



**QUEEN'S
UNIVERSITY
BELFAST**

RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning

Nguyen, K. K., Masaracchia, A., Sharma, V., Poor, H. V., & Duong, T. Q. (2022). RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning. *IEEE Journal of Selected Topics in Signal Processing*. Advance online publication. <https://doi.org/10.1109/JSTSP.2022.3172587>

Published in:
IEEE Journal of Selected Topics in Signal Processing

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights
Copyright 2022, IEEE.
This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: <http://go.qub.ac.uk/oa-feedback>

RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning

Khoi Khac Nguyen, *Student Member, IEEE*, Antonino Masaracchia, *Member, IEEE*, Vishal Sharma, *Senior Member, IEEE*, H. Vincent Poor, *Life Fellow, IEEE*, Trung Q. Duong, *Fellow, IEEE*

Abstract—Many of the devices used in Internet-of-Things (IoT) applications are energy-limited, and thus supplying energy while maintaining seamless connectivity for IoT devices is of considerable importance. In this context, we propose a simultaneous wireless power transfer and information transmission scheme for IoT devices with support from the reconfigurable intelligent surface (RIS)-aided unmanned aerial vehicle (UAV) communications. In particular, in the first phase, IoT devices harvest energy from the UAV through wireless power transfer; and then in the second phase, the UAV collects data from the IoT devices through information transmission. To characterise the agility of the UAV, we consider two scenarios: a hovering UAV and a mobile UAV. Aiming at maximising the total network sum-rate, we jointly optimise the trajectory and the power allocation of the UAV, the energy harvesting scheduling of IoT devices, and the phase-shift matrix of the RIS. We formulate a Markov decision process and propose two deep reinforcement learning algorithms to solve the optimisation problem of maximising the total network sum-rate. Numerical results illustrate the effectiveness of the UAV's flying path optimisation and the network's throughput of our proposed techniques compared with other benchmark schemes. Given the strict requirements of the RIS and UAV, the significant improvement in processing time and throughput performance demonstrates that our proposed scheme is well applicable for practical IoT applications.

Index Terms—Internet-of-Things (IoT), UAV, RIS, deep reinforcement learning, wireless power transfer.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) have recently drawn considerable attention thanks to their agile mobility and cost-effectiveness. UAVs have been used for geometry monitoring, disaster relief [1], emergency services, and wireless networks [2]. UAVs can be deployed at sporting events or in rescue missions to provide and enhance connectivity to the users in wireless networks. UAVs are also used as data collectors that fly to remote areas to collect sensor data [3]. However, restrictions regarding flying time and on-board processing ability are bottlenecks that must be dealt with in unexpected

environment and complicated missions. Very recently, UAVs have been used as an energy supplier to the energy-constrained IoT devices due to the fact that the UAV can easily be recharged at the docking station and the energy of these IoT devices is comparably smaller than UAV's capacity [4]–[9].

Reconfigurable intelligent surface (RIS) has emerged as a promising technology for future wireless networks. The arrival signal at a RIS is reflected toward the receiver by the RIS's passive elements operated by a module controller. The received signal at the users is composed of elements from the direct channel and the reflective link. It helps to increase the signal quality and reduce the interference. The RIS is usually deployed in high locations such as buildings to reduce the cost of establishing a new station. However, the optimisation of RIS performance is still challenging due to the large number of elements and the processing ability of the controller.

One area in which UAVs can be useful is in supporting Internet-of-Things (IoT) applications. Not only can they provide communication coverages, but, since many IoT devices are energy-limited, they can also be sources of power for such devices through downlink power transfer. A downlink power transfer and uplink information transmission protocol can be implemented in two phases: wireless power transfer (WPT) and wireless information transmission (WIT). In the first phase, the IoT devices harvest energy from a base station (BS) or from the UAV. The harvested energy is then used for transmitting local information to receivers or back to the UAV and the BS. By using such a downlink power transfer and uplink information transmission protocol, the IoT devices can obtain the energy to establish and maintain communication with the BS and the UAV.

Machine learning is an effective tool for optimising the performance of large-scale networks. One of the approaches is deep reinforcement learning (DRL), which is a combination of reinforcement learning and neural networks. In wireless networks, DRL algorithms are used for maximising the network performance, reducing power consumption and improving the processing time for real-time applications [2], [10], [11]. DRL algorithms are powerful in wireless networks because the agents do not need pre-collected data for training. Rather, DRL agents interact with their environment and establish training samples for the responses in those interactions. The neural networks are trained by up-to-date state transitions to adjust their parameters for maximising a designated reward. Then, the trained networks are deployed for real-time prediction.

K. K. Nguyen, A. Masaracchia, V. Sharma and T. Q. Duong are with Queen's University Belfast, Belfast BT7 1NN, UK (e-mail: {knguyen02, a.masaracchia, v.sharma, trung.q.duong}@qub.ac.uk).

H. V. Poor is with Princeton University, Princeton, NJ 08544, USA (e-mail: poor@princeton.edu)

This work was supported in part by the U.K. Royal Academy of Engineering (RAEng) under the RAEng Research Chair and Senior Research Fellowship scheme Grant RCSR2021\11\41 and in part by the EPSRC Impact Acceleration Accounts (IAA) Award 2022.

Corresponding author is Trung Q. Duong, (e-mail: trung.q.duong@qub.ac.uk).

A. State-of-the-art

UAV-assisted wireless communications have been widely used to enhance network coverage as well as network performance [1], [2], [12]. In [1], [12], the authors used the UAV for establishing the network for the disaster relief missions. A UAV can also serve as an energy source provider for device-to-device communications [2]. Recently, RIS technology has been introduced as a low-cost and easily installed technology to mitigate interference and direct transmitted signals toward their receivers [13]–[16]. In [14], the authors considered two-way communications assisted by a RIS. The reciprocal channel to maximise the signal-to-interference-plus-noise ratios (SINR) and the non-reciprocal channel with the target of maximisation of the minimum SINR were considered. The gamma approximation was used for the reciprocal channel, while the semi-definite programming relaxation and a greedy-iterative method were used for the non-reciprocal channel. In [15], an iterative algorithm with low computation complexity was proposed to solve the joint optimisation of transmit beamforming vector and the phase shift of a RIS under proper and improper Gaussian signalling. In [16], the authors optimised the beamforming matrices at the BS and the reflective vector at the RIS to minimise the total transmit power at a multiple-input-single-output (MISO) non-orthogonal multiple access (NOMA) networks. An algorithm based on the second-order cone programming-alternating direction method of multipliers was proposed to reach an optimal local problem.

By utilising both advantages of the UAV and the RIS, the received signal at the ground users is strengthened while the power consumption is reduced and the flying time of the UAV can be extended [17]–[20]. In [18], the UAV's trajectory and the RIS's passive beamforming vector were optimised to maximise the average rate in RIS-assisted UAV communications. The problem was derived into two subproblems; then, a closed-form phase shift algorithm was introduced to find the local optimal reflective matrix and the successive convex approximation was used to find the suboptimal trajectory solution. In [19], the UAV acts as a mobile relay and the RIS was used to provide ultra-reliable and low-latency communications between ground transmitter and ground IoT devices. The UAV's position, the RIS phase shift and the blocklength were optimised to minimise the total decoding error rate by using a polytope-based method, namely Nelder-Mead simplex.

