中国科学: 信息科学 2024 年 第 54 卷 第 8 期: 1871–1883
SCIENTIA SINICA Informationis
论文

基于 MASAC 强化学习算法的多无人机协同路径规划

方城亮[1], 杨飞生[1,2*], 潘泉[1]

1. 西北工业大学自动化学院, 西安 710129
2. 西北工业大学重庆科创中心, 重庆 401151
* 通信作者. E-mail: yangfeisheng@nwpu.edu.cn

收稿日期: 2024–02–07; 修回日期: 2024–04–26; 接受日期: 2024–05–11; 网络出版日期: 2024–08–08

国家自然科学基金 (批准号: 62073269)、重庆市自然科学基金面上项目 (批准号: CSTB2022NSCQ-MSX0963)、航空科学基金 (批准号: 2020Z034053002) 和广东省基础与应用基础研究基金自然科学基金面上项目 (批准号: 2023A1515011220) 资助

摘要 针对动态不确定环境下异构多无人机协同路径规划问题, 提出了一种新的多智能体深度强化学习算法. 首先, 开发了一个空域场景下多无人机到达目标地点的强化学习环境, 环境引入了无人机动力学方程, 并考虑了无人机异构的因素以及安全避障的需求. 其次, 设计了任务完成率、编队保持率、飞行时间等性能指标, 用以衡量算法的优劣. 然后, 将多无人机协同路径规划问题建模为部分可观 Markov 决策过程, 提出了一种多智能体柔性执行评价 (multi-agent soft actor critic, MASAC) 算法寻求该问题的近似最优策略. 最后, 通过仿真实验验证了所提算法的有效性和优越性.

关键词 多无人机, 路径规划, 多智能体深度强化学习, 部分可观 Markov 决策过程, MASAC 算法

1 引言

随着智能化决策与空中装备性能的快速发展, 无人机 (unmanned aerial vehicle, UAV) 作为一支蓬勃发展的空中对抗力量, 其零伤亡和卓越机动性的优势在现代战争中发挥着愈来重要的作用 [1]. 研究其相关技术对掌握作战主动权、提高作战性能有着明显的作用 [2]. 其中, 路径规划是 UAV 作战最基本的问题和前提, 如何快速且安全地到达目标地点是 UAV 路径规划的最终目标.

经典的路径规划算法原理简单, 在大多数情况下能够完成对路径规划的需求. Dijkstra [3] 提出的 Dijkstra 算法通过正向遍历所有节点得到最优路径. A* 算法是 Hart 等 [4] 对 Dijkstra 的改进, 其通过评估函数对节点进行针对性扩展. D* 算法是 Stentz [5] 对 A* 的改进, 其在计算出到达终点的路径后, 动态地利用评估函数对节点进行更新, 并基于当前状态对下一阶段的路径进行重新规划.

引用格式: 方城亮, 杨飞生, 潘泉. 基于 MASAC 强化学习算法的多无人机协同路径规划. 中国科学: 信息科学, 2024, 54: 1871–1883, doi: 10.1360/SSI-2024-0050
Fang C L, Yang F S, Pan Q. Multi-UAV collaborative path planning based on multi-agent soft actor critic (in Chinese). Sci Sin Inform, 2024, 54: 1871–1883, doi: 10.1360/SSI-2024-0050

由于路径规划是一个 NP 优化难题 [6], 经典算法在复杂的环境下计算复杂度会急剧上升, 甚至无法求解, 这也被称为维数诅咒问题. 启发式算法可以避免维数诅咒的问题, 其在路径规划中的应用也越来越广泛 [7]. 文献 [8] 将贪婪分配策略应用于为每个 UAV 分配目标点, 并将改进的基于可变信息素的蚁群优化算法应用于路径规划. 文献 [9] 提出了一种在图形处理单元上并行实现遗传算法来解决静态环境中的路径规划问题. 文献 [10] 提出了一种新的混合粒子群优化算法, 该算法通过合并模拟退火算法以解决复杂环境下路径规划问题. 无论是经典算法还是启发式算法, 以上这些算法更适合解决静态路径规划问题. 然而, 对于动态路径规划问题, 全局环境信息是未知的, 需要实时规划路径 [11].

深度强化学习 (deep reinforcement learning, DRL) [12~14] 的出现为解决动态路径规划问题提供了一种新思路, DRL 算法既具备强化学习 [15] 的决策能力又具备深度学习的感知能力, 有利于在复杂且动态的环境下进行路径规划. 文献 [16] 使用改进的深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 算法解决了滑翔机在时变洋流环境中的路径规划问题. 文献 [17] 采用竞争深度 Q 网络算法 [18], 解决了固定高度下带有碰撞规避的路径规划问题. 文献 [19] 采用双延迟 DDPG (twin delayed DDPG, TD3) 算法解决了自主水下机器人的自适应运动规划和避障技术. 文献 [20] 采用 TD3 算法使 UAV 能在具有随机性和动态性的多障碍环境中执行路径规划任务. 然而, 上述文献均用单智能体 DRL 算法处理单无人机或多无人机路径规划问题, 这种研究方法难以解决 UAV 间的协同问题.

因此, 针对多无人机协同路径规划问题, 本文采用多智能体 DRL 算法, 提出了一种多智能体柔性执行评价 (multi-agent soft actor critic, MASAC) 算法, 通过数据训练的方式摆脱了对准确模型的依赖. 该算法借鉴了 Lowe 等 [21] 提出的多智能体 DDPG (multi-agent DDPG, MADDPG) 算法思想, 将 MADDPG 的 DDPG 部分替换为探索性更强的柔性执行评价 (soft actor critic, SAC) [22]. 相较先前工作, 本文的主要贡献包括以下 3 个方面:

(1) 传统的经典算法无法避免维数诅咒, 启发式算法不适应于解决动态路径规划问题, 而单智能体 DRL 算法难以使智能体学会协同的能力. 为此, 本文使用多智能体 DRL 算法来处理多无人机协同路径规划问题.

(2) 当前开源的多无人机路径规划环境稀少, 本文开发了一个空域场景下多无人机路径规划的可视化强化学习环境. 该环境中 UAV 运动受到动力学方程的约束, 并且还考虑了无人机异构的因素以及安全避障的需求.

