doi: 10.7690/bgzdh.2026.02.022

基于 PPO 算法的多无人机编队避障控制方法

王何鹏飞[1,2]，黄　杰[2]，王　伟[1]，曾　刊[1]，王　楠[2]，洪华杰[2]
（1. 中国兵器装备集团自动化研究所有限公司系统总体部，四川 绵阳 621000；
2. 国防科技大学智能科学学院，长沙 410073）

摘要：为解决多无人机编队在复杂障碍物中执行任务时训练难度大、多机难以建模等问题，提出一种基于链式训练并含有启发式信息的近端策略优化（proximal policy optimization，PPO）算法的多无人机穿梭树林端到端运动规划方法。综合考虑无人机的动态特性和 3 维连续环境的复杂性，设计一种有效的运动规划策略的强化学习训练方法。通过模拟实验，验证了该方法在多无人机编队穿梭树林任务中的有效性和优越性。研究结果表明：该方法能够在避障的前提下保持一定的编队稳定性，到达目标点，且在保持编队稳定性和通过率方面均优于传统的人工势场法。该研究为无人机编队在复杂环境中的自主导航和路径规划提供了新的视角和解决方案。

关键词：无人机编队；编队任务；运动规划；改进 PPO 算法；自主导航；路径规划
中图分类号：V279　　**文献标志码：**A

Multi-UAV Formation Obstacle Avoidance Control Method Based on PPO Algorithm

Wang Hepengfei[1,2], Huang Jie[2], Wang Wei[1], Zeng Kan[1], Wang Nan[2], Hong Huajie[2]
(1. *Department of System General*, *Automation Research Institute Co*., *Ltd. of China South Industries Group Corporation*, *Mianyang* 621000, *China*;
2. *College of Intelligence Science and Technology*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: In order to solve the problems of difficult training and modeling of multiple UAVs in complex obstacles, an end-to-end motion planning method for multiple UAVs shuttling through the forest based on chain training and proximal policy optimization (PPO) algorithm with heuristic information is proposed. Considering the dynamic characteristics of UAV and the complexity of three-dimensional continuous environment, an effective reinforcement learning training method for motion planning strategy is designed. Simulation results show the effectiveness and superiority of the proposed method in the task of multiple UAVs formation shuttling through the forest. The results show that the method can maintain a certain formation stability and reach the target point on the premise of obstacle avoidance, and it is superior to the traditional artificial potential field method in maintaining formation stability and passing rate. This study provides a new perspective and solution for autonomous navigation and path planning of UAV formation in complex environment.

Keywords: UAV formation; formation mission; motion planning; improved PPO algorithm; autonomous navigation; path planning

0　引言

随着无人机技术的发展日新月异，其应用领域也日益广泛，特别是在复杂环境中，如森林、城市峡谷等，多无人机的自主导航和路径规划能力显得尤为重要[1]。在这样的环境中，无人机需要具备高度的自主性和灵活性，以便在执行任务时能够避开障碍物，同时保证高效的飞行路径[2]。

随着人工智能技术的快速发展，基于深度强化学习（deep reinforcement learning，DRL）[3]的控制方法引发了广泛的关注。基于深度强化学习算法的多无人机穿梭密集障碍物环境的运动规划成为了一个研究热点[4]。近端策略优化（PPO）算法[5]是一种在强化学习领域中广泛使用的算法，通过优化策略网络来提高决策的质量。通过改进 PPO 算法，可以进一步提升多无人机在复杂环境中的自主导航能力。

笔者介绍一种基于 PPO 算法的多无人机穿梭树林的端到端的运动规划方法。该方法通过结合无人机的动态特性和环境的复杂性，设计了一种有效的强化学习训练方案，得到高质量的无人机运动控制策略。通过模拟实验，验证了该方法在多无人机编队穿梭树林任务中的有效性和优越性。

1　问题描述

笔者考虑的任务是多无人机编队面对复杂障碍物能够在 3 维空间自由移动协同避障到达各自目标

收稿日期：2024-11-15；**修回日期：**2024-12-17
第一作者：王何鹏飞（2001—），男，湖北人，硕士。

点。为解决该任务训练难度大，多机难以建模实现等问题，笔者设计了基于 gymnasium 的连续 3 维虚拟丛林世界作为训练测试环境，以适配多无人机编队穿梭任务。虚拟世界环境中随机生成树林障碍物，使用基于深度强化学习算法训练 9 架无人机能避开静态障碍物(树)以及动态障碍物(其他无人机)并保持队形，整体到达设定好的距离集群中心正前方 5～6 m，上下左右偏移 2 m 内的任意位置，如图 1 所示。

