

语义通信下基于强化学习的无人机通信容错协同控制

章阳¹, 顾宏宇¹, 冯博豪², 王然¹

(1. 南京航空航天大学计算机科学与技术学院, 江苏 南京 211106; 2. 武汉理工大学计算机与人工智能学院, 湖北 武汉 430070)

摘要: 无人机集群技术近年在各类军民应用中得到了广泛应用。为了提高无人机执行任务的成功率, 无人机集群的通信与协同成为重要技术研究方向。然而, 在通信不确定的环境下, 无人机通信与协同可能受到主观或客观环境因素带来的通信干扰, 无法正确地收发信息, 导致协同任务失败。为解决这一问题, 针对通信受限环境下“长机-僚机”伴飞跟随的需求, 提出语义通信下基于强化学习的无人机通信容错协同方法。该方法在基于强化学习的僚机跟随策略基础上, 引入语义通信机制和基于近端策略优化算法的长机行为预测算法。在正常通信时, 僚机会接收长机通信信息并执行相应指令操作; 而在通信受限时, 僚机利用历史通信信息提取语义信息, 匹配自身的语义通信模型以推测长机的未来目标。同时结合对长机行为模式的学习预测模型, 决策僚机自身前进方向。在无额外通信抗干扰设备负载的条件下, 一定程度上能抵御通信干扰, 实现对长机的不间断跟随, 从而提高通信受限环境下任务协同效率。通过实验证明, 相较传统方法, 基于语义通信的强化学习无人机通信抗干扰方法能更好地适应复杂环境, 实现更佳的僚机跟随效果, 有效提升通信中断情况下的任务成功率, 为无人机在通信受限环境下的协同应用提供了可行的解决方案。

关键词: 无人机抗干扰; 通信干扰; 语义通信; 强化学习

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024025

Semantic communication aware reinforcement learning for communication fault-tolerant UAV collaborative control

ZHANG Yang¹, GU Hongyu¹, FENG Bohao², WANG Ran¹

1. College of Computer Science and Technology Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

2. School of Computer Science and Artificial Intelligence Wuhan University of Technology, Wuhan 430070, China

Abstract: Unmanned aerial vehicle (UAV) swarms have seen extensive deployment across a spectrum of military and civilian applications in recent years. The success of UAV missions is contingent upon robust communication and

收稿日期: 2023-10-10; 修回日期: 2024-03-13

通信作者: 王然, wangran@nuaa.edu.cn

基金项目: 国家自然科学基金项目 (No.62071343); 省部共建公共大数据国家重点实验室开放基金 (No.PBD2023-12)

Foundation Items: National Natural Science Foundation of China (No.62071343), Foundation of State Key Laboratory of Public Big Data (No.PBD2023-12)

引用格式: 章阳, 顾宏宇, 冯博豪, 等. 语义通信下基于强化学习的无人机通信容错协同控制[J]. 网络与信息安全学报, 2024, 10(02): 69-80.

Citation Format: ZHANG Y, GU H Y, FENG B H, et al. Semantic communication aware reinforcement learning for communication fault-tolerant UAV collaborative control[J]. Chinese Journal of Network and Information Security, 2024, 10(02): 69-80.

collaboration among the UAV, which has become a pivotal area of technical research. However, in environments rife with communication uncertainties, both subjective and objective environmental factors can disrupt UAV communication and collaboration. This interference can prevent UAV from accurately transmitting and receiving information, thereby jeopardizing the success of collaborative missions. To address this challenge, a fault-tolerant UAV collaboration method grounded in reinforcement learning and semantic communication was developed to cater to the leader-follower UAV mission pattern within environments constrained by limited communication capabilities. To enhance the follower UAV's strategy for reinforcement learning-based following, a semantic communication mechanism coupled with a Proximal Policy Optimization (PPO) method was implemented. This approach facilitated the prediction of the leader UAV's actions. Under normal communication conditions, the follower UAV received data transmitted by the leader and executed the corresponding command operations. In scenarios where communication was interfered, the follower UAV leveraged historical flight and communication data to extract semantic information. This information was then used autonomously to predict the future flight paths of the leading UAV. By integrating the learned and predicted behavior patterns of the leader, the follower UAV was able to make informed decisions. The proposed scheme, which did not necessitate additional anti-interference equipment, enabled the UAV swarm to counteract communication interference and bolster the efficiency of collaboration within a challenging and obstructed communication context. Experimental studies show that, when compared to benchmark methods, the proposed scheme not only endures complex environments with interferences but also significantly improves the efficiency of UAV leading-following operations and the overall mission success rate. This research provides valuable insights into viable solutions for future UAV swarm collaborations within communication-constrained and interfered environments.

Keywords: UAV anti-jamming, communication jamming, semantic communication, reinforcement learning

0 引言

近年来, 无人机凭借成本低廉、控制灵活等特点迅速发展, 各式各样的无人机层出不穷。日益发展的网络信息技术促进了无人机技术的高速发展^[1-2]。无人机在监控、灾害搜救、医疗、物流、建筑安防、遥感测绘等方面^[3]大放光彩, 在军事方面也不断展现出无可比拟的优势, 近年来的许多战争充分显示了其优势和军事地位, 无人机成为了许多国家的重要作战力量^[4-5]。但随着任务需求复杂度和多样性的上升, 尽管单个无人机可以完成一些复杂的任务, 但成功率不高, 且承受紧急情况的能力相对较弱, 因此无人机集群学习已经成为各国技术发展的重要方向。

无人机机群之间为了实现协同控制, 信息互通尤为重要, 无论是天然的干扰源还是敌方投放的恶意干扰信号, 都会对无人机间的通信协同工作造成破坏。但目前对于无人机机群协同工作的研究多数是建立在机群间完好通信的假设之上^[6], 部分研究讨论了无人机机群在通信受限环境下保持通信的应对方案^[7-9], 其目的是保证通信质量, 从而建立无人机机群间的完美通信。然

而, 稳定通信或者使用其他屏蔽干扰的方法(如加装抗干扰模块等)都会增加通信编码的复杂度和编码时长, 使无人机消耗更多的资源, 从而缩短了无人机工作时长, 而采用语义通信的方式则可以在提升无人机抗干扰能力的同时, 减少编码复杂度和编码时长。同时, 考虑通信过程中可能存在非法干扰和窃听等情况, 语义信息的引入可以大大降低通信数据的传输量和可破解性, 从而增强通信数据的安全性。除此之外, 传统的屏蔽干扰方法需要针对不同的环境进行更改, 这样会大大降低此类方法的适用性和灵活性。而采用语义通信作为抗干扰方法可以很好地适应不同的通信环境, 大大降低了再训练耗时, 能够更好地适应复杂、瞬变的战场环境。