Along with the development of the IoT devices is the increased power supply for each device. However, not all the nodes are equipped with fixed power providers and have solar batteries. Thus, we need to find a solution to provide power to the nodes. The downlink power transfer and uplink information transmission protocol is one of the solutions to enable the IoT devices to harvest energy from source providers and switch to information transmission in the uplink phase on demand [21]–[25]. That helps reduce the power consumption as well as the cables, wires for providing power. In [21], the authors designed a time-switching protocol for a RIS with the energy harvesting phase to charge the RIS capacitor and the signal reflecting phase to assist the transmission from

the access point (AP) to the receivers. The AP's transmit beamforming, the RIS's phase scheduling and the passive beamforming were optimised to maximise the information rate. The resultant two sub-problems were solved following the conventional semi-definite relaxation method and monotonic optimisation. In [22], the transmit precoding matrices of the BS and the RIS's passive phase shift matrix were optimised for maximising the weight sum-rate of all information receivers in power transfer scenarios.

The demand for a technique that is flexible and adaptive to changes in the environment while satisfying real-life constraints is rising, and DRL algorithms are among the most potential methods to deal with these problems in wireless networks [2], [3], [10], [11]. Recently, DRL algorithms are also used for the RIS-assisted wireless networks and have shown promising results [20], [26]–[29]. The power allocation and the phase shift optimisation were optimised for maximising the sum rate in [26]. In [27], a RIS-assisted UAV was deployed for serving ground users. The trajectory and phase shift optimisation relying on DRL for maximising the sum rate and fairness of all users was proposed. In [28], the authors used a RIS to assist the secure communications against eavesdroppers. The DRL algorithms were used to optimise the BS beamforming and the RIS's reflecting beamforming were shown to improve the secrecy rate and the quality-of-service satisfaction probability. In [29], a deep deterministic policy gradient was proposed to obtain the optimal phase shift matrix at the RIS to maximise the received signal-to-noise ratio (SNR) in a MISO system. In [20], the joint optimisation of the power and the RIS's phase shift in a multi-UAV-assisted network is considered. In Table I, we compare our work to the existing methods in the literature. While other works only consider a simple problem, our paper introduces a comprehensive approach with the joint optimisation of power allocation, energy harvesting time scheduling, UAV's trajectory design and RIS's phase shift adjustment. Furthermore, we utilise the advantage of deep neural networks in RL algorithm for fast deployment and real-time decision-making.

B. Contributions

Inspired by the aforementioned discussion, in this paper, we consider the IoT wireless networks with the support of a UAV, and one RIS, and employ the downlink power transfer and uplink information transmission protocol for maximising the total network's sum-rate. In particular, we adopt the harvest-then-transmit protocol, which means the IoT devices use all the harvested energy in the first phase for transmitting during the remaining time. Then, two DRL algorithms are deployed for solving the problem in RIS-assisted UAV communications. In summary, our main contributions are as follows:

- To characterise the agility of the UAV in supporting the energy harvesting (EH) and information transmission of IoT devices, we consider two scenarios of the UAV. Firstly, the UAV is hovering at the centre of the cluster and provides energy to the IoT devices. The RIS helps alleviate the uplink interference when the IoT devices transmit their information to the UAV. Secondly, the UAV

TABLE I
A COMPARISON WITH EXISTING LITERATURE

	[18]	[21]	[26]	[27]	[30]	[28]	[29]	[31]	[32]	[33]	Our work
UAV's trajectory design	✓			✓	✓				✓		✓
Sum-rate maximisation	✓	✓	✓	✓	✓					✓	✓
Power allocation		✓	✓			✓		✓	✓		✓
Energy harvesting time optimisation		✓									✓
RIS phase shift optimisation	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Random users			✓		✓	✓		✓	✓	✓	✓
Reinforcement learning				✓	✓	✓	✓		✓		✓
Deep neural networks				✓		✓	✓		✓	✓	✓

is deployed in an initial location and required to find a better location for communication. In each location of the UAV's flying trajectory, the UAV's power, the EH time scheduling and the RIS's phase shift matrix are optimised for maximising the network throughput performance.

- For the defined problem, we formulate a Markov decision process (MDP) [34] with the definition of the state space, action space and the reward function. Then, we propose a method based on deep deterministic policy gradient (DDPG) and proximal policy optimisation algorithm (PPO) for solving the maximisation game. Through the trial-and-learn ability of the DRL algorithm, the UAV can collect the data and train the neural networks to adapt to the dynamic environment.
- Our results suggest that with the support of the RIS, a better connection is established and the overall performance is significantly improved. Moreover, the deep neural networks after training can be used for making the real-time decision.

C. Organisations

In the remaining paper, we present the system model of the UAV-assisted wireless communication with the support of the RIS and problem formulation in Section II. The MDP for maximising the network throughput problem is introduced for the hovering UAV scenario in Section III. In Section IV, the DDPG algorithm is deployed for continuous control of the UAV's trajectory, EH time scheduling and the phase shift matrix of the RIS. The PPO technique with the clipping method is introduced in Section V for improving the network throughput. The simulation results are presented in Section VI to illustrate the efficiency of our proposed methods compared with other baseline schemes. Some existing problems and potential future research topics for the RIS and the UAV in real-life applications are discussed in Section VII.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider that the system includes one single-antenna UAV and N ground IoT devices. Here, we assume that the IoT devices are located randomly in a specific area from the centre in which the UAV is the centroid. Although the IoT devices can be mobilised before making connection with the UAV, for the sake of simplicity, it is presumed that the devices are static during the power and information transmission. However, there are some practical scenarios where IoT devices are located in a crowded area with surrounding obstacles and objects. In such

a complex environment, IoT devices suffer high attenuation and severe path loss. In this case, the RIS is also installed at the wall of a tall building to enhance the communication quality by reflecting signals from the UAV to the IoT devices. Here, we deploy a RIS composed of K elements to enhance the network performance. The 3D coordinate of the UAV at the time step t is $X_{UAV}^t = (x_{UAV}^t, y_{UAV}^t, z_{UAV}^t)$. In this paper, we consider the fixed attitude of the UAV at H_{UAV} . The location of the n th IoT devices at time step t is $X_n^t = (x_n^t, y_n^t)$ with $n = 1, \dots, N$. The position of the RIS component $k \in K$ at time step t is (x_k^t, y_k^t, z_k^t) . In this paper, we use the wireless downlink power transfer and uplink information transmission protocol for deploying the UAV and collecting data. Before establishing the connection between the UAV and RIS, the UAV will make an initial interaction with RIS via the control channel. Then the wireless power and information transmission can be carried out through the physical channel. Particularly, we have two phases during the physical channel as: wireless power transfer (WPT) and wireless information transmission (WIT). In the first phase, the downlink is activated to transfer energy to the IoT devices from the UAV during time span $\tau\mathcal{T}$. Then, the WIT phase takes place when the IoT devices transmit information to the UAV in the uplink during $(1 - \tau)\mathcal{T}$. We normalise the length of time step to $\mathcal{T} = 1$ for convenience.

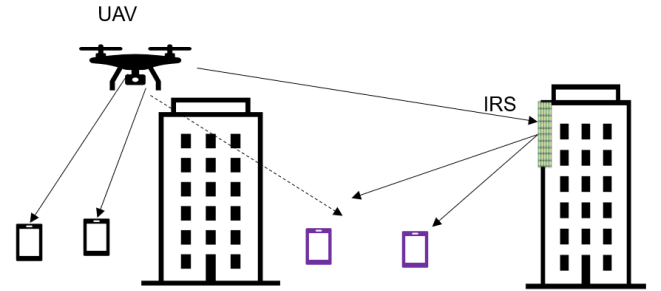


Fig. 1. An illustration of the system model of UAV-assisted IoT wireless communications with the support of a RIS.

A. Channel model

We denote the channel gain between the UAV and the RIS, between the RIS and the n th IoT device, and the direct link from the UAV to n th IoT node at time step t by $H^t \in \mathbb{C}^{1 \times K}$, $h_{RIS,n}^t \in \mathbb{C}^{1 \times K}$, and h_n^t , respectively. The small-scale fading of the direct link from the UAV to the IoT

devices is assumed to be Rayleigh fading due to the extensive scatters. The air-to-air channel is considered for the UAV and the RIS link, while the link from the RIS to the IoT devices can be modelled by the Rician fading channel. In practical applications, due to the acquisition delay and the feedback overhead incurred during the mobility of UAV, obtaining a perfect CSI of the links between UAV and RIS as well as IoT devices is a formidable challenge. In this paper, we assume that the UAV and IoT devices are static and therefore the perfect CSI's knowledge can be achieved [35].

