DOI:10.19652/j.cnki.femt.2305883

基于深度强化学习的无人机辅助物联网多目标优化*

徐钰龙[1] 李 君[1,2] 李正权[3] 胡 静[1] 张 圣[1] 王子威[1]
(1.南京信息工程大学电子与信息工程学院 南京 210044;2.无锡学院 无锡 214105;
3.北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要:无人机辅助无线供电物联网是一种创新的网络架构,利用无人机作为能量传输中介,能够解决物联网设备电力供应的限制和局限性。针对无人机辅助无线供电物联网网络中多目标控制策略学习的问题,提出了一种基于深度强化学习的多目标双延迟深度确定性策略梯度(MOTD3)算法,旨在满足偏航角、飞行速度以及发射功率约束条件下,实现总数据速率、总收获能量最大化以及能耗和悬停时间最小化的多目标联合优化,同时因需求动态变化无人机进行在线路径规划。仿真结果表明,该算法在保证良好的收敛情况和稳定性前提下,较其他算法在总数据速率、总收获能量、能耗与悬停时间方面分别提高14.7%、10.6%、6.1%和10.3%,且具有较强泛化能力,可适用于实际中不同通信场景。

关键词:物联网;无人机;深度强化学习;多目标优化;路径规划
中图分类号:TN929.5 **文献标识码**:A **国家标准学科分类代码**:510.5030

Multi-objective optimization of unmanned aerial vehicle assisted internet of things based on deep reinforcement learning

Xu Yulong[1] Li Jun[1,2] Li Zhengquan[3] Hu Jing[1] Zhang Sheng[1] Wang Ziwei[1]
(1. College of Electronic and Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China; 2. Wuxi University, Wuxi 214105, China; 3. State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The unmanned aerial vehicle (UAV)-assisted wireless power supply for the internet of things (IoT) is an innovative network architecture where UAVs serve as energy transmission intermediaries, effectively addressing the limitations and constraints of power supply for IoT devices. In addressing the challenge of multi-objective control policy learning in UAV-assisted wireless power supply for the IoT, this study proposes a multi-objective twin-delay deep deterministic policy gradient (MOTD3) algorithm based on deep reinforcement learning. The MOTD3 algorithm aims to achieve joint optimization of multiple objectives, including maximizing the total data rate and total harvested energy, while minimizing energy consumption and hover time, under constraints such as yaw angle, flight speed, and transmission power. Additionally, it adapts UAVs to dynamic demand changes through online path planning. Simulation results demonstrate that the proposed algorithm can improve the total data rate, total harvest energy, energy consumption and hover time by 14.7%, 10.6%, 6.1% and 10.3% respectively compared with other algorithms, and has strong generalization ability, which can be applied to different communication scenarios in practice.

Keywords: internet of things(IoT); unmanned aerial vehicle(UAV); deep reinforcement learning(DRL); multi objective optimization; trajectory optimization

0 引 言

物联网(internet of things,IoT)被视为未来网络的关键和具有发展前景的技术,是新一代信息技术的重要组成部分,能够相互通信、收集和交换数据,实现信息共享和智能化决策[1-2]。随着物联网的大规模推广和应用,终端设

收稿日期:2023-12-26
* **基金项目**:网络与交换技术全国重点实验室(北京邮电大学)开放课题项目(SKLNST-2023-1-13)资助

备数量呈爆炸式增长趋势。预计到2030年,将有约5 000亿台设备配备传感器接入互联网[3]。物联网设备的不断增加对通信系统也提出了更高的要求,包括更高的数据速率,采集信息的及时性[4],能源消耗与供应问题,同时常见的地面空间通信受到了一定的局限性,因此需要向三维空间进行扩展[5],此时无人机便是首要选择对象。无人机凭借其高机动性和低部署成本,被广泛应用于无线网络中,以提高通信覆盖率、系统容量和部署效率[6-11]。结合无线电力传输(wireless power transfer,WPT)无人机可以实现对广泛分布的物联网设备进行直流处理和能量传输。近年来,针对无人机辅助无线供电物联网网络的优化问题进行了大量研究工作[12-16]。然而现有研究的优化目标各不相同,文献[12-14]的目标是最大化物联网设备的上行链路吞吐量。文献[15]的目标是最大化地面终端的最小吞吐量。文献[16]的目标是最大化总吞吐量、最小化总时间和总能量。在一些环境瞬息万变的应用中,保证数据的实时性是非常重要的,而上述文献均未考虑该实际因素。文献[17-18]的目标是追求收集数据的新鲜度,最小化传感数据的信息年龄。文献[19]的目标是揭示能量消耗与任务完成时间之间的权衡关系。无人机的机动性、物联网系统的随机性和动态性对无人机辅助无线物联网网络的优化带来了巨大的挑战。面对复杂、动态的物联网网络环境,需要无人机具备对周围环境的感知能力和实时决策能力。传统的优化方法在应对这些复杂网络优化方面变得难以有效管理。为了智能且高效地处理动态复杂环境中的优化问题,基于深度强化学习(deep reinforcement learning,DRL)的方法开始被应用于解决无人机辅助物联网网络中资源优化问题,文献[20]提出了一种基于DRL的无人机控制策略,以实现在通信覆盖率、公平性和能耗方面的优化。文献[21]提出了一种基于深度确定性策略梯度(DDPG)的方法,以分布式方式控制每个无人机,最大化所有考虑的兴趣点(point of interest,PoI)的地理公平性并最小化总能耗。文献[22]提出了一种深度Q网络解决方案,以实现在传感区域收集最需要的数据。其中,文献[20-21]并没有考虑用户的各种优先级,而文献[22]虽然考虑了优先级目标,但是其在执行任务期间数据的优先级要求是确定的,并没有考虑实际应用场景中动态数据优先级的要求。由于在大多数实际应用场景中,部署了大量地面节点来观测物理过程的实时更新,因此不能忽视对传感数据的动态优先级要求。文献[23-24]提出了描述实时数据更新过程的数据生成模型,并基于DQN算法开发了无人机在线路径规划。但是其关于状态空间的设置过于庞大且复杂,导致运算要求和计算成本太高,不利于实际应用。