本文以“长机-僚机”法的编队模型为基础, 针对干扰环境导致通信链路错误的情况下的僚机, 提出了语义通信下基于强化学习的无人机通信抗干扰 (semantic communication reinforcement learning, SemComRL) 方法。该方法的优势在于无须为僚机增加额外的通信抗干扰设备, 在通信受干扰的情况下, 无法正常通信的僚机将使用历史完整通信信息作为语义信息模拟通信效果, 同时

配合对长机行为模式的学习模型猜测长机动向,以进行不间断地跟随完成协同动作,在尽可能跟随长机提升任务完成率的同时减少资源开销,并且一定程度上能抵消通信中的错误对我方无人机协同控制的影响。相比传统通信协同优化方法需要对通信环境与过程进行建模分析,强化学习可以通过历史数据和运行时与环境的交互生成合适的策略,对建模分析依赖较少。并且基于强化学习的方法无须手工设计规则便可以自动寻找最优策略,适用于复杂系统的无人机机群控制任务。

1 系统与问题描述

本文中,系统任务场景为无人机集群搜捕敌方谍机并进行打击,系统以“长机-僚机”编队方式为基础问题设定。其中长机为人为操控或预设路径的主控无人机,配备高性能抗干扰设备,具备强的通信抗干扰能力,掌握敌方谍机目标信息,任务是搜捕目标,同时向僚机提供跟随打击信息。僚机为智能无人机,主要任务为跟随长机并根据指挥对目标进行打击或损毁式攻击,受限于无人机成本、尺寸以及设计载荷等,僚机并没有充裕的空间与资源用于装备通信抗干扰设备,并且任务过程中并不掌握敌方目标信息,只有在长机规定范围内收到长机的攻击命令才会发动打击。

如图1(a)所示,长机与僚机群正保持三角队列进行探测任务。在探测任务中,长机因具备强的搜索能力而保持在队列最前方,而僚机群则保持三角队形在长机后智能伴飞;当发现攻击目标时,僚机群则会快速变换队形,围绕长机以随时对长机范围内的目标进行打击,如图1(b)所示。

从上述任务描述可知,当长机僚机的通信不受干扰时,对于僚机而言,协同控制任务相对简单,只需要遵循长机的命令,自主智能跟随长机,并在长机发出攻击命令时执行打击任务。但作战环境往往并非如此理想,可能出现通信错误,即便不考虑敌方为反击而投放的通信干扰设备,自然环境中存在的天然磁场、城市中的电磁设备等都会对长机僚机之间的通信产生干扰。当僚机通信因受干扰而丢失时,常见的处理方法有4种:一是无人机停止工作,等待回收;二是无人机原路返航;三是无人机随机飞行,以期望越

过通信干扰区域;四是为无人机配备抗干扰设备。

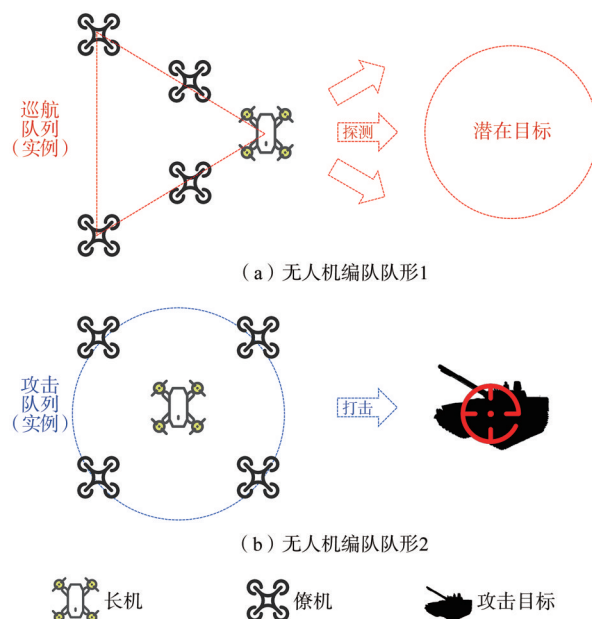


图1 长机-僚机编队
Figure 1 Leader-follower collaboration

上述方法一和二都属于放弃任务的方式,不符合系统任务要求;而随机方法在概率上存在成功的可能性,但其不稳定性决定了这并不是一种常规适用的应对方法;受限于僚机资源,增加额外的设备必将挤压原有设备空间、动力与火力储备,属于顾此失彼的方法,故僚机并不适合额外配备抗干扰设备。

基于上述论述,本文提出了语义通信下基于强化学习的无人机通信抗干扰方法。该模型并不需要增加额外的设备,只需要在僚机本身的跟随模型基础上额外维护两个模型:一是由长机的历史任务飞行数据得到的语义模型;二是僚机对长机飞行模式进行学习得到的猜测模型。所以,当长机与僚机处于正常通信的情况下,僚机会根据长机传递的信息进行跟随,并使用该信息对语义数据库进行维护更新;当僚机受电信干扰而导致间歇性通信时,僚机当前最后一次收到长机传来信息将作为语义信息匹配僚机中的语义模型,为僚机跟随长机的方向作出指导;且此时僚机将启动对长机飞行模式的猜测模型,因为僚机的任务是跟随长机,故长机与僚机飞行环境相对一致,因此通过猜测模型,可以近似得到长机在当前僚机环境下的飞行方向。僚机根据以上所得方向对

长机进行不间断地追随，提升任务的成功率。

2 系统分析建模

2.1 场景设计

本文无人机的实验场景为处于山林中的小镇，环境对象有无人机（长机和僚机）、飞行障碍物（房屋、山林、移动的车辆等）、通信干扰源（天然磁场、敌方干扰器等）以及目标。

其中，长机为人为操控的无人机，属于侦察类无人机，具备较强的通信抗干扰能力，任务是追捕目标，同时向僚机提供跟随信息。僚机为智能无人机，属于攻击类无人机，并不具备通信抗干扰能力，主要任务为跟随长机并打击目标。对于飞行障碍物，无人机无法从这些障碍物中穿过，需要绕行。通信干扰源的效用是会在其附近一定范围内对长机与僚机之间的通信产生干扰。