(3) 本文将多无人机路径规划问题建模为部分可观 Markov 决策过程 (partially observable Markov decision process, POMDP), 提出了 MASAC 算法求解 POMDP 模型以得到多个智能体的联合动作, 最终的实验结果表明了所提算法的有效性与优越性.

2 问题描述

2.1 多无人机协同路径规划任务描述

多无人机在一个有边界的空域环境中进行协同路径规划的任务示意图如图 1 所示. N 个异构无人机 (以 $N = 3$ 为例) 形成单领导者多跟随者形式的突防编队, 企图避开环境中的障碍物 (山丘、干扰区), 并在中途形成突防编队, 最终到达右上方的目标地点. 其中 UAV 拥有两种不同类型, 一种为领导者 (为图中的蓝色 UAV), 另一种为跟随者 (为图中的绿色 UAV), 不同类型的 UAV 性能不同. 领导者 UAV 有且仅有一架, 其任务是避开障碍物并到达目标地点, 而跟随者 UAV 数量可为多架, 任务则是保证时刻跟随领导者 UAV 以保持飞行编队. 由此, 可以得到多无人机协同路径规划所需要完成的

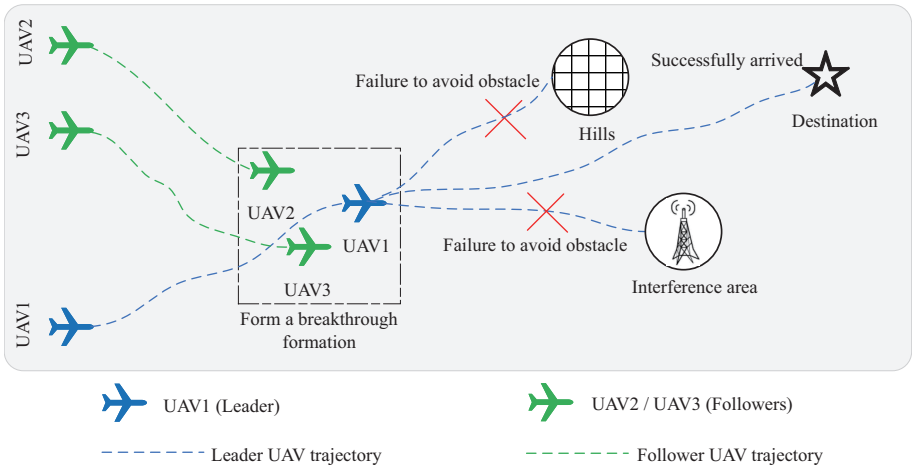

图 1 (网络版彩图) 多无人机协同路径规划的任务示意图

Figure 1 (Color online) Task illustration for multi-UAV collaborative path planning

三类子任务: (1) 避开障碍物, 避免发生碰撞检测从而导致 UAV 损毁; (2) 到达目标地点; (3) 跟随者 UAV 尽可能跟随领导者在一定范围内以保持突防编队.

2.2 任务成败的定义

2.2.1 任务失败

情况 1 (碰到障碍物, 自身损毁): 如果领导者 UAV 碰撞到山丘等障碍, 从而导致该 UAV 损毁, 则被认为是任务失败; 情况 2 (自身存活, 但没有在规定时间内到达目标地点): 如果领导者 UAV 无法做出合理的路径规划, 不能在规定时间内到达目标地点, 则被认为是任务失败.

2.2.2 任务成功

只要领导者 UAV 能够在规定的时间内到达目标地点, 并且不发生与障碍物之间的碰撞, 此时无论跟随者 UAV 是否到达目标地点, 都被认为是任务成功, 并且回合结束, 环境重置.

2.3 无人机动力学模型

现考虑有 N 架无人机, 且 $N>1$, 多无人机从时刻 t 到时刻 $t+1$ 的离散动力学方程为

$$\begin{cases} x_i(t+1)=x_i(t)+v_i(t)\Delta t\cos\psi_i(t), \\ y_i(t+1)=y_i(t)+v_i(t)\Delta t\sin\psi_i(t), \\ v_i(t+1)=v_i(t)+u_i(t)\Delta t, \\ \psi_i(t+1)=\psi_i(t)+\omega_i(t)\Delta t, \end{cases} \quad \text{s.t.} \quad \begin{cases} x_{i,\min}\leqslant x_i(t)\leqslant x_{i,\max}, \\ y_{i,\min}\leqslant y_i(t)\leqslant y_{i,\max}, \\ v_{i,\min}\leqslant v_i(t)\leqslant v_{i,\max}, \\ \psi_{i,\min}\leqslant \psi_i(t)\leqslant \psi_{i,\max}, \\ u_{i,\min}\leqslant u_i(t)\leqslant u_{i,\max}, \\ \omega_{i,\min}\leqslant \omega_i(t)\leqslant \omega_{i,\max}, \\ \sqrt{(x_i(t)-x_O)^2+(y_i(t)-y_O)^2}\geqslant R_O, \end{cases} \tag{1}$$

其中 i 为 UAV 的索引, (x,y) 为 UAV 的位置, v 是 UAV 速度, ψ 是 UAV 航向角, u 为 UAV 的加速度控制输入, ω 为角速度控制输入, Δt 为从时刻 t 到时刻 $t+1$ 的时间步长. 飞行状态约束条件: $x_{i,\min}$,

$x_{i,\max}$, $y_{i,\min}$, $y_{i,\max}$ 为第 i 架 UAV 位置上的最大、最小值, $v_{i,\min}$, $v_{i,\max}$ 为第 i 架 UAV 速度的最大、最小值, $\psi_{i,\min}$, $\psi_{i,\max}$ 为第 i 架 UAV 航向角的最大、最小值. 控制量约束条件: $u_{i,\min}$, $u_{i,\max}$ 为第 i 架 UAV 加速度控制输入的最大、最小值, $\omega_{i,\min}$, $\omega_{i,\max}$ 为第 i 架 UAV 角速度控制输入的最大、最小值. 飞行路径约束条件: 为便于处理, 本文中的山丘等障碍物设为标准的圆形区域, 其中 (x_O, y_O) 为障碍物的圆心位置, R_O 为障碍物的半径.

