Long-Lasting UAV-aided RIS Communications based on SWIPT

Haoran Peng*, Li-Chun Wang,*, Geoffrey Ye Li†, and Ang-Hsun Tsai*
*Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
†Department of Electrical and Electronic Engineering, Imperial College London, London, U.K.

***Abstract*—Reconfigurable intelligent surface (RIS) is a promising technology for energy efficient wireless communications and has drawn significant attention recently. Combining unmanned aerial vehicle with RIS (UAV-RIS) can provide on-demand deployment services in dynamic scenarios. However, reaping the benefits of UAV-RIS will be limited by the energy of the battery-powered UAV. To enhance the endurance of UAV-RISs, we develop a novel energy harvesting scheme for simultaneous wireless information and power transfer (SWIPT), resource allocation, and energy harvest from impinging radio-frequency (RF) signals. Different from the exist works, the proposed scheme creatively splits the passive reflected arrays on geometric space for transporting information and harvesting energy simultaneously. Furthermore, a deep deterministic policy gradient (DDPG) scheme is designed to continuously allocate UAV-RIS's resources on both the time and space domains to maximize the total harvested energy, while guaranteeing the communication quality for each user. As shown by our simulation results, the proposed UAV-RIS SWIPT system improves performance significantly over the benchmark.**

***Index Terms*—Unmanned Aerial Vehicle, Reconfigurable Intelligent Surface, SWIPT, Energy Harvesting**

I. Introduction

Reconfigurable intelligent surfaces (RISs) are artificial meta-surfaces of an electromagnetic material with large passive reflected arrays and each antenna element can be independently controlled [1]. The passive reflective antenna elements in RIS can be intelligently configured with amplitude, polarization, and phase shift in a programmable manner to create a desirable multipath effect, thereby enhancing the signal strength of the overall received signals or suppressing interference [2], [3]. Recent research demonstrates that RIS achieves significant success in sustainable and green wireless communications [4]. However, the static deployment of RISs, e.g., installed on buildings, limits the effectiveness of RIS in dynamic application scenarios.

Thanks to small sizes, controllability, and flexibility, unmanned aerial vehicles (UAVs) have been widely adopted for rapid networking between access point (AP) and user terminals (UTs) in communication-disabled areas recently [5]. Hence, UAV mount RIS (UAV-RIS) provides a potential solution to on-demand deploy RISs in dynamic scenarios. Nevertheless, the on-board battery capacity of UAVs limits the endurance of UAV-assisted RISs communications.

This work has been partially funded by the Ministry of Science and Technology under the Grants MOST 110-2221-E-A49-039-MY3 and MOST 110-2634-F-A49-006-, Taiwan. This work was also financially supported by the Center for Open Intelligent Connectivity from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

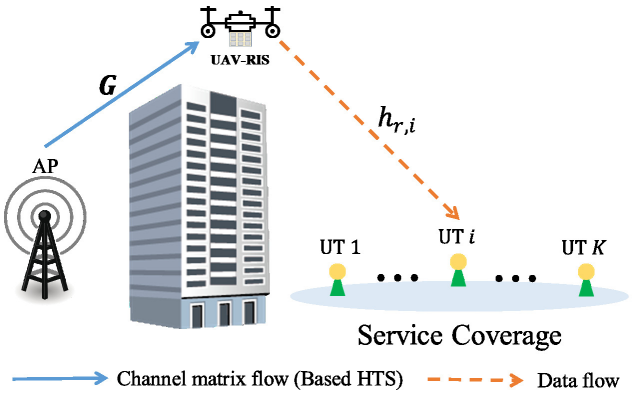

Fig. 1: The considered application scenario.

To overcome the limitation of on-board energy of UAV-RISs, energy harvest (EH) from impinging radio-frequency (RF) signals via simultaneous wireless information and power transfer (SWIPT) was proposed in [6]. One of the efficient SWIPT modes, the harvest-transmit-store (HTS) model divides each time block into two time slots for EH and information transmission [7]. In addition, when there is a small number of user terminals (UTs) in the service coverage, using all the reflect-arrays for signal transmitting may waste resources. To enhance the endurance of UAV-RISs, partial reflection units can be used to collect energy from the received RF signals, while the others reflect signals. However, maximizing the harvested energy and guaranteeing the communication quality on both the time and space domains is a non-convex problem. Hence, reaping UAV-RISs' benefits depends on how to balance the effects of EH and communication quality.

