Task Allocation Based on Improved Contract Network[C]// 2017 IEEE International Conference on Mechatronics and Automation (ICMA). Japan: IEEE, 2017: 608-613.

[25] Zhang A, Guo F. Dynamic task allocation for formation air-to-ground attack[C]// 2013 Sixth International Conference on Advanced Computational Intelligence (ICACI). China: IEEE, 2013: 119-123.

[26] Ji X T, Niu Y F, Shen L C. Robust Satisficing Decision Making for Unmanned Aerial Vehicle Complex Missions Under Severe Uncertainty[J]. PLoS One (S1932-6203), 2016, 11(11).

[27] 刘光猛, 汪卫华. 基于排队论的无人机突防概率研究[J]. 舰船电子工程, 2013, 33(5): 123-125.
Liu Guangmeng, Wang Weihua. Penetration Probability of UAV Based on Queuing Theory[J]. Ship Electronic Engineering, 2013, 33(5): 123-125.

[28] Lua C, Altenburg K, Nygard K. Synchronized Multi-point Attack by Autonomous Reactive Vehicles with Simple Local Communication[C]// 2003 IEEE Swarm Intelligence Symposium. SIS'03 (Cat. No. 03EX706). USA: IEEE, 2003: 95-102.

[29] Liu G L, Xing D J, Hou J Y, et al. Distributed Cooperative Control Algorithm for Multi Uavmission Rendezvous[J]. Transactions of Nanjing University of Aeronautics and Astronautics (S1005-1120), 2017, 34(6): 29-38.

[30] 罗德林, 徐扬, 张金鹏. 无人机集群对抗技术新进展[J]. 科技导报, 2017, 35(7): 26-31.
Luo Delin, Xu Yang, Zhang Jinpeng. New Progresses on UAV Swarm Confrontation[J]. Science & Technology Review, 2017, 35(7): 26-31.

[31] Gaertner U. UAV Swarm Tactics: An Agent-based Simulation and Markov Process Analysis[R]. Master Thesis, 2013.

[32] 罗德林, 张海洋, 谢荣增. 基于多 agent 系统的大规模无人机集群对抗[J]. 控制理论与应用, 2015, 32(11): 1498-1504.
Luo Delin, Zhang Haiyang, Xie Rongzeng. Unmanned Aerial Vehicles Swarm Conflict Based on Multi-agent System[J]. Control Theory & Application, 2015, 32(11): 1498-1504.

[33] Busoniu L, Robert Babuka, Schutter B D. A Comprehensive Survey of Multiagent Reinforcement Learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, PartC (Applications and Reviews) (S0018-9472), 2008, 38(2): 156-172.

[34] Li W H, Wang X F, Jin B, et al. Structured Diversification Emergence via Reinforced Organization Control and Hierarchical Consensus Learning[C]// 20th International Conference on Autonomous Agents and Multi Agent Systems. Richland: IFAAMAS, 2021: 773-781.

[35] Li W H, Wang X F, Jin B, et al. Dealing with Non-stationarity in Multi-agent Reinforcement Learning via Trust Region Decomposition[J/OL]. ArXiv, [2021-03-15]. https://arxiv.org/abs/2102.10616, 2021.

[36] Blume L. The Statistical Mechanics of Strategic Interaction[J]. Games and Economic Behavior (S0899-8256), 1993, 5(3): 387-424.

[37] Stanley H. Introduction to Phase Transitions and Critical Phenomena[J]. American Journal of Physics (S0002-9505), 1971, 40: 927-928.

[38] 宋怡然, 申超. 美国分布式低成本无人机集群研究进展[J]. 飞航导弹, 2016(8): 17-22.
Song Yiran, Shen Chao. Research Progress of Distributed Low-cost UAV Cluster[J]. Winged Missiles Journal, 2016(8): 17-22.

[39] Baum M, Passino K. A Search-theoretic Approach to Cooperative Control for Uninhabited Air Vehicles[C]// AIAA Guidance, Navigation, and Control Conference and Exhibit. California: AIAA 2002: 4589.

[40] Ilachinski A. Artificial War: Multiagent-based Simulation of Combat[M]. World Scientific, 2004.

[41] 刘金星. 空战指挥控制的自主决策思维属性[J]. 电光与控制, 2010, 17(6): 1-4.

Liu Jinxing. Mental Attributes of Autonomous Decision-Making in Air Combat Command and Control[J]. Electronics Optics & Control, 2010, 17(6): 1-4.

[42] He H, Boyd-Graber J, Kwok K, et al. Opponent Modeling in Deep Reinforcement Learning[C]// International Conference on Machine Learning. New York: PMLR, 2016: 1804-1813.

[43] Gupta J K, Egorov M, Kochenderfer M. Cooperative Multi-agent Control Using Deep Reinforcement Learning[C]// International Conference on Autonomous Agents and Multiagent Systems. Cham: Springer, 2017: 66-83.

[44] Peng P, Wen Y, Yang Y, et al. Multiagent Bidirectionally-Coordinated Nets for Learning to Play Starcraft Combat Games[J/OL]. ArXiv, [2021-03-15]. https://arxiv.org/abs/1703.10069.

[45] Lowe R, Wu Y, Tamar A, et al. Multi-agent Actor-critic for Mixed Cooperative-Competitive Environments[C]// 31st International Conference on Neural Information Processing Systems. USA: Curran Associates Inc, 2017: 6382-6393.

[46] Foerster J, Farquhar G, Afouras T, et al. Counterfactual Multi-agent Policy Gradients[C]// AAAI Conference on Artificial Intelligence. 2018.

[47] Jeong S H, Kang A R, Kim H K. Analysis of Game Bot's Behavioral Characteristics in Social Interaction Networks of MMORPG[J]. IEEE Transactions on Robotics & Automation (S1552-3098), 2015, 18(5): 813-825.

[48] Lasry J, Lions P. Mean Field Games[J]. Japanese Journal of Mathematics (S0289-2316), 2007, 2(1): 229-260.

[49] Caines P, Huang M, Malhamé R. Large Population Stochastic Dynamic Games: Closed-loop Mckean-vlasov Systems and the Nash Certainty Equivalence Principle[J]. Communications in Information and Systems (S1529-3181), 2006, 6(3): 221-252.

[50] Weintraub G, Benkard C L, Benjamin V R. Oblivious Equilibrium: A Mean Field Approximation for Large-scale Dynamic Games[C]// 18th International Conference on Neural Information Processing Systems. USA: MIT Press, 2005: 1489-1496.

[51] Yang J C, Ye X J, Trivedi R, et al. Deep Mean Field Games for Learning Optimal Behavior Policy of Large Populations[C]// International Conference on Learning Representations. 2018.

[52] Michael L L. Markov Games as a Framework for Multi-agent Reinforcement Learning[C]// International Conference on Machine Learning. USA: Morgan Kaufmann Publishers Inc, 1994: 157-163.

[53] Watkins C J C H, Dayan P. Q-learning[J]. Machine Learning (S0885-6125), 1992, 8(3/4): 279-292.

[54] Konda V R, Tsitsiklis J N. Actor-critic Algorithms[C]// 18th International Conference on Neural Information Processing Systems. USA: MIT Press, 1999: 1008-1014.