基于元平均场Q学习的大规模无人机组网抗干扰与隐蔽通信机制

钱玉文[1]，吴浩阳[1]，时龙[1]，曹阳[1]，王喆[2]，陈光霁[1]，马川[3]，韦康[4]

（1.南京理工大学电子工程与光电技术学院，江苏 南京 210094；2.南京理工大学计算机科学与工程学院，江苏 南京210094；3.重庆大学计算机学院，重庆 400044；4.东南大学网络空间安全学院，江苏 南京 210096）

摘 要：在恶意干扰和被动监听者并存的动态复杂电磁环境中，大规模无人机组网系统对于抗干扰能力与通信隐蔽性的需求日益凸显。为此，本文提出一种基于元强化学习的多无人机抗干扰隐蔽通信机制，通过融合实时频谱感知与历史频谱信息构建统一观测空间，并在此基础上对分布式信道选择策略进行跨场景泛化学习，以增强大规模无人机组网的抗干扰与隐蔽通信性能。具体而言，所提出机制首先为在多无人机抗干扰隐蔽通信系统设计频谱感知的时隙框架，以支撑无人机组网在干扰与监听环境下的通信；其次，设计一种平均场Q学习算法，将每个无人机用户对与所有邻居的多边交互等效为其与虚拟用户对之间的博弈，从而实现大规模无人机场景下的分布式信道接入策略学习；最后，进一步在平均场Q学习算法中嵌入元强化学习模块，构建基于元强化学习的平均场Q学习算法，使信道接入策略在不同干扰与监听场景下具备良好泛化性，从而实现抗干扰性能与隐蔽性的协同优化。仿真结果表明，在大规模无人机通信场景下，相较于传统算法，所提算法的归一化奖励值平均提升15.41%，收敛速度平均提升44.56%。

关键词：大规模无人机；频谱感知；抗干扰通信；元强化学习；隐蔽通信

中图分类号：TN929.5

文献标志码：A

Meta mean-field Q-learning based anti-jamming and covert communication mechanisms for large-scale UAV networks

Qian Yuwen[1], Wu Haoyang[1], Shi Long[1], Cao Yang[1], Wang Zhe[2], Chen Guangji[1], Ma Chuan[3], Wei Kang[4]

1. School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
3. College of Computer Science, Chongqing University, Chongqing 400044, China
4. School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China

Abstract: In dynamic and complex electromagnetic environments where malicious jammers and passive eavesdroppers coexist, large-scale multi-UAV networks impose increasing demand on anti-jamming capability and communication covertness. In this context, we propose a meta-reinforcement-learning based anti-jamming covert communication mechanism for large-scale multi-UAV networks. In the proposed mechanism, we fuse real-time spectrum-sensing results with historical spectrum information to construct a unified observation space, through which the UAV users learn distributed channel-selection policies with cross-scenario generalization to enhance the anti-jamming and covert communication performance. Specifically, we first develop a time-slotted spectrum-sensing framework in a multi-UAV anti-jamming covert

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX

通信作者：时龙，slong1007@gmail.com

基金项目：国家自然科学基金青年项目（No. 6250010332）；国家自然科学基金青年项目（No. 62202232）；国家自然科学基金青年项目（No. 62501279）；海南省重点研发计划（No. ZDYF2024GXJS292）

Foundation Items：The National Natural Science Foundation of China Youth Project (No. 6250010332), the National Natural Science Foundation of China Youth Project (No. 62202232), the National Natural Science Foundation of China Youth Project (No. 62501279), the Hainan Province Science and Technology Special Fund (No. ZDYF2024GXJS292)

communication network to support the communication process of UAV networks under jamming and eavesdropping. Second, we design a mean-field Q-learning (MFQ) algorithm from a mean-field game perspective to achieve distributed channel-access policy learning in large-scale scenarios, wherein the interactions between each UAV user and all neighbors are approximated by a game between the UAV user and a virtual user representing the average behavior of its neighbors. Finally, we propose a Meta-MFQ (MMFQ) algorithm to jointly optimize anti-jamming performance and covertness, wherein meta-reinforcement learning is embedded into MFQ to enhance generalization performance of the channel-access policy to different jamming and eavesdropping scenarios. Simulation results demonstrate that, compared with baselines, the proposed algorithm achieves average improvements of 15.41% in normalized reward and 44.56% in convergence speed over large-scale UAV communication networks.

Keywords: large-scale unmanned aerial vehicles, spectrum sensing, anti-jamming communication, meta reinforcement learning, covert communication

0 引言

无人机（Unmanned Aerial Vehicle，UAV）作为移动服务载体，灵活性高、覆盖能力强且成本低廉，其在通信领域的应用价值正不断凸显，同时也导致通信频谱资源稀缺[1]，因此高效的频谱管理具有重要的实用价值。而频谱感知作为认知无线电系统中的重要技术，可在动态频谱策略选择过程中提升无人机组网的频谱利用率，有效节省频谱资源[2]。

在无人机通信系统中，数据通过开放无线信道进行传输，使其在物理层容易遭受恶意干扰。在大规模无人机组网场景下，由于可用频谱资源稀缺，系统存在频谱竞争与信道拥塞的问题。如何在保证通信有效性与传输可靠性的前提下设计高性能的抗干扰通信机制，已成为该领域的重要研究方向。为解决上述问题，现有研究采用以下技术手段，例如：无人机组网可为缺乏地面通信基础设施的区域提供稳定的通信服务[3]，此外，该组网能够根据业务需求动态调整传输参数，从而形成可扩展的空中组网通信体系[4]。在海洋等地面基站覆盖薄弱的远距离场景，文献[5]利用无人机组网辅助海事通信；文献[6]为最小化无人机数据传输链路的功率消耗，设计基于相关向量回归的功率控制与信道选择算法；在多无人机协作抗干扰领域，文献[7]通过合理分配发射功率，能够提升频谱资源受限时无人机组网的干扰容忍能力；此外，文献[8]通过联合优化无人机与干扰机之间的功率发射策略，能够最大化通信系统的保密容量；基于上述研究，文献[9]进一步从移动模型、网络调度及路由策略维度梳理了无人机抗干扰通信的研究进展。随着无人机组网规模扩大，传统抗干扰方案的计算复杂度呈指数增长。此外，存在多无人机频谱感知节点的情况下，传统抗干扰算法难以做出有效的协同抗干扰决策，成为制约其在大规模网络中实际应用的主要瓶颈[10]。

需要指出的是，大规模组网抗干扰决策依赖对频谱占用与干扰态势的及时获取，因此频谱感知被视为支撑信道接入与协同抗干扰的前置环节，其在无人机组网中的应用仍然面临挑战。首先，固定感知节点位置的系统模型无法满足无人机频谱感知节点对高灵活性的需求[11]，且噪声干扰与时变信道可能造成频谱资源浪费；同时，无人机规模的扩大使感知计算开销显著增加，并要求感知过程实时更新以适应不同高度、速度与环境条件。为应对上述问题，文献[12]面向感知辅助的无人机通信网络，提出深度神经极化码设计以支持低时延高可靠传输；在此基础上，文献[13]通过优化前导无人机位置与感知持续时间提升有效吞吐量，文献[14]进一步通过优化感知弧度在干扰吞吐量约束下最大化有效吞吐量。

近年来，人工智能技术迅速兴起，其具备对现实世界高维特征学习[15]的能力，为突破传统算法在多无人机抗干扰通信场景下的局限性提供了新的思路。深度强化学习（Deep Reinforcement Learning, DRL）作为人工智能技术的重要分支，在系统资源管理及任务调度中展现出广阔的应用前景[16]。具体而言，根据博弈论的知识体系，可构建多智能体在对抗博弈场景下的信息交互与竞争协作机制。基于上述思路，文献[17]提出一种多智能体深度确定性策略梯度算法，以优化系统资源调度效率。另一方面，可将DRL与多种学习范式进行跨域融合，以增强其泛化和迁移性能。例如，为提升智能体面

向新任务的泛化能力，模型无关元学习（Model-Agnostic Meta-Learning, MAML）算法[18]结合元学习（Meta Learning, ML）[19]与DRL，以提取不同任务中的共享知识。类似地，文献[20]结合迁移学习（Transfer Learning, TL）与DRL，使智能体将在源任务中获得的策略迁移到新场景，并通过微调实现对新场景的泛化。针对多无人机任务异质性问题，个性化联邦强化学习框架的应用，能够在边缘计算环境下实现多智能体的自适应决策迁移[21-22]。

在无人机抗干扰通信研究中，直接序列扩频、跳频、自适应波束成形等传统方法虽具备抗干扰能力，但其会消耗大量频谱资源，无法完成多无人机的协同抗干扰决策。DRL的引入能够提升系统在动态干扰环境下的决策能力。例如，文献[23]设计了一种模式感知的DRL抗干扰算法。在干扰信道先验信息未知的情况下，该算法通过卷积神经网络分析频谱瀑布图来识别干扰模式，并结合Q学习进行实时信道选择，但计算开销过大。针对该问题，文献[24]采用相邻时隙间的频谱差异作为观测输入，减少算法对无限状态的依赖。在多用户通信中，文献[25]设计一种基于多智能体深度强化学习的抗干扰算法，在存在功率干扰的情况下改善空地一体化网络的端到端通信性能。然而，该模型仅考虑干扰机发射功率对用户信干噪比（Signal-to-Interference-plus-Noise Ratio, SINR）的影响，未能考虑监听者对系统安全性的影响。

在复杂电磁环境中，合法用户之间的通信不仅存在恶意干扰，还会受到监听者的监听。因此，在强干扰背景下，如何同时保障通信安全性和隐蔽性成为亟待解决的问题[26]。传统网络层加密机制虽可防止未经授权的访问，但其生成的高随机性密文易被识别为可疑目标[27]，从而引发更强的检测与解码攻击，严重威胁通信安全[28]。为此，研究者开始从物理层传输的角度，协同研究无人机网络的抗干扰性能和通信隐蔽性。近期研究表明，多智能体强化学习可通过功率控制与信道动态建模提升无人机通信的隐蔽性与抗干扰性能[29]；无人机作为移动通信节点可通过联合优化轨迹与发射功率降低被侦测概率[30]；此外，面向短包与有限码长约束的无人机中继协助隐蔽通信研究也表明，中继与功率的联合设计可在干扰与监听并存条件下提升隐蔽传输可靠性[31]。进一步地，人工噪声增强的短包隐蔽通信机制通过干扰监听者判决统计并提升短包可检测门限，为无人机网络中的抗干扰隐蔽传输提供了可行路径[32]。此外，中继辅助的太赫兹隐蔽通信通过功率优化增强了对多天线监听者的抗干扰能力[33]，中继节点选择与功率分配策略则进一步提高了系统在多干扰源与主动侦测场景下的安全性与隐蔽性[34]。综上，抗干扰环境下的无人机隐蔽通信研究需要在物理层安全与隐蔽性之间实现动态平衡，为基于深度强化学习的协同抗干扰与隐蔽传输提供了新的思路。然而，现有研究多聚焦于单无人机或小规模通信场景，对大规模无人机组网条件下的抗干扰隐蔽通信机制与协同决策建模仍缺乏系统性分析，这一问题亟待进一步研究。