为了更专注于解决僚机在通信受限环境下如何高效且准确地跟随长机的问题，本文对无人机实验场景进行了网格化处理，建立了笛卡尔坐标系，如图2所示。并以数组 $\text{map}(x, y)$ 对地图网格化后划分的区域进行定义，以小镇的中央为坐标起点，正东方向延伸为 x 轴，正北方向延伸为 y 轴。



图2 网格化的无人机实验场景
Figure 2 Gridded system area for UAV simulation

在网格化的实验环境中，本文做了以下定义：对于区域 (x_i, y_i) ，根据不同的状态，有6种描述标记，Clear表示该区域可通行；Blocked表

示该区域存在飞行障碍物，须绕行；Jammed表示该区域存在通信干扰源，属于通信干扰区域，无人机在此处有可能发生通信中断；Target表示该区域为目标区域；Leader表示该区域存在长机盘旋；Follower表示该区域存在僚机盘旋，如式（1）所示。

$$\text{map}(x_i, y_i) \in \{ \text{Clear}, \text{Blocked}, \text{Jammed}, \text{Target}, \text{Leader}, \text{Follower} \} \quad (1)$$

2.2 无人机模型设计

由于三维空间里无人机的动作行为较为复杂，因此本文舍掉三维空间中的高度这一维，分析无人机在二维空间下的行为，以更专注于如何解决通信干扰问题。同理，也应尽可能地简化环境中对无人机的真实条件约束，以排除不必要的变量对本研究产生干扰。

（1）动作行为建模

由上可知，本文中无人机的动作活动空间为二维平面空间。更进一步，在此次实验中，假设无人机为定速航向，且将动作离散化定义为， $\text{action} = \{\text{hover}, \text{straight}, \text{left}, \text{right}, \text{turn-round}\}$ ，如表1所示。

表1 无人机航向行为表示
Table 1 UAVs heading behavior representation

离散动作	含义
hover	悬停
straight	直行
left	左转 90°
right	右转 90°
turn-round	旋转 180°

（2）状态观测建模

在本次实验中，无人机的状态信息主要由两部分构成：一是无人机与目标的相对位置 pos_{rel} ；二是无人机当前所在区域的环境观测，如式（2）所示。

$$\text{state} = \left(\text{pos}_{\text{rel}}, \text{obs}_{(x_i, y_i)} \right) \quad (2)$$

pos_{rel} 由无人机当前位置 (x_i, y_i) 与目标位置 (x^*, y^*) 作向量差所得，如式（3）所示，采用相对坐标体系是为了增加无人机在场景参数发生变化下的性能鲁棒性。需要注意的是， pos_{rel} 中的目标是相对而言的，对于执行搜寻任务的长机，目

标是其搜寻对象, 对于执行跟随任务的僚机, 目标则是长机。

$$\text{pos}_{\text{rel}} = (x_i - x^*, y_i - y^*) \quad (3)$$

考虑长机人为操控且仅需要执行观测和通信任务, 需要为长机增加通信质量检测设备进行全

$$\text{obs}_{(x_i, y_i)} = \begin{pmatrix} \text{map}(x_i - k, y_i - k) & \cdots & \text{map}(x_i + k, y_i - k) \\ \vdots & \ddots & \vdots \\ \text{map}(x_i - k, y_i + k) & \cdots & \text{map}(x_i + k, y_i + k) \end{pmatrix} \quad (4)$$

此外, 受限于无人机体积, 僚机机群并没有充裕的空间和资源配备检测设备, 并且在任务过程中仅需要跟随长机并进行打击即可, 所以僚机机群的观测数组以长机发送信息中的环境观测数据为准。

(3) 回报函数建模

对长机的目标搜寻动作而言, 这是一个稀疏回报问题^[10], 但这并不在本文的讨论范围内。对于僚机的跟随动作和对长机的猜测动作, 一个好的回报函数是至关重要的。对于僚机的跟随动作, 其回报函数与长机当前的位置信息以及僚机本身当下的动作有关。当僚机紧随长机且保持安全距离时, 得到的是正向回报奖励 r ; 而当僚机偏离长机时, 将会得到负向回报惩罚 p 。当长机搜寻到目标且此刻僚机位于长机附近时 (默认僚机只要靠近目标就会击中目标), 任务成功, 将会得到更大的正向回报奖励 R ; 相应的, 当僚机发生碰撞或者出界时, 任务失败, 会得到更大的负向回报惩罚 P 。僚机跟随策略回报函数见表2所示。

表2 僚机跟随策略回报函数
Table 2 Reward function of the follower

状态	回报值
紧随长机	r
偏离长机	p
任务成功	R
任务失败	P

2.3 语义通信建模

语义通信的概念由 Weaver^[11]提出, Weaver将沟通分为3个层次: 符号的传递、语义交换和语义交换的效果。其中, 第一层次通信, 即符号

周观测, 即无人机360°全方位观测。为了便于传输后续观测数组, 设长机处于区域 (x, y_i) 并且设其观测范围为 $\sqrt{2}k$, 同时舍去部分位于观测范围边缘的区域, 可得到如式(4)所示的观测数组 $\text{obs}_{(x_i, y_i)}$

的传递, 已被深入研究, 并在传统通信系统中交付使用, 其容量已接近香农容量极限。然而, 在很多情况下, 通信的最终目标是交换语义信息, 符号的传递强调的是传输符号的准确性, 语义交换强调的是传输符号是否准确地表达预期含义^[12]。

与香农模式相比, 语义通信只在接收端上传输与任务相关的必要信息, 从而实现了数据流量显著减少的真正智能系统。图3将香农传统通信系统与语义通信系统进行了对比, 展示了语义通信的概念, 其中传输任务是图像识别。语义发送器不是传输代表整个图像的比特序列, 而是提取与识别物体相关的特征, 也就是“猫”, 即在通信的源头上将图像背景等无关信息省略, 在不降低系统性能的前提下, 最小化传输数据。因此, 对能量和无线资源的需求将显著降低, 从而形成更可持续的通信网络。

在传统的通信系统中, 由于传输的符号 (比特) 受到信道干扰而失真, 因此接收到的消息是混乱的。然而, 在语义通信系统中, 即使存在句法不匹配的情况, 也能准确传达语义。此外, 在发生信道干扰情况下, 即使通信发生中断, 语义通信系统可能仍然表现良好, 并且可能消耗更少的能量^[13]。

