信息年龄约束下的无人机数据采集能耗优化路径规划算法

高思华[①] 刘宝煜[①] 惠康华*[①] 徐伟峰[②③] 李军辉[①] 赵炳阳[①]

[①](中国民航大学计算机科学与技术学院 天津 300300)

[②](华北电力大学(保定)计算机系 保定 071066)

[③](河北省能源电力知识计算重点实验室 保定 071066)

摘 要：信息年龄(AoI)是评价无线传感器网络(WSN)数据时效性的重要指标，无人机辅助WSN数据采集过程中采用优化飞行轨迹、提升速度等运动策略保障卸载至基站的数据满足各节点AoI限制。然而，不合理的运动策略易导致无人机因飞行距离过长、速度过快产生非必要能耗，造成数据采集任务失败。针对该问题，该文首先提出信息年龄约束的无人机数据采集能耗优化路径规划问题并进行数学建模；其次，设计一种协同混合近端策略优化(CH-PPO)强化学习算法，同时规划无人机对传感器节点或基站的访问次序、悬停位置和飞行速度，在满足各传感器节点信息年龄约束的同时，最大限度地减少无人机能量消耗。再次，设计一种融合离散和连续策略的损失函数，增强CH-PPO算法动作的合理性，提升其训练效果。仿真实验结果显示，CH-PPO算法在无人机能量消耗以及影响该指标因素的比较中均优于对比的3种强化学习算法，并具有良好的收敛性、稳定性和鲁棒性。

关键词：无线传感器网络；信息年龄约束；协同混合近端策略优化算法；无人机路径规划；深度强化学习

中图分类号：TN926.2; V279 **文献标识码**：A **文章编号**：1009-5896(2024)10-4024-11

DOI: 10.11999/JEIT240075

Energy-Efficient UAV Trajectory Planning Algorithm for AoI-Constrained Data Collection

GAO Sihua[①] LIU Baoyu[①] HUI Kanghua[①] XU Weifeng[②③] LI Junhui[①]
ZHAO Bingyang[①]

[①](*Institute of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*)

[②](*Department of Computer, North China Electric Power University* (*Baoding*), *Baoding* 071066, *China*)

[③](*Hebei Key Laboratory of Knowledge Computing for Energy & Power, Baoding* 071066, *China*)

Abstract: The information freshness is measured by Age of Information (AoI) of each sensor in Wireless Sensor Networks (WSN). The UAV optimizes flight trajectories and accelerates speed to assist WSN data collection, which guarantees that the data offloaded to the base station meets the AoI limitation of each sensor. However, the UAV's inappropriate flight strategies cause non-essential energy consumption due to excessive flight distance and speed, which may result in the failure of data collection mission. In this paper, firstly a mathematical model is investigated and developed for the UAV energy consumption optimization trajectory planning problem on the basis of AoI-constrained data collection. Then, a novel deep reinforcement learning algorithm, named Cooperation Hybrid Proximal Policy Optimization (CH-PPO) algorithm, is proposed to simultaneously schedule the UAV's access sequence, hovering position, the flight speed to the sensor nodes or the base station, to minimize the UAV's energy consumption under the constraint of data timeliness for each sensor node. Meanwhile, a loss function that integrates the discrete policy and continuous policy is designed to increase the rationality of hybrid actions and improve the training effectiveness of the proposed algorithm.

收稿日期：2024-01-30；改回日期：2024-09-05；网络出版：2024-09-10

*通信作者： 惠康华 khhui@cauc.edu.cn

基金项目：国家自然科学基金(62173332)，中央高校基本科研业务费专项资金(3122019118)，河北省能源电力知识计算重点实验室开发基金(HBKCEP202202)

Foundation Items: The National Natural Science Foundation of China (62173332), The Fundamental Research Fundation for the Central Universities (3122019118), The Open Fundation of Hebei Key Laboratory of Knowledge Computing for Energy & Power (HBKCEP202202)

Numerical results demonstrate that the CH-PPO algorithm outperforms the other three reinforcement learning algorithms in the comparison group in energy consumption of UAV and its influencing factors. Furthermore, the convergence, stability, and robustness of the proposed algorithm is well verified.

Key words: Wireless Sensor Networks (WSN); AoI limitation; Cooperation hybrid proximal policy optimization algorithm; Unmanned aerial vehicle trajectory planning; Deep reinforcement learning

1　引言

无线传感器网络(Wireless Sensor Networks, WSN)[1]以成本低、自组织能力强等特点，广泛应用于工业生产、交通运输、实时监控等领域[2,3]。在大数据和人工智能背景下，数据的时效性已成为评判WSN服务质量的重要指标之一。然而，传统WSN中节点通信受环境等因素影响，传输过程数据质量低、时效性差等问题无法得到解决。近年来，无人机凭借机动性高、存储能力强等优点[4,5]，逐渐成为辅助WSN采集数据的重要工具[6–8]。当前，无人机运动策略与高时效性数据获取间关系的探索成为研究热点[9–11]。

2012年，Kaul等人[12]首次提出使用信息年龄(Age of Information, AoI)作为评价信息时效性的指标。张建行等人[13]使用DQN算法规划无人机对传感器节点访问次序，降低数据的平均AoI。然而，该方法忽略了无人机速度、能量等因素对优化目标的影响。Liao等人[14]通过增加无人机返回基站的频率保证各节点数据的时效性。为平衡往返基站产生的额外能耗，作者使用Bender分解法计算两者的帕累托最优。但无人机无法充分利用能量减小数据的平均AoI。Sherman等人[15]将最小化平均AoI问题抽象成序列决策问题，使用PPO算法控制无人机在每个时隙内的运动策略。作者将任务时间划分为若干时隙，通过优化数据采集位置来减少无人机飞行时间，在节约能量的同时优化各节点中数据的AoI。Sun等人[16]使用TD3(Twin Delayed Deep Deterministic)算法控制无人机在每个时隙内的运动策略，最小化AoI和能耗的加权和。然而，该算法在WSN规模较大时存在求解质量低、收敛速度慢等问题。Liu等人[17]提出使用峰值AoI评价WSN中各节点数据的时效性。作者首先将网络分成若干个簇，设计一种遗传算法求解无人机访问各簇的次序。然而，该方法中各簇边缘节点的数据传输效率低，增加了数据采集的能耗。Dai等人[18]使用蒙特卡罗树搜索算法动态调整无人机运动速度和方向，以增强WSN中数据的时效性。该方法考虑无人机运动速度和数据时效性间的联系，但访问节点的顺序具有盲目性，影响各节点数据的AoI。

从研究问题上看，现有研究主要通过平均AoI和峰值AoI对WSN数据时效性进行整体评价，关注各节点数据AoI差异[19]的研究较少。已有研究成果通过优化无人机访问传感器节点和基站次序、数据采集和卸载位置、飞行速度来保障数据时效性。然而，研究者忽略了各优化策略间的相互关联，以及各优化策略易造成无人机能量消耗过快而无法完成数据采集任务。从求解方法上看，传统优化方法和现有的深度强化学习算法表达能力有限，无法同时优化以上3种影响数据AoI的策略。本文研究的问题为规划无人机运动策略，在确保WSN中各节点数据均满足AoI约束的同时，最小化无人机能耗，主要贡献如下：