总之,现有的大多数文献虽然考虑到了优化无人机辅助物联网中的资源分配问题,但是其多数采用仍是单目标优化或多个目标分别优化的方式,且选取优化目标的同时仍缺少对关键因素的考虑,如实际场景中动态数据的优先级,数据收集的及时性,时间成本以及无人机能源消耗等。基于此,本文考虑的是多目标的联合优化 ,同时因动态需求实现无人机在线路径规划,考虑动态数据的优先级以及时间成本,而提出一种利用深度强化学习方法,提取与无人机飞行决策密切相关的少量信息组成状态向量,构建系统模型,旨在满足其约束条件下实现多目标的联合优化。

本文提出在无线供电的物联网网络中,利用无人机辅助数据收集和能量传输,实时更新物联网设备上传数据的需求,无人机采用飞-悬停通信协议,根据物联网设备的需求优先级依次访问物联网设备。

本文研究了一个多目标优化问题,该问题旨在最大限度地提高总数据速率和总收集能量,同时最小化无人机的能量消耗和悬停时间,提出了一种求解无人机飞行决策最优策略的MOTD3算法,将奖励设计为一个五维向量,其中4个元素对应4个优化目标,另一个辅助元素保证基本任务的完成,并将经典的TD3算法扩展到多维奖励。

通过实验结果表明,基于MOTD3算法的最优策略比DDPG、优势演员评论家算法(advantage actor-critic,A2C)、传统的基于规则的策略更具优势与灵活性,且在不同环境下更具有泛化性。

1 系统模型与问题构造

1.1 系统模型

本文研究的是一个无人机辅助的无线供电物联网网络,网络模型如图1所示。该网络中具有一个双天线无人机和$m\in\{1,2,\cdots,M\}$个单天线物联网设备,物联网设备随机分布在有限的地理区域中,$[x_m,y_m]$表示物联网设备m的位置。由于无人机的能量有限,每次飞行任务持续一段时间,持续时间$T>0$。无人机采用飞行-悬停通信协议,即无人机在飞行时不与地面节点通信,仅在悬停时进行直流处理和能量传输。该无人机配备了混合接入点(hybrid access point,HAP),当它悬停在特定的位置时,以全双工模式工作,利用一根天线在下行链路向物联

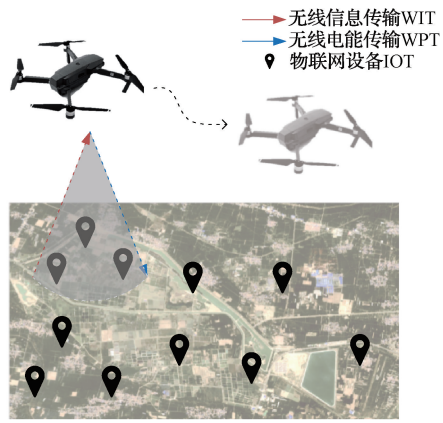


图1 系统模型

Fig. 1 System model

网设备传输能量，同时利用另一根天线在上行链路从物联网设备收集数据。

物联网设备在线监控多种物理过程，实时收集所观察到的过程状态更新数据，并将其存储于数据缓冲区中，其更新数据过程表示为：

$$D_m(t+\Delta t)=D_m(t)+\lambda_m(t)\Delta t \tag{1}$$

式中：$D_m(t)$ 为 t 时队列中等待上传的数据($0\leqslant t\leqslant T$)；Δt 为更新间隔；$\lambda_m(t)$ 为 t 时物联网设备 m 的数据生成速率，其服从泊松分布，不同设备的泊松分布参数不同[25]。由于 $D_m(t)$ 通常受到硬件限制，假设限制于 $[0,D_{max}]$，D_{max} 为最大数据缓冲区的存储容量且每个设备都相同。当数据缓冲区存储满时，可能会产生旧数据被新数据覆盖或新收集的数据上传丢失，因此迫切需要物联网设备能够及时将收集的数据信息上传到无人机。

物联网设备 m 在 t 时传输的数据表示为：

$$N_m(t)=\frac{D_m(t)}{D_{max}}N \tag{2}$$

式中：N 为最大数据缓冲区 D_{max} 对应的传输数据量。

由于不同设备数据缓冲区的长度和数据生成速率不同，上传数据的优先级也不同。设备 m 的数据上传优先级表示为：

$$q_m^u(t)=\lambda_m(t)\frac{D_m(t)}{D_{max}} \tag{3}$$

数据传输优先级不仅取决于收集到的数据占存储容量的比例，还受到数据生成速率的影响。它含有对未来优先级的预测。

无人机飞行于固定高度 $H>0$，t 时的水平位置坐标为 $[x_u(t),y_u(t)]$，此处忽略悬停高度。无人机使用飞行速度 $v(t)$ 和偏航角 $\theta(t)\in[-\pi,\pi]$ 来描述飞行控制，其中 $v(t)$ 受限于最大飞行速度 $v_{max}=20$ m/s。在飞行速度 $v(t)$ 时的推进功率消耗[26]表示为：

$$P(V)=P_0\left(1+\frac{3V^2}{U_{tip}^2}\right)+P_i\left(\sqrt{1+\frac{V^4}{4v_0^4}}-\frac{V^2}{2v_0^2}\right)^{1/2}+\frac{1}{2}d_0\rho sAV^3 \tag{4}$$

式中：P_0 为悬停状态时的叶片轮廓功率；U_{tip} 为转子叶片的尖端速度；P_i 为悬停状态时的感应功率；v_0 为悬停状态时平均旋翼感应速度；d_0 为机身阻力比；ρ 为空气密度；s 为叶片覆盖比例；A 为旋翼面积。推进功率消耗随速度的变化趋势如图 2 所示。式(4)能够计算出最低功耗，其相对应的速度称为最大续航速度 VME，$V=0$ 计算得到悬停功率 P_{hov}。