由以上状态观测建模信息可知, 长机向僚机传递的是由位置信息和观测信息组成的数据信息。假设在一次完整的任务过程中, 长机在通信干扰区域向僚机传递的信息为 n 条, 而通信干扰导致的通信丢失率为 l , 则僚机在受干扰情况下得到的有效信息期望值为 $n(1-l)$ 条。在整体通信的宏观角度下, 此次通信并未丢失, 而是信息

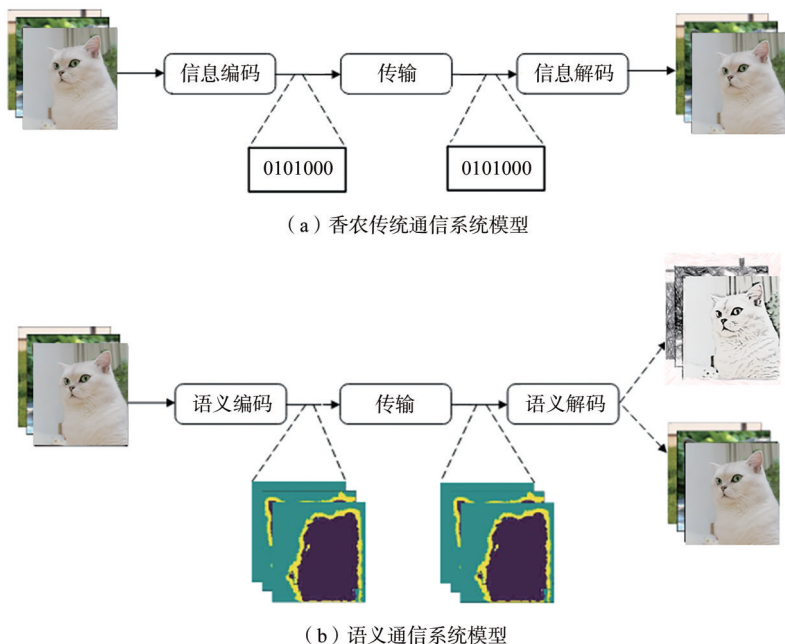


图3 香农传统通信系统模型与语义通信系统模型
Figure 3 Shannon's conventional communication system model and semantic communication system model

受损失真，只得到了部分有效信息。从传统通信层面分析，通信内容受到干扰不完整，是混乱的，所以此次通信是失败的。但在语义通信层面分析，若此次得到的部分可辨识信息能够指导僚机正确跟随长机的方向，即使信息是混乱不完整的，但仍然能够准确传达语义，则此次通信是成功的。

例如，在一次指挥过程中，长机给僚机传递了 5 条命令 I_1, I_2, I_3, I_4, I_5 ，由于受到了通信干扰，僚机只正确接收到了 3 条命令 I_5, I_1, I_2 ，并且命令次序发生了改变。从传统通信角度分析，此次通信传递失败。而在语义系统中，如果使用语义模型将这 3 条命令与所带语义数据库进行信息匹配补全得到了完整的命令序列 I_1, I_2, I_3, I_4, I_5 ，则僚机接收到的失真命令序列 I_5, I_1, I_2 仍能有效指导僚机执行任务。因此，从语义通信角度分析可知，此次通信的目的并没有受到通信干扰的影响，是有效成功的。

值得注意的是，采用语义通信的方式，主要适用于通信被干扰时产生误码导致信息错误等情况。通常情况下，干扰和攻击导致的通信错误包含信息错误和信息被更改等情况。对于信息被更改这种情况，需要对传输信息进行抓取并对其进行伪造更改后重新发送，涉及信息的理解与解密，且相对于直接传输会产生一定的时延，因此

僚机可以通过计时器或实时比较长机的动向以及信息之间的区别来判断和过滤信息更改的情况。而传输过程中的信息错误是一个更现实与常见的问题。传统通信方法中，需要添加纠错码或重复发送以减轻影响，这一过程无形中增大了信息的传输负担，可能会造成更加严重的信息堵塞及丢失。而基于语义通信的方法则可以分析内容内部与外部的相关特性，通过推理来还原出原数据，不仅能够降低传输信息量与时延，还能够抽取通信信息核心内容实现容错通信。由此，本文主要研究由干扰引起的传输错误情况下，僚机如何使用语义通信模型以及学习模型预测长机动向，在一定程度上抵消通信干扰带来的影响，实现容错通信与集群协同。

3 算法设计与实现

3.1 长机飞行算法

因本文主要讨论的是长机与僚机发生通信中断后，僚机如何自主追随长机的问题，故长机如何进行目标搜寻暂不在本次实验设计范围内，所以在此次实验环境下，长机飞行策略并不使用智能策略模型，而是使用附带随机因子的广度优先搜索算法寻找目标。

使用广度优先搜索算法的原因是在本次实验

环境下,长机的任务期望是快速定位目标区域,从任务结果角度看,其任务实质是找到航点与目标区域间的最短路径,而广度优先搜索算法就是求解两点之间最短路径问题的经典思想,在假设目标一定存在的情况下,广度优先搜索算法一定能保证找到一条通向它的最佳航路。而在广度优先搜索算法中加入随机因子是为了模拟真人操控的效果,在获得最佳路径的同时尽可能地符合人工操控行为。附带随机因子的广度优先搜索算法的长机操控模拟算法伪码如算法1所示。

算法1 长机操控模拟

```

1. 初始化探索队列L,初始化已发现标记数组 Visited
2. 初始区域加入探索队列L,且置已发现标记
3. while 探索队列非空 do:
4.   探索队列L队头指示区域 $V_0$ 出队
5.   探索区域 $V_0$ 
6.   if 符合目标区域:
7.     执行任务
8.     任务结束,结束 while
9.   else:
10.    进入随机探索步骤
11.    使用随机因子任意发现 $V_0$ 周边未探索区域 $V_1$ 
12.    if  $V_1$ 存在:
13.      将 $V_1$ 加入到探索队列L,并置已发现标记
14.      重复随机探索步骤
15.    else:
16.      continue
17.  执行失败,不存在目标区域
18. end while

```

3.2 僚机运作算法

由系统问题可知,僚机会面对两种情况:一是当长机与僚机通信正常时,僚机需要根据长机传递过来的信息进行跟随或打击目标;二是当通信受阻时,僚机使用断开通信前一次的通信数据作为语义信息匹配本机维护的语义通信模型,以推测长机的前进目标,且配合对长机行为模式的行动预测模型来得到当前的前进方向。其算法伪码如算法2所示。