(1)提出信息年龄约束下的无人机数据采集能耗优化问题，并对该问题进行数学建模和分析。设计包括节点访问次序、悬停位置和飞行速度在内的联合运动策略，指导无人机完成数据采集任务。在保证各传感器节点数据时效性的同时，最大限度地减少无人机的能耗；

(2)提出一种基于深度强化学习的协同混合近端策略优化算法(Cooperation Hybrid Proximal Policy Optimization, CH-PPO)，实现无人机的联合运动策略。该算法首次将H-PPO算法思想引入到无人机辅助WSN进行AoI约束下数据采集的工作中并加以改进，实现离散动作与多种连续动作相混合的智能体训练模式。针对该训练模式，在CH-PPO算法损失函数的设计中，使用离散策略和连续策略新旧策略比值的乘积构建演员(Actor)网络的损失函数，增强离散策略和连续策略的关联性，保证混合动作选择的合理性以及策略更新的稳定性。

2　系统模型和问题描述

2.1　传感器模型

本文假定在大小为$L\times L$的平面区域中心部署基站B，同时在该区域内随机部署n个传感器$Q=\{q_1,q_2,\cdots,q_n\}$。$\boldsymbol{p}_{\mathrm{B}}=(x_{\mathrm{B}},y_{\mathrm{B}})$为基站的坐标，$\boldsymbol{p}_i=(x_i,y_i)$为传感器节点$q_i$的坐标。本文应用文献[20]中的数据生成模型，传感器节点将获取的数据打包成大小为D的数据包。当无人机悬停在传感器节点q_i连通范围内时，该传感器可通过配备的唯一天线与之建立传输功率为P_{d}的数据传递链路，完成自身缓存区内数据包的传递工作。

2.2 无人机模型

无人机从基站出发，在WSN区域上方H高度以可变速度V飞行，逐一采集所有传感器节点的数据。当无人机确定采集节点q_i后，飞行至q_i附近悬停，通过建立的数据上传链路采集q_i缓冲区内的数据。当q_i完成数据上传工作后，无人机与其断开链路，采集WSN中其他传感器节点的数据或返回基站卸载已采集的数据。为了确保数据传输的速率和质量，无人机仅与服务半径内的传感器节点或基站建立数据传递链路。无人机悬停位置的平面投影与q_i的距离d_i应满足以下限制

$$d_i = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \le R \tag{1}$$

其中，$\boldsymbol{U}_k = (x_k, y_k)$表示无人机第$k$次悬停位置的平面投影坐标，$R$为无人机数据采集的最大服务半径。同理，无人机悬停位置的平面投影与基站间的距离应满足$d_{\mathrm{B}} \le R$，与基站连线的方位角$\theta_B \in [0, 2\pi)$。同理，无人机悬停位置的平面投影与传感器节点q_i连线的方位角$\theta_i \in [0, 2\pi)$。无人机在第k次悬停结束时已服务的传感器节点的集合记为J_k。无人机采集WSN中传感器数据的示意图如图1所示。

WSN中部署了6个传感器节点。无人机从基站正上方1号位置出发，依次飞至2号、3号和4号悬停位置，传感器节点q_3, q_2, q_1通过数据上传链路将缓存中的数据上传至无人机。为了保证已采集数据的时效性，无人机飞至5号悬停位置与基站建立数据卸载链路，卸载已采集的数据。随后，无人机依次飞至6号、7号和8号悬停位置采集传感器节点q_5, q_6, q_4的数据，并在返回9号悬停位置卸载已采集的数据后完成任务。

2.3 信道模型

信道功率增益是评价信道质量的一项重要指标，决定数据和能量传输的速率和质量。本文信道模型参考文献[21]，考虑信道可视情况和通径损失。无人机第k次悬停与传感器节点q_i建立视距信道的概率$P_i^{\mathrm{LoS}}(k)$为

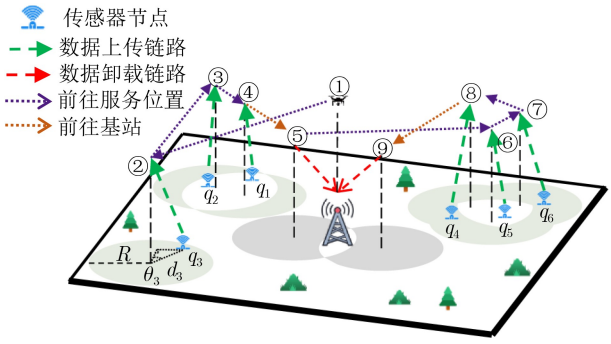

图 1 任务示意图

$$P_i^{\mathrm{Los}}(k) = \frac{1}{1 + a\exp\left[-b\left(\rho_i - a\right)\right]} \tag{2}$$

其中a，b为视距信道LoS和非视距信道NLoS的环境依赖常数。ρ_i代表传感器节点q_i与无人机之间的仰角。非视距信道概率为$P_i^{\mathrm{NLoS}}(k) = 1 - P_i^{\mathrm{LoS}}(k)$。两种信道下的通径损失分别为

$$\mathrm{PL}_i(k) = \begin{cases} \zeta\left(\sqrt{d_i^2 + H^2}\right)^{-\alpha}, & \mathrm{LoS} \\ \mu\zeta\left(\sqrt{d_i^2 + H^2}\right)^{-\alpha}, & \mathrm{NLoS} \end{cases} \tag{3}$$

其中ζ为单位距离的信道功率增益，α，μ分别表示通径损失指数和非视距信道的额外衰减系数。无人机第k次悬停与传感器节点q_i建立的信道功率增益$g_i(k)$为

$$g_i(k) = \left(P_i^{\mathrm{LoS}}(k) + \mu P_i^{\mathrm{NLoS}}(k)\right)\zeta\left(\sqrt{d_i^2 + H^2}\right)^{-\alpha} \tag{4}$$

此时无人机与传感器节点q_i的数据传输速率$c_i(k)$计算方式为

$$c_i(k) = W\log_2\left(1 + \frac{P_{\mathrm{d}}|g_i(k)|^2}{\sigma^2}\right) \tag{5}$$

其中W和σ分别表示信道的带宽和噪声功率。

2.4 无人机能耗模型

本文假设无人机的能量消耗仅发生在巡航和数据采集过程中，其他能耗忽略不计。无人机的初始能量为E_{init}，飞行过程中的牵引功率$P(V)$的计算方法参考文献[22]，其中无人机悬停时的功率为$P(0)$。无人机第k次悬停为传感器q_i提供数据采集服务的时间$T_{\mathrm{h}}(k)$为

$$T_{\mathrm{h}}(k) = \frac{D}{c_i(k)} \tag{6}$$

$c_i(k)$为无人机与传感器q_i间的数据传输速率。无人机在悬停过程中的能量消耗$E_{\mathrm{h}}(k)$表示为

$$E_{\mathrm{h}}(k) = T_{\mathrm{h}}(k) \times P(0) \tag{7}$$

由于无人机与基站间传输数据的速率较快，本文假定该过程中无人机的悬停时长和能耗均为零。无人机从第$k-1$次悬停位置(x_{k-1}, y_{k-1})运动至第k次悬停位置(x_k, y_k)过程中，飞行的距离$d(k)$表示为