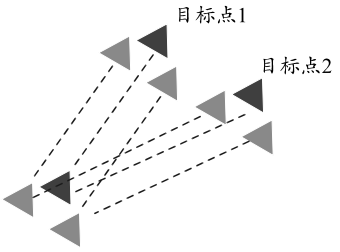

图 1　问题描述

由图 1 可知：三角形为无人机，深黑色为编队中心无人机，目标点的位置坐标作为编队中心按队形相对位置生成每个无人机的对应目标点，无人机按预设队形到达，同一组目标点不会产生交叉，强化学习每轮训练以范围内随机位置作为目标点。训练结束后，可以选取范围内任意点位作为编队中心的目标点，无人机会保持队形在避障的前提下移动到对应位置，并且任务执行中与障碍物碰撞几率相对传统的人工势场法大幅降低，同时也能在一定程度上牺牲部分路径代价来保持编队稳定性。

2　方法设计

为适配多维连续运动空间，采用基于 PPO 强化学习的链式学习方法，并设计启发式信息、状态空间、动作空间、奖励函数以及链式训练流程，对多个智能体进行训练学习。

2.1　基于 PPO 强化学习的链式学习方法

链式 PPO 训练框架采用分布式训练方式，主要解决 PPO 算法面对多智能体复杂连续环境下难以收敛的问题。多智能体中每个个体执行独立连续动作时会对整体状态产生较大影响，同时多个智能体同时训练时会使环境特别不稳定，难以训练出多维连续的运动空间与状态空间下的策略模型。笔者设计的框架采用分布式训练方式。按链式训练流程，之前已经训练的无人机按照对应训练好的 PPO 策略网络选择动作，未训练的无人机按照人工势场法避障到达目标点，只有当前训练的无人机的策略网络参与本次训练并更新参数。使用这种方法主要是因为多智能体如果同时训练动作维度太大，智能体变化也大，强化学习网络难以收敛，然而当每次只有一架无人机参与训练，动作维度大幅减小，在每次训练时，非当前训练的其他架无人机的策略相对固定，让当前策略适配其他无人机策略并达到个体相对最优。训练框架如图 2 所示。

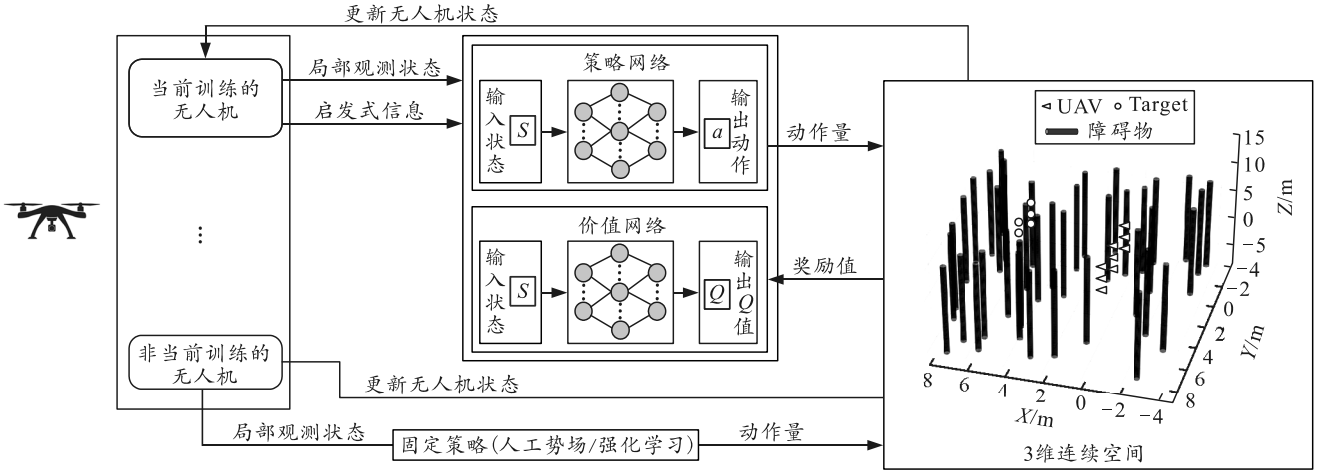

图 2　基于 PPO 多无人机训练框架

在训练过程中除了当前训练中的无人机，其他无人机保持固定策略，即策略不会参与训练调整，这样可以有效保持训练环境的稳定性。当前训练的无人机采用 PPO 算法，并增加启发式信息作为指引，策略网络和值函数网络的输入为当前无人机局部观测状态，策略网络输出动作量到训练环境中对整体状态进行更新，值函数网络输出 Q 值对当前状态进行评估。非当前训练的其他无人机基于各自状态采用已有策略分别输出各自动作量到环境更新状态。