算法2 基于强化学习方法的僚机跟随

```

1. while 僚机未完成任务/消亡:
2.   if 通信正常:

```

```

3.     接收长机传递信息
4.     if 跟随指令:
5.       解析获取长机当前位置数据作为目标信息
6.       使用跟随模型对目标信息与僚机观察组成状态数据进行求解
7.       得到跟随模型行动方向输出,执行
8.     else if 打击指令:
9.       解析获取打击目标信息
10.      确认打击命令
11.      执行打击命令
12.    else if 通信间断:
13.      从缓存中获得正常通信下长机向僚机传递的最后一条信息数据
14.      将该数据作为语义信息输入至语义通信模型中,得到长机预测目标信息
15.      使用长机行动预测模型对目标信息与僚机观察组成状态数据进行求解
16.      得到长机行动预测模型行动方向输出,执行
17.    end while

```

3.3 长机行动预测算法

由算法1可知,对于僚机来说,正常通信和间断通信下的跟随动作输出模型的输入形式是一致的,都是<目标信息,环境观测>,故僚机正常通信下的跟随模型和通信间断时的长机行动预测模型都使用近端策略优化(proximal policy optimization, PPO)^[14]算法实现,以减少切换异构模型而产生的额外资源开销。

PPO算法是一种简单稳定、性能强大且易于实现的Actor-Critic框架算法^[15]。OpenAI DOTA2智能体OpenAI Five^[16]和腾讯旗下的游戏智能体JueWu^[17]都采用PPO算法实现。如算法3所示,作为Actor-Critic框架算法,PPO算法有两个初始输入参数,一个是用于Actor的初始化策略参数 θ_0 ,另一个是用于Critic的初始化价值函数参数 ϕ_0 。对于第 k 个episode,即在第 k 次任务中,使用当前的动作运行策略 π_k 即Actor做出无人机行为决策,并且搜集该决策下飞行得到的运动轨迹 D_k , D_k 是由状态 s_t 与动作 a_t 组成的序列 $\{s_1, a_1, \dots, s_T, a_T\}$,在此次任务完成后,无论是成功还是失败,都会进行奖励计算,得到此次任务的行动奖励 $\hat{R}_t$;然后使用 $\hat{R}_t$ 基于当前值函数 V_{ϕ_t} ,即Critic计算优势估计值 $A^{\pi_0}(s_t, a_t)$,优

势估计值 $A^{\pi_{\theta_k}}(s_t, a_t)$ 是此次运行策略参数 θ_k 的性能评价, 通过使用 PPO-截断 (PPO-Clip) 将目标函数进行限制, 以保证新参数与旧参数差距不会太大; 接着通过此次性能评价更新得到下一个 episode 的策略参数 θ_{k+1} 和价值函数参数 ϕ_{k+1} , 可以看到, 对于策略参数 θ_{k+1} 使用的是随机梯度下降的方法进行优化, 对于价值函数参数 ϕ_{k+1} 则是使用平方损失函数, 即实际回报与期望回报的平方的方法进行更新。

算法 3 近端策略优化

1. 初始化策略参数 θ_0 , 初始化价值函数参数 ϕ_0
2. **for** $k = 0, 1, 2, \dots$, **do**:
3. 在环境中运行策略 $\pi_k = \pi(\theta_k)$ 收集动作轨迹 $D_k = \{\tau_i\}$
4. 计算行动奖励 $\hat{R}_t$
5. 基于当前值函数 V_{ϕ_k} 计算优势估计值 $A^{\pi_{\theta_k}}(s_t, a_t)$
6. 通过随机梯度下降方法更新策略:

$$\theta_{k+1} = \arg \max_{\theta} \frac{1}{|D_k|} \sum_{t=0}^T \min \left(\frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_k}(a_t | s_t)} A^{\pi_{\theta_k}}(s_t, a_t), g(\varepsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right)$$

$$g(\varepsilon, A^{\pi_{\theta_k}}(s_t, a_t)) =$$

$$\max \left(\min \left(A^{\pi_{\theta_k}}(s_t, a_t), 1 + \varepsilon \right) 1 - \varepsilon \right)$$

7. 通过均方误差回归拟合值函数:

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|D_k|} \sum_{\tau \in D_k} \sum_{t=0}^T (V_{\phi}(s_t) - \hat{R}_t)^2$$

8. **end for**

4 仿真实验

本次仿真实验主要是为了验证在长机与僚机发生通信干扰时, 僚机在使用 SemComRL 方法后是否能进行有效沟通且紧随长机, 并是否优于现有常用的处理方法。故在实验中, 设定一个长僚机队进行任务测试, 即一部长机带领一部僚机, 在范围大小为 40×40 忽略高度维数的平面地图中执行任务。在对实验场景进行网格化处理, 进一步对实验场景进行了数据化处理, 得到了如图 4 所示的实验环境, 并对上述式 (1) 中的区域状态标注进行了颜色差异区分, 白色表示 Clear 状态, 黑色表示 Blocked 状态, 灰色表示

Jammed 状态, 红色表示 Target 状态, 蓝色表示 Leader 状态, 绿色表示 Follower 状态。

由图 4 可知, 通信干扰区域分为两部分: 一是地图中部 ($0 \leq x < 40, 10 \leq y < 29$) 的固定干扰区域, 是通往目标所在地的必经之路, 确保智能体不会通过迂回躲避干扰而限于局部最优解; 二是随机分布于地图上的干扰块, 其目的是训练智能体在正常通信区域与干扰区域之间切换的适应能力。长僚机在两部分通信干扰区域中进行通信都会出现一定概率的通信丢失, 称该概率为干扰概率 α , 僚机接收的命令数量如式 (5) 所示。

$$\text{len}_{\text{receive}} = \begin{cases} (1 - \alpha) \text{len}_{\text{send}}, & p < \alpha \\ \text{len}_{\text{send}}, & \text{其他} \end{cases} \quad (5)$$