$$d(k) = \sqrt{(x_k - x_{k-1})^2 + (y_k - y_{k-1})^2} \tag{8}$$

该过程经历的时长$T_{\mathrm{m}}(k)$计算方法如下

$$T_{\mathrm{m}}(k) = \frac{d(k)}{v(k)} \tag{9}$$

其中$\boldsymbol{v}(k) \in (0, v_{\max}]$为无人机从第$k-1$次悬停位置运动至第$k$次悬停位置过程中的飞行速度。此过程无人机的飞行能耗为

$$E_{\mathrm{m}}(k)=P(\boldsymbol{v}(k))\times T_{\mathrm{m}}(k) \tag{10}$$

$P(v(k))$, $T_{\mathrm{m}}(k)$分别为无人机以$v(k)$飞行时的功率和飞行时间。

综上所述，无人机从离开基站至数据采集任务结束过程中，消耗的总能量为

$$E_{\mathrm{c}}=\sum_{k=1}^{K}\left(E_{\mathrm{m}}(k)+E_{\mathrm{h}}(k)\right) \tag{11}$$

2.5 信息年龄

信息年龄(AoI)表示传感器节点感知的数据从开始上传直至传递到基站所经过的时长。传感器节点q_i中数据在任意时刻M_s的AoI值$A_i(M_s)$表示为

$$A_i(M_s)=\begin{cases}M_s-C(q_i), & L(q_i)\geq M_s\\ L(q_i)-C(q_i), & L(q_i)<M_s\\ 0, & M_s<C(q_i)\end{cases} \tag{12}$$

其中，$C(q_i)$表示无人机开始采集传感器节点q_i的时刻；$L(q_i)$表示无人机将传感器节点q_i中数据卸载到基站的时刻。

图2展示了图1中各传感器节点数据的AoI，红色箭头表示无人机到达数据采集位置时刻，黄色箭头表示对某节点数据采集结束时刻，绿色箭头表示无人机返回基站卸载数据时刻。无人机从1号位置飞行$T_{\mathrm{m}}(1)$后，在M_1时刻飞至2号悬停位置，耗时$T_{\mathrm{h}}(1)$采集传感器节点q_3数据，采集过程在M_2时刻结束。随后，无人机采集传感器节点q_2和q_1的数据，并于M_7时刻返回基站卸载数据。由于数据已被卸载至基站，无人机后续任务不改变传感器节点q_3，q_2和q_1中数据的AoI。无人机在时刻M_{F}完成所有数据采集任务返回基站，传感器节点q_i中数据的AoI值表示为$A_i(M_{\mathrm{F}})$，为表达简洁，本文使用A_i表示$A_i(M_{\mathrm{F}})$。即$A_3=L(q_3)-C(q_3)=M_7-M_1$，$A_5=M_{\mathrm{F}}-C(q_5)=M_{\mathrm{F}}-M_8$。

2.6 信息年龄约束下的无人机数据采集能耗优化问题

现有的研究能有效提升数据的时效性，但存在无人机能量消耗过快无法完成数据采集任务的问题。本文分析现有成果中提高WSN数据时效性的策略，认为无人机对节点的访问次序、悬停位置和飞行速度共同影响了WSN中数据的AoI和无人机能量消耗。首先，合理的节点访问次序可减少无人机飞行距离和返回基站频次，降低在相同运动速度下的能耗。其次，合理的数据采集位置可平衡无人机采集数据的时长与移动至下个悬停位置的飞行时长，降低该过程的整体能耗。再次，无人机可通过增加飞行速度来减少运动时间，通过增加能耗来降低各节点数据的AoI。三者在无人机能耗上相互制约，在增强网络数据时效性上相互补充。

综上所述，本文在WSN中各节点AoI阈值约束下，通过联合规划无人机访问传感器节点和基站的次序、数据采集位置和飞行速度，对最小化无人机的能量消耗问题展开研究，具体描述为

$$\min\left(E_{\mathrm{c}}\right) \tag{13}$$

$$\text{s.t.}\begin{cases}\mathrm{C1}: U_0=\left(\dfrac{L}{2},\dfrac{L}{2}\right)\\ \mathrm{C2}: 0<\boldsymbol{v}(k)\leq\boldsymbol{v}_{\max},\forall k\in\{1,2,\cdots,K\}\\ \mathrm{C3}: 0\leq\theta_i<2\pi,\forall i\in\{1,2,\cdots,n\}\\ \mathrm{C4}: 0\leq d_i\leq R,\forall i\in\{1,2,\cdots,n\}\\ \mathrm{C5}: 0<A_i\leq A_i^{\mathrm{c}},\forall i\in\{1,2,\cdots,n\}\\ \mathrm{C6}: 0<E_{\mathrm{c}}\leq E_{\mathrm{init}}\\ \mathrm{C7}: J_K=Q\\ \mathrm{C8}: |J_K|=n\end{cases} \tag{14}$$

C1表示无人机的初始位置U_0为WSN中心；C2表示无人机的飞行速度须在规定区间$(0,\boldsymbol{v}_{\max}]$之内；C3和C4共同决定无人机采集每个传感器数据或返回基站时的悬停位置；C5约束WSN中各节点q_i中数据的AoI小于其阈值A_i^{c}；C6要求无人机能量消耗不超过无人机的初始能量；C7和C8表示在任务结束时所有传感器均只被服务一次。

3 基于 CH-PPO 的无人机路径规划算法

3.1 环境建模

3.1.1 状态

状态s_k由WSN中各传感器节点信息、基站位置信息和无人机信息组成，表示无人机做第k次决

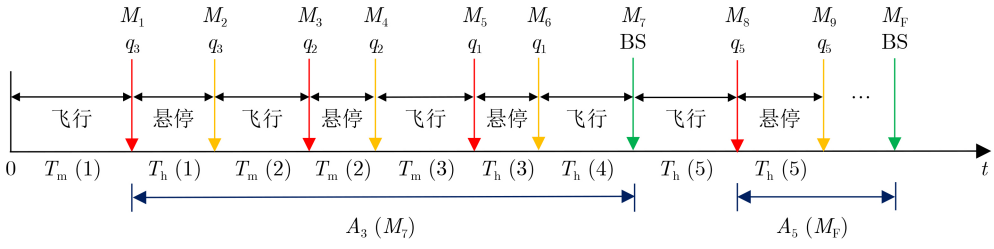

图 2 AoI示意图

策前的环境状态信息。传感器节点信息包括位置信息和缓存数据信息；无人机信息包括位置信息和剩余能量信息。$\boldsymbol{s}_k$具体表示如下

$$\begin{aligned}&S=\{\boldsymbol{s}_k|\boldsymbol{s}_k=\{(p_i,f_i,A_i(G(k-1)),A_i^{\mathrm{c}}),\\&p_{\mathrm{B}},(\boldsymbol{U}_{k-1},E(k-1))\},i\in\{1,2,\cdots,n\},\\&k\in\{1,2,\cdots,K\}\}\end{aligned}\tag{15}$$