针对上述挑战，本文面向动态复杂电磁环境下的大规模无人机组网，研究抗干扰与隐蔽通信一体化机制。通过构建多无人机自组网系统模型及频谱感知时隙框架，提出融合元学习与平均场理论的抗干扰-隐蔽通信Q学习算法，以联合构建干扰机、监听者与无人机用户的博弈过程，从而实现抗干扰性能与隐蔽性的协同提升。该研究为大规模无人机网络的安全可靠通信提供可扩展的理论与算法支撑。具体而言，本文的主要研究贡献点如下：

1. 构建面向动态复杂电磁环境的大规模无人机抗干扰隐蔽通信框架，设计频谱感知的时隙模型，在系统层面刻画多UAV用户对、恶意干扰机与被动监听者Willie的共存交互；在此基础上将多无人机抗干扰-隐蔽通信系统建模为部分可观测随机博弈（Partially Observable Stochastic Game, POSG），以最大化长期折扣累积收益为目标，为大规模分布式信道接入决策提供解决思路。

2. 针对大规模无人机网络中用户对间高维耦合导致的计算复杂难题，设计平均场Q学习算法，将每个无人机用户对与所有邻居的多边交互等效为其与邻居集合平均场之间的博弈，实现大规模场景下高效的分布式信道接入策略学习；在此基础上，设计基于元强化学习的平均场Q学习算法，通过引入元学习机制，使无人机用户对在动态抗干扰任务场景下具备良好泛化性；并在奖励函数中联合考虑抗干扰收益与隐蔽性风险，通过求解纳什均衡近似解，实现动态复杂环境下抗干扰性能与隐蔽性的协同优化决策。

3. 通过仿真验证，所提算法在收敛速度、稳定

性和新场景的泛化性上均优于传统强化学习方法，并能够有效降低监听者的检测风险。

1　多无人机抗干扰隐蔽通信的系统模型分析

1.1　系统模型

如图1所示，本文构建了一个多无人机自组网抗干扰隐蔽通信系统模型。系统中部署N个无人机用户对、L个干扰机和单个监听者Willie，其中每个无人机用户对由一架发射无人机与一架接收无人机构成。Willie的监听位置固定，拟对所有信道被动监听，并尝试解码任意用户对在某一信道上的传输。Willie也会受到干扰机发射信号的影响。在该系统中，可用频谱在中心载频附近被划分为若干条相互正交的窄带信道以支撑多无人机用户对接入。然而，在若干干扰机共存且频谱资源受限的场景下，多无人机用户对在进行信道接入决策时，不仅要在有限带宽条件下降低Willie实现成功监听的概率，还需要抑制来自干扰机及其他无人机用户对所产生的同频干扰。为构建该通信场景，本文对相关集合作出如下定义：无人机用户对的集合表示为$N=\{1,...,n,...,N\}$，可供选择的信道集合表示为$\mathcal{M}=\{1,...,m,...M\}$，干扰机集合表示为$\mathcal{L}=\{1,...,l,...,L\}$。对于第$n$个无人机用户对，其发射端的发射功率定义为$P_n$；定义该用户在时隙$t$所占用的信道为$w_n(t)\in\mathcal{M}$；第$j$个干扰机以功率$P_j$运行，并定义其在时隙$t$占用的信道为$g_j(t)\in\mathcal{M}$。信道状态随时间的变化可建模为一条有限状态马尔可夫链，其演化可由状态转移概率矩阵描述。

本文采用高斯马尔可夫随机过程对无人机运动轨迹进行建模，以描述其在时变场景下的位置信息。本文假设在单次通信任务中，各干扰机被视为静止节点，在不同任务场景下，其空间位置可发生变化；同时，为保证通信链路持续可用，各UAV接收端的移动范围被限制在对应发送端的一定距离内。设在时隙t内，UAV发送端的运动速度和方向分别为$v_n(t)$和$\theta_n(t)$，依次表示为

$$v_n(t)=\alpha_1 v_n(t-1)+(1-\alpha_1)\bar{v}_n+\sqrt{1-\alpha_1^2}\,\sigma_n \quad (1)$$

$$\theta_n(t)=\alpha_2\theta_n(t-1)+(1-\alpha_2)\bar{\theta}_n+\sqrt{1-\alpha_2^2}\,\phi_n \quad (2)$$

其中，$\alpha_{1,2}\in[0,1]$分别用于描述运动速度和方向的记忆特性，反映系统对历史状态的依赖程度；

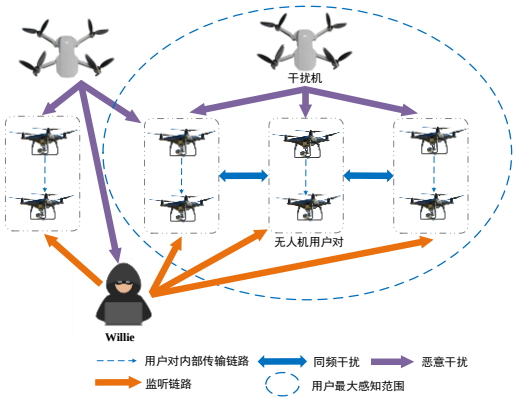

图1　多无人机自组网抗干扰隐蔽通信系统模型

$\bar{v}_n$表示无人机的平均飞行速度，σ_n服从高斯分布，表示速度随机扰动强度；同理，$\bar{\theta}_n$表示无人机的平均运动方向，ϕ_n服从高斯分布，表示方向的随机扰动强度。

1.2　信道模型

在无人机空-空自组网通信场景中，无人机在500m×250m的水平空域内机动飞行，飞行高度为80m[35]，每个用户对的接收端限制在其发射端100m范围内。在开阔低空通信场景中，信道由显著的直达分量及少量由地面反射和环境散射形成的非视距分量共同构成，呈现“强视距、少多径”的小尺度衰落统计特征[36]。为此，本文采用莱斯衰落信道对直达分量与复高斯散射分量的叠加进行建模，以刻画上述传播机理，并使信道统计特性能够随用户的三维几何关系变化。记第n个UAV发送端与对应接收端之间的信道增益为$h_n(t)$，表示为[37]

$$h_n(t)=\sqrt{10^{-\rho_0/10}d_n^{-\eta}}\left[p_n^{\mathrm{LoS}}h_n^{\mathrm{LoS}}+(1-p_n^{\mathrm{LoS}})h_n^{\mathrm{NLoS}}\right] \quad (3)$$

其中，$\rho_0=20\log\left(\frac{4\pi f_c d_0}{c}\right)$表示参考距离$d_0=1\mathrm{m}$处的路径损耗，单位为dB，$c$为光速；$d_n$表示第$n$个UAV发送方与对应接收端之间的距离，$\eta$为路径损耗系数；$h_n^{\mathrm{LoS}}$为随机相位的直达视距分量；$h_n^{\mathrm{NLoS}}$为非视距多径分量，服从圆对称复高斯分布$\mathrm{CN}(0,1)$。第$n$条链路处于直射路径下的概率记为$p_n^{\mathrm{LoS}}$，其表达式为

$$p_n^{\mathrm{LoS}}=\frac{1}{1+A\exp\left[-B\left(\theta_n-A\right)\right]} \quad (4)$$

其中，A表示自由空间路径损耗，B表示信号衰减速率，$\theta_n=\sin^{-1}\left(\Delta z_n/d_n\right)$为式(2)中UAV发送端运动方向的几何表达，定义为第n个用户对中收发无人

机连线与水平面之间的夹角，Δz_n为UAV发送方与接收方之间的高度差。

同理，第i个UAV发送方与第n个UAV接收方之间的信道可以表示为$h_{in}(t)$，第n个无人机用户对到Willie之间的信道可以表示为$h_{nw}(t)$；形式均与式(3)相同。

在上述信道模型基础上，进一步考虑同频复用引起的干扰耦合效应。在多无人机组网中，用户对之间是否存在显著干扰耦合，取决于其空间位置关系下的同频干扰强度。例如，若第i个用户对的无人机发送端在时隙t个对第n个用户对接收端产生的平均干扰功率超过预设阈值I_{th}，则定义第i个用户对为第n个用户对的“邻居”。由此，系统拓扑可表示为随无人机位置与信道占用动态变化的局部干扰耦合拓扑。基于上述分析，第n个无人机用户对的邻居集合$\mathcal{N}_{\mathrm{ne}}(n)$可定义为

$$\mathcal{N}_{\mathrm{ne}}(n)=\left\{i|i\in\mathcal{N}\backslash\{n\},d_i\leqslant d_{\mathrm{th}}\right\} \tag{5}$$

其中，$d_{\mathrm{th}}=\left(P_0\mathbb{E}\left[h_{i,n}(t)\right]/I_{\mathrm{th}}\right)^{\frac{1}{\eta}}$为同频干扰邻居距离阈值，$\mathcal{N}\backslash\{n\}$表示集合差运算，即从集合$\mathcal{N}$中移除元素$n$，得到一个新的集合。各无人机发射端的发射功率$P_i$设为定值$P_0$。

1.3　频谱感知的时隙模型

如图2所示，本文设计了一种感知-决策一体化的多无人机抗干扰隐蔽通信同步时隙模型，其运行流程包含四个阶段：在时隙起始阶段，各无人机节点依据式(1)和式(2)更新其三维空间坐标；在频谱感知阶段，各发送端UAV在全部可用信道集上进行频谱感知，获取当前时隙的频谱状态信息；在决策与数据传输阶段，发送端UAV综合前一时隙的历史信息与当前感知结果，通过预设的决策算法选择最优通信信道并执行数据传输命令；最后，在信息交互与反馈阶段，发送端UAV向其邻居节点广播自身的信道选择决策，同时接收端UAV根据接收结果向发送端回复确认或非确认信号，形成闭环反馈。需特别指出，Willie为被动监听节点，在无人机用户对进行数据传输时同步执行信息监听操作，并在之后的时隙中进行信息处理，以获得接收SINR。干扰机虽遵循相同的时隙结构，但其时钟与无人机网络及Willie保持异步，这种异步特性增加了通信环境的不确定性与抗干扰策略的设计难度。