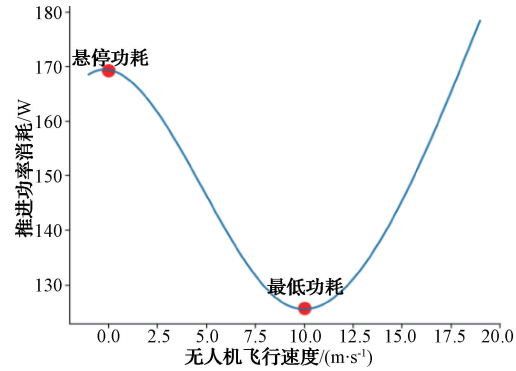


图 2 推进功率变化趋势

Fig. 2 Trend diagram of propulsion power change

1.2 问题构造

本文的目标是最大化总数据速率和总收集能量，同时最小化无人机的能量消耗和悬停时间。无人机需要感知物联网环境，实现实时路径规划。无人机飞行轨迹的确定和悬停位置的选择应考虑设备的服务质量和无人机的能耗，并按照物联网设备的实时需求优先级依次访问这些设备。在第 j 处悬停时($0\leqslant j\leqslant J$)，传输数据速率表示为：

$$R^j=W\log_2\left(1+\frac{P_u\mid h_{m^j}(t)\mid^2}{\sigma^2}\right) \tag{5}$$

式中：W 为无线通信带宽；P_u 为物联网设备上行发射功率；σ^2 为无人机的信道噪声功率。

将收集到的数据上传到无人机，其使用的悬停时间为：

$$t^j=\frac{N_{m^j}(t)}{R^j} \tag{6}$$

假设无人机的直流和能量传输范围有限，无人机仅对覆盖范围内的物联网设备传输能量和收集数据，用 R_{dC} 和 R_{eh} 表示数据收集和能量传输的最大覆盖半径。每一时刻，无人机都会选择一个物联网设备作为目标设备进行数据收集。当目标设备位于 R_{dC} 范围内，无人机进行悬停接收相应位置的信息，同时向 R_{eh} 范围内的其他设备传输能量，直到目标设备完成数据上传。物联网设备 m 的接收功率表示为：

$$P_m^r(t)=\mid h_m(t)\mid^2P_c,\quad \forall\Delta d_m(t)\leqslant R_{eh} \tag{7}$$

式中：$h_m(t)$ 为无人机与物联网设备 m 之间的信道功率增益；P_c 为物联网设备的发射功率；$d_m(t)$ 为无人机与物联网设备 m 之间的距离。

为贴合实际应用，本文采用非线性能量收集模型，考虑了电路的饱和限制，收集的能量表示为：

$$P_m^h(t)=\frac{P_{max}^{cd}e^{cd}-P_{max}^{cd}e^{-c(P_m^r(t)-d)}}{e^{cd}(1+e^{-c(P_m^r(t)-d)})} \tag{8}$$

式中：P_{max}^{cd} 为最大直流输出功率，c 和 d 分别为能量收集系统相关电路特性常数。

无人机向其能量传输覆盖范围内的设备传输能量，物联网设备 m 所收集的能量为：

$$E_m=P_m^h(t)t^j\ \forall,\quad \Delta d_m(t)\leqslant R_{eh},m\neq m^j \tag{9}$$

那么，第 j 处悬停时，所收集能量为：

$$E^j=\sum_{m,\forall\Delta d_m(t)\leqslant R_{eh},m\neq m^j}E_m \tag{10}$$

所有悬停阶段的总数据速率和总收集能量如下所示：

$$R_{sum}=\sum_{j=0}^{J}R^{j} \tag{11}$$

$$E_{sum}^{h}=\sum_{j=0}^{J}E^{j} \tag{12}$$

无人机在任务持续时间内飞行和悬停的总能量消耗为：

$$E_{sum}^{c}=\int_{0}^{T}P(v(t))\mathrm{d}t+P_{dl} \tag{13}$$

式中：P_{dl} 为下行链路发射功率，设为常数。

本文的优化目标为多目标优化，优化问题公式表示为：

$$\begin{cases}max_{v(t),\theta(t),P_u}(R_{sum},E_{sum}^{h},-E_{sum}^{c},-t^{j})\\ \text{s.t.}\quad 0\leqslant v(t)\leqslant v_{\max},-\pi\leqslant\theta(t)\leqslant\pi,\\ \quad 0\leqslant P_u\leqslant P_u^{max}\end{cases} \tag{14}$$

为了最大化总数据速率，一方面无人机应该以更高的速度飞行，以便访问更多的物联网设备。另一方面，悬停位置应靠近目标设备，以提高数据速率。为了最大化收集的能量，除了最大化悬停次数 J 之外，本文还希望每次悬停时无人机覆盖范围内的设备更多。此外，无人机与充电设备之间的距离越小越好。这与无人机直接悬停在目标设备上方以获得最大数据速率相冲突。对于无人机的能耗目标，VME 可以实现其最小化，然而速度不够快，无法收集更多的数据并为更多的设备充电。此外，飞行速度低可能导致物联网设备的数据溢出。对于无人机悬停时间目标，悬停时间缩短可能会造成收集能量减少或造成数据传输损失。这 4 个目标之间存在一定程度上的冲突，且物联网设备随机分布，其数据生成是动态的，因此寻找最优悬停位置并进行飞行决策十分复杂困难并且计算成本大。

2 算法设计

2.1 深度强化学习

本文提出的优化问题是一个非凸的问题，很难用最优化的方法解决这个问题。根据强化学习的理论，对于一个给定的马尔可夫随机过程，尤其当系统是动态时，深度强化学习可以找到最优的决策动作，解决这个多目标优化问题，从而实现最大化总速率和收集的能量，并最小化能量消耗和悬停时间。强化学习具有很强的环境交互能力，智能体与环境交互的过程可通过马尔可夫决策过程(markov decision process，MDP)进行描述，并将式(14)转换为决策问题。MDP 由一个 4 元组来构成，即 $M=\{S,A,P,R\}$。其中 S 为状态空间，A 为动作空间，P 是当前状态转移到下一个状态的状态转移概率，R 为累积回报。TD3 作为一种强化学习算法，适用于在连续区间内选择飞行速度和偏航角的无人机飞行决策问题。本文设计的状态空间、动作空间和奖励函数如下。

1)状态空间

本文将无人机作为一个智能体，状态的选取对于智能体很重要，在实际场景中，无人机无法获取全局网络信息，每个设备的实时信息是未知的，且大多数信息对于决策来说并不是必须的。收集所有物联网设备的实时服务需求依赖于无人机与物联网设备之间频繁的信息交换。因此，本文假设无人机只能观察到自己的状态和部分网络信息，即无人机可以观察到自身位置，累计飞出禁区的次数，目标设备的位置，数据丢失设备数量。最终，状态空间表示为：

$$S=\{[d_{\hat{m}}^{x}(t),d_{\hat{m}}^{y}(t)],[x_u(t),y_u(t)],N_r(t),N_d(t)\} \tag{15}$$