2.4 性能指标的设计

为对所提方法进行评估, 本文对训练后的多无人机协同路径规划决策进行了多次 Monte Carlo 测试, 选取任务完成率 (mission completion rate, MCR)、编队保持率 (formation keeping rate, FKR)、飞行时间、飞行轨迹和能量消耗作为性能指标.

根据 2.2 小节任务成败的定义, 任务完成率为 $J_{\mathrm{MCR}} = N_C/N_T$, 其中 N_C 表示任务成功完成的回合数, N_T 表示进行 Monte Carlo 测试的总回合数. 编队保持率为 $J_{\mathrm{FKR}} = T_C/T_T$, 其中 T_C 为一个回合内多无人机编队成功保持的时间步长, T_T 为这一个回合的总时间步长. 飞行时间为 $J_T = t_f$, 其中 t_f 为无人机的终端时刻. 飞行轨迹为 $J_S = \int_0^{t_f} v\mathrm{d}t$. 能量消耗为 $J_C = \int_0^{t_f} (|u| + |\omega|)\mathrm{d}t$.

3 基于多智能体深度强化学习的多无人机路径规划

3.1 动态不确定多无人机路径规划环境搭建

本文描述的动态不确定是指每次回合结束环境重置时, 所有 UAV 的初始状态 ——x 位置、y 位置、速度 v 和航向角 ψ 全部随机生成, 目标地点、障碍物的位置完全随机生成. 本文将 UAV 看成在同一高度下运动, 其运动符合二维空间下的运动学方程. 因此, 作出以下假设.

假设1 无人机的运动只考虑平面中的运动, 把现实中的三维运动以二维形式作简化处理.

假设2 目标地点位置等信息已知, 被认为由地面雷达测得并把数据告知无人机.

整个多无人机协同路径规划仿真环境被定义为一个 700 像素 × 600 像素的二维平面[1)], 该 700 像素 = 7000 m, 600 像素 = 6000 m, 环境的动画刷新频率 FPS 为 60. 无人机有两种不同的类型, 一种是领导者 UAV, 它的速度范围为每帧 10 ∼ 20 个像素, 经过换算为 100 ∼ 200 m/s, 加速度控制输入范围为 $-3 \sim 3$ m/s^2, 角速度范围为 $-0.6 \sim 0.6$ rad/s; 另一种为跟随者 UAV, 它拥有比领导者 UAV 更高的机动性, 其速度范围为 100 ∼ 300 m/s, 加速度范围为 $-6 \sim 6$ m/s^2, 角速度范围为 $-1.2 \sim 1.2$ rad/s. 领导者 UAV 一旦到达目标地点所在的圆形区域内, 即被认为规划任务成功, 且本场回合结束; 而一旦到达障碍物所在的圆形区域内, 即被认为发生碰撞, 此时规划任务失败, 且本场回合结束.

3.2 多无人机路径规划 POMDP 建模

POMDP 是环境状态部分可知下序贯决策的理想模型 [23∼25], 是对传统 Markov 决策过程的扩展 [26, 27]. 现利用 POMDP 来描述多无人机路径规划的决策模型, 该模型可以定义为一个七元组 $\langle S, A, R, P, Z, O, \gamma\rangle$, 其中, S 是状态集合, $S = \{s_1, s_2, \ldots, s_i, \ldots, s_N\}$, N 为智能体的数量, 状态集 S 对于智能体是不可知的; A 是动作集合, $A = \{a_1, a_2, \ldots, a_i, \ldots, a_N\}$, a_i 是第 i 个 UAV 的动作, $i = \{1, 2, \ldots, N\}$, 动作对于智能体是可知的; $R(s, a)$ 表示 UAV 在状态 s 下, 采取动作 a 获得的奖励; P 是状态转移概率, $P(s'|s, a)$ 表示 UAV 在状态 s 采取动作 a 转移到状态 s' 的概率, 状态转移概率对

1) https://github.com/henbudidiao/UAV-path-planning.

于智能体是不可知的; Z 是观测集合, $Z=\{z_1,z_2,\ldots,z_i,\ldots,z_N\}$, z_i 是第 i 个 UAV 的观测信息, 它对于智能体是可知的; O 是观测函数, $O(z|s,a)$ 表示在执行一个动作 $a\in A$ 后, 状态 $s\in S$ 产生一个观测 $z\in Z$ 的概率, 观测函数对于智能体是不可知的; 衰减因子 $\gamma\in[0,1]$ 表示对当前奖励与未来奖励的偏好与占比.

(1) 观测集 Z 的设计. 第 i 个 UAV 的观测信息可以用向量 z_i 表示为 $z_i=[x_i,y_i,v_i,\psi_i,m_i]$, 其中 $[x_i,y_i,v_i,\psi_i]$ 是第 i 个 UAV 的飞行状态量, m_i 是与 UAV 类型相关的观测状态信息. 对于领导者 UAV, $m_i=[x_G,y_G,O_{\text{flag}}]$, O_{flag} 为障碍物标志位. O_{flag} 的定义如下:

$$O_{\text{flag}}=\begin{cases}1,\ d_O<2R_O,\\ 0,\ d_O\geqslant 2R_O,\end{cases}\tag{2}$$

其中, $d_O=\sqrt{(x-x_O)^2+(y-y_O)^2}$ 为 UAV 与障碍物之间的距离. 当 UAV 与障碍物之间的距离小于两倍的障碍物半径时, 则认为 UAV 附近有障碍物, 此时障碍物标志位置为 1, 否则置为 0.

对于跟随者 UAV, $m_i=[x_L,y_L,v_L]$, 其中 (x_L,y_L) 为领导者 UAV 的位置, v_L 为领导者 UAV 的速度. 由于 UAV 突防编队的领导者 UAV 有且仅有一架, 所以其余的 UAV 都为跟随者 UAV, 跟随者 UAV 都有 $m_i=[x_L,y_L,v_L]$.

(2) 状态集 S 的设计. 第 i 个 UAV 的状态信息可以用向量 s_i 表示, 每个 UAV 的状态量 s_i 都包括该 UAV 的观测状态和其他 UAV 的观测状态信息, 即可以表示为 $s_i=[z_1,z_2,\ldots,z_N]$.