定义Willie在时隙t对第n个无人机用户对监听时的SINR为

$$\mathrm{SINR}_w^n(t)=\frac{P_n\left|h_{nw}(t)\right|^2}{\sum\limits_{i\in\mathcal{U}_{q_n(t)}}P_i\left|h_{iw}(t)\right|^2+N_w+I_{\mathrm{jam}}} \tag{6}$$

其中，N_w表示Willie接收端的噪声功率，$P_n\left|h_{nw}(t)\right|^2$表示第$n$个用户对的发射功率经LoS路径损耗和小尺度衰落，并考虑天线增益后的等效接收信号功率，I_{jam}表示干扰机对Willie的干扰功率，单位均为W。与第n个无人机用户对使用相同信道的其他用户对的集合表示为

$$\mathcal{U}_{q_n(t)}=\{i|w_i(t)=w_n(t),i\in\mathcal{N},i\neq n\} \tag{7}$$

其中，$w_i(t)$表示第i个无人机用户对在时隙t所选择的信道。从统计检测角度看，Willie的判决性能主要受其接收端SINR决定。SINR越低，目标信号越容易被噪声与同信道干扰淹没，通信隐蔽性就越强。

本文设计的传输时隙可以从空间、时间和抗干扰性能三个方面提升无人机组网的频谱利用效率。在空间维度上，系统通过动态频谱接入机制，实时探测并利用系统信道中的“频谱空穴”，使得无人机在每一传输时隙均可依据实时感知结果选择最优信道，避免因持续干扰或固定占用导致的信道闲置，解决了固定频谱分配模式下的资源僵化问题；在时间维度上，系统综合历史信息与当前感知结果，构建“感知-决策-反馈”的闭环学习框架，使无人机能够预测干扰机的行为，主动规避受干扰信道；在抗干扰性能方面，无论干扰机处于感知、干扰或静默状态，无人机均能在每个传输时隙前获取最新的频谱状态快照，规避即将遭受干扰的信道，从而实现快速频谱切换。

此外，在隐蔽性维度上，系统引入被动监听者Willie的监听模型，通过将Willie的监听行为纳入系统学习过程，能够在算法学习过程中动态平衡传输速率与被侦测风险，实现抗干扰与隐蔽性的联合优化。类似思路已在IRS-UAV辅助隐蔽通信系统[38]中得到验证，为多无人机场景下的隐蔽通信建模提供了参考。

1.4　模型的数学表达

对第n个无人机用户对，设其接收端的SINR判决门限为ζ_{th}。当实际接收的SINR不低于ζ_{th}时，

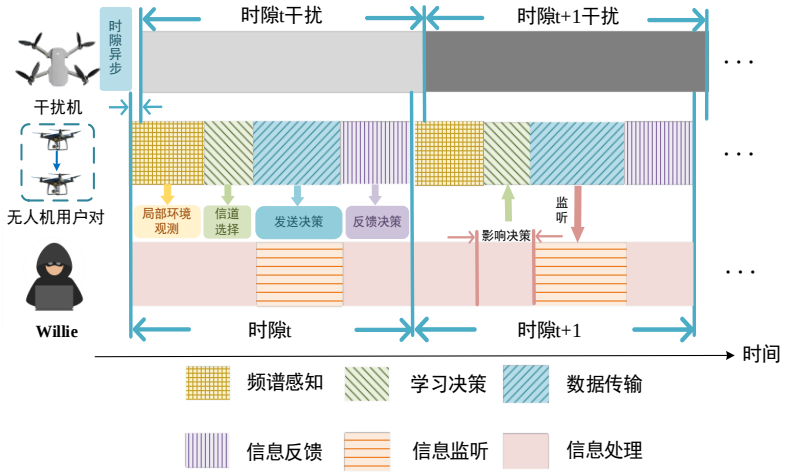

图2　无人机用户、监听者及干扰机传输时隙架构图

接收端可以成功完成信号解调并向发送端反馈确认信息；否则判定该时隙内传输失败。因此，第n个无人机用户对在第t个时隙的瞬时通信速率$C_n(t)$可表示为

$$C_n(t)=\begin{cases}B\log_2(1+\zeta_n(t)), & \text{if}\ \zeta_n(t)\geqslant\zeta_{th}\\ 0, & \text{if}\ \zeta_n(t)<\zeta_{th}\end{cases} \tag{8}$$

其中，B表示可用正交信道的带宽，$\zeta_n(t)$表示第n个无人机用户对在时隙t所选信道上的接收端信干噪比，其表达式如下

$$\zeta_n(t)=\frac{P_n\left|h_{nn}\right|^2}{I_n(t)+J_n(t)+N_n} \tag{9}$$

其中，N_n为第n个UAV接收端的噪声功率，$I_n(t)$表示来自其他无人机用户对的干扰信号功率，$J_n(t)$表示干扰机的干扰信号功率，单位均为W。其表达式依次为

$$I_n(t)=\sum_{i\in\mathcal{N}n}P_i\left|h_{in}(t)\right|^2\eta\left(w_i(t),w_n(t)\right) \tag{10}$$

$$J_n(t)=\sum_{j\in\mathcal{J}}P_j\left|h_{jn}(t)\right|^2\eta\left(g_j(t),w_n(t)\right) \tag{11}$$

其中，$\eta(\cdot)$为判别函数，用于判定某一信道是否同时被两个通信节点占据，其定义为

$$\eta(x,y)=\begin{cases}1, & \text{if}\ x=y\\ 0, & \text{if}\ x\neq y\end{cases} \tag{12}$$

其中，当$x=y$时，表示该信道被两个通信节点同时占用，此时$\eta(x,y)$取值为1；当$x\neq y$时，表示两个通信节点工作在不同信道，此时$\eta(x,y)$取值为0。

为最大化长期期望累积通信速率，各无人机用户对基于自身策略对信道选择进行优化，其优化目标表示为

$$\max_{a_n(t)\in\mathcal{M},\forall t}\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t C_n(t)\right] \tag{13}$$

其中，$\mathbb{E}[\cdot]$表示数学期望算子，$\gamma\in(0,1)$是折扣因子，T是时隙总长[39]。

2　多无人机抗干扰隐蔽通信的数学模型

2.1　部分可观测随机博弈模型

由于单个无人机用户对无法直接获取系统的全局信息，本文引入部分可观测随机博弈（Partially Observable Stochastic Game, POSG）框架，用于描述多无人机在抗干扰场景下的信道接入问题，并将其构建为一个六元组$<\mathcal{S},\mathcal{A},\mathcal{O},P,r,\gamma>$。其中，$\mathcal{S}$表示状态空间；$A=A^1\times\cdots\times A^N$表示所有无人机用户对的联合动作空间；$\mathcal{O}=O^1\times\cdots\times O^N$表示联合观测空间；$P:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$表示状态转移概率，描述系统状态在动作作用下的演化规律；r为即时奖励函数，用于评估单步决策的优劣；γ为折扣因子，用于平衡短期与长期回报。POSG各元素的详细说明如下：

1) 状态空间$\mathcal{S}$：用于描述多无人机组网在某一时隙下的整体运行状态，包括：所有无人机与干扰机的空间位置、各信道在当前时隙下的占用情况和上一时隙干扰机的信道占用情况。此外，任意单个无人机用户都无法完整获知系统的全局状态，需要借助局部观测加以推断。

2) 观测空间$\mathcal{O}$：第t个时隙下，第n个无人机用户对所获取的局部观测定义为$\mathbf{o}_t^n$，可表示为

$$\mathbf{o}_t^n = \left[F_{t-1}^n, \mathbf{J}_{t-1}^n, \mathbf{D}_t^{j,n}, \rho\left(\mathrm{SINR}_w^n(t)\right) \right] \tag{14}$$

其中，$\mathbf{o}_t^n = \left[o_t^1, o_t^2, ..., o_t^N\right]$，$F_{t-1}^n \in \{0,1\}$为第 $t-1$ 个时隙的反馈信号，若取值为0，则表示通信失败；取值为1则表示通信成功。$\mathbf{J}_{t-1}^n = \left[j_{t-1}^n(1), ..., j_{t-1}^n(m), ..., j_{t-1}^n(M)\right]$为第 $t-1$ 个时隙感知所得的干扰信道状态向量。其中，$j_{t-1}^n(m)$用于表示第 m 个信道的占用情况：若该信道被干扰机占用，则将其置为1，否则置为0。$\mathbf{D}_t^{j,n} = \left[d_t^{j_1,n_t}, d_t^{j_2,n_t}, d_t^{j_1,n_r} d_t^{j_2,n_r}\right]$记录第 t 个时隙下第 n 个无人机用户对与干扰机之间的距离，分别对应干扰机 j_1、j_2 到该UAV用户对发送端 n_t 和接收端 n_r 的距离。为将被动监听者Willie的不可检测性要求显式纳入系统学习目标，本文以Willie在时隙 t 对第 n 个无人机用户对监听时的 $\mathrm{SINR}_w^n(t)$ 表征其统计检测能力；$\rho\left(\mathrm{SINR}_w^n(t)\right)$为风险函数，表示链路在SINR条件下的风险水平。可采用logistic型映射函数[40]表示为：

$$\rho\left(\mathrm{SINR}_w^n(t)\right) = \frac{1}{1 + \exp\left[-k\left(\mathrm{SINR}_w^n(t) - \tau_{th}\right)\right]} \tag{15}$$

其中，$\rho\left(\mathrm{SINR}_w^n(t)\right) \in (0,1)$；当 $\mathrm{SINR}_w^n(t)$ 低于门限 τ_{th} 时，风险趋近于0，通信隐蔽性高；当其高于门限 τ_{th} 时，风险迅速增大，通信隐蔽性差；$k > 0$ 为Sigmoid斜率，控制风险由低到高的转折陡峭程度。