式中：$[d_{\hat{m}}^{x}(t),d_{\hat{m}}^{y}(t)]$ 为笛卡尔坐标下，目标设备与无人机之间的距离。当无人机完成对目标设备的数据收集，则根据当前系统状态选择一个新的目标设备，该元素有助于引导无人机将目标设备纳入其数据收集范围。$[x_u(t),y_u(t)]$ 为无人机的绝对坐标位置，有助于防止无人机飞出指定区域，避免资源浪费。$\mathrm{N}_r(t)$ 表示时间 t 之前无人机连续超出禁区的累计次数，$N_d(t)$ 表示数据丢失的设备数量，有助于驱使无人机及时服务高需求设备。

2)动作空间

根据观察到的状态，无人机使用飞行速度 $v(t)$ 和偏航角 $\theta(t)$ 来进行飞行控制。动作空间定义为：

$$A=\{[v(t)\cos\theta(t),v(t)\sin\theta(t)]\} \tag{16}$$

式中：$[\cos\theta(t),\sin\theta(t)]$ 表示偏航。其中飞行速度 $v(t)\in[0,v_{\max}]$ 和偏航角 $\theta(t)\in[-\pi,\pi]$ 为连续值，偏航角描述了无人机的航向偏差。与离散动作空间相比，连续动作空间增加了无人机控制的自由度也提高了控制方案的效率。

3)奖励函数

奖励函数用于衡量智能体在环境中每一步所获得的奖励，引导智能体做出适当的动作。奖励函数的设计应与智能体的目标一致，这样才能促进智能体最大化获得累积回报。对于提出的多目标优化问题，奖励函数被设计为一个五维向量。

$$R=\{r_{dc}(t),r_{eh}(t),r_{ec}(t),r_{\mathrm{ht}}(t),r_{ar}(t)\} \tag{17}$$

式中：$r_{dc}(t)$、$r_{eh}(t)$、$r_{ec}(t)$、$r_{\mathrm{ht}}(t)$ 表示 4 个优化问题，分别是最大化总数据速率、最大化总收集能量，最小化无人机能量消耗和悬停时间。当目标设备落在无人机的数据收集覆盖半径内，无人机就会悬停进行数据收集和能量传输，否则无人机就处于飞行状态。对智能体设置更高的奖励，鼓励其实现更高的数据传输速率，为更多悬停时的物联网设备提供更多的能量收集，并对其在飞行和悬停阶段的较高能量消耗和使用较多悬停时间进行惩罚。ω_{dc}、ω_{eh}、ω_{ec}、ω_{ht} 为每个属性相关联的优先级权重。

$$r_{ar}(t)=-d_{\hat{m}}^{x}(t)-d_{\hat{m}}^{y}(t)-\mathrm{N}_r(t)-N_d(t) \tag{18}$$

式中：$r_{ar}(t)$ 为辅助奖励，当无人机距离目标设备较远时，该值会很小，这有助于无人机识别目标设备的位置，从而接近目标设备。当无人机试图飞出禁区或由于无法及时收集数据导致物联网设备数据溢出，则会获得负奖励。当

无人机执行错误飞行决策时，进行惩罚，使其无论优化目标的偏好如何，都能学会完成基本任务。对应的权重 ω_{ar} 始终设置为1。

2.2 基于深度强化学习多目标优化算法

DDPG算法是一种基于Actor-Critic框架下的一种深度强化学习算法，每个Critic网络都以与Actor网络完全相同的网络结构构造，以便Critic网络对Actor网络进行评估。采用 $Q_\pi(s_t,a_t)$ 函数作为动作价值函数。在时隙 t，通过观察状态 $s_t \in S$，智能体采取动作 $a_t \in A$，与环境交互，得到奖励 r_t，Q 值函数为：

$$Q_\pi(s_t,a_t)=E[U_t \mid S_t=s_t,A_t=a_t] \tag{19}$$

$$U_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\gamma^3 r_{t+3}+\cdots=\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \tag{20}$$

式中：$E[\cdot]$ 是期望算子；$Q_\pi(s_t,a_t)$ 是回报 U_t 的条件期望；U_t 表示从时隙 t 开始，未来所有奖励的加权求和；$\gamma \in [0,1]$ 是权衡当前奖励和未来奖励重要性的折扣因子，如果 $\gamma=0$ 表示智能体只关注当前奖励；

在传统DDPG方法中，目标Critic网络的输出值表示为：

$$y_t^{DDPG}=r_t+\gamma Q'(S_{t+1},a' \mid \theta') \tag{21}$$

由于DDPG算法每次学习时并没有使用下一次交互时的真实动作，而是使用当前被认为最具有价值的动作来更新目标函数，因此会出现Q值的高估问题。

为解决这一问题，TD3算法引入了两个 Q 网络（$Q1$，$Q2$）来表示不同的 Q 值，通过选择较小的网络作为更新目标，抑制连续的高估以及提高了算法的稳定性。TD3中目标Critic网络的输出值表示为：

$$y_t^{1,2}=r_t+\gamma \min Q'_{1,2}(S_{t+1},a' \mid \theta'_{1,2}) \tag{22}$$

根据当前网络 Q 值与目标网络 Q 值，利用均方误差计算损失函数，损失函数表示为：

$$L(\theta_{1,2})=E[(y_t-Q_\pi(s_t,a_t \mid \theta))^2_{1,2}] \tag{23}$$

为了减少累计误差，降低方差以及缓解策略震荡，TD3算法采取延迟策略更新，Critic网络的更新频率调整为高于Actor网络的更新频率，提高对Q值的准确估计。网络更新方式为：

$$\theta_{i=1,2} \leftarrow \min_{\theta_i}\frac{1}{N}\sum(y_t-Q_\pi(s_t,a_t \mid \theta))^2_i \tag{24}$$

$$\nabla_{\theta^\mu}J(\theta^\mu)=\frac{1}{N}\sum \nabla_a Q_{\theta_{i=1,2}}(s_t,a_t \mid \theta_i)\mid_{a=\mu_\varphi(s_t)} \nabla_{\theta^\mu}\mu(s \mid \theta^\mu) \tag{25}$$

式中：N 为样本数量；μ 为选择动作的策略。

此外，目标网络的更新采取软更新，提高学习的稳定性。目标网络参数更新方式为：

$$\begin{cases}\theta'_{i=1,2}=\tau\theta_i+(1-\tau)\theta'_i\\ \theta^{\mu'}=\tau\theta^\mu+(1-\tau)\theta^{\mu'}\end{cases} \tag{26}$$