(3) 动作集 A 的设计. 本文所定义的 UAV 动作决策是通过选择合适的加速度和角速度执行 Δt 时间, 来达到 UAV 的期望速度和期望航向角. 第 i 个 UAV 的动作为一个连续的二维向量, 即联合动作空间可表示为 $a_i=[u_i,\omega_i]^{\mathrm{T}}$, 其中 u_i 表示无人机 i 的加速度, ω_i 表示无人机 i 的角速度, 满足式 (1) 中的约束条件.

(4) 奖励集 R 的设计. 为了更好地完成多无人机协同路径规划任务, 本文对不同类型的 UAV 设计了不同的奖惩机制. 具体地, 针对领导者 UAV (仅为一架), 首先设计边界奖励 r_{edge}, 当领导者 UAV 到达边界附近则给予 -1 的惩罚, 否则不给任何奖惩,

$$r_{\text{edge}}=\begin{cases}-1,\ \text{处于边界},\\ 0,\quad \text{否则}.\end{cases}\tag{3}$$

其次, 设计避障奖励 r_{obs}, 当领导者 UAV 与障碍物的距离小于障碍物半径的两倍, 此时 UAV 有潜在撞击障碍物的风险, 给予 -2 的惩罚; 当领导者 UAV 与障碍物的距离小于障碍物半径, 此时 UAV 与障碍物相撞, UAV 损毁, 给予 -500 的惩罚; 否则除上述情况外, 不给予 UAV 任何奖惩,

$$r_{\text{obs}}=\begin{cases}0, & d_O>2R_O,\\ -2, & R_O\leqslant d_O\leqslant 2R_O,\\ -500, & d_O<R_O.\end{cases}\tag{4}$$

然后, 设计目标奖励 r_{goal}, 当领导者 UAV 与目标点的距离小于阈值 D_1 时, 即认为到达目标点, 给予一个 1000 的正向奖励, 否则根据 UAV 与目标点的距离设计惩罚,

$$r_{\text{edge}}=\begin{cases}1000, & d_G\leqslant D_1,\\ -0.001\times d_G, & d_G>D_1.\end{cases}\tag{5}$$

再设计编队距离奖励 r_{follow}, 当领导者 UAV 与其他跟随者 UAV 的距离都保持在一定范围内时, 即认为成功组成了突防编队, 此时不给予任何的奖惩; 否则, 根据领导者 UAV 与其他跟随者 UAV 的距离设计惩罚,

$$r_{\text{follow},j}=\begin{cases}0, & d_{jL,\min}\leqslant d_{jL}\leqslant d_{jL,\max},\\ -0.001\times d_{jL}, & \text{否则},\end{cases} \tag{6}$$

$$r_{\text{follow}}=\sum_{j=0}^{N-1} r_{\text{follow},j}, \tag{7}$$

其中, j ($j=1,2,\ldots,N-1$) 为跟随者 UAV 的索引, d_{jL} 为第 j 个跟随者 UAV 与领导者 UAV 之间的距离, $d_{jL,\min}$, $d_{jL,\max}$ 为第 j 个跟随者与领导者的距离范围限制.

最后, 设计速度协同奖励 r_{speed}, 当领导者 UAV 与其他跟随者 UAV 的速度误差小于阈值 1 时, 即认为成功达到了速度协同, 此时给予 1 的奖励; 否则, 不给予 UAV 任何奖惩,

$$r_{\text{speed},j}=\begin{cases}1, & |v_L-v_j|<1,\\ 0, & \text{否则},\end{cases} \tag{8}$$

$$r_{\text{speed}}=\sum_{j=0}^{N-1} r_{\text{speed},j}. \tag{9}$$

综上, 可以得到领导者 UAV 的奖励函数为 $R_L=\omega_{L,1}r_{\text{edge}}+\omega_{L,2}r_{\text{obs}}+\omega_{L,3}r_{\text{goal}}+\omega_{L,4}r_{\text{follow}}+\omega_{L,5}r_{\text{speed}}$, 其中 $\omega_{L,1},\omega_{L,2},\omega_{L,3},\omega_{L,4},\omega_{L,5}$ 是每个部分的奖励权重.

针对跟随者 UAV (为 $N-1$ 架), 其奖励函数由编队距离奖励与速度协同奖励构成. 于是, 根据式 (6) 与 (8) 可以得到每个跟随者无人机 j 的奖励函数为 $R_{F,j}=\omega_{F,1}r_{\text{follow},j}+\omega_{F,2}r_{\text{speed},j}$, 其中 $\omega_{F,1},\omega_{F,2}$ 是每个奖励的部分权重.

总奖励集 R 由一个领导者 UAV 的奖励 R_L 和 $N-1$ 个跟随者 UAV 的奖励 $R_{F,j}$ 组成, $R=[R_L,R_{F,1},R_{F,2},\ldots,R_{F,N-1}]$, $r\in R$.

3.3 基于 MASAC 的多无人机协同路径规划策略求解

3.3.1 MASAC 算法

如图 2 所示, MASAC 算法依然采用了集中式训练分布式执行的框架, 每一个智能体都拥有一个独立的 SAC 算法. 不同于文献 [28] 中的 MASAC 算法将全局状态信息输入给策略网络的做法, 本文所提的 MASAC 算法对智能体策略网络的输入为局部状态信息. 在算法训练过程中, 智能体 i 使用局部观测信息 z_i 作为策略网络 actor 的输入, 输出则为动作 a_i. 环境 (environment, Env) 执行所有智能体的动作所组成的动作集 A 得到下一时刻的状态信息集 S' 和当前时刻奖励集 R, 再将当前时刻状态信息集 S、动作集 A、奖励集 R 和下一时刻的状态信息集 S' 存入到经验池中. 从经验池中随机抽取一批数据 (s,a,r,s'), 智能体 i 使用状态与动作拼接后的信息 (s_i,a_i) 作为 critic 网络的输入来指导价值网络训练.

3.3.2 MASAC 的网络更新

如图 3 所示, 对于智能体 i, 其 critic 网络、actor 网络和熵网络的更新过程与传统 SAC 算法类似. 用权重参数为 θ 的 actor 网络来拟合策略网络, 用权重参数为 w 的 critic 网络拟合价值网络.