由于单次训练只更新当前强化学习网络，机间协同难以实现，需要多轮训练，使得策略之间互相

适配。第一轮训练时，采用人工势场法模拟非当前训练的其他无人机避障穿梭运动控制策略。

选择第一架无人机训练后，选择编队最近邻的无人机作为下一个训练的无人机，训练框架如图 3 所示。

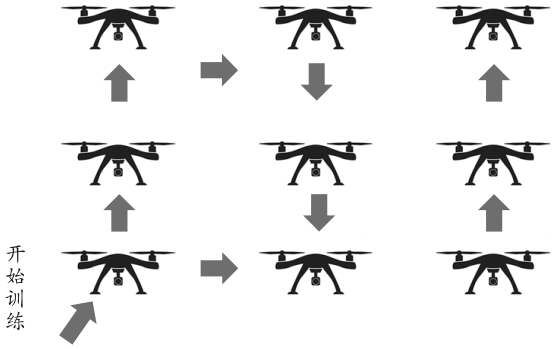

图 3 链式训练流程

图 3 中箭头指向的是下一个参与训练的无人机，当训练完成一架无人机后，为该无人机加载训练后的模型，参与之后的训练。一轮训练后，如果实验结果中任务完成率(重复多次实验无碰撞)小于 95%，无人机加载各自模型进行多轮重复训练。

2.2 穿梭任务 MDP 建模

针对链式 PPO 训练方法，需要设计动作空间、局部观测状态空间、奖励函数以及启发式信息，满足无人机能够避开密集树林障碍物保持编队状态到达目标点的任务要求。

2.2.1 局部观测状态空间

多架无人机执行该任务时，由于空间是 3 维连续动作空间，即有 x, y, z 轴，环境相对一般设定的 2 维离散的环境更为复杂。同时多架无人机在同一场景协同行动也给环境带来不稳定性。

无人机无法直接观察到环境的完整状态，只能通过接收到的观测信息来推断当前环境的状态，即局部观测状态空间[6]。局部观测状态空间可以有效提高策略泛化性，同时状态空间的设计需要与期望效果联系，反映真实情况。

设当前训练无人机为第 i 架无人机，笔者使用当前位置朝自身目标点的方向向量 dir_i 作为状态空间前 3 维的值。为降低路径代价即缩短到目标点的距离，前 3 维状态量反映当前无人机最短路径的速度方向。最近邻 2 架无人机位置相对当前无人机位置的向量作为其中 6 维的状态量 $\mathrm{pos}_1, \mathrm{pos}_2$，距离当前无人机距离小于 1.5 m 的静态障碍物信息作为其中 18 维状态量$\{\mathrm{obs}_1\}$，同时设计当前无人机速度信息 speed_i、与预期编队的相对位置 $\mathrm{formation}_i$ 作为状态量。局部观测状态空间设计：

$$S_i=\{\mathrm{dir}_i, \mathrm{pos}_p, \mathrm{pos}_z, \mathrm{obs}_1, ..., \mathrm{speed}_i, \mathrm{formation}_i\}。\quad (1)$$

2.2.2 动作空间

为满足无人机在 3 维连续空间$\{x, y, z\}$中运动避障，动作空间考虑连续的 3 维速度$\{V_x, V_y, V_z\}$，$\{V_x, V_y, V_z\}$分别控制当前训练无人机在此状态下各方向速度，各方向速度约束条件为$|v|\leqslant 0.4$ m/s，每个时间步采用当前策略生成的动作量对无人机位置信息等状态进行更新。