3) 动作空间 $\mathcal{A}$：在时隙 t 内，第 n 个无人机用户对从可用信道集合 $\mathcal{M}$ 中选取一个信道进行通信，该决策动作记为 $a_t^n \in \mathcal{M}$。由于每个用户对的收发无人机采用相同的信道，因此在时隙 t 内所有无人机用户对构成的联合动作向量表示为 $\boldsymbol{a}_t^n = \left[a_t^1, a_t^2, ..., a_t^N\right]$。

4) 奖励函数 r：为在最大化系统长期累积通信速率的同时降低被Willie监听的风险，将第 n 个无人机用户对在第 t 个时隙获得的即时奖励定义为通信速率与隐蔽性惩罚项之差。因此，第 n 个无人机用户对在时域 $[0,T]$ 内的累积奖励 $\tilde{r}_t^n$ 可表示为

$$\tilde{r}_t^n = \sum_{t=0}^{T} \gamma^t \left[C_n(t) - \lambda \rho\left(\mathrm{SINR}_w^n(t)\right) \right] \tag{16}$$

其中，$C_n(t)$ 是式(8)定义的通信速率，$\lambda > 0$ 为隐蔽性权重系数。在此基础上，整个无人机集群在时隙 t 的联合奖励向量可表示为 $\tilde{\boldsymbol{r}}_t^n = \left[r_t^1, r_t^2, ..., r_t^N\right]$。将 $\rho\left(\mathrm{SINR}_w^n(t)\right)$ 作为隐蔽性惩罚项并入时隙奖励，使无人机在信道接入与资源分配时能够同步权衡通信收益与被监测风险。

5) 策略：第 n 个无人机用户对在时隙 t 内的信道选择策略为 $\pi_t^n: O \to \Omega\left(a_t^n\right)$，表示将观测 O 映射到动作空间 a_t^n 上的概率分布集合 $\Omega\left(a_t^n\right)$。系统的联合策略向量记为 $\boldsymbol{\pi}_t^n \triangleq \left[\pi_t^1, \pi_t^2, ..., \pi_t^N\right]$。

2.2　优化目标

在分布式学习框架下，各无人机用户对作为自主决策的智能体，能够通过不断优化信道选择策略使长期累积收益最大化。各用户对之间同频复用干扰的存在，使任意一对用户的性能不仅依赖于自身策略，还会受到联合策略 $\boldsymbol{\pi}_t^n$ 的影响。因此，在联合策略 $\boldsymbol{\pi}_t^n$ 下，第 n 个无人机用户对的价值函数可定义为累积奖励，表示为

$$V_{\boldsymbol{\pi}_t^n}^n\left(\boldsymbol{o}_t^n\right) = V^n\left(\boldsymbol{o}_t^n, \boldsymbol{\pi}_t^n\right) = \sum_{t=0}^{T} \gamma^t \mathbb{E}_{\boldsymbol{\pi}_t^n, p}\left[\tilde{r}_t^n \middle| o_0^n = o, \pi_t^n\right] \tag{17}$$

其中，γ 为折扣因子，p 为状态转移概率，o 为初始观测结果。

因此，第 n 个无人机用户对的最优策略优化问题可写为

$$\left(\boldsymbol{\pi}_t^n\right)^* = \arg\max_{\pi_t^n} V^n\left(\boldsymbol{o}_t^n, \boldsymbol{\pi}_t^n\right) \tag{18}$$

若每个无人机用户对均选择了自身的最优响应策略，且任一用户对单独偏离该策略都不能获得更高回报，则认为该POSG已达到纳什均衡状态。若存在联合策略 $\left(\boldsymbol{\pi}_t^n\right)^* \triangleq \left[\left(\pi_t^1\right)^*, \left(\pi_t^2\right)^*, \cdots, \left(\pi_t^n\right)^*\right]$，使得对任意观测 $\boldsymbol{o}_t^n$ 和任意可行策略 $\boldsymbol{\pi}_t^n$，均满足[41]

$$V_{\left(\boldsymbol{\pi}_t^n\right)^*}^n\left(\boldsymbol{o}_t^n\right) = V_{\left(\boldsymbol{\pi}_t^n\right)^*, \left(\boldsymbol{\pi}_t^{-n}\right)^*}^n\left(\boldsymbol{o}_t^n\right) \geqslant V_{\boldsymbol{\pi}_t^n, \left(\boldsymbol{\pi}_t^{-n}\right)^*}^n\left(\boldsymbol{o}_t^n\right) \tag{19}$$

其中，$\left(\boldsymbol{\pi}_t^{-n}\right)^* \triangleq \left[\left(\pi_t^1\right)^*, \cdots, \left(\pi_t^{n-1}\right)^*, \left(\pi_t^{n+1}\right)^*, \cdots, \left(\pi_t^N\right)^*\right]$ 表示除第 n 个无人机用户对外所有无人机用户对的联合策略。

纳什均衡定理指出，在一个有限博弈中，至少存在一个纳什均衡解。尽管POSG中各智能体仅能获得局部观测，但在状态、动作和时域均受限的条件下，可基于观测历史的随机策略将POSG等价于一个有限动态不完全信息博弈，使之仍然符合纳什均衡定理[42]。因此，在包含 N 个无人机用户对的

POSG 框架下，至少存在一个稳定的纳什均衡解。式(19)刻画了 POSG 下的纳什均衡条件，即当任一用户对在给定他人均衡策略时不存在单边改进空间时，认为联合策略达到局部纳什均衡。

3 大规模无人机抗干扰隐蔽通信算法设计

在大规模用户群进行动态抗干扰隐蔽通信的情况下，最优信道接入策略难以在 POSG 框架下直接求解。在干扰机位置节点不变的前提下，针对多无人机之间复杂耦合关系导致的计算复杂难题，本文引入平均场 Q 学习（Mean Field Q-Learning，MFQ），将“个体–多邻居”交互近似为“个体–平均场”博弈，在保持分布式决策的同时有效降低维度与复杂度；进一步地，考虑干扰机节点位置在不同场景下动态变化，本文引入元强化学习机制，从任务分布视角学习具备快速适应能力的初始化策略，从而提高对新场景的泛化能力。综合上述两点，本文将 MFQ 和元学习结合的抗干扰-隐蔽通信联合信道选择算法命名为元平均场抗干扰-隐蔽通信 Q 学习算法（Meta Mean-Field Q-Learning for Anti-Jamming and Covert Communication, MMFQ-ACC）。

3.1 平均场 Q 学习算法

作为研究多体系统的数学框架，平均场理论的核心在于不再逐一刻画个体之间复杂的耦合关系，而是将其归结为个体与一个代表整体影响的“平均场”之间的作用。平均场理论能够将原本高维、耦合的多体问题降维成可处理的单体问题，从而降低计算复杂度。这一思想在无人机组网中被广泛应用于分布式强化学习[43]。在多无人机抗干扰通信中，可将其余用户对的整体影响等效为一个平均场。每个用户对只需与该平均场交互，而不必显式处理与其他用户对之间的两两博弈。在这一建模方式下，各无人机用户对能够依据学习得到的信道接入策略自适应调整自身动作，从而在大规模集群场景下实现高效分布式学习和实时决策。

基于式(17)，在系统联合策略已知的前提下，第 n 个无人机用户对的 Q 函数可定义为

$$Q^n\left(\boldsymbol{o}_t^n,\boldsymbol{a}_t^n\right)=\tilde{r}_t^n\left(\boldsymbol{o}_t^n,\boldsymbol{a}_t^n\right)+\gamma\mathbb{E}_p\left[V_{\boldsymbol{\pi}_t}^n\left(\boldsymbol{o}_{t+1}^n\right)\right] \tag{20}$$

其中，$\boldsymbol{o}_{t+1}^n$ 表示第 n 个无人机用户对在时隙 t 的观测结果。

基于式(20)，可用 Q 函数表示局部观测 $\boldsymbol{o}_t^n$ 的价值函数，表示为

$$V_{\boldsymbol{\pi}_t}^n\left(\boldsymbol{o}_t^n\right)=\mathbb{E}_{\boldsymbol{a}_t^n\sim\boldsymbol{\pi}_t^n}\left[Q^n\left(\boldsymbol{o}_t^n,\boldsymbol{a}_t^n\right)\right] \tag{21}$$

需要指出的是，随着无人机用户对数量的增加，组网的联合动作空间维度会急剧上升，此时若直接在全局动作空间上进行 Q 函数的计算和更新，将引发高昂的时间及样本复杂度开销。另一方面，由于各无人机用户对之间没有合作关系，因此在实际系统中难以使每个用户对都学习全局 Q 函数。基于上述考虑，本文将第 n 个无人机用户对与其邻居用户对交互的动作对 Q 函数进行分解，表示为

$$Q^n\left(\boldsymbol{o}_t^n,\boldsymbol{a}_t^n\right)=\frac{1}{N_{\mathrm{ne}}^n}\sum_{i\in\mathcal{N}_{\mathrm{ne}}(n)}Q^n\left(\boldsymbol{o}_t^n,a_t^n,a_t^i\right) \tag{22}$$

其中，$N_{\mathrm{ne}}^n=\left|\mathcal{N}_{\mathrm{ne}}(n)\right|$ 表示第 n 个无人机用户对的邻居数量。第 i 个邻居的动作可表示为

$$a_t^i=\bar{a}_t^n+\delta a_t^{n,i} \tag{23}$$

其中，$\bar{a}_t^n=\dfrac{1}{N_{\mathrm{ne}}^n}\sum_i a_t^i$ 表示邻居动作均值，$a_t^{n,i}$ 表示第 n 个无人机用户对的第 i 个邻居用户在时隙 t 的动作，$\delta a_t^{n,i}$ 表示邻居 i 相对该平均动作的微小波动。

在假设 $Q^n\left(\boldsymbol{o}_t^n,a_t^n,a_t^i\right)$ 关于 a_t^i 二阶可导的条件下，可对式(22)进行泰勒展开得

$$\begin{aligned}Q^n\left(\boldsymbol{o}^n,\boldsymbol{a}^n\right)=&\frac{1}{N_{\mathrm{ne}}^n}\sum_i\Big[Q^n\left(\boldsymbol{o}^n,a^n,\bar{a}^n\right)\\&+\nabla_{\bar{a}^n}Q^n\left(\boldsymbol{o}^n,a^n,\bar{a}^n\right)\delta a^{n,i}\\&+\frac{1}{2}\delta a^{n,i}\nabla_{\tilde{a}^{n,i}}^2Q^n\left(\boldsymbol{o}^n,a^n,\tilde{a}^{n,i}\right)\delta a^{n,i}\Big]\\=&Q^n\left(\boldsymbol{o}^n,a^n,\bar{a}^n\right)+\frac{1}{2N_{\mathrm{ne}}^n}\sum_i R_{\boldsymbol{o}^n,a^n}^n\left(a^i\right)\\\approx&Q^n\left(\boldsymbol{o}^n,a^n,\bar{a}^n\right)\end{aligned} \tag{24}$$