其中，p_i表示传感器节点q_i的位置信息；f_i为q_i中数据的状态标识，分别对应数据存储在传感器节点、已传递至无人机和卸载到基站3种状态；p_{B}表示基站的位置信息；$G(k-1)$为无人机第$k-1$次悬停结束时刻，$A_i(G(k-1))$表示传感器节点q_i中数据在该时刻的AoI；A_i^{c}表示传感器节点q_i中数据AoI的阈值。$\boldsymbol{U}_{k-1}$和$E(k-1)$分别表示无人机第$k-1$次悬停后的位置信息和剩余能量。

3.1.2 动作

动作a_k由离散动作和连续动作混合而成，前者表示无人机选择前往传感器节点或基站，后者表示无人机悬停位置与前往目标节点的相对位置关系和运动速度。$\boldsymbol{a}_k$具体表示为

$$\begin{aligned}&A=\{\boldsymbol{a}_k|\boldsymbol{a}_k=\{i,(l(k),\theta(k),v(k))\},\\&i\in\{0,1,2,\cdots,n\},l(k)\in[0,R],\theta(k)\in[0,2\pi),\\&v(k)\in(0,v_{\max}]\}\end{aligned}\tag{16}$$

其中，$i=0$表示无人机将前往的目标节点为基站，其他取值分别对应WSN中各传感器节点；$l(k)$为无人机第k次悬停位置的水平投影与目标节点的距离；$\theta(k)$为该水平投影与目标节点连线的方位角；$v(k)$为无人机飞往第k次悬停位置的速度。

3.1.3 奖励

无人机任务执行过程分为数据采集阶段、数据卸载阶段和任务结束阶段，根据无人机执行动作a_k后所处阶段，为无人机分别设计奖励r_k。

数据采集阶段：r_k由能量消耗奖励$r_{\mathrm{e}}(k)$和AoI阈值限制惩罚$r_{\mathrm{c}}(k)$组成。$r_{\mathrm{e}}(k)$指导无人机集中采集某一个区域的数据，减少无效飞行，具体表示为

$$r_{\mathrm{e}}(k)=\frac{(N(k)+1)^2}{E_{\mathrm{m}}(k)+E_{\mathrm{h}}(k)}\tag{17}$$

其中，$N(k)$表示无人机在第k次悬停位置时，可服务范围内未被采集的传感器节点数量。本文通过$r_{\mathrm{c}}(k)$对无人机携带已超过阈值的数据进行惩罚，指导无人机适时返回基站，确保已采集数据的时效性，具体表示为

$$r_{\mathrm{c}}(k)=\sum_{i\in N_{\mathrm{c}}(k)}\min\{A_i^{\mathrm{c}}-A_i(G(k)),0\}\tag{18}$$

其中，$N_{\mathrm{c}}(k)$表示无人机到达第k次悬停位置时携带数据的集合。因此，本阶段的奖励$r_k=\tau_1\times r_{\mathrm{e}}(k)+\tau_2\times r_{\mathrm{c}}(k)$。

数据卸载阶段：相较于数据采集阶段，本阶段新增数据卸载奖励$r_{\mathrm{d}}(k)$。若卸载的数据均满足AoI阈值限制，则给予无人机正向奖励；否则，奖励值为0。该奖励指导无人机充分利用已采集数据AoI阈值，减少返回基站次数和高速运动产生的额外能耗，具体表示为

$$r_d(k)=\begin{cases}\displaystyle\sum_{i\in N_{\mathrm{c}}(k)}\frac{1}{(A_i^{\mathrm{c}}-A_i(G(k)))+1},\forall A_i\le A_i^{\mathrm{c}}\\0\ ,\qquad\qquad\qquad\qquad\text{其它}\end{cases}\tag{19}$$

此时，$r_k=\tau_1\times r_{\mathrm{e}}(k)+\tau_2\times r_{\mathrm{c}}(k)+\tau_3\times r_{\mathrm{d}}(k)$。根据无人机能耗模型，此时$r_{\mathrm{e}}(k)$中的$E_{\mathrm{h}}(k)=0$。

任务结束阶段：相较于数据卸载阶段，本阶段新增任务评价奖励r_{F}，用于对本轮无人机的飞行策略进行整体评价。若网络中所有数据均被传输至基站，r_{F}对本次飞行给予奖励$E_{\mathrm{init}}-E_{\mathrm{c}}$，并根据是否存在数据AoI超过阈值限制，分别设置折扣率β和λ。

$$r_{\mathrm{F}}=\begin{cases}\beta\times(E_{\mathrm{init}}-E_{\mathrm{c}}),\forall A_i\le A_i^{\mathrm{c}}\\\lambda\times(E_{\mathrm{init}}-E_{\mathrm{c}}),\exists A_i\ge A_i^{\mathrm{c}}\end{cases}\tag{20}$$

其中$i\in\{1,2,\cdots,n\}$。若无人机在能量耗尽时网络中尚有d_{num}数据未传输至基站，则$r_{\mathrm{F}}=-\tau_4\times d_{\mathrm{num}}$。此时，$r_k=\tau_1\times r_{\mathrm{e}}(k)+\tau_2\times r_{\mathrm{c}}(k)+\tau_3\times r_{\mathrm{d}}(k)+r_{\mathrm{F}}$。其中$\tau_1,\tau_2,\tau_3,\tau_4$为用于调整各阶段奖励量级的常数。

综上所述，奖励函数能够鼓励无人机减少飞行能耗，充分利用网络中数据的阈值，在时效性限制下采集数据并传输至基站。

3.2 协同混合近端策略优化算法

3.2.1 网络结构

协同混合近端策略优化算法采用演员-评论家(Actor-Critic)框架，Actor网络中离散动作网络和连续动作网络共享两层全连接层提取的状态信息。离散动作网络输出长度为$n+1$的一维张量，表示WSN中各传感器节点和基站被无人机选择服务的概率，并根据概率分布随机采样获得离散动作。连续动作网络采用随机高斯策略，其双分支结构分别输出各连续动作取值的均值(mean)和标准差(mu)，二者共同构建正态分布，并从中采样获得连续动作。最终，离散动作和连续动作组合得到混合动作$\boldsymbol{a}_k$。Critic网络使用两层全连接层构建状态价值函数$V(\boldsymbol{s}_k;\omega)$，计算对当前状态$\boldsymbol{s}_k$的评价$\mathrm{val}_k$。CH-PPO的网络结构如图3所示。

3.2.2 算法执行流程

CH-PPO算法根据输入的状态$\boldsymbol{s}_k$计算状态价值

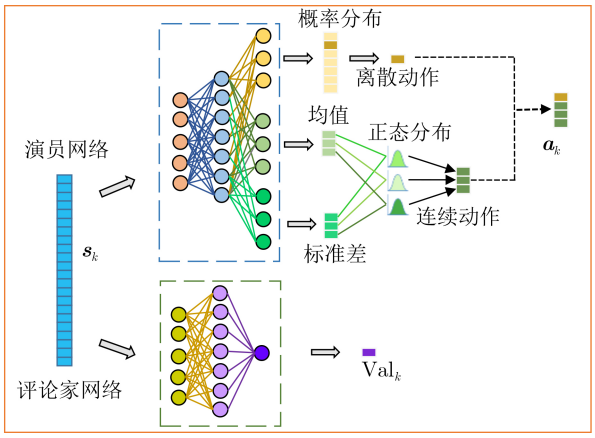

图 3 网络结构图

val_k和混合动作$\boldsymbol{a}_k$。无人机执行动作$\boldsymbol{a}_k$后，得到环境反馈的瞬时奖励r_k，并进入状态$\boldsymbol{s}_{k+1}$。算法将轨迹$(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1})$存入经验池(memory pool)。