The distance between UAV and the k th RIS in time step t is given by

$$d_k^t = \sqrt{(x_{UAV}^t - x_k^t)^2 + (y_{UAV}^t - y_k^t)^2 + (z_{UAV}^t - z_k^t)^2}. \quad (1)$$

Similarly, we denote the distance between the UAV and the n th IoT device and between the k th RIS element and the n th IoT node by d_n^t and $d_{k,n}^t$, respectively.

The channel gain between the UAV and the n th IoT device is given by

$$h_n^t = \sqrt{\beta_0(d_n^t)^{-\kappa_1}} \hat{h}, \quad (2)$$

where β and κ_1 are the path loss at reference distance $1m$ and the path loss exponent for the UAV and the IoT devices link, respectively; $\hat{h}$ represents the small-scale fading modelled by complex Gaussian distribution with zero-mean and unit-variance $\mathcal{CN}(0, 1)$.

Similarly, the channel gain between the UAV and the RIS is an air-to-air channel dominated by the line-of-sight (LoS) links. Thus, the channel of the UAV-RIS link in time step t is denoted as follows:

$$H^t = \sqrt{\beta_0(d_k^t)^{-\kappa_2}} [1, e^{-j\frac{2\pi}{\lambda}d \cos(\phi_{AoA}^t)}, \dots, e^{-j\frac{2\pi}{\lambda}(K-1)d \cos(\phi_{AoA}^t)}]^T, \quad (3)$$

where the right term is the array signal from the UAV to the RIS, $\cos(\phi_{AoA}^t)$ is the cosine of the angle of arrival (AoA) from the UAV to RIS; κ_2 , d and λ are the path loss exponent for the UAV and the RIS link, the element spacing and the carrier wavelength, respectively.

The channel gain between the RIS and the n th IoT device following the Rician fading is expressed as

$$h_{RIS,n}^t = \sqrt{\beta_0(d_{k,n}^t)^{-\kappa_3}} \left(\sqrt{\frac{\beta_1}{1+\beta_1}} h_{RIS,n}^{LoS} + \sqrt{\frac{1}{\beta_1+1}} h_{RIS,n}^{NLoS} \right), \quad (4)$$

where the deterministic LoS component is denoted by $h_{RIS,n}^{LoS} = [1, e^{-j\frac{2\pi}{\lambda}d \cos(\phi_{AoD}^t)}, \dots, e^{-j\frac{2\pi}{\lambda}(K-1)d \cos(\phi_{AoD}^t)}]$ and the non-line-of-sight (NLoS) component is the Rayleigh fading that follows the complex Gaussian distribution with zero mean and unit variance; $\cos \phi_{AoD}$ is the angle of departure (AoD) from the RIS to IoT devices; β_1 is the Rician factor, and κ_3 is the path loss exponent for the RIS and IoT devices link.

B. Power transfer phase

The achievable signal at the n th IoT device is composed of the direct signal from the UAV and the reflected signal from the RIS at time step t as

$$y_{1n}^t = (h_n^t + H^t \Phi^t h_{RIS,n}^t) \sqrt{P_0} x + \varrho^2, \quad (5)$$

where ϱ^2 is the noise signal following the complex Gaussian distribution $\mathcal{CN}(0, \alpha^2)$, x is the energy signal and P_0 is the transmission power at the UAV; $\Phi^t = \text{diag}[\phi_1^t, \phi_2^t, \dots, \phi_K^t]$ is the diagonal matrix at the RIS, where $\phi_k^t = e^{j\theta_k^t}$, $\forall k = 1, 2, \dots, K$ and $\theta_k^t \in [0, 2\pi]$ denotes the phase shift of the k th element in the RIS at time step t .

In the WPT phase, the UAV transfers energy to the IoT devices during time span τ^t at time step t . Thus, the received power at the n th IoT device at time step t is given by

$$p_n^t = \tau^t \eta P_0 |h_n^t + H^t \Phi^t g_n^t|^2, \quad (6)$$

where η is the power transfer efficiency. It is important to note that although we employ the linear EH model in this paper the non-linear counterpart should have been adopted in realistic scenarios.

C. Information transmission phase

We assume that the IoT devices do not have fixed energy sources and use all the harvested energy for the WIT phase. The signal received at the UAV from the n th IoT device is given by

$$y_{2n}^t = (h_n^t + H^t \Phi^t h_{RIS,n}^t) \sqrt{p_n^t} u_n + \varrho^2, \quad (7)$$

where u_n is the symbol signal from the n th IoT device to the UAV. The received SINR at the UAV from transmission of the n th IoT device at time step t can be formulated as follows:

$$\gamma_n^t = \frac{p_n^t |h_n^t + H^t \Phi^t g_n^t|^2}{\sum_{m \neq n}^N p_m^t |h_m^t + H^t \Phi^t g_m^t|^2 + \alpha^2}, \quad (8)$$

The sum-rate from the IoT devices at time step t is formulated as follows:

$$R_{total}^t = \sum_{n=1}^N (1 - \tau^t) B \log_2(1 + \gamma_n^t), \quad (9)$$

where B is the bandwidth.

Our objective is to maximise the achieved sum-rate performance by optimising the phase shift matrix Φ at the RIS, the UAV's trajectory Γ , power allocation P_0 and the EH time τ as

$$\begin{aligned} & \max_{\tau, P_0, \Phi, \Gamma} \sum_{n=1}^N (1 - \tau^t) B \log_2(1 + \gamma_n^t), \\ & \text{s.t.} \quad 0 < \tau^t < 1, \\ & \quad \theta_k^t \in [0, 2\pi], \forall k \in K, \\ & \quad v^t \leq v_{max}, \\ & \quad 0 \leq P_0 \leq P_{max}, \\ & \quad X_{UAV}^t \in Z, \end{aligned} \quad (10)$$

where Z represents the flying area restricted in the vertical and horizontal dimensions; P_{max} is the maximum transmission power at the UAV; v^t and v_{max} are the UAV's velocity at time step t and the UAV's maximum flying velocity, respectively.

III. HOVERING UAV FOR DOWNLINK POWER TRANSFER AND UPLINK INFORMATION TRANSMISSION IN RIS-ASSISTED UAV COMMUNICATIONS

Besides WPT, the UAV uses most of the energy for its movement. Thus, to extend the operating time, the UAV is considered to hover at a fixed position at the center of the cluster. Firstly, we present the mathematical background of the DRL algorithm and then we apply the DRL algorithm for solving the sum-rate maximisation problem in RIS-assisted UAV communications.

A. Preliminaries

To search for optimal policy in the DRL algorithm, we have two main directions by searching directly the parameters of policy and indirectly through the value function. In the value function approach, we estimate the value in a given state to adjust the policy parameter. The state-value function V is defined by following the policy π at the state $s \in \mathcal{S}$ as follows:

$$V^\pi = \mathbb{E}[\mathcal{R}|s, \pi], \quad (11)$$

where $\mathbb{E}$ is the expectation operation; $\mathcal{S}$ and $\mathcal{R}$ are the state space and the reward function, respectively.

The optimal policy π^* corresponds to the optimal state-value function $V^*(s)$ as

$$V^*(s) = \max_{\pi} V^\pi(s), s \in \mathcal{S}. \quad (12)$$

The agent at the state s chooses the action a from the action space $\mathcal{A}$ to maximise the cumulative reward following the optimal policy π^* by making use of the Bellman equation [34]

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}[r(s, a)] + \zeta \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}(a) V^*(s') \right\}, \quad (13)$$

where $\mathcal{P}_{ss'}(a)$ is the transition probability with $s = s^t$, $s' = s^{t+1} \in \mathcal{S}$ and $a \in \mathcal{A}$; $0 \leq \zeta \leq 1$ is the discount factor.