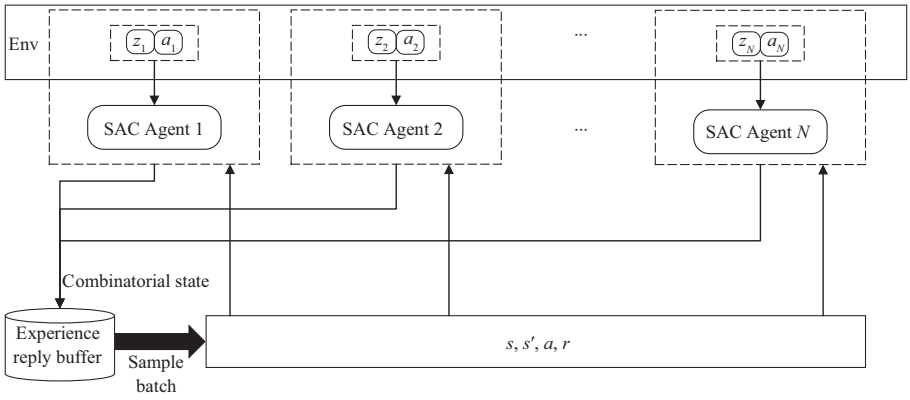

图 2 MASAC 算法训练总体框图

Figure 2 Training process of MASAC algorithm

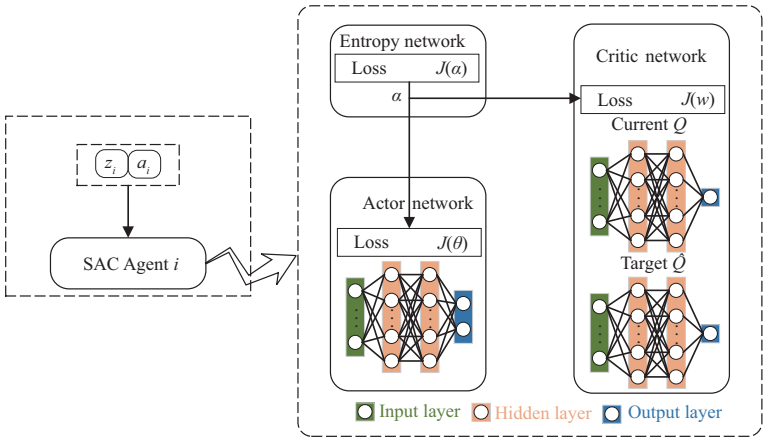

图 3 (网络版彩图) 第 i 个智能体的网络更新

Figure 3 (Color online) Network update for the i-th agent

Critic 网络的 $Q_i(s_t, a_t)$ 表示第 i 个智能体在状态 s_t 时, 采取动作 a_t 后得到的当前 Q 值. 为了能够估计连续和高维行动空间中的目标 $\hat{Q}_i(s_t, a_t)$ 值, 使用由参数 w 表示的神经网络来逼近 $\hat{Q}_i(s_t, a_t)$, 即 $Q_i(s_t, a_t) \approx \hat{Q}_i(s_t, a_t)$, 使用均方误差作为损失函数, 用 adaptive moment estimation (adam) 算法来优化更新 MASAC 网络参数. 因此, 第 i 个智能体的 critic 网络损失函数可表示为

$$J_{Q,i}(w) = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim D}\left[\frac{1}{2}\left(Q_i(s_t,a_t) - \hat{Q}_i(s_t,a_t)\right)^2\right], \tag{10}$$

其中 D 是经验池, 负责收集训练数据, 为训练提供样本. 由于 D 的存在, MASAC 可以利用以前的数据进行训练, 从而提高了样本的效率, 打破了样本之间的关联性. 目标 Q 网络的参数由 $\bar{w}$ 表示, 用式 (11) 的方式进行软更新,

$$\overline{w} = \tau w + (1-\tau)\overline{w}, \tag{11}$$

其中 τ 是软更新率.

Actor 网络是状态到行动的映射, MASAC 通过最小化 KL 散度来寻找到一个新的策略 π_{new}, 满

算法 1 基于 MASAC 的多无人机协同路径规划训练算法

初始化 N 个 actor 网络、N 个 critic 网络和 N 个熵网络;
初始化经验池 D 与 OU 噪声;
for episode = 1, Maxepisode **do**
  环境初始化并得到当前时刻的状态 s;
  for step = 1, Maxstep **do**
    使用 OU 噪声进行动作选择, 对于每个无人机 i, 根据状态 s 向 actor 网络输入观测信息 z_i, 得到动作 $a_i = [u_i, \omega_i]$;
    动作限幅至 $[-1, 1]$;
    根据联合动作 $a = [a_1, a_2, \ldots, a_N]$, 执行 step 函数, 环境返回奖励 r, 下一时刻状态 s', 回合结束标志位 done;
    存储数据 (s, a, r, s') 于经验池 D;
    if 经验池 D 存满 **then**
      从 D 中随机抽取一批数据, 对于每个无人机 i, 根据式 (10) 执行 critic 网络的梯度下降;
      根据式 (12) 执行 actor 网络的梯度下降;
      根据式 (13) 执行熵网络的梯度下降;
      Critic 网络的权重参数根据式 (11) 软更新, actor 网络与熵网络采用固定步长的硬更新;
    end if
    更新状态 $s \leftarrow s'$;
    if 回合结束标志位 done 为真 **then**
      Break;
    end if
  end for
end for

足新策略下的 Q 值大于等于老策略下的 Q 值. 因此, 第 i 个智能体的策略 π 损失函数可表示为

$$J_{\pi,i}(\theta) = D_{\mathrm{KL}}\left(\pi_i\left(\cdot \mid s_t\right) \| \exp\left(\frac{1}{\alpha} Q_i\left(s_t, \cdot\right) - \log \mathcal{Z}_i\left(s_t\right)\right)\right), \tag{12}$$

其中, α 为温度系数, $\mathcal{Z}_i(s_t)$ 是用于归一化分布的函数, 它取决于智能体 i 的状态, 对策略网络的参数梯度没有影响.

由于熵网络具有自动调整温度系数 α 的能力 [29], 从而保持了探索和利用之间的平衡, 智能体 i 的熵网络在时刻 t 的损失函数可表示为

$$J_i(\alpha) = \mathbb{E}_{a_t \sim \pi}\left[-\alpha \log \pi\left(a_t \mid s_t\right) - \alpha \mathcal{H}_0\right], \tag{13}$$

其中, $\mathcal{H}_0$ 是预定义的最小策略熵阈值 (文献 [22] 中推荐 $\mathcal{H}_0 = -\dim(A)$; 例如, 四维连续动作空间环境的 $\mathcal{H}_0$ 为 -4).