其中, p 是 0~1 均匀分布的, len_{send} 为长机发送命令的数量。

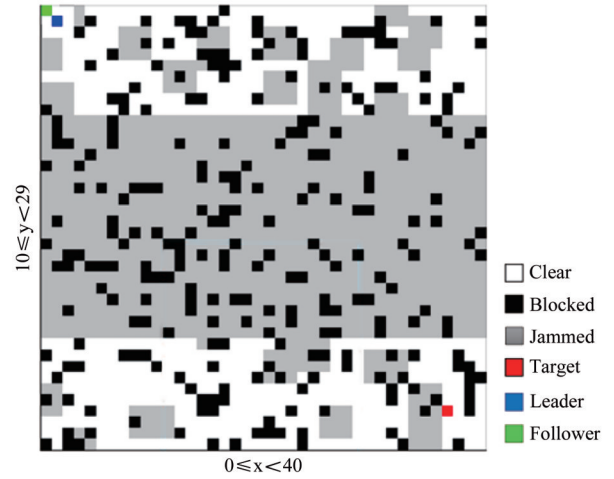


图 4 网格化的实验环境
Figure 4 Gridded experimental environment

4.1 模拟环境与轨迹演示

本次实验的评价标准为: 在发生通信中断情况下长机到达任务区域时, 僚机到达长机区域的最短路径距离。因地图中存在障碍物, 两者间的绝对距离并不能表达僚机到达长机区域的真实距离, 而最短路径距离能更恰当地表示两者之间的实际距离。

本实验在通信干扰区域干扰概率为 0.5 的情况下验证了 SemComRL 方法的有效性, 并与预定的 4 种基准方法进行了对比, 在通信中断情况下僚机执行的 4 种基准方法如下。

- (1) HOVER: 悬停待命, 当发生通信中断时,

僚机悬停以待通信恢复后再继续前行进行任务。

(2) STATS: 统计, 当发生通信中断时, 使用对长机行为选择的统计模型选择前进方向以继续进行任务。

(3) RANDOM: 随机, 当发生通信中断时, 僚机随机选择前进方向以继续进行任务。

(4) KEEP: 保持航向, 当发生通信中断时, 不改变原有前进方向, 继续按原有方向前进以继续进行任务。

使用 HOVER 方法是为了验证使用了 SemComRL 方法是否会导致比不动作更差的结果。STATS 方法是采用传统的统计方法对长机的行为模式进行建模, 统计长机在不同区域中采取的前进方向, 受到通信干扰时, 僚机使用统计模型推测长机动向以进行跟随, 此方法是为了验证使用近端策略优化算法所得预测模型的先进性。RANDOM 方法是在通信中断期间随机选择前进方向, KEEP 方法则是当丢失通信时, 僚机不对方向进行选择, 按照惯性, 继续按原有方向前进。通过与 RANDOM、KEEP 两种方法的对比, 可以验证使用语义通信方法配合近端策略优化算法所得预测模型的优越性。

本文选取其中一次实验并且对这 5 种方法的

飞行轨迹进行了跟踪, 以深蓝色轨迹代表长机飞行路径, 浅蓝色、红色、黄色、绿色、紫色轨迹分别代表在通信中断情况下僚机使用 SemComRL 方法、HOVER 方法、STATS 方法、RANDOM 方法和 KEEP 方法的飞行路径, 其飞行轨迹如图 5 所示。

图 5 中, 图 5 (a) 是所有方法飞行轨迹的叠加; 图 5 (b) 是僚机在干扰区域丢失通信概率为 0.5 的情况下使用 SemComRL 方法执行任务的飞行轨迹; 图 5 (c) ~ 5 (f) 分别是僚机在干扰区域通信丢失概率为 0.5 的情况下使用 HOVER 方法、STATS 方法、RANDOM 方法以及 KEEP 方法执行任务的飞行轨迹。

根据实际的飞行轨迹进行了表格化处理, 如表 3 所示。通过统计长机总移动次数 N 与僚机准确预测次数 C , 计算 5 种方法的偏航率, 如式 (6) 所示。其中偏航率越小, 表示该方法的性能越好。

$$\lambda = 1 - \frac{C}{N} \quad (6)$$

在本次任务中, 长机要到达的任务区域坐标为 (36, 36), 从图 5 (a) 可知, 位于长机初始位置的东南方向, 可见长机飞行具有东南向的趋势。在未进入中部密集干扰区域前, 长机经过了 3 个零

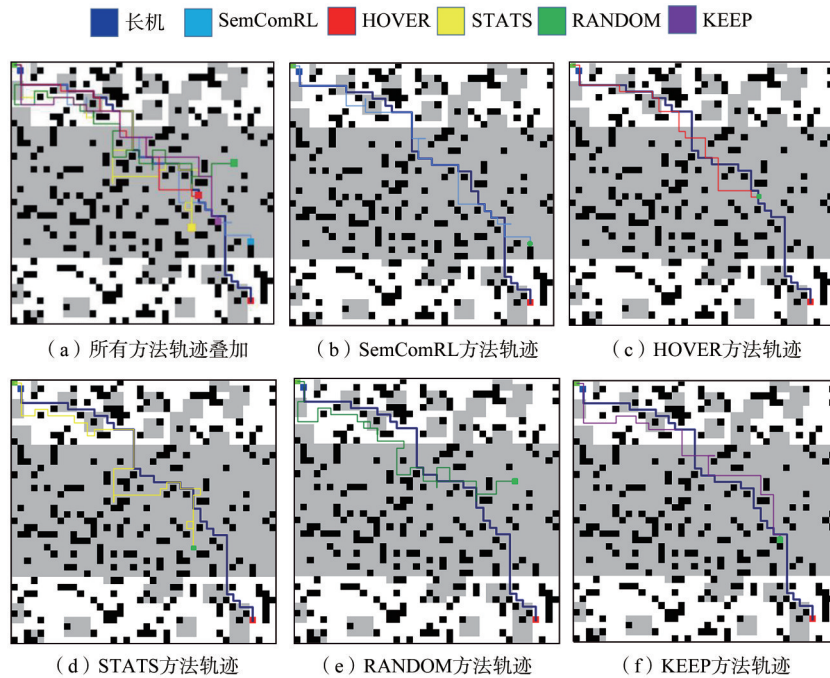


图5 丢失概率为0.5时的飞行轨迹
Figure 5 UAV trajectories with probability of loss of 0.5

散的通信干扰块,在这一部分路程中,可见STATS方法、RANDOM方法以及KEEP方法都受到了一定程度的干扰,偏离了长机的轨迹;HOVER方法由于此段路程长机航向变化较少,看似能较好地贴合长机飞行路径,但从后面的飞行轨迹可知,由于遇到通信中断时悬停待命,这里贴合的只是长机的历史飞行轨迹,而不是实时紧密跟随的轨迹;对比图5(b)可知SemComRL方法在此区域下能实时贴合长机飞行路径,即紧密跟随长机。