每一轮任务结束后，CH-PPO算法使用梯度下降方法以学习率η对网络参数进行更新，并每隔一定轮数将旧Actor网络的参数替换为新Actor网络的参数。在训练过程中，CH-PPO算法首先使用经验池中的奖励信息计算状态$\boldsymbol{s}_k$的折扣回报和优势函数估计值$\hat{A}_k$如下：

$$\hat{A}_k = \sum_{i=k}^{K} \gamma^{i-k} r_i - \mathrm{val}_k \tag{21}$$

其次，CH-PPO算法分别计算离散策略和连续策略的新旧策略比值，记作$r_k^{\mathrm{d}}(\theta_{\mathrm{d}})$和$r_k^{\mathrm{c}}(\theta_{\mathrm{c}})$。离散策略使用离散动作在概率分布中的概率取值表示，连续策略使用连续动作在高斯分布中被采样的概率表示。$r_k^{\mathrm{d}}(\theta_{\mathrm{d}})$和$r_k^{\mathrm{c}}(\theta_{\mathrm{c}})$计算方法如下

$$r_k^{\mathrm{d}}(\theta_{\mathrm{d}}) = \frac{\pi_{\theta_{\mathrm{d}}}\left(\boldsymbol{a}_k^{\mathrm{d}} | \boldsymbol{s}_k\right)}{\pi_{\theta_{\mathrm{d,old}}}\left(\boldsymbol{a}_k^{\mathrm{d}} | \boldsymbol{s}_k\right)} \tag{22}$$

$$r_k^{\mathrm{c}}(\theta_{\mathrm{c}}) = \frac{\pi_{\theta_{\mathrm{c}}}\left(\boldsymbol{a}_k^{\mathrm{c}} | \boldsymbol{s}_k\right)}{\pi_{\theta_{\mathrm{c,old}}}\left(\boldsymbol{a}_k^{\mathrm{c}} | \boldsymbol{s}_k\right)} \tag{23}$$

文献[23]分别使用$r_k^{\mathrm{d}}(\theta_{\mathrm{d}})$和$r_k^{\mathrm{c}}(\theta_{\mathrm{c}})$构建离散网络和连续网络的损失函数，并将两个损失函数相加作为Actor网络的整体损失。该损失函数的设计存在以下两个问题：首先，分开考虑离散策略和连续策略，忽略了两者之间的内在联系，影响混合动作选择的合理性；其次，在训练过程中，某一策略趋于收敛而另一策略尚未收敛时，已收敛策略仍然产生损失，易造成Actor网络训练不稳定，收敛速度慢等问题。CH-PPO算法将离散策略和连续策略相结合，设计使用$r_k^{\mathrm{d}}(\theta_{\mathrm{d}})$与$r_k^{\mathrm{c}}(\theta_{\mathrm{c}})$的乘积构建Actor网络的损失函数$\mathrm{LA}_k(\theta)$如下

$$\begin{aligned}\mathrm{LA}_k(\theta) = \widehat{E}_k[&\min((r_k^{\mathrm{d}}(\theta_{\mathrm{d}}) \times r_k^{\mathrm{c}}(\theta_{\mathrm{c}})) \times \widehat{A}_k,\\ &\mathrm{clip}((r_k^{\mathrm{d}}(\theta_{\mathrm{d}}) \times r_k^{\mathrm{c}}(\theta_{\mathrm{c}})), 1-\varepsilon, 1+\varepsilon) \times \widehat{A}_k)]\end{aligned} \tag{24}$$

相较于分别使用$r_k^{\mathrm{d}}(\theta_{\mathrm{d}})$和$r_k^{\mathrm{c}}(\theta_{\mathrm{c}})$构建损失函数，本文将$r_k^{\mathrm{d}}(\theta_{\mathrm{d}})$与$r_k^{\mathrm{c}}(\theta_{\mathrm{c}})$的乘积作为使用近端策略优化算法计算损失的策略比值，构建损失函数。该方法增加离散策略和连续策略的关联性，使得混合动作更加合理。算法训练过程中某一策略趋于收敛而另一策略尚未收敛时，已收敛策略的新旧策略比值趋近于1，不再对$\mathrm{LA}_k(\theta)$产生影响，增强策略更新的稳定性。最后，本文定义Critic网络的损失函数$\mathrm{LC}_k(\omega)$如下：

$$\mathrm{LC}_k(\omega) = \hat{E}_k[(\hat{A}_k)^2] \tag{25}$$

当损失函数$\mathrm{LC}_k(\omega)$趋于0时，表明Critic网络对当前状态的评价接近真实奖励值。CH-PPO算法的伪代码如算法1所示。

4　仿真实验

4.1　仿真环境

本节通过仿真实验就算法收敛性、无人机能量

算法 1　CH-PPO算法

(1)输入：训练轮数EP，参数更新次数M，学习率η，裁剪系数ε；
(2)初始化网络参数：θ，θ_{old}和ω；
(3)循环训练：$i = 1, 2, \cdots, \mathrm{EP}$：
(4) done $\neq 0$时：
(5)　计算离散动作$\chi_{\mathrm{d}}\left(\boldsymbol{s}_k; \theta_{\mathrm{d,old}}\right)$；
(6)　计算连续动作$\chi_{\mathrm{c}}\left(\boldsymbol{s}_k; \theta_{\mathrm{c,old}}\right)$；
(7)　得到混合动作$\boldsymbol{a}_k = \{i, (l(k), \theta(k), \boldsymbol{v}(k))\}$；
(8)　智能体在状态$\boldsymbol{s}_k$下执行动作$\boldsymbol{a}_k$，获得奖励r_k，并进入下一状态$\boldsymbol{s}_{k+1}$；
(9)　将$(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1})$存储在经验池；
(10) 直到done $= 0$；
(11) 循环参数更新：$j = 1, 2, \cdots, M$：
(12)　从经验池中获得所有经验$(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1})_{k\in\{1,2,\cdots,K\}}$；
(13)　计算经验池中所有状态的状态价值$\mathrm{val}_1, \cdots, \mathrm{val}_K$；
(14)　计算优势函数的估计值$\hat{A}_1, \cdots, \hat{A}_K$；
(15)　分别计算离散策略和连续策略的新旧策略比值：$r_k^{\mathrm{d}}(\theta_{\mathrm{d}}) = \frac{\pi_{\theta_{\mathrm{d}}}(\boldsymbol{a}_k^{d}|\boldsymbol{s}_k)}{\pi_{\theta_{\mathrm{d,old}}}(\boldsymbol{a}_k^{\mathrm{d}}|\boldsymbol{s}_k)}$，$r_k^{c}(\theta_{\mathrm{c}}) = \frac{\pi_{\theta_{\mathrm{c}}}(\boldsymbol{a}_k^{\mathrm{c}}|\boldsymbol{s}_k)}{\pi_{\theta_{\mathrm{c,old}}}(\boldsymbol{a}_k^{\mathrm{c}}|\boldsymbol{s}_k)}$；
(16)　分别计算Actor网络和Critic网络的损失：$\mathrm{LA}_k(\theta)$，$\mathrm{LC}_k(\omega)$；
(17)　分别计算Actor网络和Critic网络的梯度：$\nabla_\theta l_k^X(\theta)$，$\nabla_\omega l_k^V(\omega)$；
(18)　更新参数$\theta = \theta - \eta\,\nabla_\theta l_k^X(\theta)$，$\omega = \omega - \eta\nabla_\omega l_k^V(\omega)$；
(19) 直到$j = M$；
(20) 更新旧Actor网络的参数：$\theta_{\mathrm{old}} = \theta$；
(21) 清空经验池；
(22)直到$i = \mathrm{EP}$，训练结束。