The state-action value Q is obtained when the agent at the state s takes action $a \in \mathcal{A}$ following the policy π as follows:

$$Q^\pi(s, a) = \mathbb{E}[r(s, a)] + \zeta \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}(a) V(s'). \quad (14)$$

From (13) and (14), we have

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a), \quad (15)$$

where $Q^*(s, a)$ is the action-value function with optimal policy π^* .

Instead of finding the optimal policy π^* through the value function, we directly find the policy by adjusting the policy parameter in the policy search method. The policy gradient is one of the most popular methods due to its efficient sampling. Our objective is to find the optimal policy π^* to maximise the reward, which is defined by

$$J(\theta_\pi) = \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi(a|s) r(s, a), \quad (16)$$

where θ_π and $d^\pi(s)$ are the policy parameter and the stationary distribution of Markov chain, respectively. We adjust the policy parameter θ_π by using the gradient descent relying on

$$\begin{aligned} \nabla_{\theta_\pi} &= \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_{\theta_\pi} \pi(a|s) Q^\pi(s, a) \\ &= \mathbb{E}[\nabla_{\theta_\pi} \ln \pi(s, a) Q^\pi(s, a)] \end{aligned} \quad (17)$$

One of the most popular algorithms based on policy search is REINFORCE algorithm, in which we adjust the policy parameter θ_π by using the Monte-Carlo methods and episode samples learning. The optimal policy can be obtained by

$$\theta_\pi^* = \operatorname{argmax}_{\theta_\pi} \mathbb{E} \left[\sum_{a \in \mathcal{A}} \pi(a|s; \theta_\pi) r(s, a) \right] \quad (18)$$

We use the gradient descent method to adjust the policy parameters θ_π as

$$\theta_\pi \leftarrow \theta_\pi - \iota \nabla_{\theta_\pi}, \quad (19)$$

where $0 \leq \iota \leq 1$ is the step-size parameter and the gradient is defined as

$$\nabla_{\theta_\pi} = \mathbb{E}[\nabla_{\theta_\pi} \ln \pi(a|s; \theta_\pi) r(s, a)] \quad (20)$$

At the state s we can choose the best action a^* by the maximum probability as

$$a^* = \operatorname{argmax}_{a \in \mathcal{A}} \pi(a|s; \theta_\pi) \quad (21)$$

B. The DDPG method

The DDPG algorithm is a hybrid model composed of the value function and policy search methods. Thus, the DDPG algorithm is suitable for large-scale action and state spaces. Based on the current policy, the actor function $\mu(s; \theta_\mu)$ maps the states to a specific action with θ_μ being the actor network parameters, while the critic function $Q(s, a)$ evaluates the quality of the action taken. In the DDPG algorithm, we use *experience replay buffer* and *target network* technique to improve the convergence speed and avoid excessive calculations.

The agent iteratively interacts with the environment by executing the action a^t and receiving the response with instant reward r^t and the next state s^{t+1} . The tuple of (s^t, a^t, r^t, s^{t+1}) is then stored in a replay buffer D for training the actor and critic network. The buffer D is updated by adding new samples and discarding the oldest ones due to its finite size setting. After achieving enough samples, the agent takes a batch G of transitions for training the network. Particularly, we train the actor and critic network using stochastic gradient descent (SGD) over a mini-batch G samples.

Let us denote the parameters of the critic network and the target critic network by θ_q and $\theta_{q'}$, respectively. The critic network is updated by minimising

$$L = \frac{1}{G} \sum_i \left(y^i - Q(s^i, a^i; \theta_q) \right)^2, \quad (22)$$

with

$$y^i = r^i(s^i, a^i) + \zeta Q'(s^{i+1}, a^{i+1}; \theta_{q'})|_{a^{i+1}=\mu'(s^{i+1}; \theta_{\mu'})}. \quad (23)$$

where the action at time step $(i+1)$ can be obtained by running the target actor network μ' with the state s^{i+1} ; $\theta_{\mu'}$ denotes the parameters of the target actor network and ζ is the discounting factor.

The actor network parameters are updated by

$$\nabla_{\theta_{\mu}} J \approx \frac{1}{G} \sum_i \nabla_{a^i} Q(s^i, a^i; \theta_q) |_{a^i = \mu(s^i)} \nabla_{\theta_{\mu}} \mu(s^i; \theta_{\mu}). \quad (24)$$

Moreover, we duplicate the actor network and the critic network after a number of episodes to create a target actor and a target critic network. It helps reduce the excessive calculations by using only one network to estimate the target value. The target actor network parameters θ_q and the target critic network parameter $\theta_{\mu'}$ are updated by using soft target updates associated with $\varkappa \ll 1$, as

$$\theta_{q'} \leftarrow \varkappa \theta_q + (1 - \varkappa) \theta_{q'}, \quad (25)$$

$$\theta_{\mu'} \leftarrow \varkappa \theta_{\mu} + (1 - \varkappa) \theta_{\mu'}. \quad (26)$$

For *explorations* and *exploitations* purpose, we add a noise process of $\mathcal{N}(0, 1)$ as follows [36]:

$$\mu'(s^t) = \mu(s^t; \theta_{\mu}^t) + \psi \mathcal{N}(0, 1), \quad (27)$$

where ψ is a hyper-parameter.

C. Game solving

For the hovering scenario, we formulate the MDP [34] by a 4-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$. Then, we formulate a game to solve the problem in (10).

- *Agent*: The UAV's centralised processor will act as an agent. The agent interacts with the environment to find an optimal policy π^* for maximising the total sum-rate. After training, the action-making schemes will be deployed to the UAV to predict the proper EH time scheduling τ and the RIS can choose the phase shift matrix Φ .
- *State space*: The channel is composed of both direct link and the reflective channel. Thus, we define the state space as

$$\mathcal{S} = \{h_1 + H\Phi g_1, h_2 + H\Phi g_2, \dots, h_N + H\Phi g_N\}, \quad (28)$$

In time step t , the UAV has the state $s^t = \{h_1^t + H^t \Phi^t g_1^t, h_2^t + H^t \Phi^t g_2^t, \dots, h_N^t + H^t \Phi^t g_N^t\}$

- *Action space*: The UAV hovers at a fixed position; thus, we optimise the EH time τ , the transmission power P_0 , and the RIS's phase shift Φ . The action space is defined as

$$\mathcal{A} = \{\tau, P_0, \theta_1, \theta_2, \dots, \theta_K\} \quad (29)$$

At the state s^t , the UAV takes the action $a^t = \{\tau^t, P_0^t, \theta_1^t, \theta_2^t, \dots, \theta_K^t\}$ and moves to the next state $s' = s^{t+1}$.

- *Reward function*: The UAV interacts with the environment to find the maximum obtained reward. In our work, we formulate the reward function to obtain the maximum total sum-rate performance as

$$\mathcal{R} = \sum_{n=1}^N (1 - \tau^t) B \log_2(1 + \gamma_n^t) \quad (30)$$

The UAV is hovering at X_{UAV} and chooses the action a^t based on the achieved channel state information (CSI). Then, the UAV using the power P_0 transfers the energy during τ to the IoT devices and the RIS controller adjusts the phase shift for each element. During the remaining time $(1 - \tau)$, the RIS will not change the phase shift while the IoT devices transmit information in the uplink to the UAV. It is challenging since the RIS plays a crucial role in mitigating the interferences. Thus, we need to find an intelligent scheme for the RIS to maximise the network performance in the downlink power transfer and uplink information transmission protocol. We propose a DRL, namely DDPG algorithm, to find an optimal policy for the UAV and the RIS.