表3 丢失概率为0.5时各方法的偏航率
Table 3 Yaw rate of different methods with interference probability 0.5

方法	偏航率
SemComRL	45.71%
HOVER	65.71%
STATS	68.57%
RANDOM	74.29%
KEEP	84.29%

在中部干扰密集区域,在进入长时间的通信中断情况下,各方法的差异愈发明显。从图5(b)可知对于SemComRL方法,除了在少数几个地方出现小幅度偏航外,基本上能贴合长机飞行路线,但同样受信息交流间断的影响,当长机到达目标区域(36,36)时,僚机到达区域(36,27),与长机的最短路径距离为11。图5(c)的HOVER方法,从轨迹路线看似僚机能较好地跟随长机,但从僚机的最终区域可见,僚机与长机发生通信中断时,两者的偏航程度累积越来越大,导致当长机到达目标区域时,僚机还处于干扰密集区域的中部。从图5(d)可知,STATS方法在此区域发生了较大的反向偏航,并且出现原地打转的现象,这是在统计模型训练阶段的地图参数与测试地图参数不同,导致统计模型失效。从图5(e)可见,对于RANDOM方法,僚机飞行出现了较多的迂回轨迹,这是因为在通信中断情况下同一方向的随机选择概率会呈现指数式下降,并且由于随机性,还会出现在两区域间徘徊的现象。对于KEEP方法,从图5(f)可知,在通信干扰概率为0.5的情况下,使用该方法似乎也能得到较为贴合长机飞行轨迹的轨迹路线,但该方法有两个缺陷:一是当保持的航向前方出现

障碍物时会退化成HOVER方法;二是如果出现了连续长时间的通信中断并且前进方向不存在障碍物时,容易严重偏离航线。

4.2 算法性能检验

除此之外,为了检验在不同通信丢失概率下5种方法的性能,分别对0~1间隔为0.1的11个干扰概率进行了测试,以验证SemComRL在不同情况下均有良好的性能。

为了去除误差,在同一测试环境下随机选取10个目标,各目标执行20次,即对0~1,间隔为0.1的11个干扰概率分别进行了200个episodes的测试。

以干扰概率为横轴,取相同间隔0.1,以长机到达任务区域时,僚机到达长机区域的最短路径距离为纵轴,建立性能评价曲线图,纵轴往下增大是为了更直观地表示僚机与长机发生通信干扰而导致性能下降的情况,如图6所示。

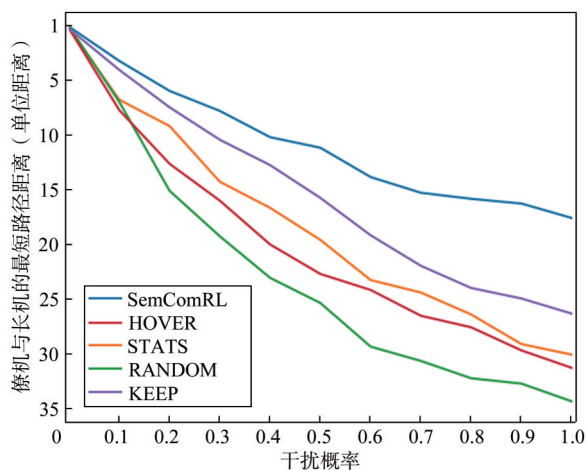


图6 干扰概率为0~1情况下长机僚机的最终距离
Figure 6 The final distance between the leader and the follower given the interference probability 0~1

由图6可知,在横坐标为0,即不存在干扰的情况下,当长机到达目标区域时,图中5种方法下所对应的纵坐标都为1,即僚机到达长机目标区域的最短路径距离为1,故在无干扰情况下,僚机的跟随任务是成功紧密跟随的,进一步证明了该实验是真实有效的。

随着干扰概率的不断提升,从图6可以看到,当长机到达目标区域时,图中5种方法所对应的纵坐标也在变大,即僚机到达长机区域的期望最短路径距离也随之变大,证明通信干扰确实会对僚机的性能造成影响,并且干扰概率越大,

影响越大。但即使如此, 本文提出的 SemComRL 方法在衡量标准上也明显优于其他 4 种基准方法。可以看到 RANDOM 方法在实验中性能表现最差, 这是实验地图中多有障碍物且构成死路, 随机方法会导致僚机进入闭合路线或者往回走, 甚至原地转圈。STATS 方法性能上略优于 HOVER 方法, 证明长机行为有一定的规律性, 但并不容易统计得出。KEEP 方法性能表现差于 SemComRL 方法, 却优于其他 3 种基准方法, 这说明长机行动方向具有趋向性。

此外, 对实验结果数据进行了表格化处理, 对比了不同干扰概率下当长机到达目标区域时, 采用不同方法后僚机到达长机区域的期望最短路径距离, 即干扰概率为 0~1 下各方法的僚机跟随性能表现, 如表 4 所示。其中最短路径距离对应的值越大, 表示僚机与长机离得越远。每种干扰概率下性能最优的方法均用加粗表示。

丢失通信时, 无人机相对保守的动作是悬停等待, 故以 HOVER 所得结果数据为性能基线, 对比了 5 种方法的僚机跟随性能, 如表 5 所示。通过对比可得 SemComRL 方法在这 11 个干扰概率值下, 相对 HOVER 方法、STATS 方法、RANDOM 方法和 KEEP 方法分别最高达到了 50.69%、

41.62%、66.81%、28.39% 的性能提升。从实验结果可知, 当通信性能越差, 干扰概率越高的情况下, SemComRL 方法越能体现其性能优越性。

5 结束语

为解决在通信不确定环境下无人机机群通信受干扰而造成任务成功率下降的问题, 本文以“长机—僚机”法的编队模型为基础, 针对通信受限环境下的僚机提出了语义通信下基于强化学习的无人机通信抗干扰方法。并通过仿真实验结果验证了其性能优越性。主要贡献有: ①在无人机机群控制中引入了语义通信概念, 利用历史通信数据缓存为发生通信中断的无人机提供了有效方向指导; ②利用强化学习方法近端策略优化算法对人为操控的无人长机行为模式进行分析建模, 得到有效的长机行为模式预测模型; ③将语义通信概念与行为预测模型结合, 使得在发生通信干扰概率为 1 的情况下, 僚机依然能有效跟随长机执行任务。

本文的研究提升了无人机机群在通信不确定环境下发生通信中断时继续执行作战任务的成功率, 但仍需累积大量实战数据以驱动完善策略模型, 下一步的工作将针对在大规模无人机机群对