In the literature, some interesting research results in balancing the energy-efficient and communication quality of UAV-aided RIS wireless communications have been reported [1], [8], [9]. Further in [10], a static RIS-assisted SWIPT system was proposed to transfer power and information from the access point (AP) to the energy receivers and the information receivers, respectively. In [6], the RIS is equipped with an energy storage system (ESS) to manage the EH process. Then, the overall energy efficiency of the RIS-assisted cellular

978-1-6654-4266-4/22/$31.00 ©2022 IEEE

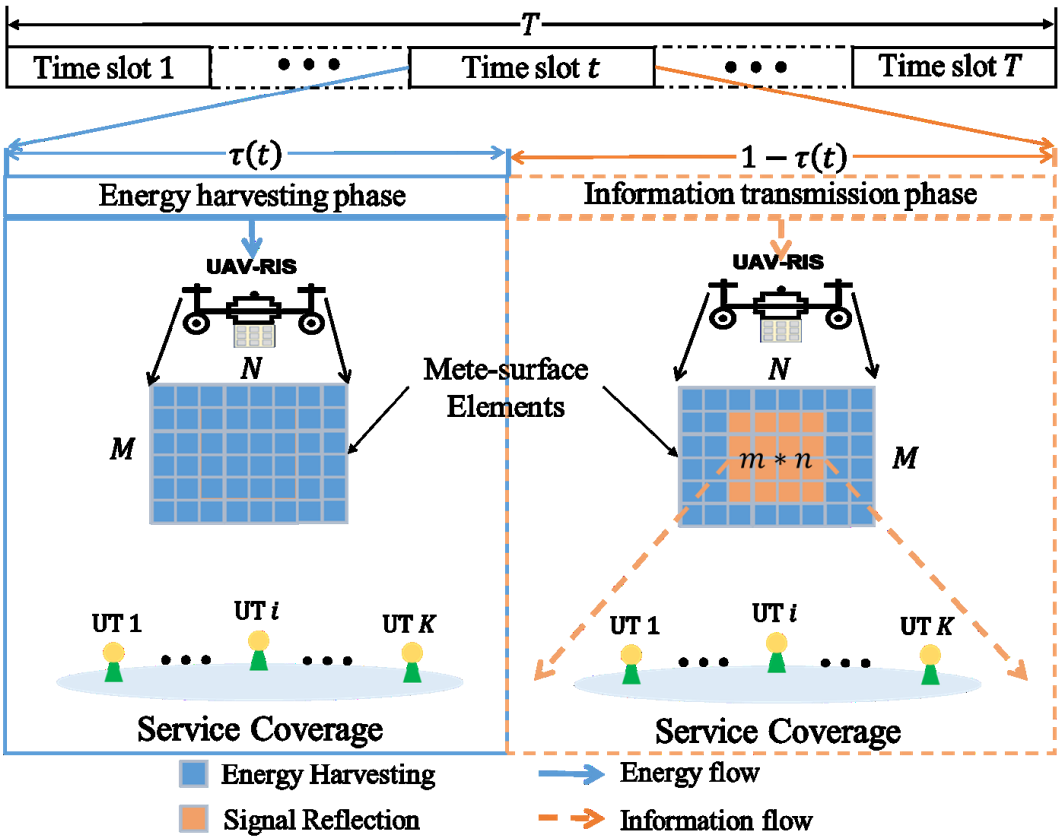

Fig. 2: A resources allocation joined with harvest-transmit-store model for the UAV-assisted RIS communication system.

network is improved by harvesting energy from the received RF signals. In [11], UAVs are integrated with RISs to flexibly deploy RISs in dynamic scenarios while other approaches install RISs on a static building.

