有人机/无人机智能协同目标搜索和轨迹规划算法

卢卓，吴启晖，周福辉
（南京航空航天大学电子信息工程学院，江苏 南京 210016）

摘 要：基于有人机/无人机智能协同平台，针对多个位置未知的干扰信号源搜索及轨迹规划进行了研究。考虑到搜索过程的实时性和动态性，提出了一种基于多智能体深度强化学习的有人机/无人机智能协同目标搜索和轨迹规划（MUICTSTP）算法。各无人机通过感知接收干扰信号强度在线决策轨迹规划，同时将感知信息和决策动作传给有人机来获得全局评估。仿真结果表明，该算法相比其他算法在长期接收干扰信号强度、碰撞等方面表现出更好性能，且获得更优的学习策略。

关键词：有人机/无人机；智能协同；多智能体深度强化学习；轨迹规划；接收干扰信号强度
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024003

Algorithm for intelligent collaborative target search and trajectory planning of MAV/UAV

LU Zhuo, WU Qihui, ZHOU Fuhui
College of Electronic and Information Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China

Abstract: Based on the manned aerial vehicle (MAV) / unmanned aerial vehicle (UAV) intelligent cooperation platform, the search of multiple interfered signal sources with unknown locations and trajectory planning were studied. Considering the real-time and dynamic nature of the search process, a MAV/UAV intelligent collaborative target search and trajectory planning (MUICTSTP) algorithm based on multi-agent deep reinforcement learning (MADRL) was proposed. Each UAV made online decision on trajectory planning by sensing the received interference signal strength (RISS) values, and then transmitted the sensing information and decision-making actions to the MAV to obtain the global evaluation. The simulation results show that the proposed algorithm exhibits better performance in long-term RISS, collision, and other aspects compared to other algorithms, and the learning strategy is better.

Keywords: MAV/UAV, intelligent collaborative, MADRL, trajectory planning, RISS

0 引言

随着无线网络迅速发展，无线电用频设备日益增加，同时非法用频手段也趋向于多样化，频谱非法使用给无线通信造成了极大困扰[1-2]。因此，快速有效地确定非法干扰信号源位置并对其实现高效管控，对保障无线电通信安全至关重要。传统定位技术分为有源定位技术和无源定位技术[3-4]，其中，有源定位技术需要自身发射电磁波，容易暴露自身位置致使安全性能较低，而无源定位技术不需要自身发射电磁信号，具有很好的隐蔽性和防电磁干扰能力[5-7]。因此，本文考虑应用无源定位技术来确定非法干扰信号源的位置。

收稿日期：2023−07−04；**修回日期：**2023−09−20
通信作者：吴启晖，wuqihui2014@sina.com
基金项目：江苏省基础研究计划自然科学基金资助项目（No.BK20222013）；江苏省科研与实践创新计划基金资助项目（No.KYCX22_0358）
Foundation Items: Basic Research Program Natural Science Foundation of Jiangsu Province (No.BK20222013), The Postgraduate Research and Practice Innovation Program of Jiangsu Province (No.KYCX22_0358)

得益于可视距通信优势和高移动性[8]，无人机能有效避开干扰信号源附近障碍物对信号的遮挡，从而减少多径效应。因此，近年来有不少研究人员在基于无人机（UAV, unmanned aerial vehicle）平台的无源定位方法上展开了研究并取得一定成果[9-16]。文献[13]利用携带到达角阵列天线的无人机进行了移动终端定位实验，该无人机是高空长航时平台的开发原型。文献[14]提出一种分布式控制算法，利用多架配置全球定位系统（GPS, global position system）和到达角传感器的无人机进行协同定位，无人机使用最优递归估计技术来最小化目标定位误差，使用微分几何技术来生成无人机航行轨迹。文献[15]针对无人机试图使用无源有效载荷传感器对发射器进行地理定位的问题，提出了一种路径规划算法。文献[16]给无人机配备了一个定向天线，放置在倾斜定位器上以测量和收集用于到达方向（DOA, direction of arrival）估计的接收信号强度（RSS, received signal strength），提出了一种基于图像处理的DOA 估计方法。上述研究均针对单纯的无人机系统，而有人机（MAV, manned aerial vehicle）/无人机自主协同系统中人类智能与机器智能的交互融合更有利于实现有人机与无人机的互补，执行复杂任务时能够更好地适应以人类目标为导向的优化[17]。本文利用有人机/无人机智能协同感知干扰信号源位置来进行轨迹规划，其中，无人机可以根据自身的观测做出自主决策，并根据有人机的评估对决策进行优化，有人机会根据无人机提供的观测和动作信息对自身的评估进行优化更新，两者实现互补。

在有人机控制多架无人机协同搜索多个干扰信号源的过程中，为了提升搜索效率，需要根据感知的不同干扰信号强度，将干扰任务合理分配给各无人机。在干扰信号源位置未知的情况下，无人机可以利用无源定位技术通过感知接收干扰信号强度（RISS, received interference signal strength）来定位各个信号源的位置[18]，在此基础上进行轨迹规划。有人机/无人机协同规划是一个具有多个参数和约束条件的非确定性多项式（NP, nondeterministic polynomial），属于 NP-hard 问题[19]。它的解决方案空间会随着目标数量呈指数级增长[20]。传统方法需要较多先验参数信息，而深度强化学习不需要先验信息，能通过与环境进行交互来优化自身策略[21]。

因此，本文利用多智能体深度强化学习（MADRL, multi-agent deep reinforcement learning）方法来解决基于有人机/无人机智能协同平台的干扰信号源搜索及轨迹规划的问题，该方法能够在干扰信号源位置未知和高动态环境下，使无人机通过感知不同 RISS 进行轨迹规划，来实现对干扰信号源的搜索定位。本文的主要研究工作如下。

1) 为了适应有人机/无人机协同搜索任务，将有人机/无人机协同目标搜索和轨迹规划（MUTSTP, MAV/UAV collaborative target search and trajectory planning）的问题转化为多智能体协作完全任务，建立了非完全信息决策的部分可观测马尔可夫决策过程（POMDP, partially observable Markov decision processes）。在所设计的奖励函数中，综合考虑最大化长期 RISS 的同时还需避免航迹中的碰撞带来的负奖励。

2) 为了适应有人机协同无人机探索目标的训练环境的不稳定性，构建了适用于有人机/无人机协同场景的集中训练与分布式执行的决策网络架构。基于此架构，单架无人机在做出决策时不受其他无人机的影响，可以通过直接与有人机进行信息交换获得更精准的全局策略评估。同时，减少了无人机之间因信息交换带来的时延和损耗，更符合有人机直接控制无人机决策场景的需求。