引入目标策略平滑处理，在目标策略的输出上加入一定的噪声，以减少对目标 Q 值的高估计，使估计更平滑，处理过程表示为：

$$\begin{cases}y_t^{1,2}=r_t+\gamma \min Q'_{1,2}(S_{t+1},a' \mid \theta'_{1,2}+\varepsilon)\\ \varepsilon \sim clip(N_0(0,\sigma),-c,c)\end{cases} \tag{27}$$

式中：ε 为随机噪声。随机噪声需要在一定范围内截断，以保证目标策略网络的输出值在一定范围内。MOTD3训练框图如图3所示。

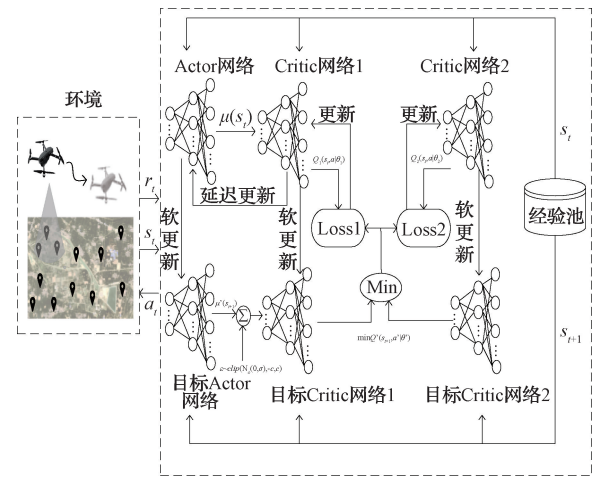


图3 MOTD3训练框图

Fig. 3 MOTD3 training block diagram

算法MOTD3的训练过程如下。

算法 MOTD3算法

初始化两个Critic网络参数 θ_1、θ_2 和Actor网络参数 θ^μ，目标Critic网络参数 θ'_1、θ'_2，目标Actor网络参数 $\theta^{\mu'}$

初始化权重参数 $\omega=[\omega_{dc},\omega_{eh},\omega_{ec},\omega_{ht},\omega_{ar}]$，回合数 E，经验池 D，学习率 α，批量样本数 K

输入：状态，即目标设备与无人机之间的距离 $[d^x_{\hat{m}}(t),d^y_{\hat{m}}(t)]$，无人机的绝对坐标位置 $[x_u(t),y_u(t)]$，累计飞出禁区的次数 $N_r(t)$ 和数据丢失的设备数量 $N_d(t)$

for $e=1,2,\cdots,E$

初始化状态 s_t

for $t=1,2,\cdots,T$

智能体在状态 s_t 下，选择动作 $a_t \sim \mu(s_t)+\varepsilon$，$\varepsilon \sim N(0,\sigma)$

执行动作 a_t，根据式(17)获得奖励 r_t，并进入下一个状态 s_{t+1}

将获得的经验 (s_t,a_t,r_t,s_{t+1}) 存储至经验池 D 中

从经验池 D 中根据批量样本数 K 抽取样本 (s_t,a_t,r_t,s_{t+1}) 用于训练神经网络

根据式(22)计算目标Critic1,2网络输出值 $y_{1,2}$

计算损失函数 $L(\theta_{1,2})$，反向传播更新Critic1,2网络参数 $\theta_{1,2}$

if t % A_I ==0：

根据式(25)更新Actor网络参数 θ^μ

15　　根据式(26)更新目标网络参数 $\theta'_{i=1,2}$、$\theta^{\mu'}$
16　结束
17　结束

3　仿真结果与分析

3.1　仿真环境设置

在无人机辅助的无线物联网网络中，设定 100 个物联网设备随机分布在 500 m×500 m 的正方形区域内。每次任务开始时，无人机在指定区域的随机位置开始任务，飞行高度为 10 m，最大飞行速度 $v_{\max}=20$ m/s。无人机的收集数据半径 R_{dC} 和能量传输半径 R_{eh} 分别设置为 10 和 30 m。数据缓冲区最大容量 D_{max} 为 5 000，对应的传输数据大小 Q 为 10 Mbits。在 MOTD3 算法中，神经网络隐藏层的激活函数均采用 ReLU，而 Actor 网络的最终输出层设置为约束动作的 tanh 层。所有实验采用 NVIDIA GeForce RTX 4060Ti(32 GB 内存)和 Intel i5-13600KF CPU，使用 Python3.7、TensorFlow 2.2.0 和 CuDA 12.1 在 Windows10 系统上实现了实验代码。其他系统参数如表 1 所示。

表 1　训练超参数设置

Table 1　Training hyperparameter settings

参数	值
带宽 B	1 MHz
物联网设备的发射功率 P_c	40 dBm
叶片轮廓功率 P_0	79.86
感应功率 P_i	88.63
转子叶片的尖端速度 U_{tip}	120 m/s
平均旋翼感应速度 v_0	4.03
机身阻力比 d_0	0.6
空气密度 ρ	1.225 km/m^3
叶片覆盖比例 s	0.05
旋翼面积 A	0.503 m^2
最大直流输出功率 P_{max}^{cd}	9.079 μW
经验池大小 D	16 000
批量样本数量 B	128
Actor 学习率 α_A	0.001
Critic 学习率 α_C	0.001
软更新参数 τ	0.001

3.2　仿真结果与分析

将本文提出的 MOTD3 算法与其他深度强化学习算法(DDPG、A2C)进行实验对比，其中为了进一步验证算法性能，将 DDPG、A2C 参数与本文算法参数设置一致。图 4 所示为本文算法和 DDPG、A2C 的训练结果，并通过应用平滑窗口进行了处理，以展示其收敛情况。从图 4 可以