消耗、无人机飞行距离、任务时间，将CH-PPO算法与DQN算法[13]、PPO算法[15]、H-PPO算法[23]和变长染色体遗传算法(Variable-Length Chromosome Genetic Algorithm,VLC-GA)[24]进行对比。其中，用于处理离散动作空间问题的DQN算法和连续动作空间问题的PPO算法是解决无人机路径规划问题的两类典型强化学习算法；H-PPO算法是用于求解混合动作空间问题的强化学习算法；VLC-GA算法是适用于求解次序问题的传统优化算法。在仿真实验的环境设置中，WSN区域为$A \in \{200\times200\ \mathrm{m}^2, 300\times300\ \mathrm{m}^2, 400\times400\ \mathrm{m}^2\}$，传感器节点数量为$n\in\{20,40,60\}$。无人机从位于区域A中心的基站起飞，完成数据采集任务。实验参数参照文献[21]，如表1所示。

表 1 仿真参数

参数	取值	参数	取值
无人机初始能量E_{init}	1×10^5 J	无人机最大飞行速度$v_{\max}$	30 m/s
无人机飞行高度H	10 m	传感器节点的数据量D	5×10^4 Byte
无人机最大连通半径R	30 m	LoS和NLoS依赖常数a,b	10, 0.6
带宽W	1 MHz	信道功率P_{d}	−20 dBm
非视距信道额外衰减系数μ	0.2	单位信道功率增益ζ	−30 dB
噪声功率σ^2	−90 dBm	通径损失指数α	2.3

CH-PPO算法的网络框架使用tensorflow2.0搭建，离散网络和连续网络共享两层全连接层，神经元个数分别为128,64，激活函数为ReLU。离散网络对应一个神经元个数为n+1的分支输出层，连续网络对应两个神经元个数为3的分支输出层。CH-PPO算法的训练参数如表2所示。

4.2 收敛性验证与分析

图4展示了4种深度强化学习算法在20 000轮训练中获得的回合累计奖励，从算法收敛结果、收敛速度和收敛稳定性3个方面进行分析。从算法收敛结果看，CH-PPO算法和H-PPO算法的无人机同时学习对传感器节点的访问次序、悬停位置和飞行速度的选择策略，收敛值均较高。与H-PPO算法相比，CH-PPO算法中离散策略和连续策略的关联度更高，运动策略更合理，使得奖励收敛值更高；DQN算法和PPO算法存在局限性，前者仅指导无人机学习对传感器节点或基站的访问次序，而忽略了悬停位置和飞行速度对能量消耗的影响；后者仅指导无人机学习悬停位置和飞行速度，而忽略了访问次序对数据AoI的影响，因此二者获得的累计奖励较低。从算法收敛速度看，CH-PPO算法、H-PPO算法和PPO算法收敛速度均慢于DQN算法。CH-PPO算法和H-PPO算法需要同时优化离散动作和多维连续动作，动作空间大于DQN算法，因此CH-PPO算法和H-PPO算法收敛速度较慢。PPO算法仅指导无人机选择悬停位置，在该实验环境下任务完成难度较高，收敛速度慢。从算法收敛稳定性看，CH-PPO算法稳定性略高于H-PPO算法，但相对于DQN算法存在波动，而PPO算法在10 000轮训练之前奖励波动较大，算法收敛后稳定性强。CH-PPO算法和H-PPO算法中离散策略和连续策略相互影响，导致算法稳定性相对较差，但CH-PPO算法中损失函数的设计使得趋近收敛的策略不会对整体策略产生影响，因此算法稳定性略高于H-PPO算法。DQN算法随着探索率的降低和Q值的收敛，逐步趋于稳定；PPO算法策略收敛后稳定性较强。

表 2 网络参数

参数	取值
训练轮数EP	20 000
学习率η	1×10^{-4}
奖励折扣率γ	0.99
裁剪系数ε	0.2

4.3 能量消耗对比实验

表3展示了5种算法下无人机的能量消耗情况。实验结果显示在不同任务规模下，CH-PPO算法指

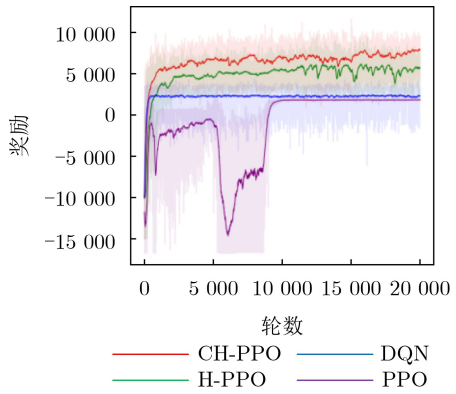

图 4 奖励收敛效果图

导的无人机能耗均少于其他4种算法，且与DQN算法、PPO算法和VLC-GA算法之间的差距随问题规模增加逐渐扩大。CH-PPO算法和H-PPO算法均通过指导无人机对传感器节点和基站的访问次序和悬停位置来实现飞行距离的缩短；通过指导无人机根据AoI阈值限制和当前飞行路线选择飞行速度实现任务时间的优化。CH-PPO算法中无人机飞行路线选择更优，速度选择更合理，因此能量消耗更少。飞行距离和飞行速度对能量消耗影响的分析见4.4节。DQN算法和VLC-GA算法均无法对无人机悬停位置和飞行速度进行优化，PPO算法在较大问题规模下难以找到最优悬停位置，因此，三者的能量消耗明显多于CH-PPO算法和H-PPO算法。另外，DQN算法和VLC-GA算法能耗差距较小，二者均能较好地在AoI限制下完成数据采集任务。

4.4 影响无人机能量消耗的指标对比实验

4.4.1 飞行距离对无人机能量消耗影响

表4展示了5种算法下无人机的飞行距离。实验结果显示在不同任务规模下，CH-PPO算法指导的无人机飞行距离均少于其他4种算法，且与DQN算法、PPO算法和VLC-GA算法之间的差距随问题规模增加逐渐扩大。为获取较多的数据采集阶段奖励，CH-PPO算法和H-PPO算法均指导无人机集中采集某一个区域的数据，减少无效飞行距离。同时，两种算法合理规划无人机对节点的访问次序，在保证数据时效性的同时减少返回基站次数。相较于H-PPO算法，CH-PPO算法中无人机对悬停位置的选择更好地综合了后续访问节点次序，缩短了飞行距离。DQN算法、PPO算法和VLC-GA算法的局限性导致无人机的飞行距离明显多于CH-PPO算法和H-PPO算法；其中，DQN算法和VLC-GA算法仅优化无人机对节点的访问次序，PPO算法则仅优化无人机数据采集或卸载位置。