At time step t , the agent has the information of the channel (28) and uses the actor network to choose the action a^t (29). By executing the action a^t , the agent receives the response following the reward function (30) from the environment. The critic action takes part to justify the efficiency of the chosen action a^t . After storing enough samples in buffer D , the agent trains the networks over a mini-batch G by using the SGD with Adam optimiser [37].

In this section, we assume that the UAV is hovering at a fixed position to reduce the flying energy consumption. It is a trade-off game with the energy and total achievable sum-rate. In the next section, we propose a joint optimisation of trajectory, power allocation, EH time and the phase shift to maximise the network throughput in a short operation time.

IV. JOINT TRAJECTORY, TRANSMISSION POWER, EH TIME SCHEDULING AND THE RIS PHASE SHIFT OPTIMISATION USING DEEP REINFORCEMENT LEARNING

Given a short flying time of the UAV, to maximise total achievable sum-rate, we propose a joint optimisation scheme between the UAV's trajectory, transmission power, EH time scheduling of IoT, and the RIS's phase shift. We define the state space and the reward function as in Section III. We modify the action space as follows:

$$\mathcal{A} = \{v, \varsigma, \tau, P_0, \theta_1, \theta_2, \dots, \theta_K\} \quad (31)$$

At the state s^t , the UAV takes the action $a^t = \{v^t, \varsigma^t, \tau^t, P_0^t, \theta_1^t, \theta_2^t, \dots, \theta_K^t\}$ and moves to the next state $s' = s^{t+1}$. Particularly, the position of the UAV at time step $(t + 1)$ is represented as follows:

$$X_{UAV}^{t+1} = \begin{cases} x_{UAV}^{t+1} = x_{UAV}^t + v^t \cos \varsigma^t + \Delta x^{t+1} \\ y_{UAV}^{t+1} = y_{UAV}^t + v^t \sin \varsigma^t + \Delta y^{t+1} \\ H_{UAV}^{t+1} = H_{UAV}^t + \Delta H^{t+1}, \end{cases} \quad (32)$$

where Δx^{t+1} , Δy^{t+1} , and ΔH^{t+1} are the environmental noise on the UAV at time step $(t + 1)$. The UAV is flying from the position X_{UAV}^t to X_{UAV}^{t+1} but still needs to satisfy the flying zone constraint $X_{UAV} \in \mathcal{Z}$. Moreover, the velocity of the UAV is set to satisfy the requirement $v \leq v_{max}$ and the flying angle is set to satisfy a constraint, $\varsigma \in [0, 2\pi]$.

Our objective is to find the optimal policy π^* for maximising the expected reward $\mathcal{R}$. The agent has the local knowledge and interacts with the environment to receive the reward. Base on

Algorithm 1 Deep deterministic policy gradient algorithm for joint trajectory design, transmission power, EH time and phase shift optimisation in RIS-assisted UAV communications.

- 1: Initialise the actor network $\mu(s; \theta_\mu)$, target actor network μ' and the critic network $Q(s, a; \theta_q)$, the target critic networks Q' .
- 2: Initialise replay memory pool $\mathcal{D}$
- 3: **for** episode = 1, ..., M **do**
- 4: Initialise an action exploration process $\mathcal{N}$
- 5: Receive initial observation state s^0
- 6: **for** iteration = 1, ..., T **do**
- 7: Find the action a^t for the state s^t
- 8: Execute the action a^t
- 9: Update the reward r^t according to (30)
- 10: Observe the new state s^{t+1}
- 11: Store transition (s^t, a^t, r^t, s^{t+1}) into replay buffer $\mathcal{D}$
- 12: Sample randomly a mini-batch of G transitions (s^i, a^i, r^i, s^{i+1}) from $\mathcal{D}$
- 13: Update critic parameter by SGD using the loss (22)
- 14: Update the actor policy parameter (24)
- 15: Update the target networks as in (25) and (26)
- 16: Update the state $s_i^t = s_i^{t+1}$
- 17: **end for**
- 18: **end for**

the received reward, the agent adjusts the policy π and executes a new action at a new state. The agent can find a better policy with a better reward by the iterative interactions. After each execution of the action, the UAV will move to a new position and receive responses from the environment. By interacting iteratively with the environment, the agent can choose the proper velocity and the flying direction for the UAV in each time step based on the achieved CSI. Simultaneously, the EH scheduling τ , the power allocation P_0 and the phase shift matrix are also optimised for maximising network performance. Here, M and T are the number of the maximum episodes and time steps, respectively. The details of our DDPG algorithm-based technique for joint trajectory design, power allocation, EH time and phase shift matrix optimisation in RIS-assisted UAV communications are presented in Algorithm 1.

V. PROXIMAL POLICY OPTIMISATION TECHNIQUE FOR JOINT TRAJECTORY, POWER ALLOCATION, EH TIME AND THE PHASE SHIFT OPTIMISATION

For the continuous state and action space as in our problem, we propose an on-policy algorithm, namely the PPO algorithm, for the joint optimisation of trajectory, power allocation, EH time and the phase shift of the RIS. Instead of using critic network based on value function in the Algorithm 1, we search directly for the optimal policy π^* in the PPO algorithm. We define the policy by π with the parameter θ_π . Here, we train the policy and adjust the parameter to find an optimal policy π^* by running the SGD over a mini-batch of G transitions (s^i, a^i, r^i, s^{i+1}) . The policy parameters are updated

for optimising the objective function as follows:

$$\theta_\pi^{i+1} = \operatorname{argmax}_{\theta_\pi} \frac{1}{G} \sum_i \nabla_{a^i} \mathcal{L}(s^i, a^i; \theta_\pi). \quad (33)$$

In the PPO algorithm, the agent interacts with the environment to find the optimal policy π^* with the parameter θ_{π^*} that maximises the reward as

$$\mathcal{L}(s, a; \theta_\pi) = \mathbb{E} \left[p_\theta^t A^\pi(s, a) \right], \quad (34)$$

where $p_\theta^t = \frac{\pi(s, a; \theta_\pi)}{\pi(s, a; \theta_{old})}$ is the probability ratio of the current policy and previous policy; $A^\pi(s, a)$ is the advantage function [38].

Here, if we use only one network for the policy, the excessive modification occurs during the training stage. Thus, we use the clipping surrogate method as follows [39]:

$$\mathcal{L}^{\text{clip}}(s, a; \theta_\pi) = \mathbb{E} \left[\min \left(p_\theta^t A^\pi(s, a), \right. \right. \\ \left. \left. \text{clip}(p_\theta^t, 1 - \epsilon, 1 + \epsilon) A^\pi(s, a) \right) \right], \quad (35)$$

where ϵ is a small constant. In this paper, the advantage function $A^\pi(s, a)$ [40] is formulated as follows:

$$A^\pi(s, a) = r^t + \zeta V^\pi(s^{t+1}) - V^\pi(s^t). \quad (36)$$

The policy is then trained by a mini-batch B and the parameters are updated by

$$\theta_\pi^{i+1} = \operatorname{argmax}_{\theta_\pi} \mathbb{E} \left[\mathcal{L}^{\text{clip}}(s, a; \theta_\pi) \right]. \quad (37)$$

The details of our PPO algorithm-based technique for joint trajectory design, transmission power, EH time and phase shift matrix optimisation in RIS-assisted UAV communications are presented in Algorithm 2.