其中，$R_{\boldsymbol{o}^n,a^n}^n\left(a^i\right)$ 表示泰勒展开式的余项，$\tilde{a}^{n,i}=\bar{a}^n+\sigma^{n,i}\delta a^{n,i}$，$\sigma^{n,i}\in[0,1]$。考虑到 $\sum_i\delta a^{n,i}=0$，可消去式(24)中的一阶项。特别地，当 $Q^n\left(\boldsymbol{o}^n,a^n,a^i\right)$ 为线性函数时，余项 $R_{\boldsymbol{o}^n,a^n}^n\left(a^i\right)$ 可收敛为零，此时泰勒展开式可近似为 $Q^n\left(\boldsymbol{o}^n,a^n,\bar{a}^n\right)$。为简洁起见，式(24)中省略了时隙下标 t。

引入平均场近似后，可将第 n 个无人机对与其所有邻居之间的复杂多边交互等效为其与一个虚拟代理之间的双边交互，该虚拟代理反映所有邻居对第 n 个用户对的平均影响。由式(24)给出的这一近

似形式称为MFQ函数。

3.2　元平均场Q学习算法

值得注意的是，MFQ算法能够有效处理干扰机节点位置静止的情况，但难以实现在干扰机位置动态变化下的场景泛化。在假设干扰机的位置在三维空间动态变化的背景下，由于干扰机的运动轨迹不同，无人机面对的抗干扰决策场景也会发生变化。传统强化学习面向一组固定的干扰机部署与运动轨迹进行策略优化，其获得的最优策略缺乏跨场景泛化能力，因此在动态干扰条件下难以实现快速有效的策略迁移。而元学习可以提供快速适应新任务的策略，当干扰机的移动轨迹发生变化时，无人机用户对可通过元强化学习框架实现对新场景的快速泛化与策略迁移[44]。此外，本文在平均场强化学习框架中引入了基于Willie检测风险的奖励约束项，在动态干扰环境下考虑通信隐蔽性，使无人机在决策过程中自动规避被监听概率高的信道。

需要说明的是，MMFQ-ACC算法采用的训练数据来自POSG仿真环境中由无人机用户对与环境逐时隙交互在线生成的经验样本。这些交互样本被持续写入经验池，并以小批量随机抽取的方式用于网络参数更新，从而保证训练和适应阶段的数据生成与策略更新在同一闭环过程中完成。

在MMFQ-ACC框架中，引入参数为θ的神经网络对平均场Q函数进行逼近，其输出可写为$Q^n\left(\theta;\boldsymbol{o}^n,a^n,\bar{a}^n\right)$。基于该网络输出，无人机用户对构造玻尔兹曼策略，形式为

$$\pi^n\left(a^n|\boldsymbol{o}^n,\bar{a}^n\right)=\frac{\exp\left(-\eta Q^n\left(\theta^n;\boldsymbol{o}^n,a^n,\bar{a}^n\right)\right)}{\sum\limits_{a^{n\prime}\in\mathcal{A}^n}\exp\left(-\eta Q^n\left(\theta^n;\boldsymbol{o}^n,a^{n\prime},\bar{a}^n\right)\right)}\tag{25}$$

其中，η是玻尔兹曼温度常量，用于调节策略的随机性。η越大，动作分布越平坦、探索性越强；当η逐渐减小时，策略则趋于贪婪。

为使上述策略收敛，本文通过最小化损失函数来更新无人机用户对的网络参数，表示为

$$\begin{aligned}L\left(\theta^n\right)=\mathbb{E}_{<\boldsymbol{o}_t^n,a_t^n,r_t^n,\boldsymbol{o}_{t+1}^n,\bar{a}_t^n>}\Big[&\tilde{r}_t^n+\gamma\max_{a^n}\\&\hat{Q}^n\left(\theta^{n\prime};\boldsymbol{o}^n,a^n,\bar{a}^n\right)-Q^n\left(\theta^n;\boldsymbol{o}^n,a^n,\bar{a}^n\right)\Big]\end{aligned}\tag{26}$$

其中，$\hat{Q}^n(\cdot)$表示目标神经网络输出的最大Q值，$\theta^{n\prime}$为目标神经网络的参数。目标网络可在训练过程中抑制价值函数的波动，提升参数更新的稳定性。

图3展示了MMFQ-ACC算法的整体流程，可分为训练阶段和适应阶段两部分。

本文通过构造具有显著分布差异的任务集合来形成跨场景泛化测试环境，从而验证MMFQ-ACC在复杂电磁环境下的适应与泛化能力。为此，上述算法流程在训练与适应两个阶段设置不同的干扰移动路线配置：训练阶段采用较为简单的直线路径，适应阶段引入更复杂的随机圆形路径，以考察干扰策略对态势变化所带来的场景差异的适应性与泛化能力。

在训练阶段，干扰机移动路径为直线，所有直线路径构成任务集合$p(\tau)$。每轮训练从$p(\tau)$中随机选取任务$\mathcal{T}_x$，且不同的训练轮次能够选择相同的移动路径。首先，各UAV发送端获得时隙t下的局部观测$\boldsymbol{o}_t^n$和时隙t - 1下的平均动作$\bar{a}_{t-1}^n$，在网络参数θ_t^n下根据玻尔兹曼策略对动作空间进行采样，生成动作a_t^n。随后，各无人机对在所选信道上执行数据传输操作，获取即时奖励r_t^n和时隙$t+1$下的观测$\boldsymbol{o}_{t+1}^n$，进而更新平均动作$\bar{a}_{t+1}^n$。然后，各无人机对将更新后的训练样本$<\boldsymbol{o}_t^n,a_t^n,r_t^n,\boldsymbol{o}_{t+1}^n,\bar{a}_t^n>$存入经验池$\mathcal{D}$，直至经验池容量达到上限。最后，从$\mathcal{D}$中随机

抽取一批样本，利用式(26)推导的损失函数将网络参数更新为：

$$\theta_{\text{meta}}\leftarrow\theta-\frac{\alpha}{B}\sum_{i=1}^{n_s}\nabla_\theta L_i(\theta)\tag{27}$$

其中，α为训练阶段的学习率，n_s为训练阶段抽取的样本数量。

训练阶段的输出为可迁移的网络参数初始化及其对应的信道选择策略，该初始化参数刻画了训练任务分布上的跨任务知识先验。适应阶段以该初始化为起点，在新任务上利用少量交互样本进行更新，得到适应后的网络参数及信道接入策略。

在适应阶段，干扰机的移动路径为圆形，所有圆形路径构成任务集合$p'(\tau)$。在每一轮训练开始时，算法从集合$p'(\tau)$中随机抽取一个新任务$'_x$，并基于该任务的观测信息和网络参数θ_{meta}，根据玻尔兹曼策略生成动作a_t^n，并将交互产生的数据样本$(\boldsymbol{o}_t^n,a_t^n,r_t^n,\boldsymbol{o}_{t+1}^n,\bar{a}_t^n)$存入经验池$'$，直至经验池容量达到上限。从$'$中随机抽取一批样本，利用式(26)推导的损失函数将网络参数更新为

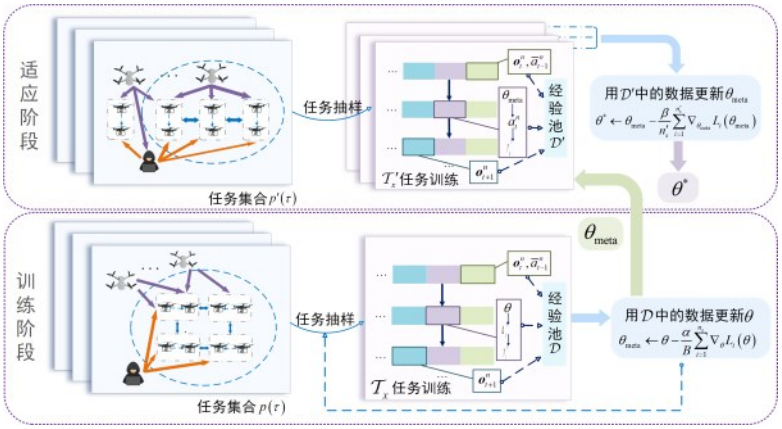

图3　MMFQ-ACC 算法框架

$$\theta^* \leftarrow \theta_{\text{meta}} - \frac{\beta}{n_s'} \sum_{i=1}^{n_s'} \nabla_{\theta_{\text{meta}}} L_i\left(\theta_{\text{meta}}\right) \tag{28}$$

其中，θ^*表示训练结束后的最优参数，β为适应阶段的学习率，n_s'为适应阶段抽取的样本数量。

4　仿真结果与分析

4.1　仿真参数设置

为验证所提 MMFQ-ACC 算法的有效性，本文通过仿真实验将其与三种基准算法进行对比，分别为概率 Q 学习[45]（Probabilistic Q-learning, PQ）、独立 Q 学习[46]（Independent Q-Learning, IQL）和平均场 Q 学习（MFQ）。与 MMFQ-ACC 相比，MFQ 未设置适应阶段，因此，当干扰环境发生变化时，MFQ 需要对其神经网络进行完整的重新训练，而无法实现跨场景泛化。PQ 将 Q 函数建模为概率分布，并能够依据概率分布对动作进行随机选取；IQL 则进行策略更新与决策时不考虑其他用户对的策略，仅将其视作环境的一部分。