2.2.3 奖励函数

笔者结合多维复杂的训练环境以及编队避障任务的要求，将奖励函数设计为：碰撞惩罚、路径代价以及编队稳定性奖励，这 3 部分之和作为奖励来指引无人机学习到最优策略。

1) 碰撞惩罚。

由于任务障碍物分为动态障碍物与静态障碍物 2 种，笔者设计的碰撞惩罚设计为 2 部分：

$$\mathrm{num\_epsiode}=k;\quad (2)$$

$$R^i_{\mathrm{obs}}=\begin{cases}0 & \mathrm{if}\left\|\mathrm{traj}_i-\mathrm{tree}\right\|\in(0.35,+\infty)\\ -10 & \mathrm{if}\left\|\mathrm{traj}_i-\mathrm{tree}\right\|\in[0.22,0.35]\\ -50 & \mathrm{if}\left\|\mathrm{traj}_i-\mathrm{tree}\right\|\in(0,0.22)\end{cases};\quad (3)$$

$$R^i_{\mathrm{obst}}=\begin{cases}0 & \mathrm{if}\left\|\mathrm{traj}_i-\mathrm{UAV}_S\right\|\in(0.3,+\infty)\\ -10 & \mathrm{if}\left\|\mathrm{traj}_i-\mathrm{UAV}_S\right\|\in[0.1,0.3]\\ -50 & \mathrm{if}\left\|\mathrm{traj}_i-\mathrm{UAV}_S\right\|\in(0,0.1)\end{cases};\quad (4)$$

$$R_{\mathrm{obs}}=\sum_{i=1}^{k}R^i_{\mathrm{obs}}+R^i_{\mathrm{obst}}。\quad (5)$$

式中：num_epsiode 为单步沿途采样点数量；k 为每步移动轨迹均分份数，用于计算沿途碰撞惩罚；R^i_{obs} 和 R^i_{obst} 分别为第 i 个采样时刻当前无人机此时与静态障碍物和动态障碍物的碰撞惩罚；traj_i 为第 i 个采样点无人机位置；tree 为静态障碍物(密集树木)；UAV_S 为动态障碍物(同组其他无人机)，当无人机与障碍物距离处于不同区间时，选择不同的离散数值作为惩罚，表示不同距离的危险度；R_{obs} 为所有采样点碰撞惩罚之和，作为这一步的总碰撞惩罚。

2) 路径代价。

要使无人机轨迹能满足任务需求并最优，训练时需要设计每步的路径代价避免非必要绕路，即路径代价为：

$$R_{\mathrm{pc}}=-m*||\mathrm{action}||,\ m>0。\quad (6)$$

式中：R_{pc} 为当前步路径代价；action 为动作空间；

m 为一正系数。

3) 编队稳定性奖励。

无人机编队因其在执行复杂多变任务(如搜索救援、资源勘探、侦察监视等)方面的卓越功能和灵活性而受到广泛关注[7]。为保持队形稳定性，笔者采用虚拟 Leader-Follow 模式进行稳定性控制，Leader-Follow 模式是一种常见的控制策略，笔者使用位于队形中心的一架虚拟无人机作为领航者(Leader)，其他无人机作为跟随者(Followers)保持队形飞行。由于动作约束设定为每个方向速度绝对值小于 0.4 m/s，设计虚拟领航者无人机以 0.4 m/s 的速度到达目标点附近，当前虚拟无人机位置通过向量变换获得当前训练无人机预期位置。为保持多架无人机集群时队形的稳定性，设计奖励函数使得无人机更偏好在预设轨迹上飞行：

$$R_{\mathrm{fc}}=\frac{n}{\left\|\mathrm{cur}_{\mathrm{pos}}-\mathrm{pre}_{\mathrm{pos}}\right\|+0.01},\quad n\in(0,1)。\tag{7}$$

式中：R_{fc} 为计算无人机当前位置与预期位置距离的倒数作为保持编队的奖励；$\mathrm{cur}_{\mathrm{pos}}$ 为当前位置；$\mathrm{pre}_{\mathrm{pos}}$ 为预期位置。当无人机与预期位置越接近，该奖励越大。

$$R_{\mathrm{vs}}=1.5\ \text{ if}\left\|\mathrm{cur}_v-\mathrm{leader}_v\right\|<0.05。\tag{8}$$

式中：cur_v 为当前速度；leader_v 为领航者速度；R_{vs} 为稀疏奖励，鼓励训练无人机速度与虚拟领航者速度相似，保持编队稳定性。

4) 启发式信息。

由于 3 维连续空间较大，如果不给予启发式信息，无人机能够到达目标点这样的正样本相对较少，稀疏的正样本难以复现，训练难度高，同时由于碰撞惩罚较大，无人机训练时会习得踌躇不前的决策，难以获得较优的策略模型，笔者采用一个启发式速度叠加到 action 动作量上，启发式速度 V_h^i：

$$V_h^i=l*\left(\mathrm{cur\_pose}_i-\mathrm{tar\_pos}_i\right)\Big/\left\|\mathrm{cur\_pose}-\mathrm{tar\_pos}\right\|。\tag{9}$$