VI. SIMULATION RESULTS

TABLE II
SIMULATION PARAMETERS

Parameters	Value
Bandwidth (W)	1 MHz
Maximum UAV transmission power	5 W
UAV maximum speed per time step	5 m
UAV's coverage	500 m
The initial UAV's position	(0, 0, 200)
The RIS's position	(200, 0, 50)
Path-loss parameter	$\kappa_1 = 4, \kappa_2 = 2, \kappa_3 = 2.2$
Channel power gain	$\beta_0 = -30$ dB
Rician factor	$\beta_1 = 4$
Noise power	$\alpha^2 = -134$ dBm
Max number of episodes	$M = 2000$
Max number of time step	$T = 400$
Clipping parameter	$\epsilon = 0.2$
Discounting factor	$\zeta = 0.9$
Max number of IoT devices	20
Initial batch size	$K = 32$

Algorithm 2 Our proposed approach based on the PPO algorithm for the RIS-assisted UAV communications.

- 1: Initialise the policy π with the parameter θ_π
- 2: Initialise the penalty method parameters ϵ
- 3: **for** episode = 1, ..., M **do**
- 4: Receive initial observation state s^0
- 5: **for** iteration = 1, ..., T **do**
- 6: Find the action a^t based on the current state s^t by following the current policy
- 7: Execute the action a^t
- 8: Update the reward r^t according to (30)
- 9: Observe the new state s^{t+1}
- 10: Update the state $s_i^t = s_i^{t+1}$
- 11: Collect a set of partial trajectories with G transitions
- 12: Estimate the advantage function as (36)
- 13: **end for**
- 14: Update policy parameters using SGD with a mini-batch B of the collected samples

$$\theta^{i+1} = \underset{\theta_\pi}{\operatorname{argmax}} \frac{1}{G} \sum \mathcal{L}^{\text{clip}}(s, a; \theta_\pi) \quad (38)$$

15: **end for**

In our work, we use the Tensorflow 1.13.1 [41] for implementing our algorithms. We deploy the UAV at $(0, 0, 200)$, the RIS at $(200, 0, 50)$ and assume $d/\lambda = 1/2$ for convenience. All other parameters are provided in Table II. In order to compare our proposed model with other baseline schemes, in this paper, we consider the techniques as follows:

- **Optimisation with the hovering UAV:** the UAV is maintained at a fixed position at the centre of the cluster $(0, 0, H_{\text{UAV}})$. We optimise the EH time τ , the transmission power P_0 and the phase shift matrix at the RIS. We use the DDPG algorithm (H-DDPG) and the PPO algorithm (H-PPO) for the problem in the hovering UAV scenario.
- **Our proposed model with mobile UAV:** For the game formulated as in Section IV, we use the DDPG algorithm (F-DDPG) and the PPO algorithm (F-PPO) for solving the problem of joint optimisation of trajectory, power allocation, EH time scheduling and the phase shift matrix at the RIS.
- **Random selection scheme (RSS):** The value of Φ is selected randomly and we use the DDPG algorithm (RSS-HDDPG) for optimising the EH time τ and the power P_0 in the hovering UAV scenario.
- **Without RIS:** We do not deploy the RIS in this scenario and optimise the EH time τ , the power P_0 in the hovering UAV scenario using the PPO algorithm (WithoutRIS-HPPO).
- **Fixed EH time (FEH):** The EH time τ is fixed to 0.5 and the flying path, power allocation P_0 and the phase shift Φ are optimised to maximise the performance. We use the DDPG algorithm (FEH-DDPG) for optimisation.
- **Maximal transmission power (MTP):** The transmission power at the UAV is set to $P_0 = P_{\max}$. We use the

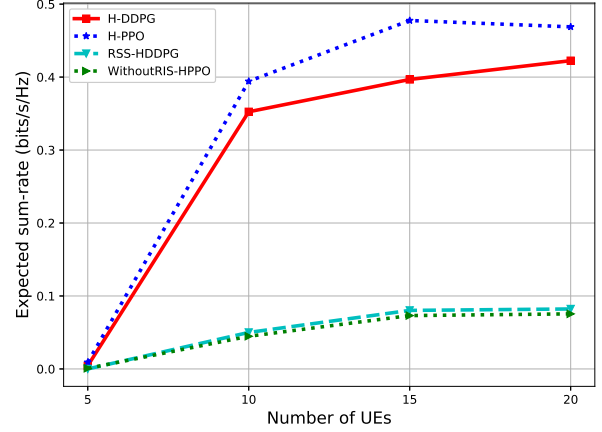


Fig. 2. The sum-rate performance in the hovering UAV scenario with different numbers of IoT devices, N .

PPO algorithm (MTP-PPO) for optimising the UAV's trajectory, EH time scheduling and the RIS's phase shift matrix.

Firstly, we consider the hovering UAV scenario and compare the performance versus the different number of IoT devices, N with the number of RIS, $K = 20$ in Fig. 2. We take the average of over 500 episodes for each scheme to draw the figures. When using the H-PPO algorithm, the total expected throughput is higher than in other schemes, including the ones using the H-DDPG algorithm, the RSS-HDDPG and WithoutRIS-HPPO technique. The results suggest that with the EH time and the RIS's reflecting coefficient optimisation, the PPO algorithm is adequate irrespectively of the number of IoT devices.

Next, we present the achieved sum-rate of the PPO and DDPG algorithm in the hovering UAV scenario comparing with the RSS and without RIS case while the number of IoT devices is fixed at $N = 10$ in Fig. 3. The H-PPO again shows the effective results with the different number of RIS elements, K . The sum-rate performance of the H-PPO algorithm improves from 0.2 to 1.4 (bits/s/Hz) following the increase of the RIS elements. The RIS is a passive reflector; thus, the reflected signal is diverse and not towards the destinations if we cannot control the coefficient of the RIS and select the phase shift randomly.

Furthermore, in Fig. 4, we investigate the performance while using the PPO algorithm in the hovering UAV scenario with the different number of RIS elements, K . When we increase the number of RIS elements K , the performance is enhanced. The results suggest that the PPO algorithm can handle the large optimising variables and effective in the hovering UAV scenario.

In Fig. 5, we compare the performance of optimising for the mobile UAV to the hovering UAV with different number of IoT devices, N . When we optimise the UAV's trajectory, the performance is enhanced using both DDPG and PPO

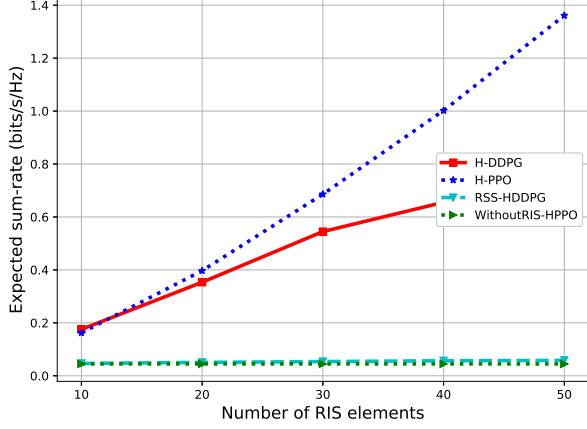


Fig. 3. The sum-rate performance in the hovering UAV scenario with varying number of RIS elements, K .

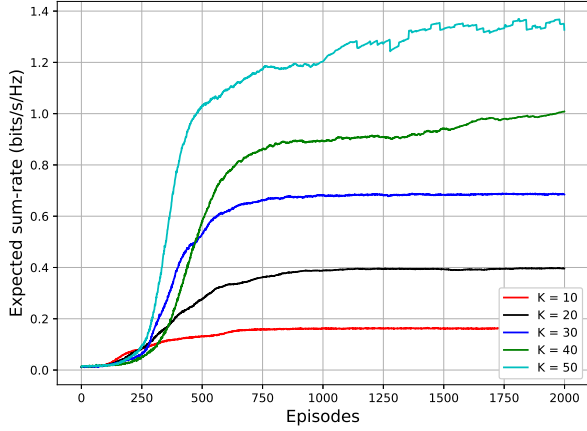


Fig. 4. The performance while using the PPO algorithm in the hovering UAV scenario with different number of RIS elements, K .

algorithms. The F-PPO approach reaches the best-expected result with different number of IoT devices, N .