本文的仿真参数设置如下：系统以 2GHz 为中心载频[35]，共部署 200 架无人机，并随机配对为 100 个独立的无人机用户对；配置 80 条可用通信信道，通信过程中同时存在 2 台恶意干扰机和 1 个被动监听者。本文设置可用信道数小于用户对数量，旨在刻画频谱资源有限条件下的信道复用、频谱竞争与拥塞现象。为了模拟无人机真实的运动特性，各节点的移动轨迹通过高斯马尔可夫随机过程生成，并约束发送端与接收端的最大间距为 50m。训练过程包括 $N_e = 200$ 个回合，每回合包含 $N_t = 2000$ 个时隙，隐藏层神经元数 $N_t = 50$。训练数据由无人机用户对与仿真环境逐时隙交互在线生成，每条样本包含当前状态、动作、即时奖励、下一状态以及平均动作等数据项，整个训练过程累计生成 4000 万条在线交互样本，能够为模型训练与收敛分析提供充分的数据支撑。其他参数设置见表 1。表 1 中噪声功率单位是 dBm，但在仿真中统一换算为 W，其转换公式为

$$P_{\text{W}} = 10^{\frac{P_{\text{dBm}} - 30}{10}} \tag{29}$$

其中，P_{dBm}表示以 dBm 为单位的功率值，P_{W}表示以 W 为单位的功率值。

表 1　系统参数设置

参数名称	参数描述	参数值
d_{th}	邻居距离阈值	100m
H	无人机飞行高度	80m
f_c	载波频率	2GHz
W_{ch}	单信道带宽	1MHz
p_j	干扰机干扰功率	1W
p_0	无人机发射功率	1W
N_n	系统噪声功率	-115dBm
N_w	Willie 接收端噪声功率	-110dBm
ζ_{th}	SINR 门限	5dB
ρ_0	参考距离 1m 处的路径损耗	-30dB
λ	隐蔽性权重系数	0.1~1
α	训练阶段学习率	0.01
β	适应阶段学习率	0.01
γ	折扣因子	0.95
γ_w	Willie 判决阈值	0.3

本文通过仿真实验对MMFQ-ACC的收敛性进行评估，并从信道资源与无人机集群规模两个维度分析其对系统性能的影响，以此与不同基准算法的表现进行对照。同时，针对隐蔽通信场景，本文还引入相关隐蔽性指标进行检验，以验证MMFQ-ACC在隐蔽通信约束下的有效性。

4.2　不同无人机规模下的性能展示与分析

图4展示了不同无人机规模下各算法的性能对比，纵坐标为归一化奖励值，可表示为 $r_t' = (r_t - r_{\min})/(r_{\max} - r_{\min})$。实验中设定无人机数量分别为20、60、100和200，并随机将无人机两两配对形成用户对。实验设置可用信道数与无人机对数量之比为1:2，以模拟频谱资源受限场景。

由图4可得以下结论：总体来说，随着无人机数量由20增至200，所有算法的奖励值均上升，说明各算法在无人机规模扩展时仍能持续改善信道接入收益。当无人机规模较小（20和60）时，各算法性能相对接近，MMFQ-ACC与其他三种算法的差距并不显著，此时平均场近似与元学习技术带来的性能增益相对有限。随着无人机规模变大（100以上），MMFQ-ACC的奖励值稳定高于其他对比算法，并且领先幅度随规模扩大而逐渐增大。这表明在大规模复杂干扰环境下，MMFQ-ACC通过元强化学习的跨场景泛化能力与平均场博弈建模，能更有效地解决大规模决策中的拥塞与干扰问题，从而获得更高的长期回报。

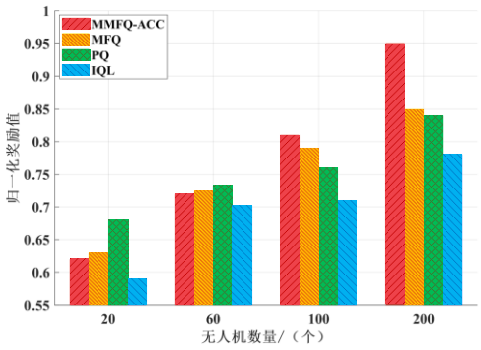

图4　无人机数量对算法性能的影响

4.3　不同频谱资源下的性能展示与分析

图5展示了MMFQ-ACC、MFQ与IQL三种算法在不同可用信道配置下的算法性能对比。为模拟无人机组网通信场景，仿真中共设置N=100个无人机用户对。整体而言，随着可用信道数量的增加，三种算法的奖励值均呈上升趋势，但增益逐渐趋于平缓。在所有信道设置下，MMFQ-ACC始终取得最高的奖励值，其相对优势在信道资源受限时尤为显著；而在信道数量较多时，MMFQ-ACC与MFQ之间的性能差距则有所缩小；IQL则明显低于另外两种算法。具体来说，当可用信道数为40时，各算法的奖励值均处于较低水平，但MMFQ-ACC获得的奖励值分别较MFQ和IQL高出14.5%和51%，说明其在频谱资源匮乏的场景下具备更高的信道利用效率。当可用信道数为80时，系统可用频谱增加使所有算法的奖励值显著提升，此时MMFQ-ACC获得的奖励值相较于MFQ和IQL分别高出2.5%和26%，说明在频谱资源增加的条件下，MMFQ-ACC仍能保持稳定的性能优势。当可用信道数为100时，MMFQ-ACC和MFQ的归一化奖励值均接近1，说明频谱资源已被充分利用，而MMFQ-ACC的奖励值仍略高于MFQ。此外，由于IQL受制于其独立学习机制，难以有效应对复杂干扰环境，性能始终较低。值得注意的是，可用信道数量从80增加到100的过程中，MMFQ-ACC的奖励提升斜率明显变缓，表明其在有限信道条件下即可逼近系统容量的性能极限，在保障传输效率的同时实现频谱资源的高效利用。

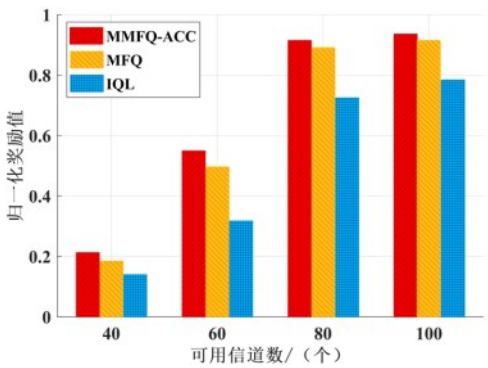

图5　不同算法的奖励值随可用信道数变化的对比

4.4　隐蔽性性能展示与分析

图6展示了在MMFQ-ACC算法学习过程中，监听者Willie的SINR随训练轮次的变化趋势。在前20个训练轮次中，Willie的SINR从约4dB快速下降，波动明显。这表明算法在初始阶段仍在探索最优策略，系统的干扰控制和功率分配尚未稳定。当算法轮次在20到60之间时，SINR快速下降至0dB以下。这表明MMFQ-ACC算法有效降低了

Willie的接收能力，算法逐步学习到了更优的UAV资源分配策略。在第80轮训练之后，SINR曲线趋于平稳，波动幅度减小，数值基本维持在约-2dB附近。这表明算法已经收敛，系统的干扰策略和功率分配策略趋于稳定，Willie的监听被有效抑制。

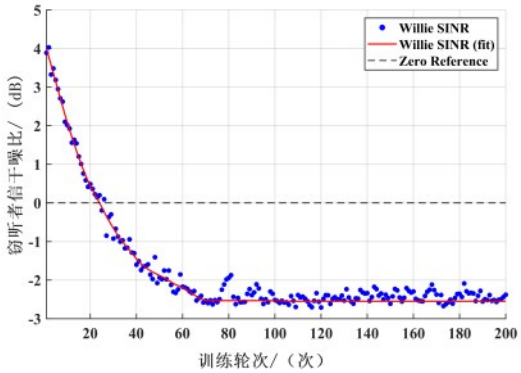

图6　MMFQ-ACC算法下Willie的SINR随训练轮次变化的趋势

图7展示了UAV用户与监听者Willie的平均归一化指标随训练轮次变化的对比曲线。该指标来源于强化学习过程中对系统效用函数的归一化处理，其物理含义如下：对于UAV用户，纵轴表示平均归一化吞吐率，用于评估合法通信链路的平均

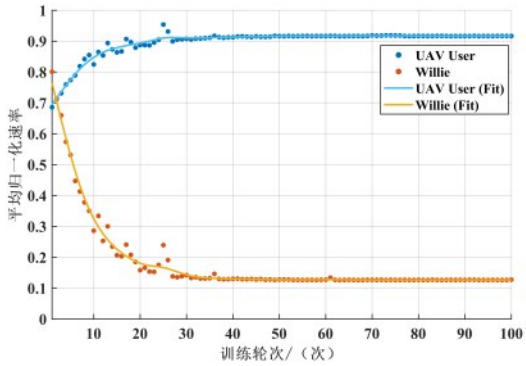

图7　UAV用户与Willie平均归一化指标随训练轮次变化的对比

传输质量，数值越高表示UAV端的信号质量与数据传输效率越好。对于Willie，纵轴表示平均归一化检测风险概率，用于评估监听端对通信行为的平均感知能力，数值越高表示Willie越容易检测到UAV的传输行为，系统隐蔽性越差。

具体地，首先将UAV侧瞬时通信速率$C_n(t)$进行线性归一化，定义为

$$\tilde{C}_n(t)=\frac{C_n(t)-C_{\min}}{C_{\max}-C_{\min}} \tag{30}$$

其中$C_{\min}$为瞬时速率下界，表示通信失败时的瞬时速率；$C_{\max}$为瞬时速率上界，表示在最有利链路条件下的瞬时速率理论最大值。

随后，在每个训练轮次内对所有用户对与时隙求平均，得到UAV侧平均归一化吞吐率为

$$\bar{C}_{\mathrm{UAV}}^{\mathrm{nor}}=\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\tilde{C}_n(t) \tag{31}$$

对应地，将Willie侧由风险映射得到的检测风险进行归一化，定义为

$$\tilde{P}_{w,n}(t)=\frac{\rho\left(\mathrm{SINR}_w^n(t)\right)-\rho_{\min}}{\rho_{\max}-\rho_{\min}} \tag{32}$$

其中$\rho_{\min}$与$\rho_{\max}$为归一化所采用的检测风险上下界。

最后，在每个训练轮次内对所有用户对与时隙取平均，得到Willie侧平均归一化检测风险概率为

$$\bar{P}_w^{\mathrm{nor}}=\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\tilde{P}_{w,n}(t) \tag{33}$$