本文提出的基于 MASAC 多无人机协同路径规划策略训练算法步骤如算法 1 所示. 采用可以产生时序相关探索的 Ornstein-Uhlenbeck (OU) 噪声作为策略网络的探索噪声.

4 仿真验证

4.1 MASAC 算法参数设置

表 1 中列出了 MASAC 算法用到的所有参数和超参数. 设置 OU 噪声的 3 个参数大小为 $\sigma_{\mathrm{OU}} = 0.1, \theta_{\mathrm{OU}} = 0.1, dt_{\mathrm{OU}} = 0.01$, 并且仅在前 20 回合加入 OU 噪声. 一些主要安装包的版本信息为 pygame 版本: 2.1.2; gym 版本: 0.19.0; pytorch 版本: 1.10.0+cu113; numpy: 1.23.1.

表 1 MASAC 的参数和超参数

Table 1 Parameters and hyperparameters of MASAC

MASAC parameter	MASAC hyperparameter	Actor network	Critic network
Actor learning rate: 1e−3	Maxstep: 1000	Input: 7	Input: $7 \times N + 2$
Critic learning rate: 3e−3	Batch size: 128	Hidden: (256, 256)	Hidden: (256, 256)
Entropy learning rate: 3e−4	Maxepisode: 500	Output: 2	Output: 1
Soft update rate: 1e−2	Replay buffer: 20000	Activation: relu, tanh	Activation: relu
Discount factor: 0.9	Optimizer: adam	Weights: (0, 0.1)	Weights: (0, 0.1)

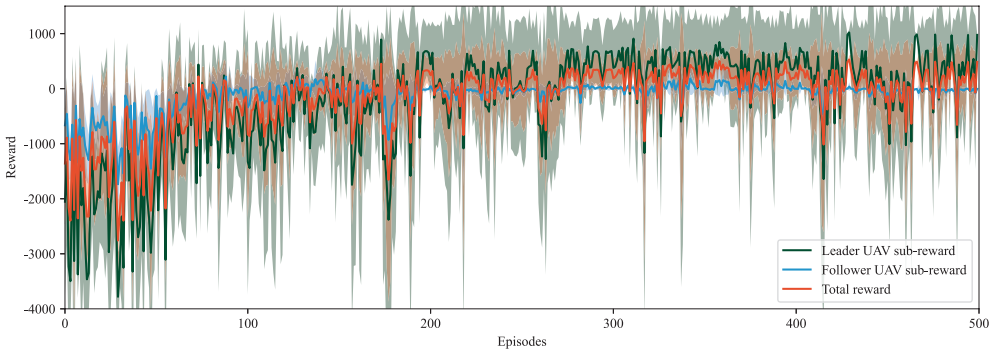

图 4 (网络版彩图) 500 回合训练下 MASAC 算法奖励曲线图

Figure 4 (Color online) Reward curves of the MASAC algorithm under 500 training episodes

如表 1 所示, actor 网络是一个具有 7-256-256-2 结构的全连接层, 激活函数由 relu 和 tanh 构成, 神经网络的初始权重在均值为 0, 方差为 0.1 的正态分布上随机取得, 学习率为 1e−3. Critic 网络将状态与动作的增广信息作为输入, 输入层节点数目是 $7 \times N + 2$, 激活函数仅由 relu 构成, 神经网络的初始权重在均值为 0, 方差为 0.1 的正态分布上随机取得, 学习率为 3e−3. 熵网络没有全连接层结构, 其对数熵权重在训练过程中自动调整, 以平衡探索和利用, 学习率为 3e−4.

4.2 训练过程

在训练阶段, 为了加速训练过程, 将时间步长 Δt 设置为 1 以节省训练的时间花费, 整个仿真实验在具有 16 GB RAM (RTX-3050 显卡) 的标准英特尔酷睿 i5-11260H 上运行. 考虑二机编队协同来进行训练, 即一架领导者 UAV 和一架跟随者 UAV. 如果任务失败, 那么这回合的训练会被重置, 并继续开启下一回合的训练, 训练的总回合数设置为 500. 计算智能体每个回合所取得单步奖励的累加值, 画出 MASAC 算法的领导者 UAV 子奖励曲线图、跟随者 UAV 子奖励曲线图和总奖励曲线图, 并重复多次 500 回合的训练计算平均值与方差, 得到的图像如图 4 所示.

图 4 中红线为总奖励曲线, 总奖励在前 20 回合奖励约为 −2000, 此时智能体们在 OU 噪声的作用下随机地探索环境; 20 回合后, OU 噪声取消, 总奖励逐渐升高; 150 回合后, 总奖励趋于稳定, 奖励值大致收敛在 200 左右. 绿线为领导者 UAV 子奖励曲线, 它由边界奖励、避障奖励、目标奖励、编队距离奖励和速度协同奖励组成; 该奖励在 150 回合前逐渐升高, 在 150 回合后收敛到 500 左右. 蓝线为跟随者 UAV 子奖励曲线, 它由编队距离奖励和速度协同奖励组成, 奖励值由最初的 −1000 逐渐收敛在 0 左右. 上述实验结果说明了 MASAC 算法收敛, 智能体们找到了一种近似最优策略.

表 2　不同无人机数量下性能指标

Table 2　Evaluation metrics with different numbers of UAVs

Number of UAVs	J_{MCR} (%)	J_{FKR} (%)	J_T	J_S	J_C
2	92.0	64.62	269.53	148.46	301.62
3	92.0	53.24	251.67	145.07	281.25
4	91.0	43.39	264.12	146.51	299.76
5	90.0	35.66	249.73	141.17	281.39

表 3　MASAC 算法与其他策略的对比

Table 3　Comparison between MASAC and other strategies[a)]

Strategy	J_{MCR} (%)	J_{FKR} (%)	J_T	J_S	J_C
Random strategy	2.0	0.00	937.79	459.82	940.85
MADDPG	83.0	0.61	333.09	197.79	321.35
MASAC	**90.0**	**35.66**	**249.73**	**141.17**	**281.39**

a) The bold indicates that it is the optimal.