In Fig. 6, we present the advantage of optimising the UAV's trajectory in comparison to the hovering UAV scenario. While increasing the number of RIS elements, the performance is increased. The sum-rate obtained by using the PPO algorithm is higher than the ones using the DDPG algorithm in both hovering UAV and mobile UAV scenarios. The results in Fig. 5 and Fig. 3 suggest that the optimisation of the UAV's trajectory is effective.

In Fig. 7, we compare the total sum-rate in the mobile UAV with the number of RIS elements $K = 20$ and different numbers of IoT devices, N . The method based on the F-PPO algorithm shows impressive results over other schemes. When using the F-PPPO algorithm, we can achieve a total throughput of around 0.8 bits/Hz with number of IoT devices set to 20. The performance obtained by the MTP-PPO algorithm is slightly lower than the F-PPO algorithm. Moreover, the FEH-DDPG and F-DDPG algorithm are not goods and unstable.

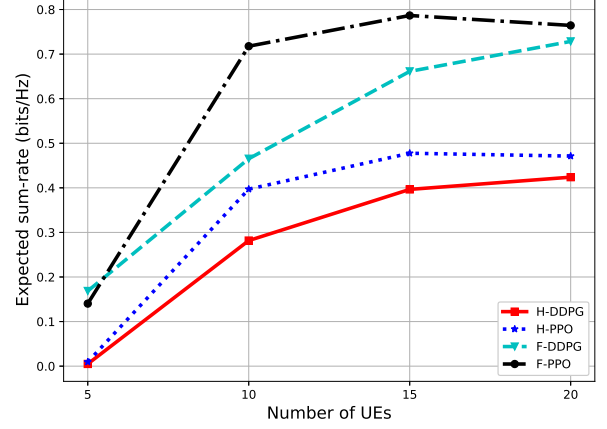


Fig. 5. The performance with different number of IoT devices, N .

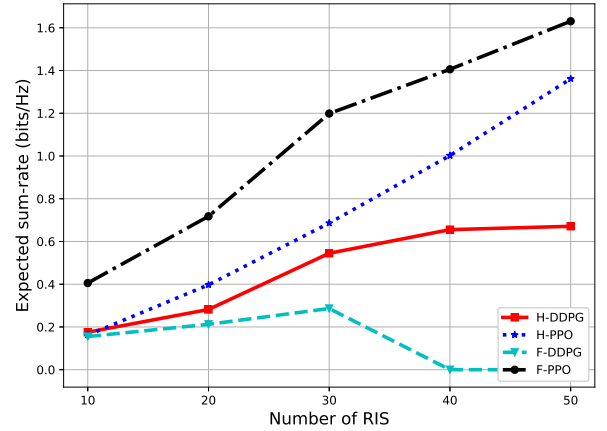


Fig. 6. The performance with different number of RIS elements, K .

The reason is that the exploration-exploitation scheme for DDPG algorithm is based on the initial value and the training scheme for the trajectory problem is affected by the noise of the training samples.

We consider the different number of RIS elements K and compare the performance of our proposed algorithms with REH schemes and hovering UAV scenario, as shown in Fig. 8. The F-PPO algorithm-based technique outperforms other schemes while it reaches around 1.6 (bits/s/Hz). Following the performance using the F-PPO algorithm is the ones using the MTP-PPO algorithm. The achievable sum-rate when we optimise the UAV's trajectory, transmission power, EH time and phase shift matrix is higher than when we only consider the fixed EH time in FEH or when we use the maximum power in MTP.

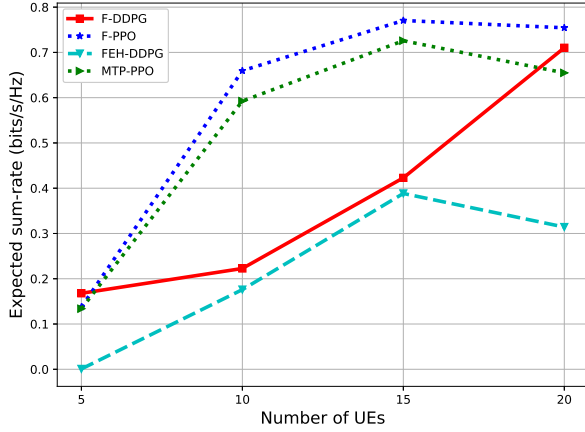


Fig. 7. The sum-rate performance with different number of IoT devices, N .

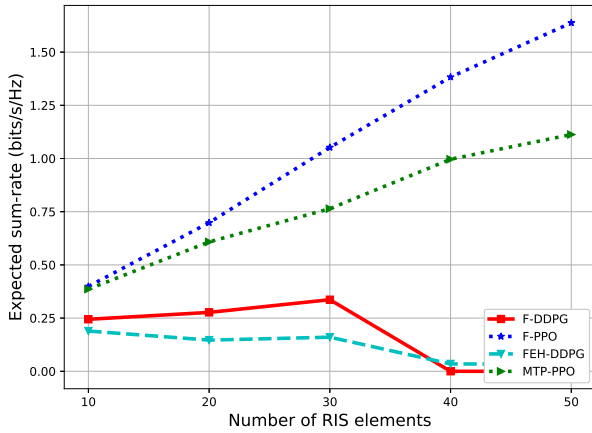


Fig. 8. The sum-rate performance with different numbers of RIS elements, K .

VII. CONCLUSION

In this paper, we have introduced a new system model for RIS-assisted UAV communications with the downlink power transfer and uplink information transmission protocol. By utilising the UAV's mobility, the flexibility of the RIS, and the effectiveness of the protocol, the RIS-assisted UAV network is a promising technique for practical applications. We have proposed two DRL techniques for jointly optimising the UAV's trajectory, power allocation, IoT's EH time scheduling and the phase shift matrix of the RIS to maximise the network's throughput. The results suggest that the systems learned by the DRL algorithm can deal with large-scale scenario and satisfy some power restrictions and processing time in RIS-assisted UAV communications. In the future, we plan to extend our work to include a distributed model and cooperative communications with multiple UAVs.

ACKNOWLEDGEMENT

This work was supported in part by the U.K. Royal Academy of Engineering (RAEng) under the RAEng Research Chair and Senior Research Fellowship scheme Grant RCSR2021\11\41 and in part by the EPSRC Impact Acceleration Accounts (IAA) Award 2022.