4.4.2 飞行速度对无人机能量消耗影响

由于无人机的飞行速度不断变化，因此本文通过无人机能耗、飞行距离和任务时间综合分析无人机速度对能耗的影响。表5展示了无人机在5种算法下的任务时间。根据表3和表4所示，DQN算法、PPO算法和VLC-GA算法在所有任务规模下的飞行距离均较长且能量消耗较高，导致无人机任务时间多于CH-PPO算法，此时飞行速度对无人机能耗影响不明显。在区域范围为300×300 m^2，WSN规模分别为40和60以及区域范围为400×400 m^2，WSN规模为60的三个任务规模下，CH-PPO算法的任务

表 3　不同任务规模下的无人机能量消耗(1×10^4 J)

区域边长	网络规模	CH-PPO	H-PPO	DQN	PPO	VLC-GA
200	20	1.71	1.75	2.32	2.82	2.18
	40	4.01	4.55	4.91	4.73	4.95
	60	6.10	6.72	6.82	7.08	6.93
300	20	1.96	2.07	2.61	4.02	2.54
	40	4.89	5.21	6.84	9.28	6.91
	60	9.18	9.45	11.43	12.59	11.51
400	20	2.02	2.11	2.92	4.46	2.84
	40	6.37	7.60	7.62	12.76	7.64
	60	10.03	10.70	12.04	14.78	12.13

表 4　不同任务规模下的无人机飞行距离(m)

区域边长	网络规模	CH-PPO	H-PPO	DQN	PPO	VLC-GA
200	20	1488	1560	2398	2805	2249
	40	3600	4308	5102	4596	5155
	60	5922	6566	7025	6818	7175
300	20	1785	1837	2734	4115	2654
	40	4471	4802	7270	9641	7356
	60	8483	8785	12173	12526	12322
400	20	1898	1921	3066	4623	2991
	40	6001	7242	8121	10767	8177
	60	9367	9964	12880	13135	13019

表 5 不同任务规模下的任务时间(s)

区域边长	网络规模	CH-PPO	H-PPO	DQN	PPO	VLC-GA
200	20	100	105	138	172	130
	40	241	259	291	293	293
	60	323	345	406	440	412
300	20	121	125	154	241	150
	40	**297**	**290**	399	550	403
	60	**550**	**489**	669	684	673
400	20	126	131	171	264	167
	40	355	424	443	583	445
	60	**584**	**549**	699	762	704

时间长于H-PPO算法，但无人机飞行距离略短于H-PPO算法。此时CH-PPO算法在保证数据时效性的前提下，通过对飞行速度的优化使得无人机能够充分利用AoI阈值，减少不必要的高速飞行，能量消耗少于H-PPO算法。

4.4.3 AoI阈值对无人机能量消耗影响

表6展示了在不同任务规模下AoI阈值对无人机能量消耗的影响。实验结果显示，无人机的能量消耗与AoI阈值成反比。

4.5 超参数分析

4.5.1 学习率

图5展示了在WSN规模为40，区域范围为300×300 m^{2}的环境中，CH-PPO算法在学习率$\eta\in\{1\times10^{-3},1\times10^{-4},1\times10^{-5}\}$的奖励收敛情况。实验结果显示CH-PPO算法在以上3种学习率下均能收敛，且在$\eta=1\times10^{-4}$时奖励收敛值最高。$\eta=1\times10^{-3}$时，算法的奖励上升速度较快，但稳定性较差；$\eta=1\times10^{-5}$时，算法收敛上升速度慢，但是收敛曲线平滑稳定性强；综合来看$\eta=1\times10^{-4}$时，收敛效果最好。

4.5.2 裁剪系数

图6展示了在WSN规模为40，区域范围为300×300 m^{2}的环境中，CH-PPO算法在裁剪系数$\varepsilon\in\{0.1,0.2,0.3\}$的奖励收敛情况。实验结果显示，CH-PPO算法在3种裁剪系数下均能以较快速度收敛，稳定性较强，且奖励收敛值相近，验证了该算法的鲁棒性。

表 6 不同任务规模和AoI阈值下的无人机能量消耗(1×10^4 J)

AoI阈值 / 区域边长	网络规模	[60,80]	[90,110]	[120,140]
200	20	1.90	1.71	1.45
	40	5.77	4.01	3.83
	60	8.24	6.10	5.89
300	20	2.19	1.96	1.55
	40	7.79	4.89	4.29
	60	10.47	9.18	7.96
400	20	2.29	2.02	1.95
	40	8.36	6.37	5.80
	60	11.34	10.03	9.57

5 结束语

针对信息年龄约束下的无人机数据采集能耗优化问题，本文提出CH-PPO算法规划无人机对传感

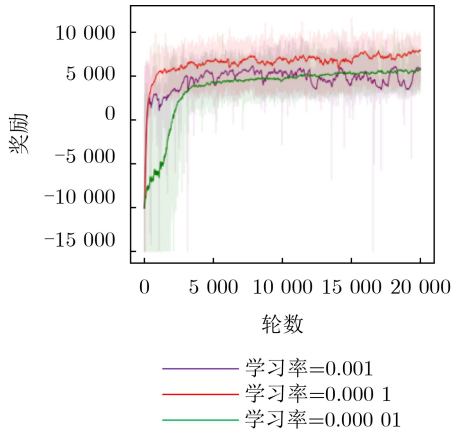

图 5 不同学习率下的奖励收敛图

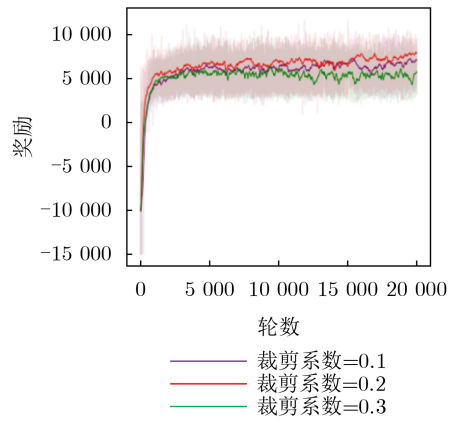

图 6 不同裁剪系数下的奖励收敛图

器节点或基站的访问次序、悬停位置和飞行速度，在满足各传感器节点中数据的AoI阈值限制条件下，最小化无人机能量消耗。仿真实验结果显示，在不同WSN规模和区域范围下，CH-PPO算法下的无人机能量消耗均少于H-PPO算法、DQN算法和PPO算法。此外，本文通过飞行距离、飞行速度和AoI阈值3个方面对影响无人机能耗的因素展开分析。最后，CH-PPO算法通过不同学习率和裁剪系数下的奖励收敛结果，验证了算法的可行性和鲁棒性。未来的工作会尝试完善无人机飞行动力学模型和环境模型，并在实际应用中验证该算法。

参 考 文 献

[1] AKYILDIZ I F, SU W, SANKARASUBRAMANIAM Y, *et al*. Wireless sensor networks: A survey[J]. *Computer Networks*, 2002, 38(4): 393–422. doi: 10.1016/S1389-1286(01)00302-4.