3) 考虑到搜索环境的动态性以及未知性，根据有人机/无人机智能协同架构提出基于 MADRL 的有人机/无人机智能协同目标搜索和轨迹规划（MUICTSTP, MAV/UAV intelligent collaborative target search and trajectory planning）算法，并制定了 POMDP 的关键要素，同时对该算法计算复杂度进行了分析。在相同条件设置下，进一步探索了适合的学习率，并将 MUICTSTP 算法与其他基准算法在总奖励、单架无人机奖励、RISS 和碰撞等性能上进行了对比，来验证 MUICTSTP 算法能为有人机/无人机智能协同目标搜索和轨迹规划提供更优的学习策略。

1 系统模型

如图 1 所示，本文构建的有人机/无人机智能协同系统中有人机负责评估多架无人机的自主决策，有人机派出几架无人机对多个干扰信号源进行感知探测，并将探测后信息快速传给有人机，有人机根据无人机反馈信息给无人机下达评估指令，无人机之间不需要进行通信就可通过与有人机交互信息获得对全局信息的观测评估。本文系统考虑有人机数量为 M，无人机数量为 U，干扰信号源数量为 N。各干扰信号源发射频率不一致，各无人机通

过使用不同的窄带接收机对不同频率的干扰信号源进行感知接收，最终协同飞往合理分配的干扰信号源。系统参数及其含义如表 1 所示。

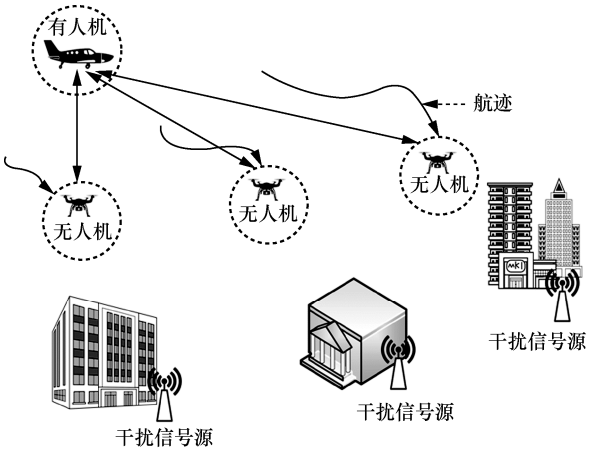

图 1　有人机/无人机智能协同系统

表 1　　系统参数及其含义

参数	含义
M,U,N	有人机数量、无人机数量、干扰信号源数量
$d_{ij}(t)$	无人机 i 与干扰信号源 j 在时刻 t 的距离
$P_{ij}(t)$	无人机 i 与干扰信号源 j 在时刻 t 的瞬时功率
$L,G^{\mathrm{T}},G^{\mathrm{R}}$	路径损耗、发射天线增益、接收天线增益
$\theta_i^{\pi},\theta_{m,i}^{Q}$	actor 网络权重、critic 网络权重
$\theta_i^{\pi'},\theta_{m,i}^{Q'}$	target actor 网络权重、target critic 网络权重
B	经验回放池大小
N_{b}	样本大小
φ	样本索引
τ	软更新率
γ	折扣因子

1.1　RISS 估计模型

无人机感知接收干扰信号强度模型如图 2 所示。

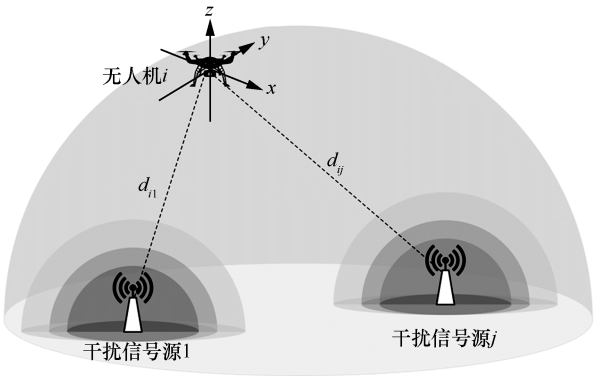

图 2　无人机感知接收干扰信号强度模型

无人机 i 在时刻 t 坐标位置为 $u_i(t)=[x_{\mathrm{uav}}^i(t),y_{\mathrm{uav}}^i(t),z_{\mathrm{uav}}^i(t)]\in\mathbb{R}^3,i\in U,t\in T$，干扰信号源 j 在时刻 t 坐标位置为 $\psi_j(t)=[x_{\mathrm{ins}}^j(t),y_{\mathrm{ins}}^j(t),z_{\mathrm{ins}}^j(t)]\in\mathbb{R}^3$，$j\in N,t\in T$。因此，无人机 i 与干扰信号源 j 在时刻 t 的距离为 $d_{ij}(t)=\sqrt{\left(u_i(t)-\psi_i(t)\right)^2}$。

该系统中无人机均配备了全向天线来测量各干扰信号源的 RISS。因此，在时刻 t 无人机 i 接收干扰信号源 j 的瞬时功率值可以表示为

$$P_{ij}(t)=P_j^{\mathrm{T}}G_j^{\mathrm{T}}G_{i,j}^{\mathrm{R}}\left(h_{ij}(t)\right)^2+\left(\eta_i(t)\right)^2 \tag{1}$$

其中，P_j^{T} 表示干扰信号源 j 的发射功率，G_j^{T} 表示干扰信号源 j 的发射增益，$G_{i,j}^{\mathrm{R}}$ 表示无人机 i 对干扰信号源 j 的接收天线增益，$h_{ij}(t)$ 表示 t 时刻无人机 i 与干扰信号源 j 之间的信道功率增益，$\eta_i(t)\sim\mathbb{N}(0,\sigma^2)$ 表示加性白高斯噪声[22]。

在自由空间传播模型中，$\left(h_{ij}(t)\right)^2$ 可以表示为

$$(h_{ij}(t))^2=\frac{\lambda^2}{(4\pi d_{ij}(t))^2L} \tag{2}$$

其中，λ 为波长，$L\geqslant 1$ 为自由空间路径损耗，计算式如下

$$\begin{aligned}&L=32.44+20\lg d_{ij}(t)+20\lg f_j\\&i\in U,j\in N\end{aligned} \tag{3}$$

其中，f_j 为干扰信号源 j 的工作频率。

1.2　问题描述

本文系统目标是通过有人机/无人机智能协同感知干扰信号强度来进行轨迹规划，在最大化长期 RISS 的同时还需要避免和其他无人机发生碰撞。可以将其描述为一个复杂组合优化问题，包括 2 个子问题，即目标搜索（TS, target search）和轨迹规划（TP, trajectory planning）。该问题可以表示为

$$\begin{aligned}&\mathrm{P1}:\max_{i\in U,j\in N}\sum_{t=0}^{T}P_{ij}(t)\\&\text{s.t.}\quad \mathrm{C1}:\left\|u_i(t+1)-u_i(t)\right\|=V\delta_t,\forall t\in T\\&\qquad \mathrm{C2}:\left\|u_x(t)-u_y(t)\right\|\geqslant\delta_d,\forall t\in T,\\&\qquad\quad (x,y)\in U,x\neq y\\&\qquad \mathrm{C3}:u_i(t)\in\mathbb{R}^3,\forall t\in T\\&\qquad \mathrm{C4}:\psi_{j_1}(t)\neq\psi_{j_2}(t),\forall t\in T,\\&\qquad\quad (j_1,j_2)\in N,j_1\neq j_2\end{aligned} \tag{4}$$