随着训练轮次的增加，UAV用户的归一化吞吐率逐渐上升并趋于稳定，说明强化学习策略能够自

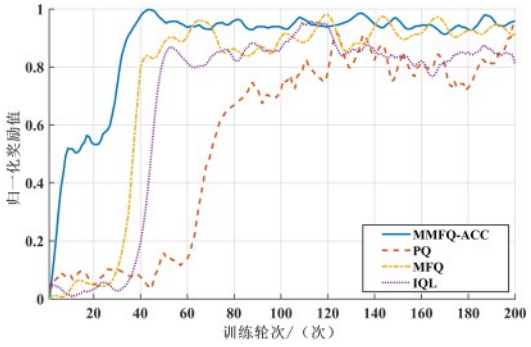

图8　不同算法收敛性比较

适应地选择受干扰影响小的信道，以提升通信质量。同时，Willie的检测概率逐步下降，表明系统在维持通信可靠性的同时，能够选择一条不易被Willie检测到的信道，最终两个指标趋于收敛。定量分析表明，相较于初始阶段，UAV端的平均归一化吞吐率提升约25.09%，Willie端的平均检测风险概率降低约79.73%，该结果验证了所设计强化学习机制在通信质量与隐蔽性权衡方面的有效性。

4.5　收敛性性能展示与分析

图8展示了MMFQ-ACC算法与其他三种基准算法的收敛性能对比。由图可知，MMFQ-ACC的

奖励曲线在约第50个回合附近趋于平稳并完成收敛，其收敛速度明显快于PQ和IQL，略快于MFQ。这表明在训练早期，MMFQ-ACC能更快形成有效的信道接入策略，从而较早适应干扰环境的变化。就收敛后的性能水平而言，MMFQ-ACC的奖励值稳定保持在0.9以上；MFQ的奖励值虽然整体接近0.9，但仍存在一定幅度的上下波动；PQ和IQL收敛后的奖励值则长期停留在较低区间。由此可见，MMFQ-ACC不仅具有更快的收敛速度，学习到的策略质量也更高。

进一步比较收敛后的曲线波动范围可以发现，MMFQ-ACC的稳定性明显优于MFQ，即其奖励在稳态阶段的起伏更小。这一优势主要来自元学习机制：算法能够利用过往任务中积累的经验，在面对新干扰轨迹时快速完成策略调整，从而在时变干扰场景下维持更稳定的收益表现。同时，相较于忽视智能体交互的IQL，MMFQ-ACC通过平均场交互刻画多无人机耦合效应，因此在大规模干扰场景中体现出更强的整体性能。

图9展示了第100次训练回合中，各算法归一化奖励值随时隙的变化趋势。为便于观察，在2000

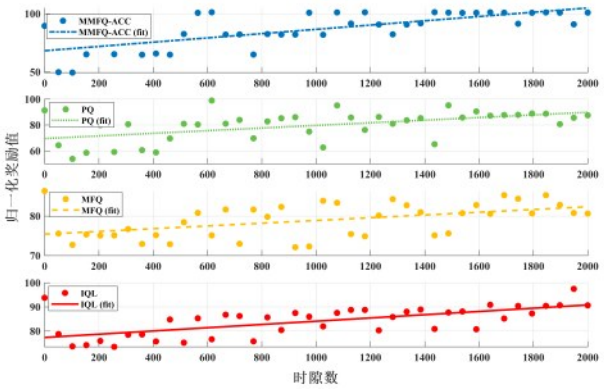

图9　奖励值随时隙数变化情况

个时隙内均匀抽取40个时隙作为采样点，并基于这些采样点得到线性回归拟合直线。MMFQ-ACC、MFQ、PQ和IQL的回归线斜率分别为1.6×10^{-3}、2×10^{-4}、1.9×10^{-4}和8×10^{-5}。由此可知，MMFQ-ACC的回归曲线斜率最大，意味着其在单个训练回合内单位时隙的收益增长速度更快，这与前文所体现的快速收敛特性一致。

相比之下，MFQ与PQ的斜率差异不大，说明在单训练回合中，仅依赖平均场近似对学习速率的优化相对有限。结合采样点分布进一步分析可知：MMFQ-ACC在回合初段的采样点离散程度较高，表明算法早期进行了广泛探索以搜索更好的选择策略，因而奖励波动较为明显；随着训练推进，算法后期采样点波动减弱，说明策略更新趋于稳定。综上所述，MMFQ-ACC能借助已学习到的元模型在新干扰环境下更高效地完成策略搜索与收敛，从而获得更快的回合内收益提升。

4.6　泛化性性能展示与分析

为验证MMFQ-ACC算法的泛化性能，对其在不同抗干扰任务下的表现进行了对比。训练阶段设置多条直线路径，不同斜率和起点代表不同任务；适应阶段采用圆形路径，通过改变圆心位置和半径生成新任务。直线路径与圆形路径分别表示训练阶段和适应阶段中干扰机的两类移动轨迹模式。训练任务由直线轨迹的斜率与起点生成，适应任务由圆形轨迹的圆心与半径生成，两者在任务空间中形成轨迹分布差异，以此评估MMFQ-ACC算法在不同抗干扰任务上的泛化性能。根据任务规模与复杂度，将实验划分为三类：

1) A类任务在训练阶段设置30条直线干扰路径，在适应阶段设置2条圆形干扰路径；

2) B类任务在训练阶段设置50条直线干扰路径，在适应阶段设置2条圆形干扰路径；

3) C类任务在训练与适应阶段分别设置50条直线干扰路径和50条圆形干扰路径。

三类任务场景的主要区别在于训练阶段与适应阶段所设置的干扰路径数量不同。随着训练阶段和适应阶段路径数量的增加，三类任务的空间覆盖范围和场景变化程度相应增大，复杂性由低到高依次为A类、B类和C类。

上述三类实验中，训练阶段的第q条直线路径参数记为

$$\boldsymbol{\psi}_q^{\mathrm{tr}}=(k_q,x_{0,q},y_{0,q}) \tag{34}$$

其中，斜率$k_q\sim\mathcal{U}_d(\mathcal{K})$，$\mathcal{U}_d(\cdot)$表示离散均匀分布，$\mathcal{K}=\{-0.5,-0.4,\cdots,0.4,0.5\}$；起点坐标$(x_{0,q},y_{0,q})$从集合$\mathcal{X}_0\times\mathcal{Y}_0$中等概率随机采样得到，其中$\mathcal{X}_0=\{0,500\}$、$\mathcal{Y}_0=\{25,50,75,\cdots,225\}$，上述坐标单位均为m。

因此，第q条直线路径对应的轨迹可表示为

$$y=k_q(x-x_{0,q})+y_{0,q},(x,y)\in[0,500]\times[0,250] \tag{35}$$

适应阶段的第q条圆形路径参数记为

$$\psi_q^{\mathrm{ad}} = (x_{c,q}, y_{c,q}, r_q) \tag{36}$$

其中，半径$r_q \sim \mathcal{U}_d(\mathcal{R})$，$\mathcal{R} = \{40,60,80\}$，单位为m；圆心坐标$(x_{c,q}, y_{c,q})$从集合$\mathcal{X}_c \times \mathcal{Y}_c$中等概率随机采样得到，其中$\mathcal{X}_c = \{100,150,200,250,300,350,400\}$、$\mathcal{Y}_c = \{80,125,170\}$，上述坐标单位均为m。

因此，第q条圆形路径对应的轨迹可表示为

$$(x - x_{c,q})^2 + (y - y_{c,q})^2 = r_q^2, (x,y) \in [0,500] \times [0,250] \tag{37}$$

需要说明的是，A、B类任务通过将适应阶段路径数压缩为2条，构造新场景样本稀缺的测试条件，用以检验所提出算法在少量交互下的快速适应能力，突出跨任务先验对新场景泛化的作用。

图10展示了MMFQ-ACC算法与其他传统算法的泛化性比较。在A、B类任务中，MFQ与PQ由于缺少元知识，在适应阶段的样本路径较少的情况下，奖励值始终较低且无明显增长，而MMFQ-ACC凭借训练阶段获得的元知识，能在新任务中迅速取得较高奖励。在C类任务中，MMFQ-ACC算法展现出更快的收敛特性，且其性能曲线波动显著降低，这表明通过引入多样化训练任务有效提升了模型对新干扰场景的泛化性能。MFQ和PQ由于获得了足够的适应阶段训练，奖励值也随训练轮次增长，但其收敛速度和最终收敛的奖励值均低于MMFQ-ACC。总体来看，MMFQ在多任务干扰环境下展现出突出的跨场景泛化优势。

5　结束语

本文针对动态复杂环境下的大规模多无人机抗干扰隐蔽通信问题，构建了多无人机自组网系统模型，并提出了一种元平均场抗干扰－隐蔽通信Q学习算法（MMFQ-ACC）。该算法在强化学习阶段引入被动监听者Willie的风险分析，对抗干扰性和隐蔽性进行联合优化。通过融合实时频谱感知与历史状态信息，MMFQ-ACC在复杂博弈环境中可快速求解近似纳什均衡。

仿真结果验证了该算法在收敛速度、稳定性及对新场景的泛化性方面均优于传统的MFQ、PQ和IQL算法，并能在大规模组网条件下实现高效分布式决策，实现了抗干扰性与隐蔽性之间的动态平衡。研究结果为无人机网络在动态复杂环境下的安全通信提供了新的理论与算法支撑，对军事与公共安全等敏感场景中的隐蔽通信具有重要参考价值。

需要说明的是，本文当前阶段主要聚焦于理论建模、算法设计和仿真验证，未来我们将基于实物仿真平台开展实验验证，以实现与本文仿真结果的对照与校验。与此同时，未来可进一步探索多监听者和多干扰机场景下的协同博弈机制，以及异构无人机网络中的跨层隐蔽优化问题，以实现更完善的多维安全通信体系。

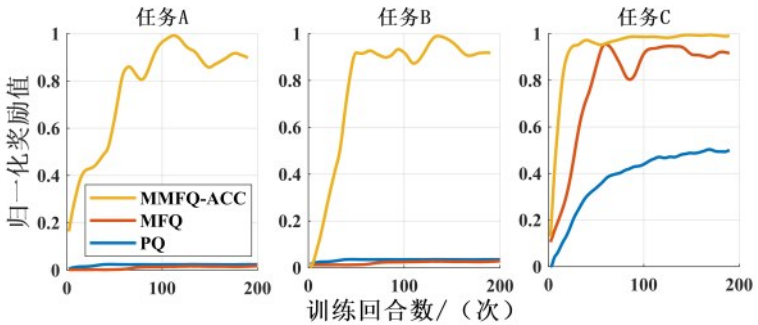

图10　不同任务场景下奖励值随训练回合数变化情况

参考文献:

[1] 陈好, 梁浩宇, 吴俊, 曹炜威. 感知时延约束下认知无人机网络的快速频谱感知[J]. 航空工程进展, 2025, 16 (05): 42-50.
Chen H, Liang H, Wu J, et al. Quick spectrum sensing for delay-constraint cognitive UAV networks[J]. Advances in Aeronautical Science and Engineering, 2025, 16(5): 42-50.