式中：l 为一较小正系数；$\mathrm{cur\_pose}_i$ 为当前无人机位置；$\mathrm{tar\_pos}_i$ 为当前无人机的目标点位置；cur_pose 为所有无人机的坐标矩阵；tar_pos 为对应目标点坐标矩阵；$l*\left(\mathrm{cur\_pose}_i-\mathrm{tar\_pos}_i\right)$ 为当前无人机指向目标点的向量；$\left\|\mathrm{cur\_pose}-\mathrm{tar\_pos}\right\|$ 为所有无人机指向各自目标点的向量矩阵的范数。当无人机过于领先其他无人机时，启发式速度值较小，提高更多探索度；当落后时，较大的启发式速度指引无人机跟上队形。

3　多无人机穿梭树林仿真实验

3.1　实验设置

笔者基于 gymnasium 仿真平台搭建仿真环境，环境中包含 9 架无人机，40 棵互不冲突的树作为静态障碍物。在距离编队位置 x 轴方向 5～6 m，y 轴和 z 轴偏移量-2～2 m 的范围内选取一个目标点，障碍物在区域内随机生成，通过链式训练实现各无人机在路径短、无碰撞以及保持编队稳定性的条件下到达各自目标点。

3.2　仿真结果以及分析

单次训练平均奖励变化曲线如图 4 所示，到达目标点平均步长数曲线变化如图 5 所示。

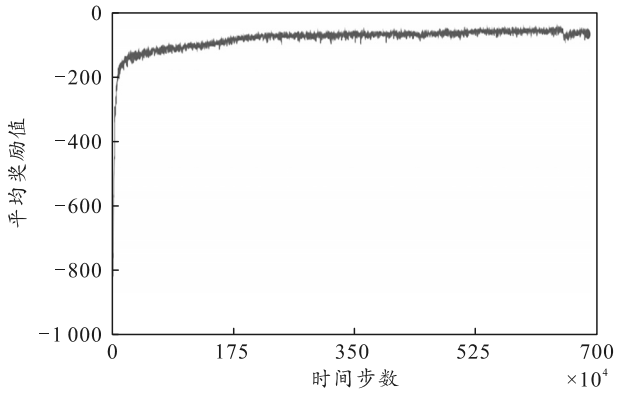

图 4　平均奖励变化

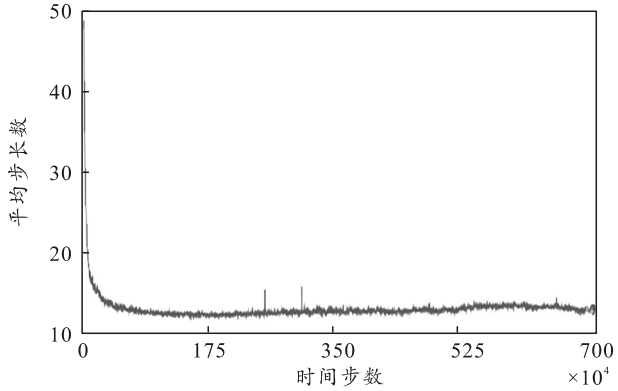

图 5　平均步长数变化

平均奖励变化曲线纵坐标表示每 2 048 个时间步的平均回合累计奖励，平均步长曲线纵坐标表示每 2 048 个时间步的平均每回合步长。训练过程中平均步长先大幅下降以减小路径代价，再略微上升以符合编队稳定性。当奖励到达-50 左右模型基本收敛。单架无人机结果如图 6 所示，多架无人机效果如图 7 所示。