In this paper, we design an EH method that combines the HTS-based SWIPT with resource allocation of passive reflect-arrays to enhance the endurance of UAV-assisted RISs communications systems. Motivated by the successful application of deep reinforcement learning (DRL) in wireless communications [3], [6], [11], [12], we suggest utilizing DRL in handling complicated control problems of resource allocation. As such, we develop a framework based on the deep deterministic policy gradient (DDPG) [13] to maximize the collected energy while guaranteeing the communication quality of each UT. Different from the previous works, this paper creatively designs a space splitting mechanism for the passive reflect-arrays for transporting information and harvesting energy simultaneously. To the best of our knowledge, this is the first work that enhances the endurance of UAV-aided RIS communications systems via harvesting energy on both the time and space domains while meeting the required constraints of quality of service (QoS).

The rest of this paper is organized as follows. The system model is introduced in Section II. In Section III, we formulate an optimization problem to maximize the total harvested energy of the UAV-RIS. Section IV develops the DDPG-based EH framework for UAV-RISs. In Section V, the proposed joint method of SWIPT and resource allocation is verified by simulation results. Our concluding remarks and future work are given in Section VI.

II. System Model

As shown in Fig. 1, a UAV-RIS is deployed to assist signal transmission from the AP to K single-antenna UTs denoted by $\mathcal{K} = \{1, 2, \cdots, K\}$, where some obstacles block the line-of-sight (LoS). In this work, the AP with D antennas transmits signal to a UAV-RIS consisting of $M \times N$ meta-surfaces. We assume that UTs can only receive the signals reflected by the UAV-RIS. The meta-surface element at the i-th row and the j-th column is denoted by $\mathcal{R}_{i,j}$. Without loss of generality, we denote the meta-surfaces element array of the UAV-RIS as $\mathcal{R} = \{\mathcal{R}_{i,j}\}_{i,j=1}^{M,N}$. In addition, the RIS can exchange channel state information (CSI) with the AP via the attached smart controller.

A. Resources Allocation based HTS Model

We present an HTS-based model to enhance the UAV's endurance via harvesting energy on time and space domains in this section. The UAV-RIS is equipped with a rechargeable battery, which stores the harvested energy and converts into electrical power [6]. We assume that the linear transmit precoding is used at the AP for simplicity of implementation

As shown in Fig. 2, the UAV-RIS often operates in the communication disabled area to extend the coverage of the AP. For reflecting signals and EH, we divide the whole time period into T equal-time slots, denoted as $\mathcal{T} = \{1, 2, \cdots, t, \cdots, T\}$. Each slot contains two phases: the EH phase and the information transmission phase. At the t-th time slot, the length of the EH phase is denoted by $\tau(t)$. During the EH phase, $\tau(t)$, all the reflecting units only harvest energy. After the EH phase, the information transmission phase starts immediately. Besides, all the meta-surfaces are used to reflect signals during the information transmission phase. Similar to [14], [15], we consider the normalized unit time slot in the sequel. The length of the information transmission phase at the t-th time slot is $(1 - \tau(t))$. There are $m \times n$ $(0 \le n \le N, 0 \le m \le M)$ meta-surfaces in the center of the UAV-RIS, which are used to reflect signals at the information transmission phase. The rest of the meta-surfaces in the UAV-RIS are used to harvest energy. For simplicity, we denote the array elements in the center of the UAV-RIS for signal reflection as $\mathcal{L} = \left\{\mathcal{R}_{i,j}\right\}_{i,j=1}^{m,n}, \mathcal{L} \subseteq \mathcal{R}$. Denote the ratio of the area of signal reflection units to that of all meta-surface units as $\lambda = \frac{m \times n}{M \times N}$. Following [10], the transmitted signals from the AP can be represented as

$$\boldsymbol{G} = \sum_{k \in \mathcal{K}} \boldsymbol{V}_k \mathcal{S}_k, \tag{1}$$

where $\boldsymbol{V}_k \in \mathbb{C}^{D \times 1}$ and $\mathcal{S}_k$ are the precoding vectors and the signals for the k-th UT, respectively. In addition, $\mathcal{S}_k$ is a circularly symmetric complex Gaussian (CSCG) random variable with zero mean and unit variance, that is $\mathcal{S}_k \sim C\mathcal{N}(0, 1)$ [16]. Thus, the total transmit power at the AP is given by

$$\mathbb{E}(\boldsymbol{G}^H \boldsymbol{G}) = \sum_{k \in \mathcal{K}} \|\boldsymbol{V}_k\|^2 \le p_{max}, \tag{2}$$