[2] Du L, An P, Tan Y, et al. Fast Opportunistic Spectrum Sensing and Throughput Maximization for Cognitive Unmanned Aerial Vehicle Communications[J]. IEEE Transactions on Vehicular Technology, 2025, 74(9): 14169-14181.

[3] 柏艳昆. 无线通信技术在自组网与无人机作业中的应用[J]. 中国新通信, 2024, 26(20): 4-6.
Bai Y. Application of wireless communication technology in ad hoc networks and UAV operations[J]. China New Telecommunications, 2024, 26(20): 4-6.

[4] Toh C K. Wireless ATM and ad-hoc networks: Protocols and architectures[M]. Springer Science & Business Media, 2012.

[5] Wang J, Liu W, Yang C. UAV-aided Maritime Communication Over the Pacific Ocean Using the Elevated Duct toward Future Wireless Networks[J]. Scientific Reports, 2025, 15: 11920.

[6] 张文秋, 丁文锐, 刘春辉. 一种无人机数据链信道选择和功率控制方法[J]. 北京航空航天大学学报, 2017, 43 (03): 583-591.
Zhang W, Ding W, Liu C. A channel selection and power control method of UAV data link[J]. Journal of Beijing University of Aeronautics and Astronautics, 2017, 43(3): 583-591.

[7] Wang X, Lei M, Zhao M, et al. Cooperative anti-jamming strategy and outage probability optimization for multi-hop ad-hoc networks[C]//2017 IEEE 86th Vehicular Technology Conference. IEEE, 2017: 1-5.

[8] Li A, Zhang W. Mobile jammer-aided secure UAV communications via trajectory design and power control[J]. China Communications, 2018, 15 (8): 141-151.

[9] Zhang J, Chen T, Zhong S, et al. Aeronautical Ad-Hoc networking for the Internet-above-the-clouds[J]. Proceedings of the IEEE, 2019, 107 (5): 868-911.

[10] Zhang Y, Zhang B, Yi X. Adaptive data sharing algorithm for aerial swarm coordination in heterogeneous network environments[C]//International Conference on Collaborative Computing: Networking, Applications and Worksharing. Cham: Springer International Publishing, 2018: 202-210.

[11] Tian J, Li E C. An improved rrt-connect algorithm used for uav 3d trajectory planning[J]. Advances in Aeronautical Science and Engineering, 2018, 9(4): 514-522.

[12] Wang Y K, Xiang L P, Liu J, et al. Deep Neural Polar Codes for Integrated Data and Energy Communication Networks Enabled by Sensing-Aided UAVs[J]. Journal of Communications and Information Networks, 2025, 10(4): 399-413.

[13] Liang X, Xu W, Gao H, et al. Throughput optimization for cognitive UAV networks: A three-dimensional-location-aware approach[J]. IEEE Wireless Communications Letters, 2020, 9(7): 948-952.

[14] Liu X, Guan M, Zhang X, et al. Spectrum sensing optimization in an UAV-based cognitive radio[J]. IEEE Access, 2018, 6: 44002-44009.

[15] Sarker I H. Machine learning: Algorithms, real-world applications and research directions[J]. SN computer science, 2021, 2(3): 160.

[16] Liu Q, Shi L, Sun L, et al. Path planning for UAV-mounted mobile edge computing with deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(5): 5723-5728.

[17] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.

[18] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International conference on machine learning. PMLR, 2017: 1126-1135.

[19] Hospedales T, Antoniou A, Micaelli P, et al. Meta-learning in neural networks: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(9): 5149-5169.

[20] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2009, 22(10): 1345-1359

[21] Wei Y T, Wu S, Ji Z, et al. Multi-UAV Collaborative Edge Computing Algorithm for Joint Task Offloading and Channel Resource Allocation [J]. Journal of Communications and Information Networks, 2024, 9(2): 137-150.

[22] Chai J, Wang Z, Ma C, et al. Personalized Federated Reinforcement Learning for Multi-AAV Assisted Edge Computing[J]. IEEE Wireless Communications Letters, 2025.

[23] Liu S, Xu Y, Chen X, et al. Pattern-aware intelligent anti-jamming communication: A sequential deep reinforcement learning approach[J]. IEEE Access, 2019, 7: 169204-169216.

[24] Chang X, Li Y, Zhao Y, et al. An improved anti-jamming method based on deep reinforcement learning and feature engineering[J]. IEEE Access, 2022, 10: 69992-70000.

[25] 廖程建, 刘思懿, 赵晨羽, 等. 基于多智能体强化学习的空地网络抗干扰传输方法研究[J]. 移动通信, 2024, 48(01): 71-78.
Liao C, Liu S, Zhao C, et al. Research on anti-jamming transmission method for air-ground networks based on multi-agent reinforcement learning[J]. Mobile Communications, 2024, 48(1): 71-78.

[26] 白恒志, 王海超, 李国鑫, 等. 无人机隐蔽通信网络研究综述[J]. 电信科学, 2023, 39(8):1-16.
Bai H, Wang H, Li G, et al. Review on unmanned aerial vehicle covert communication network[J]. Telecommunications Science, 2023, 39(8): 1-16.

[27] Makhdoom I, Abolhasan M, Lipman J. A comprehensive survey of covert communication techniques, limitations and future challenges[J]. Computers & Security, 2022, 120: 102784.

[28] Zhang L, Tan C, Yu F. Fast Decryption of Excel Document Encrypted by RC4 Algorithm[C]. International Conference on Communication Technology. IEEE, 2020: 1572-1576.

[29] Seong H, Kim T, Song J, et al. Hierarchical Multi-Agent Reinforcement Learning-Based UAV Control for Wireless Covert Communications[C]//2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC). IEEE, 2025: 1-6.

[30] Zhou X, Yan S, Hu J, et al. Joint optimization of a UAV's trajectory and transmit power for covert communications[J]. IEEE Transactions on Signal Processing, 2019, 67(16): 4276-4290.

[31] Hu J S, Wu L M, Shu F, et al. UAV-relay assisted covert communication with finite block-length[J]. Journal of Electronics & Information Technology, 2022, 44(3): 1006-1013.

[32] Zhou X B, Peng X, Yu H, et al. Artificial noise enhanced short-packet covert communications for UAV networks[J]. Journal of Signal Pro-

cessing, 2022, 38(8): 1601-1609.
[33] Pi X, Yang B, Shen Y, et al. UAV Relay-Enabled THz Covert Communications Against Colluding Detection[J]. Transactions on Emerging Telecommunications Technologies, 2025, 36(6): e70165.
[34] Ryu J Y, Lee J H. Relay Selection for Covert Communication with an Active Warden[J]. Sensors, 2025, 25(13): 3934.
[35] Hua B, Han L, Zhu Q, et al. Ultra-Wideband Nonstationary Channel Modeling for UAV-to-Ground Communications[J]. IEEE Transactions on Wireless Communications, 2025, 24(5): 4190-4204.
[36] Qi W, Bian J, Wang Z, et al. A Novel UAV Air-to-Air Channel Model Incorporating the Effect of UAV Vibrations and Diffuse Scattering[J]. Drones, 2024, 8(5): 194.
[37] Qiu Z, Chu X, Calvo-Ramirez C, et al. Low altitude UAV air - to - ground channel measurement and modeling in Semiurban environments[J]. Wireless Communications and Mobile Computing, 2017, 2017(1): 1587412.
[38] Qian Y, Yang C, Mei Z, et al. On joint optimization of trajectory and phase shift for IRS-UAV assisted covert communication systems[J]. IEEE Transactions on Vehicular Technology, 2023, 72(10): 12873-12883.
[39] Luong N C, Hoang D T, Gong S, et al. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133-3174.
[40] Shi Y, Konar A, Sidiropoulos N D, et al. Learning to Beamform for Minimum Outage[J]. IEEE Transactions on Signal Processing, 2018, 66(19): 5180-5193.
[41] Guan Y, Zhang Q, Tsiotras P. Learning nash equilibria in zero-sum stochastic games via entropy-regularized policy approximation[J]. arXiv preprint arXiv:2009.00162, 2020.
[42] Hu G, Zhu Y, Li H, et al. FM3Q: factorized multi-agent MiniMax Q-learning for two-team zero-sum Markov game[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(6): 4033-4045.
[43] Song F, Wang Z, Li J, et al. Dynamic Trajectory and Power Control in Ultra-Dense UAV Networks: A Mean-Field Reinforcement Learning Approach[J]. IEEE Transactions on Wireless Communications, 2025.6.
[44] Rao N, Xu H, Qi Z, et al. Adaptive jamming decision-making against FHSS communications via inexpert demonstrations assisted meta reinforcement learning[J]. IEEE Communications Letters, 2024.
[45] Chen C, Dong D, Li H X, et al. Fidelity-based probabilistic Q-learning for control of quantum systems[J]. IEEE transactions on neural networks and learning systems, 2013, 25(5): 920-933.
[46] Matignon L, Laurent G J, Le Fort-Piat N. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems[J]. The Knowledge Engineering Review, 2012, 27(1):1-31.

钱玉文（1975−），男，江苏南京人，博士，南京理工大学副教授，主要研究方向为无线、量子隐蔽通信和智能通信。

吴浩阳（2003−），男，江苏南京人，南京理工大学硕士研究生，主要研究方向为隐蔽通信和语义通信。

时龙（1985−），男，安徽阜阳人，博士，南京理工大学教授，主要研究方向为无线通信网络、分布式人工智能和网络安全。

曹阳（1994−），女，河北故城人，博士，南京理工大学副教授，主要研究方向为智能超表面、通感一体化和物理层安全。

王喆（1986−），女，河南郑州人，博士，南京理工大学教授，主要研究方向为无线通信网络、移动边缘计算、区块链技术和人工智能。

陈光霁（1993−），男，安徽安庆人，博士，南京理工大学副教授，主要研究方向为 6G 无线通信、智能反射面、可重构天线系统；通信感知一体化。

韦康（1991−），男，江苏南京人，博士，东南大学副教授，主要研究方向为人工智能安全与隐私、无线通信网络和云边端网络优化。

马川（1990–），男，河南郑州人，博士，重庆大学副教授，主要研究方向为分布式计算和人工智能安全。