[2] HAYAT S, YANMAZ E, and MUZAFFAR R. Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint[J]. *IEEE Communications Surveys & Tutorials*, 2016, 18(4): 2624–2661. doi: 10.1109/COMST.2016.2560343.

[3] MOTLAGH N H, BAGAA M, and TALEB T. UAV-based IoT platform: A crowd surveillance use case[J]. *IEEE Communications Magazine*, 2017, 55(2): 128–134. doi: 10.1109/MCOM.2017.1600587CM.

[4] HU Jie, WANG Tuan, YANG Jiacheng, *et al*. WSN-assisted UAV trajectory adjustment for pesticide drift control[J]. *Sensors*, 2020, 20(19): 5473. doi: 10.3390/s20195473.

[5] 周彬, 郭艳, 李宁, 等. 基于导向强化*Q*学习的无人机路径规划[J]. 航空学报, 2021, 42(9): 325109. doi: 10.7527/S1000-6893.2021.25109.

ZHOU Bin, GUO Yan, LI Ning, *et al*. Path planning of UAV using guided enhancement *Q*-learning algorithm[J]. *Acta Aeronautica et Astronautica Sinica*, 2021, 42(9): 325109. doi: 10.7527/S1000-6893.2021.25109.

[6] ZHOU Conghao, WU Wen, HE Hongli, *et al*. Delay-aware IoT task scheduling in space-air-ground integrated network[C]. 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, USA, 2019: 1–6. doi: 10.1109/GLOBECOM38437.2019.9013393.

[7] LIU Dianxiong, XU Yuhua, WANG Jinlong, *et al*. Opportunistic utilization of dynamic multi-UAV in device-to-device communication networks[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2020, 6(3): 1069–1083. doi: 10.1109/TCCN.2020.2991436.

[8] 张广驰, 何梓楠, 崔苗. 基于深度强化学习的无人机辅助移动边缘计算系统能耗优化[J]. 电子与信息学报, 2023, 45(5): 1635–1643. doi: 10.11999/JEIT220352.

ZHANG Guangchi, HE Zinan, and CUI Miao. Energy consumption optimization of unmanned aerial vehicle assisted mobile edge computing systems based on deep reinforcement learning[J]. *Journal of Electronics & Information Technology*, 2023, 45(5): 1635–1643. doi: 10.11999/JEIT220352.

[9] LUO Chuanwen, CHEN Wenping, LI Deying, *et al*. Optimizing flight trajectory of UAV for efficient data collection in wireless sensor networks[J]. *Theoretical Computer Science*, 2021, 853: 25–42. doi: 10.1016/j.tcs.2020.05.019.

[10] ZHU Yuchao and WANG Shaowei. Efficient aerial data collection with cooperative trajectory planning for large-scale wireless sensor networks[J]. *IEEE Transactions on Communications*, 2022, 70(1): 433–444. doi: 10.1109/TCOMM.2021.3124950.

[11] ZHAN Cheng and ZENG Yong. Completion time minimization for multi-UAV-enabled data collection[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(10): 4859–4872. doi: 10.1109/TWC.2019.2930190.

[12] KAUL S, YATES R, and GRUTESER M. Real-time status: How often should one update?[C]. 2012 Proceedings IEEE INFOCOM, Orlando, USA, 2012: 2731–2735. doi: 10.1109/INFCOM.2012.6195689.

[13] 张建行, 康凯, 钱骅, 等. 面向物联网的深度Q网络无人机路径规划[J]. 电子与信息学报, 2022, 44(11): 3850–3857. doi: 10.11999/JEIT210962.

ZHANG Jianhang, KANG Kai, QIAN Hua, *et al*. UAV trajectory planning based on deep Q-network for internet of things[J]. *Journal of Electronics & Information Technology*, 2022, 44(11): 3850–3857. doi: 10.11999/JEIT210962.

[14] LIAO Yuan and FRIDERIKOS V. Energy and age pareto optimal trajectories in UAV-assisted wireless data collection[J]. *IEEE Transactions on Vehicular Technology*, 2022, 71(8): 9101–9106. doi: 10.1109/TVT.2022.3175318.

[15] SHERMAN M, SHAO Sihua, SUN Xiang, *et al*. Optimizing AoI in UAV-RIS-assisted IoT networks: Off policy versus on policy[J]. *IEEE Internet of Things Journal*, 2023, 10(14): 12401–12415. doi: 10.1109/JIOT.2023.3246925.

[16] SUN Mengying, XU Xiaodong, QIN Xiaoqi, *et al*. AoI-energy-aware UAV-assisted data collection for IoT networks: A deep reinforcement learning method[J]. *IEEE Internet of Things Journal*, 2021, 8(24): 17275–17289. doi: 10.1109/JIOT.2021.3078701.

[17] LIU Juan, TONG Peng, WANG Xijun, *et al*. UAV-aided data collection for information freshness in wireless sensor networks[J]. *IEEE Transactions on Wireless Communications*, 2021, 20(4): 2368–2382. doi: 10.1109/TWC.2020.3041750.

[18] DAI Zipeng, LIU C H, YE Yuxiao, *et al*. AoI-minimal UAV crowdsensing by model-based graph convolutional reinforcement learning[C]. IEEE INFOCOM 2022-IEEE Conference on Computer Communications, London, United Kingdom, 2022: 1029–1038. doi: 10.1109/INFOCOM48880.2022.9796732.

[19] LIU Kai and ZHENG Jun. UAV trajectory optimization for time-constrained data collection in UAV-enabled environmental monitoring systems[J]. *IEEE Internet of Things Journal*, 2022, 9(23): 24300–24314. doi: 10.1109/JIOT.2022.3189214.

[20] SUN Yin, UYSAL-BIYIKOGLU E, YATES R D, *et al*. Update or wait: How to keep your data fresh[J]. *IEEE Transactions on Information Theory*, 2017, 63(11): 7492–7508. doi: 10.1109/TIT.2017.2735804.

[21] YU Yu, TANG Jie, HUANG Jiayi, *et al*. Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm[J]. *IEEE Transactions on Communications*, 2021, 69(9): 6361–6374. doi: 10.1109/TCOMM.2021.3089476.

[22] ZENG Yong, XU Jie, and ZHANG Rui. Energy minimization for wireless communication with rotary-wing UAV[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(4): 2329–2345. doi: 10.1109/TWC.2019.2902559.

[23] FAN Zhou, SU Rui, ZHANG Weinan, *et al*. Hybrid actor-critic reinforcement learning in parameterized action space[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 2019.

[24] HA V P, DAO T K, PHAM N Y, *et al*. A variable-length chromosome genetic algorithm for time-based sensor network schedule optimization[J]. *Sensors*, 2021, 21(12): 3990. doi: 10.3390/s21123990.

高思华：男，讲师，研究方向为强化学习理论、最优化理论、无线传感器网络和无人机系统.

刘宝煜：男，硕士生，研究方向为强化学习理论、无人机路径规划.

惠康华：男，副教授，研究方向为计算机视觉.

徐伟峰：男，讲师，研究方向为计算机视觉和空管系统.

李军辉：男，硕士生，研究方向为强化学习理论、无人机路径规划.

赵炳阳：男，硕士生，研究方向为强化学习理论、无人机路径规划.

责任编辑：陈 倩