where $\|\cdot\|$ represents the vector's Euclidean norm and p_{max} is the upper limit of the AP's transmit power. At the t-th time slot, the UAV-RIS harvested energy is given by

$$\begin{aligned} E(t) = \tau(t) \sum_{i=1}^{M} \sum_{j=1}^{N} \eta \|\hat{h}_{i,j}^H g_{i,j} \boldsymbol{G}\|^2 + \\ (1 - \tau(t)) \sum_{i=1}^{M} \sum_{j=1}^{N} \omega_{i,j} \eta \|\hat{h}_{i,j}^H g_{i,j} \boldsymbol{G}\|^2, \end{aligned} \tag{3}$$

where $\hat{h}_{i,j}^H$ is the channel between the AP and the meta-surface element $\mathcal{R}_{i \times j}$ and follows the complex Gaussian random distribution, $C\mathcal{N}(0, 1)$, $\eta \in (0, 1)$ is the EH efficiency. $\omega_{i,j} = 0$ denotes the fact the element $\mathcal{R}_{i,j}$ is adopted to reflecting signals and $\omega_{i,j} = 1$ otherwise. Hence, $\omega_{i,j}$ can be given by

$$\omega_{i,j} = \begin{cases} 0, & \mathcal{R}_{i,j} \in \mathcal{L}. \\ 1, & \mathcal{R}_{i,j} \notin \mathcal{L}. \end{cases} \tag{4}$$

The channel power gain, $g_{i,j}$, from the AP to each reflection element, $\mathcal{R}_{i,j}$, can be expressed as

$$\begin{aligned} g_{i,j} = \left(P_{i,j}(LoS) + \left(1 - P_{i,j}(LoS)\right) \varphi\right) \\ \times \left(\sqrt{x_{i,j}^2 + y_{i,j}^2 + H_{i,j}^2}\right)^{-\alpha}, \end{aligned} \tag{5}$$

where $H_{i,j}$ is the altitude of meta-surface element $\mathcal{R}_{i,j}$ of the Cartesian coordinate system where the AP is located at the origin, $(x_{i,j}, y_{i,j})$ is the position of the meta-surface element $\mathcal{R}_{i,j}$, α is the path loss exponent from $\mathcal{R}_{i,j}$ to the AP, and φ is the additional attenuation factor caused by the non-line-of-sight (NLoS) connection. $P_{i,j}(LoS)$ is the line-of-sight (LoS) probability between the AP and meta-surface element $\mathcal{R}_{i,j}$. Following [7], we can calculate the LoS probability $P_{i,j}(LoS)$ by (6).

$$P_{i,j}(LoS) = \frac{1}{1 + \mathcal{A} \times exp\left(-\mathcal{B}\left(\theta_{i,j} - \mathcal{A}\right)\right)}, \tag{6}$$

where $\mathcal{A}$ and $\mathcal{B}$ are constants depending on the environments [17]. The elevation angle, $\theta_{i,j}$, between the AP and the meta-surface element $\mathcal{R}_{i,j}$ is given by

$$\theta_{i,j} = \frac{180}{\pi} \sin^{-1}\left(\frac{H_{i,j}}{\sqrt{x_{i,j}^2 + y_{i,j}^2 + H_{i,j}^2}}\right). \tag{7}$$

B. AP-RIS-UT Channel Model

At the information transmission phase in time slot t, $\boldsymbol{Z} \in \mathbb{C}^{\mathcal{L} \times D}$ and $\boldsymbol{h}_{i,j}^H(k) \in \mathbb{C}^{1 \times \mathcal{L}}$ represent the baseband equivalent channels from the AP to the UAV-RIS and from the UAV-RIS element $\mathcal{R}_{i,j}$ to the k-th UT, respectively. Furthermore, the UAV-RIS passively reflects the received information signals via controlling $\mathcal{L}$ reflecting phase shifts. Following [10], we define a diagonal matrix $\boldsymbol{\Theta}$ as the reflection-coefficients matrix of the UAV-RIS by

$$\boldsymbol{\Theta} = \text{diag}(\beta_1 e^{j\theta_1^r}, \cdots, \beta_{\mathcal{L}} e^{j\theta_{\mathcal{L}}^r}), \tag{8}$$