其中，约束 C1 保证无人机的 t+1 时刻的位置状态是由 t 时刻的状态和 t 时刻的移动动作生成的，约束 C2 保证任意 2 架无人机 x 与 y 在航行过程中无碰撞，约束 C3 保证无人机 i 在任意时刻的航行边界在一定范围内，约束 C4 保证一架无人机只飞往一个干扰信号源。

2 基于 MADRL 的有人机/无人机智能协同目标搜索和轨迹规划算法

2.1 MUICTSTP 构建

本文系统的目标函数属于非凸优化的 NP-hard 问题，用传统的优化算法很难解决，本文采用 MADRL 方法解决该问题。该方法不需要先验信息，智能体通过与环境进行交互，根据环境反馈奖励来不断优化自身的学习策略，在不需要将非凸问题转化为凸问题的情况下最大化长期累积奖励期望，这种学习策略可以很好地解决本文研究问题。

本文将有人机/无人机协同目标搜索和轨迹规划问题建模为多智能体完全协作任务，考虑到实际环境下每架无人机在执行任务过程中的非完全信息决策属性，将其定义为 POMDP[23]，用 (S,A,R,P,γ) 表示，其中，S 表示全局环境的状态空间，A 表示所有智能体的动作集合，R 表示所有智能体的环境奖励集合，P 表示状态转移概率，γ 表示环境奖励的折扣因子，所有智能体在当前状态下采取动作获得下一个状态。在每个时隙下环境状态为 $s(t)$，智能体 i 只能接收局部观测 $o_i(t)=b_i(s(t))$，并根据局部观测选择动作 $a_i(t)=\pi_i(o_i(t))$，其中 b_i 和 π_i 表示智能体 i 的观测函数和策略。在选择动作后，智能体 i 会根据奖励函数设置从环境中获得奖励 $r(t)=\{r_1(t),r_2(t),\cdots,r_M(t)\}$，然后环境根据状态转移函数 p_i 转移到下一个状态 $s(t+1)$。

MUICTSTP 算法框架如图 3 所示，该系统中智能体包括有人机和无人机。无人机 i 有一个 actor 网络 $\pi_i(o_i;\theta_i^{\pi})$ 用于分布执行，其中，θ_i^{π} 为 actor 网络的权重，无人机 i 的局部观测 o_i 为 actor 网络输入，动作 a_i 为 actor 网络输出。无人机 i 的 actor 网络有对应的 target 网络 $\pi_i'(o_i';\theta_i^{\pi'})$ 与其共享相同网络架构；有人机 m 有一个 critic 网络 $Q_{m,i}\ (O_{m,i},A_{m,i};\theta_{m,i}^{Q})$ 用于集中训练，其中，$O_{m,i}=\{o_1,\cdots,o_i\}$，$A_{m,i}=\{a_1,a_2,\cdots,a_i\}$，$m\in M,i\in U$。通过 Q 值来评估 actor 网络的输出，$\theta_{m,i}^{Q}$ 表示有人机 m 对无人机 i 进行评估的 critic 网络权重。有人机的 critic 网络有对应的 target 网络 $Q'_{m,i}\left(O'_{m,i},A'_{m,i};\theta_{m,i}^{Q\,\prime}\right)$ 与其共享相同网络架构。每个时刻 t 每架无人机均会将经验值 $(o(t),a(t),r(t),o(t+1))$ 存储到有人机的大小为 B 的经验回放池 D 中，存储空间已满则用新的经验元组替换旧的经验元组。从经验回放池中批量采样对无人机的 actor 网络和有人机的 critic 网络进行训练，其中，随机样本打破了序列样本间的相关性，能有效减少训练振荡。

有人机 m 的 critic 网络用最小化均方误差（MSE, mean squared error）损失来更新，即

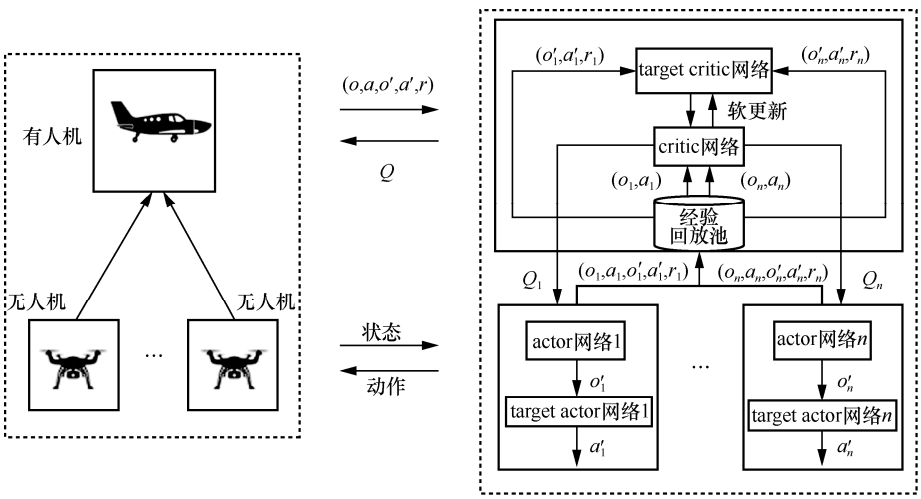

图 3 MUICTSTP 算法框架

$$L(\theta_{m,i}^{Q})=\frac{1}{N_{\mathrm{b}}}\sum_{\varphi=1}^{N_{\mathrm{b}}}\left[y_{m,i}^{\varphi}-Q_{m,i}(O_{m,i}^{\varphi},A_{m,i}^{\varphi};\theta_{m,i}^{Q})\right]^{2},i\in U \quad (5)$$

其中，N_{b} 为 batchsize，φ 为样本索引，且

$$y_{m,i}^{\varphi}=r_i^{\varphi}+\gamma Q_{m,i}'\left(O_{m,i}^{\varphi}{}',\left\{\pi_i'\left(O_{m,i}^{\varphi}{}';\theta_i^{\pi'}\right)\right\};\theta_{m,i}^{Q}{}'\right) \quad (6)$$

无人机 i 的 actor 网络可以通过最小化损失来更新

$$L(\theta_i^{\pi})=\frac{1}{N_b}\sum_{\varphi=1}^{N_b}-Q_{m,i}(O_{m,i}^{\varphi},A_{m,i}^{\varphi};\theta_{m,i}^{Q}) \quad (7)$$