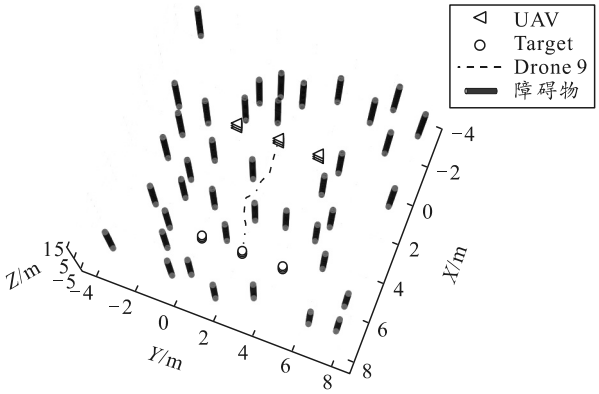

图 6 单无人机训练结果

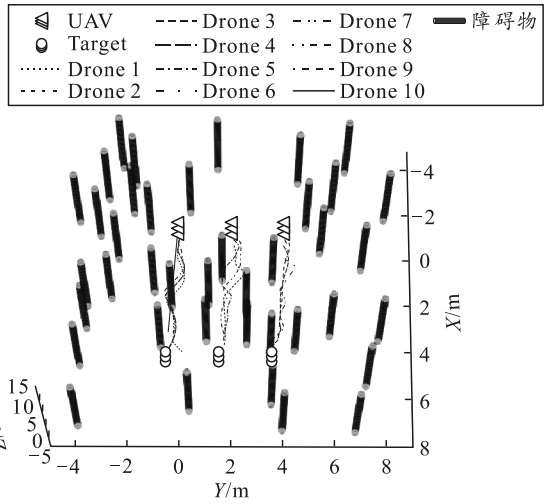

图 7 多无人机训练结果

由图 7 可知：实验结果中不同无人机基于各自观测状态选择合适动作避开障碍物，并尽量保持队形以最短的路径到达目标点。

3.3 对比分析

笔者分析本文中方法与传统人工势场算法在面对复杂问题时的性能，分别设计 50 次目标点以及障碍物随机的重复对比实验，结果如表 1 所示。

表 1 实验对比结果

算法	通过率（无碰撞）/%	平均路径代价
强化学习	96	52.25
人工势场	86	51.84

平均路径代价为 50 次重复实验下所有无人机平均移动距离。由于强化学习算法加入了队形约束，平均路径代价略高于使用人工势场法得到的路径代价。在面对复杂问题时，基于 PPO 算法的强化学习训练结果有更好的适应性，通过率相比增加 10%，同时也更好地保持编队队形，如图 8 所示。

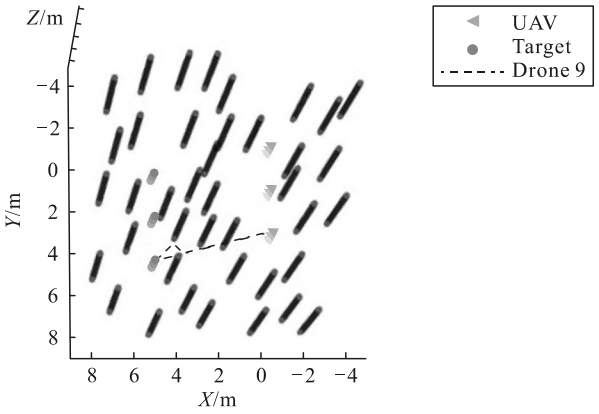

图 8 编队队形稳定性控制实验结果

4 结论

该方法设计链式训练框架，利用无人机各自局部观测状态预估动作量，在有启发式信息和奖励函数的指导下，能够在避障的前提下保持一定的编队稳定性到达目标点。实验结果表明：加入启发式信息的训练难度减小，收敛速度快且奖励不易振荡，在近似路径代价的条件下，基于 PPO 的强化学习算法相较于传统的人工势场算法具有更高的抗风险能力，同时能够保持一定的编队稳定性，具有一定可行性。如果需要更理想的结果，训练时还需要对各参数进行调整。

参考文献：

[1] ZHOU X, WEN X Y, WANG J P, et al. Swarm of micro flying robots in the wild[J]. Sci.Robot, 2022, 7(66): 5954.

[2] SUTTON R S, BARTO A G. Reinforcement learning: Anintroduction[M]. MIT press, 2018.

[3] BOUHAMED O, GHAZZAI H, BESBES H, et al. Autonomous uav navigation: A ddpg-based deep reinforcement learning approach[J]. IEEE International Symposium on Circuits and Systems (ISCAS). arXiv, 2020.

[4] CHIKHAOUI K, GHAZZAI H, MASSOUD Y. PPO-based Reinforcement Learning for UAV Navigation in Urban Environments[C]. 2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2022.

[5] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv, 2017, 17(7): 6347.

[6] 王冠政. 基于分布式深度强化学习的无人机集群感知规避[D]. 长沙: 国防科技大学, 2021.

[7] YANG Y H, XIONG X Z, YAN Y H. UAV Formation Trajectory Planning Algorithms: A Review[J]. Drones, 2023, 7(62): 62.