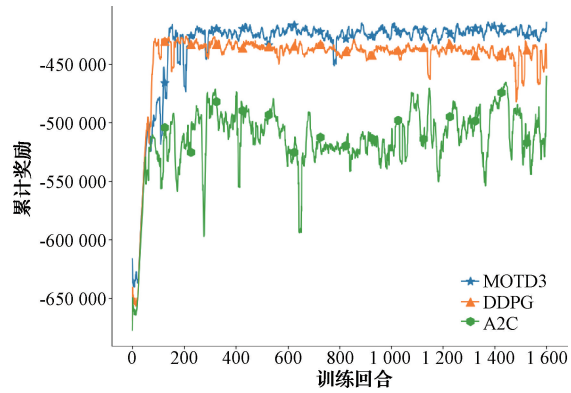
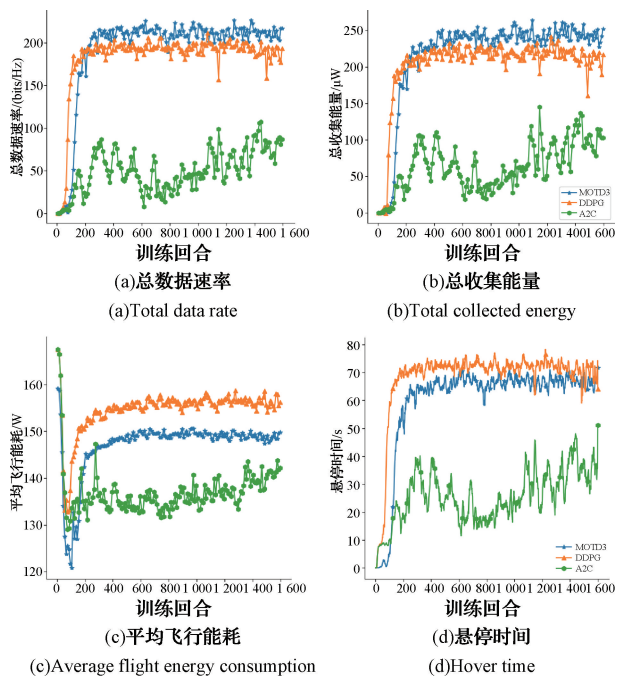


图 4　MOTD3 与 DDPG、A2C 收敛性能对比

Fig. 4　Comparison of convergence performance of MOTD3, DDPG and A2C

看出，随着迭代次数的增加，3 种深度强化学习算法的所获得的累计奖励都有所提升。但是在相同的训练回合内，本文提出的 MOTD3 算法相较于其他深度强化学习算法整体累计奖励更高，最终能收敛到的累计奖励值更大，故本文算法具有更显著的优势。其原因在于本文算法通过引入两个独立的 Q 网络来评估动作的价值以及使用目标策略平滑处理，减少了过估计问题，采取延迟策略更新，也有助于让 Q 网络更加稳定。在复杂的无人机辅助物联网网络环境中，过估计问题更为突出，而本文算法的改进有助于减少过估计，提高奖励的准确性和稳定性。

在训练过程中，3 种深度强化学习算法对于优化目标的性能对比如图 5 所示。从图 5 可以观察到，不同阶段各



(a)总数据速率　(b)总收集能量

(a)Total data rate　(b)Total collected energy

(c)平均飞行能耗　(d)悬停时间

(c)Average flight energy consumption　(d)Hover time

图 5　MOTD3 与 DDPG、A2C 多目标优化性能对比

Fig. 5　Comparison of multi-objective optimization performance of MOTD3, DDPG and A2C

算法在优化目标上的表现差异，初始阶段无人机为了提高总数据速率和总收集能量，在能量消耗与悬停时间方面消耗较多但 DDPG 算法在这个阶段产生的能耗与悬停时间最高。A2C 算法在能耗和悬停时间方面表现较优但其所获得的数据速率和收集能量最少。然而随着智能体在不断的训练过程中，逐渐学习到一种能够在最大化总数据速率，总收集能量和最少悬停时间的同时，最小化能耗的多目标优化策略。本文算法所获得的总数据速率和总收集能量更高于 DDPG 与 A2C 算法，且能够保证能耗更低，悬停时间更少，这为实际应用中的能源消耗和任务执行时间等关键指标提供了更好的保障。相比之下，本文所提出的算法在无人机辅助物联网网络环境中进行多目标优化时表现更为优越。

3 种深度强化学习算法在训练过程中针对优化结果的性能对比如图 6 所示。从图 6 可以看出，为了提高总数据速率，无人机在学习过程中将更多的目标设备纳入数据收集范围 R_{dC}，以激活数据收集。在此过程中平均数据速率也很快达到了最大值，随着能量收集设备数的增加平均速率会有略微降低，但本文算法相较于 DDPG 与 A2C 算法，能够在提高能量收集设备数的同时获得更高的平均数据速率。此外，本文算法训练过程中的平均无人机数据丢失的设备数相较于其他算法数量更少，这表明无人机能够更及时的为高需求设备提供服务。这对于保障通信质量和满足设备需求至关重要。进一步说明了本文算法在无人机辅助物联网网络环境中对于多目标优化结果的表现更为良好。

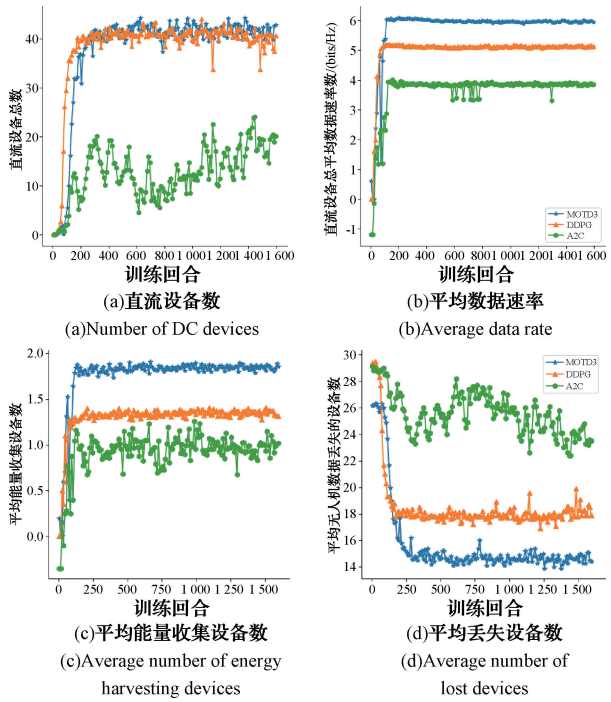


图 6　MOTD3 与 DDPG、A2C 多目标优化结果性能对比

Fig. 6　Performance comparison of multi-objective optimization results of MOTD3, DDPG and A2C