其中，$A_{m,i}^{\varphi}=\pi_i\left(O_{m,i}^{\varphi};\theta_i^{\pi}\right)$。

目标网络参数通过跟踪学习网络更新，即

$$\theta_i^{\pi'}\leftarrow\tau\theta_i^{\pi}+(1-\tau)\theta_i^{\pi'} \quad (8)$$

$$\theta_{m,i}^{Q}{}'\leftarrow\tau\theta_{m,i}^{Q}+(1-\tau)\theta_{m,i}^{Q}{}' \quad (9)$$

其中，$\tau\ll 1$。

2.2 MUICTSTP 详细描述

根据本文研究的有人机/无人机智能协同目标搜索和轨迹规划，将 POMDP 的元素具体定义如下。

2.2.1 观测空间

本文系统中的无人机之间是合作关系，各无人机只能获得局部观测。无人机可以通过 GPS 来获得自身位置信息，通过与有人机进行信息交互来获得其他无人机的位置信息。为了驱使无人机飞往未知位置坐标的干扰信号源，无人机的观测信息还包含在任意时刻 t 对各干扰信号源感知的 RISS。

因此，无人机 i 的观测空间为

$$o_i(t)\triangleq\left\{u_i(t),u_j(t),p_{i,k}(t)\right\},(i,j)\in U,i\neq j,k\in N,t\in T \quad (10)$$

有人机 m 的观测空间是当前时刻下所有无人机的状态和动作信息，可以表示为

$$o_m(t)\triangleq\left\{O_{m,i}(t),A_{m,i}(t)\right\},i\in U,m\in M,t\in T \quad (11)$$

2.2.2 动作空间

无人机通过感知 RISS 来进行轨迹规划飞往目标干扰源位置，通过改变各方向上的速度来改变轨迹。

对 xy 平面和正 z 轴方向的物理移动动作分别进行归一化处理，然后进一步获得 xy 平面速度 $v_{xy}^{i}(t)$ 和正 z 轴方向速度 $v_z^{i}(t)$。其中，$-v_{xy}^{\max}(t)<v_{xy}^{i}(t)<v_{xy}^{\max}(t)$ 和 $-v_z^{\max}(t)<v_z^{i}(t)<v_z^{\max}(t)$，$v_{xy}^{\max}(t)$ 和 $v_z^{\max}(t)$ 分别是 xy 平面和正 z 轴方向的最大速度。因此，无人机 i 的动作空间可以表示为

$$a_i(t)\triangleq\left\{v_{xy}^{i}(t),v_z^{i}(t)\right\} \quad (12)$$

因此，无人机 i 的位置状态和物理移动动作的关系可以表示为

$$\left(x_{\mathrm{uav}}^{i}(t+1),y_{\mathrm{uav}}^{i}(t+1)\right)=\left(x_{\mathrm{uav}}^{i}(t),y_{\mathrm{uav}}^{i}(t)\right)+v_{xy}^{i}(t)\delta(t) \quad (13)$$

$$z_{\mathrm{uav}}^{i}(t+1)=z_{\mathrm{uav}}^{i}(t)+v_z^{i}(t)\delta(t) \quad (14)$$

其中，$\delta(t)$ 表示时隙 t 的时间间隔长度。

而有人机 m 的动作是对每架无人机状态−动作的评估，可以表示为

$$a_{m,i}(t)\triangleq\left\{Q_{m,i}(t)\right\},i\in U,t\in T \quad (15)$$

2.2.3 奖励函数

无人机 i 在搜索目标过程中需要根据环境奖励来进行轨迹规划，其获得的 RISS 奖励可以定义为

$$r_i^{\mathrm{RISS}}(t)=k_pP_{ij}^{(\theta,\omega)}(t) \quad (16)$$

其中，k_p 是一个正常数，用于调节 RISS 奖励部分，这里 $k_p=10^8$。

如果无人机 i 移动超出 xy 平面和正 z 轴方向设定的边界范围则 d^{bound} 则获得负奖励，可以表示为

$$\text{if}\quad\left|[u_i(t)]-\left[x^d(t),y^d(t),z^d(t)\right]\right|\leqslant d^{\mathrm{bound}} \quad (17)$$

$$r_i^{\mathrm{bound}}(t)=-c^{\mathrm{bound}} \quad (18)$$

其中，$[x^d(t),y^d(t),z^d(t)]$ 分别代表边界坐标，c^{bound} 为根据场景设置的边界奖励常数，这里 $c^{\mathrm{bound}}=100$。

任意 2 架无人机之间的距离 $\left\{d_{i,j}\right\}_{(i,j)\in M,i\neq j}$ 小于设定的安全距离 d^{safe} 获得的负奖励可以表示为

$$r_i^{\mathrm{safe}}(t)=-c^{\mathrm{safe}} \quad (19)$$

其中，c^{safe} 为根据场景设置的安全奖励常数，这里 $c^{\mathrm{safe}}=100$。

因此，无人机 i 在任意时刻 t 获得的总奖励表示为

$$r_i(t)=r_i^{\text{RISS}}(t)+r_i^{\text{bound}}(t)+r_i^{\text{safe}}(t) \tag{20}$$

2.3 训练算法

本节给出了用来解决 MUTSTP 问题的 MUICTSTP 算法伪代码，如算法 1 所示。

算法 1 MUICTSTP 算法伪代码

1) 初始化网络，包括 U 架无人机的 actor 网络 $\pi_i(o_i;\theta_i^{\pi}),i\in U$ 和对应的权重为 $\theta_i^{\pi'}=\theta_i^{\pi}$ 的 target actor 网络，以及有人机的 critic 网络 $Q_{m,i}=(O_{m,i},A_{m,i};\theta_{m,i}^{Q}),i\in M$ 和对应的权重为 $\theta_{m,i}^{Q}{}'=\theta_{m,i}^{Q}$ 的 target critic 网络。

2) 初始化经验回放池 D。

3) 循环每一个回合。

4) 初始化有人机、无人机以及干扰信号源的三维位置坐标。

5) 循环回合中的每一步。

6) 无人机获得自身位置信息、当前时隙感知 RISS 以及有人机交互获得其他无人机位置信息来构成自身观测状态 $o_i(t)$，有人机获得所有无人机的状态 $O_{m,i}(t)=\{o_1(t),o_2(t),\cdots,o_i(t)\},i\in U$ 和对应的动作信息 $A_{m,i}(t)=\{a_1(t),a_2(t),\cdots,a_i(t)\},i\in U$。

7) 每架无人机 i 的动作选择 $a_i(t)=\pi_i(o_i(t);\theta_i^{\pi})+\mathbb{N}_t$ 都是根据当前的学习策略和观测得到的，其中，$\mathbb{N}_t$ 是探测噪声。有人机 m 的动作是对无人机的评估 $a_{m,i}(t)=Q_{m,i}\ (O_{m,i}(t),A_{m,i}\ \ (t);\theta_{m,i}^{Q})$。