4.3　测试过程与结果分析

4.3.1　MASAC 算法的性能指标分析

在测试阶段, 为了清晰地观察多无人机的飞行路径, 将时间步长 Δt 设置为 0.1. 使用 4.2 小节已经训练好的神经网络权重参数在环境中进行仿真测试, 在这个权重参数的基础上将 UAV 的数量扩展为 3, 4, 5 架. 选择 2.4 小节定义的任务完成率 J_{MCR}、编队保持率 J_{FKR}、飞行时间 J_T、飞行轨迹 J_S 和能量消耗 J_C 作为衡量算法优劣的性能指标, 分别进行 100 次 Monte Carlo 测试, 测试的数据结果如表 2 所示.

从表 2 可见, 无论 UAV 的数量如何变换, 其成功到达目标地点的路径规划任务完成率都在 91% 左右, 飞行时间都在 255 左右, 飞行轨迹的路程长度都在 145 左右, 能量消耗都在 300 左右. 这说明该算法能灵活应用于不同数量的 UAV 中. 其中, 多无人机的编队保持率随着 UAV 数量的增大而减少, UAV 数量越多, 其编队保持率就越低. 这是因为编队保持率是领导者 UAV 与其他所有跟随者 UAV 共同的编队保持率, 当 UAV 数量增多时, 其编队保持成功条件就越严格, 从而导致编队保持率降低. 上述测试实验验证了 MASAC 算法对多无人机协同路径规划问题的有效性.

4.3.2　MASAC 算法与其他策略的对比分析

现在将 MASAC 算法与随机策略、MADDPG 算法进行对比. 对这些策略的描述如下:

(1) 随机策略: 该随机策略的随机性符合正态分布, 对于每个智能体 i 在每一个时刻 t 下, 随机地选择控制量 (加速度 u_i 与角速度 ω_i) 的大小, 智能体 i 通过执行动作 $a_i = [u_i, \omega_i]^{\mathrm{T}}$ 从而达到移动的目的.

(2) MADDPG 算法: 如文献 [21] 所述.

在与其他策略的对比测试中, 以 UAV 数量 $N = 5$ 为例. 依然以任务完成率、编队保持率、飞行时间、飞行轨迹和能量消耗作为衡量算法优劣的性能指标. 表 3 展示了 100 个 Monte Carlo 测试回合下每种策略所取得的性能指标平均值.

如表 3 所示, 任务完成率最高的是本文所提出的 MASAC 算法, 达到了 90% 的任务完成率, 其次

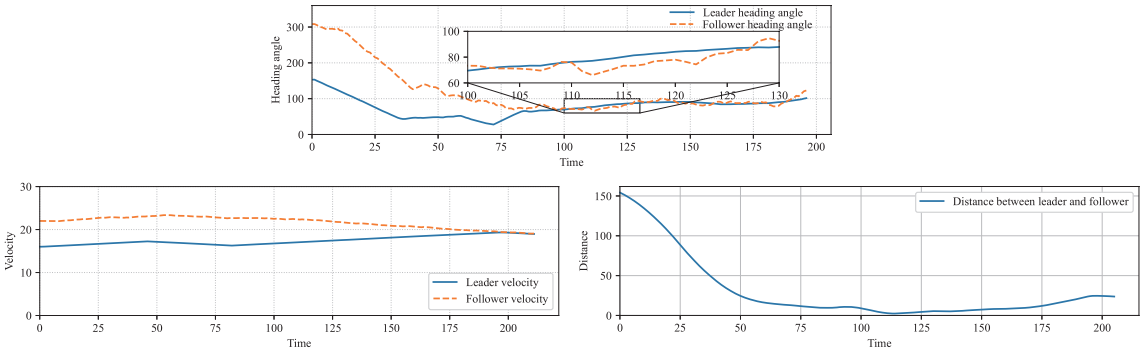

图 5　(网络版彩图) 领导者与第 i 个跟随者 UAV 的航向角、速度和距离变化曲线

Figure 5　(Color online) Variation curves of heading angle, velocity, and distance between the leader and the i-th follower UAV

是 MADDPG 算法的 83%, 而随机策略的任务完成率远低于 MASAC 与 MADDPG 算法. 编队保持率最高的是 MASAC 算法, 达到了 35.66%, 而随机策略与 MADDPG 算法的编队保持率不足 1%. 对于飞行时间、飞行轨迹和能量消耗这 3 个指标, MASAC 算法相比随机策略、MADDPG 算法取得了最短的飞行时间 249.73、最短的飞行轨迹 141.17 以及最小的能量消耗 281.39. 显然, MASAC 算法能以最小的指标成本取得更高的任务完成率与编队保持率, 比其他策略更具优越性.

4.3.3　MASAC 算法的多无人机协同性分析

图 5 画出了某测试回合中, 领导者 UAV 与第 i 个跟随者 UAV 的航向角、速度和距离随时间的变化曲线. 可以看出, 两架 UAV 的航向角、速度和距离起初相差较大, 随着时间的变化航向角和速度趋于一致, 距离则趋于 0, 说明无人机达到了航向角和速度的一致性, 两架 UAV 的距离越来越近. 这意味着跟随者 UAV 能够向领导者 UAV 靠拢, 多无人机此时做同步运动. 在此过程中, 无人机展现出显著的编队保持协同性, 即能够在维持整体编队队形的同时, 顺利完成路径规划任务.

5　结论

本文以空域场景下异构多无人机飞往目标地点为研究背景, 开发了一个多无人机协同路径规划的可视化强化学习环境, 提出了一种多智能体 DRL 算法 —— MASAC 来训练多无人机进行协同路径规划. 结果表明, 训练后的多无人机可以在编队保持、安全避障和界内飞行的情形下到达目标地点, MASAC 算法不仅可以灵活应用于不同数量的无人机中, 还能取得比随机策略、MADDPG 算法更高的性能指标, 并实现了协同作用. 在未来工作中, 我们将去除二维空间的假设, 考虑无人机在三维空间中的运动. 此外, 本文工作尚在仿真阶段, 后续可尝试应用到实际的多无人机系统中, 以解决在实际中可能出现的挑战.