为了进行性能评估，本文将对 MOTD3 算法产生的策略（记为 PMOTD3）与 P_DDPG 以及两个控制策略（P_Vmax，P_VME）进行比较[27]。其中，P_Vmax 控制策略是指无人机在悬停位置和目标设备之间以最大速度 v_{max} = 20 m/s 飞行，在目标设备上方悬停收集数据，其控制策略可以实现总数据速率的最大化。P_VME 控制策略是指无人机在悬停位置和目标设备之间以最大续航速度 $v = VME = 10.2$ m/s，在目标设备上方悬停收集数据，其控制策略可以实现能耗的最小化。本文算法与 P_DDPG、P_A2C、P_Vmax、P_VME 控制策略多目标优化性能的对比如图 7 所示。由图 7 可以看出，虽然 P_Vmax 控制策略获得的总数据速率最高，但其高能耗和长悬停时间显著限制了其在实际应用中的可行性。而 P_VME 控制策略虽然能量消耗最低，但是其在收集能量和数据速率性能方面表现较差。此外，本文算法产生的策略 PMOTD3 就收集能量性能而言，总收集能量远高于其他 3 种控制策略，在总数据速率和能量消耗性能方面仅次于基线策略且花费悬停时间少，更节省资源，提高任务执行效率。

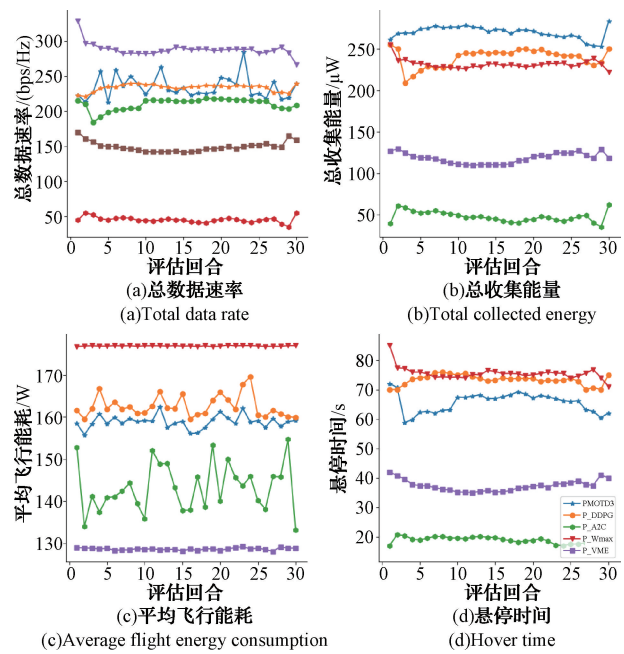


图 7　5 种控制策略多目标优化性能对比

Fig. 7　Comparison of multi-objective optimization performance of five control strategies

本文为了验证算法的泛化能力以及实际应用能力，后续分别对改变目标物联网设备数，悬停次数进行实验。在实际应用中，无人机辅助物联网设备网络的环境是复杂的且需求往往是动态变化的，目标物联网设备数和悬停次数的改变会影响系统的性能和资源利用等问题，因此本文从实际应用的角度出发，考虑并验证算法在环境动态变化过程中性能的表现以及合理地制定策略能力，对比如图 8、9

所示。由图8可知，P_Vmax控制策略在总数据速率方面表现良好，随着目标物联网设备的增多，其获得的数据速率略高于本文算法的控制策略，而P_A2C控制策略表现最为不理想。然而，由图9可知，尽管P_Vmax控制策略在数据速率上取得了良好的表现，但其平均收集能量明显低于本文算法的控制策略。从而也说明了本文算法在不同目标物联网设备数和不同悬停次数的情况下，资源分配和制定策略能力表现显著，具有更优的泛化能力。

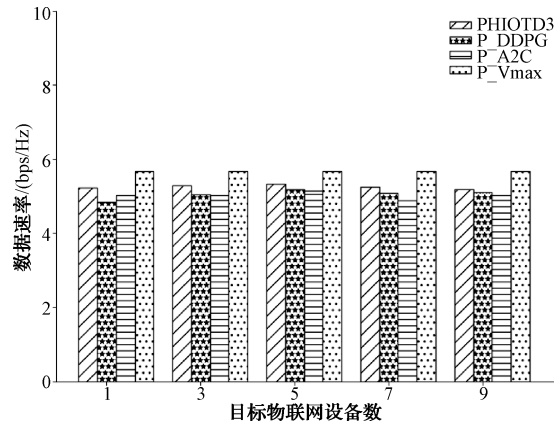


图8 不同目标物联网设备数时数据速率对比

Fig. 8 Data rate comparison of different target IOT devices

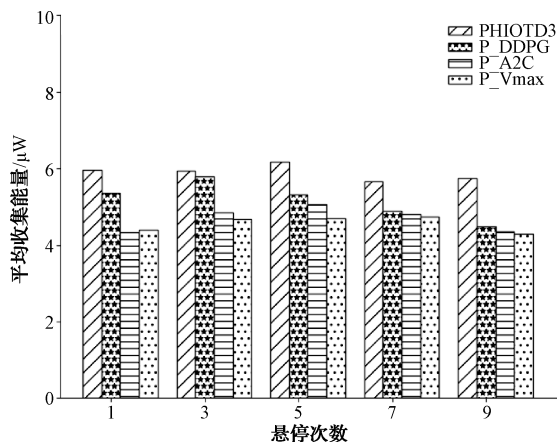


图9 不同悬停次数时平均收集能量对比

Fig. 9 Comparison of average collected energy at different hover times

4 结 论

本文研究了无线供电物联网网络中无人机辅助数据收集和能量传输的多目标优化问题。由于物联网网络的不确定性和动态性，提出了一种基于深度强化学习MOTD3算法来实现对无人机的飞行控制，在满足约束偏航角、飞行速度、发射功率条件下，同时优化了总数据速率、总收获能量、能耗以及悬停时间。仿真结果表明，本文算法相对于DDPG、A2C和传统基于规则的策略具有更优越的性能，且适用于任意目标数量的多目标优化问题，具有良好的泛化能力。无人机群通过多架无人机协同完成复杂任务具有很大的优势。在未来研究中，将进一步考虑无人机群辅助无线供电物联网网络的协同任务与资源分配以及多无人机路径规划与避撞等问题，以促进无线供电物联网网络的发展。

参 考 文 献

[1] SHARMA V, YOU I, ANDERSSON K, et al. Security, privacy and trust for smart mobile-internet of things (M-IoT): A survey[J]. IEEE Access, 2020, 8: 167123-167163.

[2] SHAFIQUE K, KHAWAJA B A, SABIR F, et al. Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios[J]. IEEE Access, 2020, 8: 23022-23040.

[3] ANANTRASIRICHAI N, BULL D. Artificial intelligence in the creative industries: A review[J]. Artificial Intelligence Review, 2022,55: 589-656.