8) 每架无人机 i 的动作都会根据奖励函数设置获得环境奖励 $r_i(t)=\{r_1(t),r_2(t),\cdots,r_M(t)\}$ 以及下一步状态 $o_i(t+1)$。

9) 将每一步中每架无人机 i 的经验信息 $(o_i(t),a_i(t),r_i(t),o_i(t+1))$ 存入经验回放池 D 中。

10) 训练每架无人机，开始循环无人机。

11) 从经验回放池 D 中随机抽取少量样本 N_{b}。

12) 根据式(6)计算有人机的 critic target。

13) 通过式(5)最小化 MSE 损失来更新有人机的 critic 网络 $\theta_{m,i}^{Q}$。

14) 通过式(7)最小化损失来更新无人机的 actor 网络 θ_i^{π}。

15) 目标网络的软更新

16) $\theta_{m,i}^{Q}{}'\leftarrow\tau\theta_{m,i}^{Q}+(1-\tau)\theta_{m,i}^{Q}{}'$。

17) $\theta_i^{\pi'}\leftarrow\tau\theta_i^{\pi}+(1-\tau)\theta_i^{\pi'}$。

18) 循环结束。

19) 循环结束。

20) 循环结束。

在训练阶段，每个回合的开始有人机和无人机的位置、干扰信号源位置都被随机初始化。在时刻 t，无人机 i 通过 actor 网络 $\pi_i(o_i(t);\theta_i^{\pi}(t))$ 选择动作 a_i，添加探测噪声 $\mathbb{N}_t$ 来防止智能体陷入局部最优策略。然后，无人机 i 获得下一个状态 $o_i(t+1)$ 和环境奖励 $r_i(t)$，并在经验回放池 D 中存储相应的经验元组 $(o_i(t),a_i(t),r_i(t),o_i(t+1))$，再从 D 中统一采样批次的经验元素，并通过最小化 actor 与 critic 的损失来更新网络。最后，根据算法中第 16)～17)行来软更新目标网络。

2.4 计算复杂度分析

MUICTSTP 算法计算复杂度由 critic 网络和 actor 网络共同决定。假设 critic 网络隐藏层是数量为 H_{c} 的全连接层，第 h 层的隐藏层含 n_h^{c} 个神经元。输入层神经元个数由全体无人机的观测状态 $\{o_i,i\in U\}$ 和动作 $\{a_i,i\in U\}$ 的总维度 $\{U\text{Dimen}(o_i,a_i),i\in U\}$ 决定，为 $U(U+N+3)$。输出层神经元个数为 1。因此，critic 网络总神经元个数为 $U(U+N+3)n_1^{\text{c}}+\sum_{h=2}^{H_{\text{c}}}n_{h-1}^{\text{c}}n_h^{\text{c}}+n_{H_{\text{c}}}^{\text{c}}$。actor 网络隐藏层是数量为 H_{a} 的全连接层，第 h 层的隐藏层含 n_h^{a} 个神经元。输入层由无人机自身局部观测状态维度 $\{\text{Dimen}(o_i),i\in U\}$ 决定，为 $U+N$。输出层神经元个数由动作维度 $\{\text{Dimen}(a_i),i\in U\}$ 决定，为 3。因此，actor 网络总神经元个数为 $(U+N)n_1^{\text{a}}+\sum_{h=2}^{H_{\text{a}}}n_{h-1}^{\text{a}}n_h^{\text{a}}+3n_{H_{\text{a}}}^{\text{a}}$。若训练一个神经元权重的计算复杂度为 W_i，则 MUICTSTP 算法的计算复杂度为 $O(W_i[U(U+N+3)n_1^{\text{c}}+\sum_{h=2}^{H_{\text{c}}}n_{h-1}^{\text{c}}n_h^{\text{c}}+n_{H_{\text{c}}}^{\text{c}}+(U+N)n_1^{\text{a}}+\sum_{h=2}^{H_{\text{a}}}n_{h-1}^{\text{a}}n_h^{\text{a}}+3n_{H_{\text{a}}}^{\text{a}}])$。因此，MUICTSTP 算法的计算复杂度与无人机数量 U 和干扰信号源数量 N 呈正相关。

3 仿真分析

本节通过仿真实验评估所提出的基于 MADRL 的 MUICTSTP 算法的有效性。

3.1　仿真设置

本节实验的仿真模拟训练环境中设置了一架有人机、3 架无人机和 3 个干扰信号源。每个回合的有人机、无人机以及干扰信号源的水平位置坐标在 2 000 m×2 000 m 区域内随机生成。有人机和干扰信号源设置为固定高度，无人机 i 飞行的高度上限 $z_i^{\max}=130$，下限 $z_i^{\min}=100$。无人机最大飞行速度 $v_i^{\max}=20\ \text{m/s}$，最大加速度 $a_i^{\max}=5\ \text{m/s}^2$。无人机和干扰信号源均设置全向天线，干扰信号源发射功率为 20 W，发射天线增益为 1，波长为 3 m，无人机接收天线增益为 1。实验中，Adam 优化器学习速率 $\alpha=0.005$或0.01，目标网络软更新率 $\tau=0.01$，折扣因子 $\gamma=0.95$；经验回放池大小 $B=200\,000$，batchsize = 1 024，每隔 100 个时间步长采样一次；每个回合的步长为 100，总回合数为 50 000。

3.2　网络架构

无人机 actor 网络架构如图 4 所示。2 个隐藏层神经元数量均为 64。输入层维度包括智能体自身的 xy 平面速度维度（为 2）和位置维度（为 2）、正 z 轴方向的速度维度（为 1）和高度维度（为 1），以及其他智能体相对距离维度（为 2×(2+1)=6），感知的各干扰信号源 RISS 的维度（为 3）。因此，无人机 i 的 actor 网络的输入状态总维度为 2+2+1+1+6+3=15。输出层维度为无人机 i 的三维物理移动动作维度（为 5+2=7），包括 xy 平面的动作（上、下、左、右、停止）以及正 z 轴方向的动作（上、下）。其中，输入层和隐藏层的激活函数为 ReLU，输出层的激活函数为 Softmax。

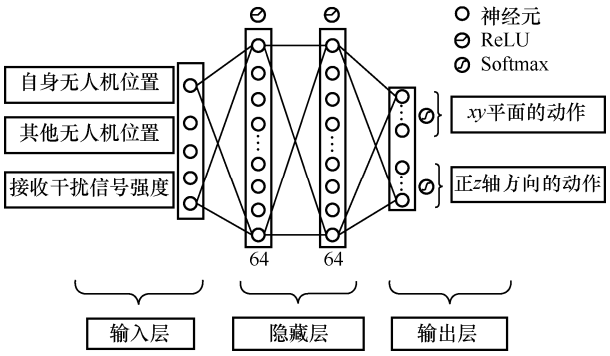

图 4　无人机 actor 网络架构

有人机 critic 网络架构如图 5 所示，其中隐藏层和 actor 网络一致，但是输入层的维度是所有无人机的状态维度（3×15）和动作维度（3×7），输出为评估每架无人机状态−动作的 Q 值。