REFERENCES

- [1] L. D. Nguyen, A. Kortun, and T. Q. Duong, "An introduction of real-time embedded optimisation programming for UAV systems under disaster communication," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 5, no. 17, pp. 1–8, Dec. 2018.
- [2] K. K. Nguyen, N. A. Vien, L. D. Nguyen, M.-T. Le, L. Hanzo, and T. Q. Duong, "Real-time energy harvesting aided scheduling in UAV-assisted D2D networks relying on deep reinforcement learning," *IEEE Access*, vol. 9, pp. 3638–3648, Dec. 2021.
- [3] K. K. Nguyen, T. Q. Duong, T. Do-Duy, H. Claussen, and L. Hanzo, "3D UAV trajectory and data collection optimisation via deep reinforcement learning," *IEEE Transactions on Communications*, pp. 1–1, 2022, early access.
- [4] L. Xie, J. Xu, and R. Zhang, "Throughput maximization for UAV-enabled wireless powered communication networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1690–1703, Apr. 2019.
- [5] Z. Chen, K. Chi, K. Zheng, G. Dai, and Q. Shao, "Minimization of transmission completion time in UAV-enabled wireless powered communication networks," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1245–1259, Feb. 2020.
- [6] L. Xie, J. Xu, and Y. Zeng, "Common throughput maximization for UAV-enabled interference channel with wireless powered communications," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3197–3212, May 2020.
- [7] S. M. Hashir, A. Mehrabi, M. R. Mili, M. J. Emadi, D. W. K. Ng, and I. Krikidis, "Performance trade-off in UAV-aided wireless-powered communication networks via multi-objective optimization," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13430–13435, Dec. 2021.
- [8] Y. Liu, K. Xiong, Y. Lu, Q. Ni, P. Fan, and K. B. Letaief, "UAV-aided wireless power transfer and data collection in rician fading," *IEEE J. Select. Areas Commun.*, vol. 39, no. 10, pp. 3097–3113, Oct. 2021.
- [9] M.-N. Nguyen, L. D. Nguyen, T. Q. Duong, and H. D. Tuan, "Real-time optimal resource allocation for embedded UAV communication systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 225–228, Feb. 2019.
- [10] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and L. D. Nguyen, "Distributed deep deterministic policy gradient for power allocation control in D2D-based V2V communications," *IEEE Access*, vol. 7, pp. 164 533–164 543, Nov. 2019.
- [11] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and N. M. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100 480–100 490, Jul. 2019.
- [12] L. D. Nguyen, K. K. Nguyen, A. Kortun, and T. Q. Duong, "Real-time deployment and resource allocation for distributed UAV systems in disaster relief," in *Proc. IEEE 20th International Workshop on Signal Processing Advances in Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.
- [13] E. Basar, M. D. Renzo, J. D. Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.
- [14] S. Atapattu, R. Fan, P. Dharmawansa, G. Wang, J. Evans, and T. A. Tsiftsis, "Reconfigurable intelligent surface assisted twoway communications: Performance analysis and optimization," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6552–6567, Oct. 2020.
- [15] H. Yu, H. D. Tuan, A. A. Nasir, T. Q. Duong, and H. V. Poor, "Joint design of reconfigurable intelligent surfaces and transmit beamforming under proper and improper Gaussian signaling," *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2589–2603, Nov. 2020.
- [16] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 664–674, Jan. 2021.
- [17] L. Ge, P. Dong, H. Zhang, J.-B. Wang, and X. You, "Joint beamforming and trajectory optimization for intelligent reflecting surfaces-assisted UAV communications," *IEEE Access*, vol. 8, pp. 78 702–78 712, Apr. 2020.

- [18] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. D. Renzo, "Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 716–720, May 2020.
- [19] A. Ranjha and G. Kaddoum, "URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, Mar. 2021.
- [20] K. K. Nguyen, S. Khosravirad, D. B. D. Costa, L. D. Nguyen, and T. Q. Duong, "Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning," *IEEE J. Select. Topics Signal Process.*, pp. 1–1, 2021, early access.
- [21] Y. Zou, S. Gong, J. Xu, W. Cheng, D. T. Hoang, and D. Niyato, "Wireless powered intelligent reflecting surfaces for enhancing wireless communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12 369–12 373, Oct. 2020.
- [22] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.
- [23] C. Pan *et al.*, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.
- [24] H. Yang, X. Yuan, J. Fang, and Y.-C. Liang, "Reconfigurable intelligent surface aided constant-envelope wireless power transfer," *IEEE Trans. Signal Process.*, vol. 69, pp. 1347–1361, Feb. 2021.
- [25] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, M. D. Renzo, and F. Chen, "Reconfigurable intelligent surfaces with reflection pattern modulation: Beamforming design and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 741–754, Feb. 2021.
- [26] Y. Chen, B. Ai, H. Zhang, Y. Niu, L. Song, Z. Han, and H. V. Poor, "Reconfigurable intelligent surface assisted device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2792–2804, May 2021.
- [27] L. Wang, K. Wang, C. Pan, W. Xu, and N. Aslam, "Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: A deep reinforcement learning approach," 2020. [Online]. Available: <https://arxiv.org/abs/2007.08380>
- [28] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.
- [29] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.
- [30] S. Yin, S. Zhao, Y. Zhao, , and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8227–8231, Aug. 2019.
- [31] Z. Yang, M. Chen, W. Saad, W. Xu, M. Shikh-Bahaei, H. V. Poor, and S. Cui, "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 665–679, Jan. 2022.
- [32] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Select. Areas Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.
- [33] X. Cao, B. Yang, C. Huang, C. Yuen, M. D. Renzo, D. Niyato, and Z. Han, "Reconfigurable intelligent surface-assisted aerial-terrestrial communications via multi-task learning," *IEEE J. Select. Areas Commun.*, vol. 39, no. 10, pp. 3035–3050, Oct. 2021.
- [34] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific Belmont, MA, 1995, vol. 1, no. 2.
- [35] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.
- [36] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. 4th International Conf. on Learning Representations (ICLR)*, 2016.
- [37] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [38] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc. 4th International Conf. Learning Representations (ICLR)*, 2016.
- [39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [40] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 1928–1937.
- [41] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Sym. Opr. Syst. Design and Imp. (OSDI 16)*, Nov. 2016, pp. 265–283.



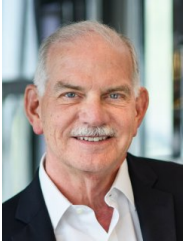
Khoi Khac Nguyen (Student Member, IEEE) was born in Bac Ninh, Vietnam. He received his B.S. degree in information and communication technology from the Hanoi University of Science and Technology (HUST), Vietnam in 2018. He is working towards his Ph.D. degree with the School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, U.K. His research interests include machine learning and deep reinforcement learning for real-time optimisation in wireless networks, reconfigurable intelligent surfaces, unmanned air vehicle (UAV) communication and massive Internet of Things (IoT).



Antonino Masaracchia (Member, IEEE) received the Ph.D. degree in electronics and telecommunications engineering from the University of Palermo, Italy, in 2016. From 2017 to 2018, he was a Post-doctoral Researcher at the Sant' Anna School of Advanced Studies, the BioRobotics Institute. Since September 2018, he has been a Research Fellow with the Centre for Wireless Innovation, Queens University Belfast, U.K. His research interests include fifth generation (5G) and beyond 5G networks, convex optimization and applied machine learning techniques to wireless communications, reconfigurable intelligent surfaces (RIS), UAV-enabled networks, and ultra-reliable and low-latency communications (URLLC).



Vishal Sharma is working as an Assistant Professor (~ UK Lecturer) in the School of Electronics, Electrical Engineering and Computer Science (EECS) at the Queen's University Belfast, Northern Ireland, United Kingdom. He received a PhD in computer science and engineering from Thapar University in 2016. From Nov'16 to Mar'19, he worked in the Information Security Engineering Department at Soonchunhyang University, South Korea. He also held a joint postdoctoral position with Soongsil University, South Korea. He was affiliated with the industry-Academia Cooperation Foundation and the Mobile Internet Security lab at Soonchunhyang University. Before this, he worked as a lecturer in the Computer Science and Engineering Department at Thapar University, India. His areas of research and interests are autonomous systems, UAV communications and security, secure computing, and blockchain.



H. Vincent Poor (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. During 2006 to 2016, he served as the dean of Princetons School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge.

His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications*. (Cambridge University Press). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.



Trung Q. Duong (Fellow, IEEE) received his Ph.D. degree in telecommunications systems from Blekinge Institute of Technology, Sweden in 2012. He is a Chair Professor of Telecommunications at Queen's University Belfast (UK) and holds a prestigious Research Chair of Royal Academy of Engineering. He currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE WIRELESS COMMUNICATIONS LETTERS, and an Executive Editor for

IEEE COMMUNICATIONS LETTERS. He is the recipient of prestigious Royal Academy of Engineering Research Fellowship (2015-2020) and has won a prestigious Newton Prize 2017. His research interests are in the areas of signal processing, optimisation, machine learning, and their applications in wireless communications and networks.