表4 干扰概率 0~1 下各方法的僚机跟随性能表现
Table 4 Performance of follower tracking for different methods with interference probability 0~1

方法	干扰概率										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
HOVER	1.39	8.7	13.51	16.81	20.73	23.37	24.81	27.12	28.15	30.22	31.78
SemComRL	1.35	4.29	6.98	8.77	11.12	12.07	14.69	16.1	16.64	17.07	18.36
SATAS	1.37	7.77	10.13	15.11	17.45	20.32	23.92	25.04	27	29.65	30.6
RANDOM	1.33	7.96	15.93	20	23.72	25.97	29.88	31.16	32.72	33.19	34.78
KEEP	1.14	5.1	8.45	11.35	13.63	16.56	19.91	22.64	24.63	25.57	26.93

表5 以 HOVER 方法为基线各方法的僚机跟随性能对比
Table 5 Performance comparison of follower tracking with HOVER method as a baseline

方法	干扰概率										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
HOVER	1	1	1	1	1	1	1	1	1	1	1
SemComRL	102.88%	150.69%	148.33%	147.83%	146.36%	148.35%	140.79%	140.63%	140.89%	143.51%	142.23%
SATAS	101.44%	110.69%	125.02%	110.11%	115.82%	113.05%	103.59%	107.67%	104.09%	101.89%	103.71%
RANDOM	104.32%	108.51%	82.09%	81.02%	85.58%	88.87%	79.56%	85.10%	83.77%	90.17%	90.56%
KEEP	98.56%	41.38%	137.45%	132.48%	134.25%	129.14%	119.75%	116.52%	112.50%	115.39%	115.26%

抗环境下发生通信中断时无人机机群如何仍能有效执行任务进行研究。

参考文献:

- [1] MOHSAN S A H, KHAN M A, NOOR F, et al. Towards the unmanned aerial vehicles (UAVs): a comprehensive review[J]. Drones, 2022, 6(6): 147.
- [2] 孙晨, 莫国美, 舒坚. 基于强化学习的无人机自组网路由研究综述[J]. 计算机应用研究, 2023, 40(7): 1937-1946.
SUN C, MO G M, SHU J. Review of research on flying Ad hoc network routing based on reinforcement learning[J]. Application Research of Computers, 2023, 40(7): 1937-1946.
- [3] HU P, ZHANG R R, YANG J X, et al. Development status and key technologies of plant protection UAVs in China: a review[J]. Drones, 2022, 6(11): 354.
- [4] 王昱, 任田君, 范子琳. 基于引导 Minimax-DDQN 的无人机空战机动决策[J]. 计算机应用, 2023, 43(8): 2636-2643.
WANG Y, REN T J, FAN Z L. Air combat maneuver decision of unmanned aerial vehicle based on guided Minimax-DDQN[J]. Journal of Computer Applications, 2023, 43(8): 2636-2643.
- [5] CHAMOLA V, KOTESH P, AGARWAL A, et al. A comprehensive review of unmanned aerial vehicle attacks and neutralization techniques[J]. Ad Hoc Networks, 2021(111): 102324.
- [6] ZHU X M, WANG L L, LI Y M, et al. Path planning of multi-UAVs based on deep Q-network for energy-efficient data collection in UAVs-assisted IoT[J]. Vehicular Communications, 2022(36): 100491.
- [7] 王志宏, 冷楚鹏, 熊凯. 面向无人机集群协同感知的多智能体资源分配策略[J]. 物联网学报, 2023, 7(1): 18-26.
WANG Z H, LENG S P, XIONG K. Multi-agent resource allocation strategy for UAV swarm-based cooperative sensing[J]. Chinese Journal on Internet of Things, 2023, 7(1): 18-26.
- [8] 金俊宇, 张婷婷. 通信带宽有限条件下无人机集群自主行为决策研究[J]. 指挥控制与仿真, 2022, 44(6): 7-15.
JIN J Y, ZHANG T T. Research on autonomous behavior decision of UAV cluster with limited communication bandwidth[J]. Command Control & Simulation, 2022, 44(6): 7-15.
- [9] CHEN J W, DING R J, WU W, et al. Multi-agent learning based packet routing in multi-hop UAV relay network[C]//Proceedings of the ICC 2022 - IEEE International Conference on Communications. Piscataway: IEEE Press, 2022: 1-6.
- [10] FU Q X, QIU T H, PU Z Q, et al. A cooperation graph approach for multiagent sparse reward reinforcement learning[C]//Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2022: 1-8.
- [11] WEAVER W. Recent contributions to the mathematical theory of communication[J]. ETC: A Review of General Semantics, 2017, 74(1/2): 136-157.
- [12] XIE H Q, QIN Z J, LI G Y, et al. Deep learning enabled semantic communication systems[J]. IEEE Transactions on Signal Processing, 2021(69): 2663-2675.
- [13] LUO X W, CHEN H H, GUO Q. Semantic communications:

overview, open issues, and future research directions[J]. IEEE Wireless Communications, 2022, 29(1): 210-219.

- [14] GU Y, CHENG Y H, YU K, et al. Anti-martingale proximal policy optimization[J]. IEEE Transactions on Cybernetics, 2023, 53(10): 6421-6432.
- [15] ZHANG J W, ZHANG Z H, HAN S, et al. Proximal policy optimization via enhanced exploration efficiency[J]. Information Sciences: An International Journal, 2022, 609(C): 750-765.
- [16] SHARMA K, SINGH B, HERMAN E, et al. Maximum information measure policies in reinforcement learning with deep energy-based model[C]//Proceedings of the 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). Piscataway: IEEE Press, 2021: 19-24.
- [17] YE D H, CHEN G B, ZHANG W, et al. Towards playing full MOBA games with deep reinforcement learning[C]//Proceedings of the Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 621-632.

[作者简介]



章阳 (1986-), 男, 湖北孝感人, 南京航空航天大学副教授, 主要研究方向为无人系统与边缘智能。



顾宏宇 (2000-), 男, 江苏苏州人, 南京航空航天大学硕士生, 主要研究方向为无人系统与边缘计算。



冯博豪 (1994-), 男, 广东江门人, 武汉理工大学硕士生, 主要研究方向为无人系统的协同与优化和强化学习。



王然 (1989-), 男, 黑龙江哈尔滨人, 南京航空航天大学副教授, 主要研究方向为计算机网络。