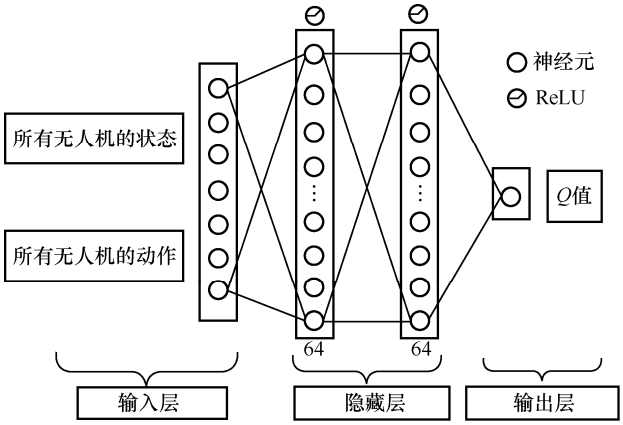

图 5　有人机 critic 网络架构

环境奖励和无人机 i 感知的当前 RISS、无人机与其他无人机之间的碰撞以及无人机是否超出飞行边界设置相关。

3.3　仿真结果

将所提出的 MUICTSTP 算法与其他 2 种基准测试算法进行比较，对比算法如下。

1) 深度 Q 网络（DQN, deep Q network）：受到文献[24]启发，单架无人机只关注自身获得的奖励。

2）深度确定性策略梯度（DDPG, deep deterministic policy gradient）：受到文献[25]启发，有人机只根据单架无人机的状态和动作信息做出评估。

在相同场景和固定随机种子设置下，通过训练将 MUICTSTP 算法与其他 2 种基准算法进行了性能对比分析。在所有仿真中，仿真结果取 100 个回合的平均值。

不同学习率下不同算法的总奖励值与回合数的关系如图 6 所示。从图 6 的总奖励值来看，相同场景设置下学习率为0.01相比学习率为0.005更适用于 3 种算法的场景训练，能获得更高奖励回报。在相同学习率的情况下，MUICTSTP 算法在经过充分训练后总奖励值增加幅度均大于其他 2 种基准算法。当学习率为 0.01 时，MUICTSTP 算法前 10 000 个回合的总奖励值为 533 169.17，最后 10 000 个回合的总奖励值为 550 571.62，总奖励值增加了 3.26%；DDPG 算法前 10 000 个回合的总奖励值为 518 306.88，最后 10 000 个回合的总奖励值为 510 837.53，总奖励值减少了 1.44%。DQN 算法前 10 000 个回合的总奖励值为 469 765.46，最后 10 000 个回合的总奖励值为 480 371.53，总奖励值增加了 2.25%。其原因是 MUICTSTP 算法中有人

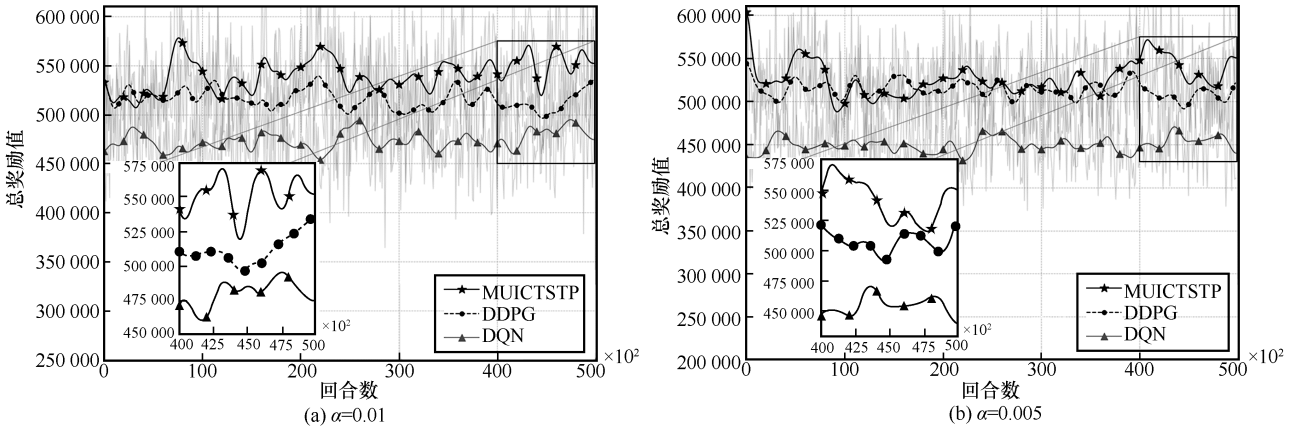

图 6　不同学习率下不同算法的总奖励值与回合数的关系

机对无人机动作的评估依据全局信息，更有利于无人机的学习策略更新。DDPG 算法的评估只根据单架无人机的信息，使评估不够全面，无法给无人机更全面的指导。DQN 算法中每个智能体只能关注自身的奖励，导致所有智能体的总奖励损失。

总奖励值变化是由各个智能体在求解目标搜索和航迹规划问题中的学习策略所决定的。因此，进一步分析了不同算法中不同智能体的奖励收敛情况。在学习率为 0.01 的情况下，不同算法中不同无人机的奖励值如图 7 所示。从图 7(a)可以看出，MUICTSTP 算法下无人机 1 和无人机 2 的奖励值在第 7 000 个回合左右出现了明显增加，且增加幅度相近，而无人机 3 的奖励值在第 10 000 个回合左右