参考文献

1　Shen C, Li L, Wu Y, et al. Research on the capabality of the U.S. manned/unmanned autonomous collaborative operations. Tactical Missile Technol, 2018, 6: 16–21 [申超, 李磊, 吴洋, 等. 美国空中有人/无人自主协同作战能力发展研究. 战术导弹技术, 2018, 6: 16–21]

2　Qiao Z, Li S L, Wang J Z, et al. UAV path planning based on PER-PDDPG. Unmanned Syst Technol, 2022, 5: 12–23 [乔哲, 黎思利, 王景志, 等. 基于 PER-PDDPG 的无人机路径规划研究. 无人系统技术, 2022, 5: 12–23]

3 Dijkstra E W. A note on two problems in connexion with graphs. Numer Math, 1959, 1: 269–271

4 Hart P E, Nilsson N J, Raphael B. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cyber, 1968, 4: 100–107

5 Stentz A. Optimal and efficient path planning for partially-known environments. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1994. 3310–3317

6 Dewangan R K, Shukla A, Godfrey W W. Three dimensional path planning using grey wolf optimizer for UAVs. Appl Intell, 2019, 49: 2201–2217

7 Han Z, Chen M, Shao S, et al. Path planning of unmanned autonomous helicopter based on hybrid satisficing decision-enhanced swarm intelligence algorithm. IEEE Trans Cogn Dev Syst, 2023, 15: 1371–1385

8 Li J, Xiong Y, She J. UAV path planning for target coverage task in dynamic environment. IEEE Internet Things J, 2023, 10: 17734–17745

9 Roberge V, Tarbouchi M, Labonte G. Fast genetic algorithm path planner for fixed-wing military UAV using GPU. IEEE Trans Aerosp Electron Syst, 2018, 54: 2105–2117

10 Yu Z, Si Z, Li X, et al. A novel hybrid particle swarm optimization algorithm for path planning of UAVs. IEEE Internet Things J, 2022, 9: 22547–22558

11 Zhu D, Yang S X. Bio-inspired neural network-based optimal path planning for UUVs under the effect of ocean currents. IEEE Trans Intell Veh, 2022, 7: 231–239

12 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533

13 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359

14 Yin S, Xiang Z. Adaptive operator selection with dueling deep Q-network for evolutionary multi-objective optimization. Neurocomputing, 2024, 581: 127491

15 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: The MIT Press, 2018

16 Lan W, Jin X, Chang X, et al. Path planning for underwater gliders in time-varying ocean current using deep reinforcement learning. Ocean Eng, 2022, 262: 112226

17 Wang X, Gursoy M C, Erpek T, et al. Learning-based UAV path planning for data collection with integrated collision avoidance. IEEE Internet Things J, 2022, 9: 16663–16676

18 Zhang Y, Chadli M, Xiang Z. Prescribed-time formation control for a class of multiagent systems via fuzzy reinforcement learning. IEEE Trans Fuzzy Syst, 2023, 31: 4195–4204

19 Hadi B, Khosravi A, Sarhadi P. Deep reinforcement learning for adaptive path planning and control of an autonomous underwater vehicle. Appl Ocean Res, 2022, 129: 103326

20 Zhang S, Li Y, Dong Q. Autonomous navigation of UAV in multi-obstacle environments based on a deep reinforcement learning approach. Appl Soft Computing, 2022, 115: 108194

21 Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of Conference on Neural Information Processing Systems, 2017

22 Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of International Conference on Machine Learning, 2018. 1861–1870

23 Ragi S, Chong E K P. UAV path planning in a dynamic environment via partially observable Markov decision process. IEEE Trans Aerosp Electron Syst, 2013, 49: 2397–2412

24 Zhang T T, Yang X J. Autonomous coordination saturation attacks method for loitering munitions in urban scenarios based on reinforcement learning. J Command Control, 2023, 9: 457–468 [张婷婷, 杨学军. 基于强化学习的城市场景下巡飞弹自主协同饱和攻击方法. 指挥与控制学报, 2023, 9: 457–468]

25 Wang L, Wang K, Pan C, et al. Multi-agent deep reinforcement learning-based trajectory planning for multi-UAV assisted mobile edge computing. IEEE Trans Cogn Commun Netw, 2020, 7: 73–84

26 Bellman R. A Markovian decision process. J Math Mech, 1957, 6: 679–684

27 Bertsekas D. Reinforcement Learning and Optimal Control. Belmont: Athena Scientific, 2019
28 Enders T, Harrison J, Pavone M, et al. Hybrid multi-agent deep reinforcement learning for autonomous mobility on demand systems. In: Proceedings of Learning for Dynamics and Control Conference, 2023. 1284–1296
29 Haarnoja T, Tang H, Abbeel P, et al. Reinforcement learning with deep energy-based policies. In: Proceedings of International Conference on Machine Learning, 2017. 1352–1361

Multi-UAV collaborative path planning based on multi-agent soft actor critic

Chengliang FANG[1], Feisheng YANG[1,2*] & Quan PAN[1]

1. *School of Automation, Northwestern Polytechnical University, Xi'an* 710129*, China*;
2. *Innovation Center NPU Chongqing, Northwestern Polytechnical University, Chongqing* 401151*, China*
* Corresponding author. E-mail: yangfeisheng@nwpu.edu.cn

Abstract This paper proposes a novel multi-agent deep reinforcement learning algorithm for the collaborative path planning problem of heterogeneous unmanned aerial vehicles (UAVs) in a dynamic uncertain environment. Firstly, a reinforcement learning environment for UAVs is developed to reach a target location in an airspace scenario, where the environment introduces the UAV dynamics equations and considers the UAV heterogeneity as well as the requirement for safe obstacle avoidance. Secondly, evaluation metrics including task completion rate, formation maintenance rate, flight time, flight trajectory, and energy consumption are designed to evaluate the algorithm performance. Then, the multi-UAV collaborative path planning problem is modeled as a partially observable Markov decision process and a multi-agent soft actor critic algorithm is proposed to seek the approximate optimal strategy for the problem. Finally, the effectiveness and superiority of the proposed algorithm are demonstrated through simulations.

Keywords multi-UAV, path planning, multi-agent deep reinforcement learning, partially observable Markov decision process, multi-agent soft actor critic algorithm