[4] SEO H, LEE H, SON K, et al. Private and fresh real-time status updating[J]. IEEE Communications Letters, 2021, 26(2): 239-243.

[5] 彭艺，唐剑，杨青青，等. 基于强化学习的应急无人机通信中继选择策略[J]. 电子测量与仪器学报，2022，36(7)：9-15.
PENG Y, TANG J, YANG Q Q, et al. Communication relay selection strategy for emergency UAV based on reinforcement learning [J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(7): 9-15.

[6] LIAO Z, MA Y, HUANG J, et al. HOTSPOT: A UAV-assisted dynamic mobility-aware offloading for mobile-edge computing in 3-D space [J]. IEEE Internet of Things Journal, 2021, 8(13): 10940-10952.

[7] BABU N, PAPADIAS C B, POPOVSKI P. Energy-efficient 3-D deployment of aerial access points in a UAV communication system [J]. IEEE Communications Letters, 2020, 24(12): 2883-2887.

[8] QI H, HU Z, HUANG H, et al. Energy efficient 3-D UAV control for persistent communication service and fairness: A deep reinforcement learning approach[J]. IEEE Access, 2020, 8: 53172-53184.

[9] 谭建豪，马小萍，李希. 无人机3D航迹规划及动态避障算法研究[J]. 仪器仪表学报，2022，40(12)：224-233.
TAN J H, MA X P, LI X. Research on 3D flight path planning and dynamic obstacle avoidance

algorithm for UAV [J]. Chinese Journal of Scientific Instrument, 2022, 40(12): 224-233.

[10] 姚昌华，韩贵真，安蕾. 多无人机协同侦察时间资源分配优化[J]. 电子测量技术，2022，45(18)：106-113.
YAO CH H, HAN G ZH, AN L. Optimization of time resource allocation for multi-UAV cooperative reconnaissance [J]. Electronic Measurement Technology, 2022, 45(18): 106-113.

[11] 翟璐璐. 基于能量效率和覆盖率优化的 UAV 部署算法[J]. 国外电子测量技术，2021,40(12)：24-29.
ZHAI L L. UAV Deployment algorithm based on energy efficiency and coverage optimization [J]. Foreign Electronic Measurement Technology, 2021, 40(12): 24-29.

[12] BLISS M, MICHELUSI N. Power-constrained trajectory optimization for wireless UAV relays with random requests [C]. 2020 IEEE International Conference on Communications (ICC). IEEE, 2020: 1-6.

[13] XIE L, XU J, ZENG Y. Common throughput maximization for UAV-enabled interference channel with wireless powered communications [J]. IEEE Transactions on Communications, 2020, 68(5): 3197-3212.

[14] LIU X, LAI B, GOU L, et al. Joint resource optimization for UAV-enabled multichannel internet of things based on intelligent fog computing[J]. IEEE Transactions on Network Science and Engineering, 2020, 8(4): 2814-2824.

[15] PARK J, LEE H, EOM S, et al. UAV-aided wireless powered communication networks: Trajectory optimization and resource allocation for minimum throughput maximization [J]. IEEE Access, 2019, 7: 134978-134991.

[16] YE H T, KANG X, JOUNG J, et al. Optimization for full-duplex rotary-wing UAV-enabled wireless-powered IoT networks [J]. IEEE Transactions on Wireless Communications, 2020, 19(7): 5057-5072.

[17] HU H, XIONG K, QU G, et al. AoI-minimal trajectory planning and data collection in UAV-assisted wireless powered IoT networks [J]. IEEE Internet of Things Journal, 2020, 8(2): 1211-1223.

[18] ABD-ELMAGID M A, DHILLON H S. Average peak age-of-information minimization in UAV-assisted IoT networks [J]. IEEE Transactions on Vehicular Technology, 2018, 68(2): 2003-2008.

[19] SONG Q, JIN S, ZHENG F C. Completion time and energy consumption minimization for UAV-enabled multicasting [J]. IEEE Wireless Communications Letters, 2019, 8(3): 821-824.

[20] QI H, HU Z, HUANG H, et al. Energy efficient 3-D UAV control for persistent communication service and fairness: A deep reinforcement learning approach [J]. IEEE Access, 2020, 8: 53172-53184.

[21] DING R, GAO F, SHEN X S. 3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach [J]. IEEE Transactions on Wireless Communications, 2020, 19(12): 7796-7809.

[22] QIN Z, ZHANG X, ZHANG X, et al. The UAV trajectory optimization for data collection from time-constrained IoT devices: A hierarchical deep Q-network approach [J]. Applied Sciences, 2022, 12(5): 2546.

[23] 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.
ZHANG J H, KANG K, QIAN H, et al. Deep Q-network (DQN) drone path planning for internet of things [J]. Journal of Electronics & Information Technology, 2022, 44(11): 3850-3857.

[24] ZHANG J, YU Y, WANG Z, et al. Trajectory planning of UAV in wireless powered IoT system based on deep reinforcement learning[C]. 2020 IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2020: 645-650.

[25] BLISS M, MICHELUSI N. Adaptive scheduling and trajectory design for power-constrained wireless UAV relays[J]. arXiv preprint arXiv:2007.01228, 2020.

[26] ZENG Y, XU J, ZHANG R. Energy minimization for wireless communication with rotary-wing UAV [J]. IEEE Transactions on Wireless Communications, 2019, 18(4): 2329-2345.

[27] YE H T, KANG X, JOUNG J, et al. Optimization for full-duplex rotary-wing UAV-enabled wireless-powered IoT networks [J]. IEEE Transactions on Wireless Communications, 2020, 19(7): 5057-5072.

作者简介

徐钰龙，硕士研究生，主要研究方向为无线通信、深度强化学习、资源分配。
E-mail:1372433241@qq.com

李君(通信作者)，教授，主要研究方向为无线通信、信道编译码、资源分配。
E-mail:07a0303105@cjlu.edu.cn

李正权，教授，主要研究方向为无线通信、信号处理、

信道编译码。
E-mail:lzq722@jiangnan.edu.cn

胡静,硕士研究生,主要研究方向为移动边缘计算、资源分配、深度强化学习。
E-mail:2848621825@qq.com

张圣,硕士研究生,主要研究方向为无线通信、深度强化学习。
E-mail:236883554@qq.com

王子威,硕士研究生,主要研究方向为无线网络优化、图神经网络。
E-mail:1151821037@qq.com