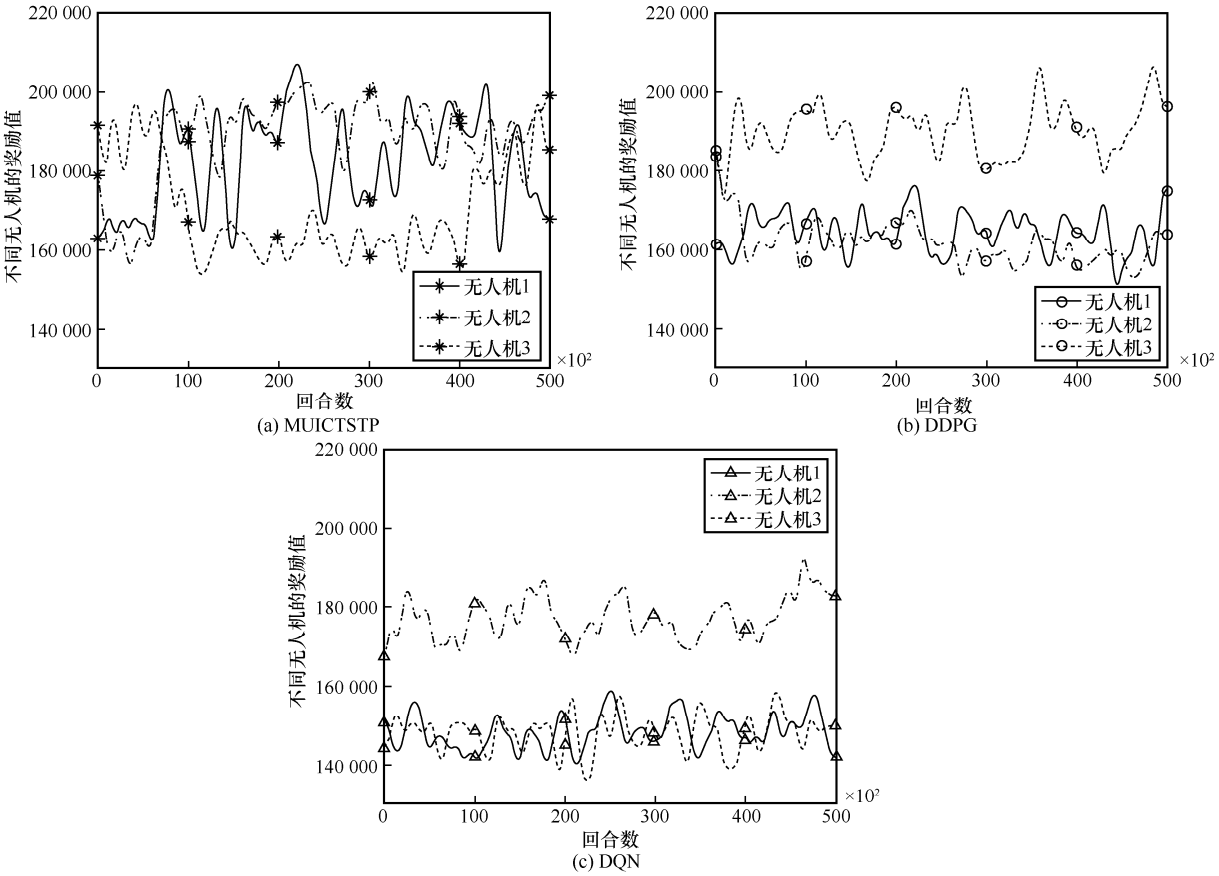

图 7　不同算法中不同无人机的奖励值

减少，直到第 40 000 个回合才开始增加，且奖励值最终趋近于无人机 1 和无人机 2。虽然不能确保所有无人机在任意时刻都获得较好的学习策略，但是能确保该算法为大多数的无人机优化学习策略。图 7(b)中，DDPG 算法下无人机 1 和无人机 3 的奖励值没有明显增加，无人机 2 的总奖励在第 3 000 个回合左右明显减少，且后面没有再增加。图 7(c)中，无人机 1 和无人机 3 的奖励值均没有明显增加，无人机 2 在最后阶段奖励值出现小幅增加，但不明显。因此，可以得出 DDPG 算法和 DQN 算法均不能为无人机训练出好的学习策略，无法有效提升每架无人机的学习能力。

该系统的最终目标是最大化长期 RISS，同时在轨迹规划过程中避免与其他无人机产生碰撞以及避免轨迹超出一定的范围边界。α=0.01 时，不同算法的 RISS 值如图 8 所示。MUICTSTP 算法在整个训练过程中获得的 RISS 整体大于其他 2 种算法，反映出在相同场景中，MUICTSTP 算法相比其他 2 种算法的无人机能更快速地飞往干扰信号源目标位置。MUICTSTP 算法在前 10 000 个回合 RISS 的平均值为 559 258.67，最后 10 000 个回合 RISS 的平均值为 576 686.95，RISS 增加了 3.12%；DDPG 算法在前 10 000 个回合的 RISS 的平均值为 544 550.37，最后 10 000 个回合的 RISS 的平均值为 537 668.72，RISS 减少了 1.26%；DQN 算法在前 10 000 个回合的 RISS 的平均值为 499 764.48，最后 10 000 个回合的 RISS 的平均值为 509 863.01，RISS 增加了 2.02%。因此，MUICTSTP 算法相比 DDPG 算法和 DQN 算法能提供更好的学习策略。

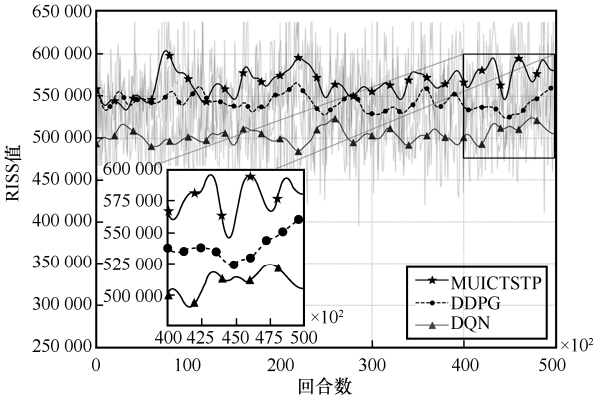

图 8　不同算法的 RISS 值

除了考虑正奖励，该系统的目标奖励设置还需考虑因碰撞或者超出边界带来的负奖励。出于对无人机的飞行安全的考虑，本文对每架无人机之间以及无人机和有人机之间都设置了安全距离，一旦无人机超出了安全距离将会收到一定的负奖励，从而尽可能地避免碰撞带来的危害。同时，为了确保实验的有效性，还将整个实验场景控制在一定活动范围内进行研究，如果无人机超过了这个设定范围也会收到负奖励。

α=0.01 时，不同算法的整体负奖励如图 9 所示。从图 9 可以看出，在相同场景设置下，MUICTSTP 算法相比其他 2 种算法均能获得更少的负奖励，也意味着它可以更有效地规避风险，因此，MUICTSTP 算法的学习策略更好，能更有效地应对动态环境。

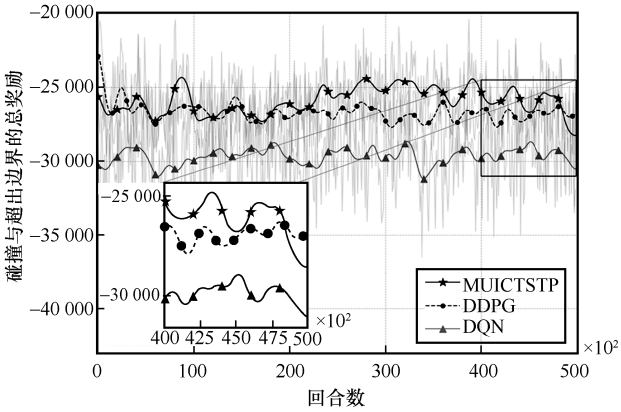

图 9　不同算法的整体负奖励

4　结束语

本文对基于有人机/无人机智能协同目标搜索和轨迹规划进行研究，考虑到干扰信号源位置的未知性，以及搜索过程中环境的动态性，利用 MADRL 建立了有人机/无人机的智能协同框架，提出了一种 MUICTSTP 算法来解决 MUTSTP 问题，使有人机/无人机在智能协同过程中能获得更优的学习策略，从而最大化长期 RISS。同时，还分析了该算法的计算复杂度。仿真结果表明，本文提出的 MUICTSTP 算法在总奖励、单架无人机奖励、RISS 和碰撞等性能方面均比现有基准算法提供更好的学习策略。

参考文献：

[1] 姚富强, 张余, 柳永祥. 电磁频谱安全与控制[J]. 指挥与控制学报, 2015, 1(3): 278-283.
YAO F Q, ZHANG Y, LIU Y X. Security and control for electromagnetic spectrum[J]. Journal of Command and Control, 2015, 1(3): 278-283.

[2] 孙佳琛, 王金龙, 丁国如, 等. 频谱知识图谱：面向未来频谱管理的智能引擎[J]. 通信学报, 2021, 42(5): 1-12.
SUN J C, WANG J L, DING G R, et al. Spectrum knowledge graph: an intelligent engine facing future spectrum management[J]. Journal on Communications, 2021, 42(5): 1-12.

[3] 孙仲康, 郭福成, 冯道旺, 等. 单站无源定位跟踪技术[M]. 北京: 国防工业出版社, 2008.
SUN Z K, GUO F C, FENG D W, et al. Passive location and tracking technology by single observer[M]. Beijing: National Defense Industry Press, 2008.

[4] 刘聪锋. 无源定位与跟踪[M]. 西安: 西安电子科技大学出版社, 2011.
LIU C F. Passive location and tracking[M]. Xi'an: Xidian University Press, 2011.

[5] BARTON D K. A half century of radar[J]. IEEE Transactions on Microwave Theory and Techniques, 1984, 32(9): 1161-1170.

[6] SKOLNIK M I. Fifty years of radar[J]. Proceedings of the IEEE, 1985, 73(2): 182-197.

[7] 唐小明, 何友, 夏明革. 基于机会发射的无源雷达系统发展评述[J]. 现代雷达, 2002, 24(2): 1-6.
TANG X M, HE Y, XIA M G. An overview of development of passive radar system based on transmitters of opportunity[J]. Modern Radar, 2002, 24(2): 1-6.

[8] 陈新颖, 盛敏, 李博, 等. 面向 6G 的无人机通信综述[J]. 电子与信息学报, 2022, 44(3): 781-789.
CHEN X Y, SHENG M, LI B, et al. Survey on unmanned aerial vehicle communications for 6G[J]. Journal of Electronics & Information Technology, 2022, 44(3): 781-789.

[9] ASGHAR S S A, SOLTANIZADEH H. Optimal trajectories for two UAVs in localization of multiple RF sources[J]. Transactions of the Institute of Measurement and Control, 2016, 38(8): 908-916.

[10] ASGHAR S S A, SOLTANIZADEH H. Single- and multi-UAV trajectory control in RF source localization[J]. Arabian Journal for Science and Engineering, 2017, 42(2): 459-466.

[11] WANG Z, CHEN G, BLASCH E, et al. Jamming emitter localization with multiple UAVs equipped with smart antennas[J]. Proceedings of SPIE-The International Society for Optical Engineering, 2010, 7696: 1-9.

[12] WANG J, HINTON J, LIU J C L. RF based target search and localization with microUVA[C]//Proceedings of the 2016 International Conference on Computational Science and Computational Intelligence (CSCI). Piscataway: IEEE Press, 2016: 1077-1082.

[13] TSUJI H, GRAY D, SUZUKI M, et al. Radio location estimation experiment using array antennas for high altitude platforms[C]// Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. Piscataway: IEEE Press, 2007: 1-5.

[14] PACK D, YORK G, FIERRO R. Information-based cooperative control for multiple unmanned aerial vehicles[C]//Proceedings of the IEEE International Conference on Networking, Sensing and Control. Piscataway: IEEE Press, 2006: 446-450.

[15] DOGANCAY K. UAV path planning for passive emitter localization[J]. IEEE Transactions on Aerospace and Electronic Systems, 2012, 48(2): 1150-1166.

[16] WANG L Y, HUANG Y. UAV-based estimation of direction of arrival: an approach based on image processing[C]//Proceedings of the 2020 International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway: IEEE Press, 2020: 1165-1169.

[17] 陈杰, 辛斌. 有人/无人系统自主协同的关键科学问题[J]. 中国科学: 信息科学, 2018, 48(9): 1270-1274.
CHEN J, XIN B. Key scientific problems in the autonomous cooperation of manned-unmanned systems[J]. Scientia Sinica (Informationis), 2018, 48(9): 1270-1274.

[18] LI Z, BRAUN T, ZHAO X H, et al. A narrow-band indoor positioning system by fusing time and received signal strength via ensemble learning[J]. IEEE Access, 2018, 6: 9936-9950.

[19] VANSTEENWEGEN P, SOUFFRIAU W, OUDHEUSDEN D V. The orienteering problem: a survey[J]. European Journal of Operational Research, 2011, 209(1): 1-10.

[20] 叶媛媛, 闵春平, 沈林成. 多 UCAV 任务分配的混合遗传算法与约束处理[J]. 控制与决策, 2006, 21(7): 781-786.
YE Y Y, MIN C P, SHEN L C. Hybrid genetic algorithm and constraint handling for multiple UCAV mission assigning[J]. Control and Decision, 2006, 21(7): 781-786.

[21] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.

[22] WU S J. Illegal radio station localization with UAV-based Q-learning[J]. China Communications, 2018, 15(12): 122-131.

[23] JAAKKOLA T, SINGH S P, JORDAN M I. Reinforcement learning algorithm for partially observable Markov decision problems[C]// Proceedings of Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 1995: 7.

[24] SHI W S, LI J L, WU H Q, et al. Drone-cell trajectory planning and resource allocation for highly mobile networks: a hierarchical DRL approach[J]. IEEE Internet of Things Journal, 2021, 8(12): 9800-9813.

[25] BOUHAMED O, GHAZZAI H, BESBES H, et al. Autonomous UAV navigation: a DDPG-based deep reinforcement learning approach[C]// Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS). Piscataway: IEEE Press, 2020: 1-5.

[作者简介]

卢卓（1994− ），女，湖南长沙人，南京航空航天大学博士生，主要研究方向为无人机群智能协同通信、多智能体强化学习、无人机轨迹规划等。

吴启晖（1970− ），男，安徽歙县人，博士，南京航空航天大学特聘教授，主要研究方向为认知信息论、电磁空间频谱智能管控、天地一体化信息网络、无人机集群智能通信等。

周福辉（1988− ），男，江西抚州人，博士，南京航空航天大学教授，主要研究方向为认知智能、频谱智能管控、资源智能优化、语义通信等。