where $j = \sqrt{-1}$ is the imaginary unit, $\theta_l^r \in [0, 2\pi), \forall l \in \mathcal{L}$ represents the phase-shift of l-th reflection unit, and $\beta_l \in [0, 1], \forall l \in \mathcal{L}$ represents the amplitude reflection coefficient. We assume that the RIS can be controlled to have an accurate phase shift as we desire in this work. The design of passive beamforming will be studied in future work. Besides, β_l ideally sets to unit since each meta-surface element's antenna can be independently controlled to maximize the signal reflection for simplicity [10]. From (1), the received RF signal at the k-th UT via the AP-RIS-UT channel can be expressed as

$$\mathcal{Y}_k = \sum_{i=1}^{M} \sum_{j=1}^{N} (1 - \omega_{i,j}) \boldsymbol{h}_{i,j}^H(k) \boldsymbol{\Theta} \boldsymbol{Z} \boldsymbol{G} + \nu_k, k \in \mathcal{K}, \tag{9}$$

where $\nu_k \sim C\mathcal{N}(0, \sigma_k^2)$ represents the additive white Gaussian noise (AWGN) at the k-th UT with noise power σ_k^2. Rician fading is considered in the wireless channel model from the UAV-RIS to UTs, while the small-scale fading is ignored in this work. The path-loss between the UAV-RIS and UTs is given by $\kappa \left(\frac{d_k^{ru}}{d'}\right)^{-\bar{\alpha}}$, where $\bar{\alpha}$ represents the path-loss exponent for RIS-UT links, d_k^{ru} is the distance between the UAV-RIS and the k-th UT, and κ corresponds to the path-loss at the reference distance of $d' = 1m$. Following [16], we assume that each UT can perfectly cancel the interference from other RIS-UT links before decoding desirable signal $\mathcal{S}_k$. The received

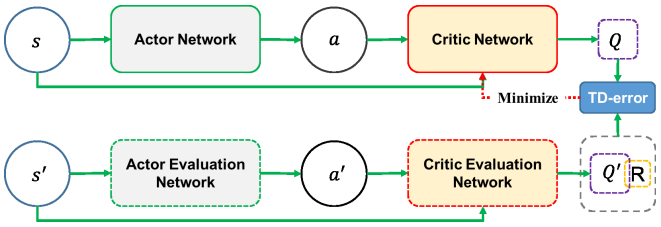

Fig. 3: The architecture of DDPG.

signal-to-noise ratio (SNR) at each UT $k, k \in \mathcal{K}$ is given by

$$\mathrm{SNR}_k = \frac{|\sum_{i=1}^{M}\sum_{j=1}^{N}(1-\omega_{i,j})\boldsymbol{h}_{i,j}^H(k)\boldsymbol{\Theta}\boldsymbol{Z}V_k|^2}{\sum\limits_{k'\neq k,k'\in\mathcal{K}}|\sum_{i=1}^{M}\sum_{j=1}^{N}(1-\omega_{i,j})\boldsymbol{h}_{i,j}^H(k')\boldsymbol{\Theta}\boldsymbol{Z}V_{k'}|^2+\sigma_k^2}. \tag{10}$$

The received SNR at each UT is required to be greater than or equal to a given SNR_{min} within the finite time horizon to maintain service quality, that is,

$$SNR_k(t) \geq SNR_{min}, \forall k \in \mathcal{K}, t \in \mathcal{T}. \tag{11}$$

III. Problem Formulation

This work aims to maximize the total harvested energy $\sum_{t=1}^{T} E(t)$ of the UAV-RIS within the finite time horizon T while satisfying the required minimal SNR constraints. Without loss of generality, the total transmit power at the AP shall also satisfy a constraint. The optimization problem is formulated as follows

$$\begin{aligned}
(\mathrm{P1}):\ \bar{E} = &\max_{\tau(t),p,\lambda}\sum_{t=1}^{T}E(t),\\
s.t.\ C1\ &:\ SNR_k(t)\geq SNR_{min},\ \forall k\in\mathcal{K}, t\in\mathcal{T},\\
C2\ &:\ 0\leq\tau(t)\leq 1, \forall t\in\mathcal{T},\\
C3\ &:\ 0\leq p=\sum_{k\in\mathcal{K}}\|\boldsymbol{V}_k\|^2\leq p_{max},\\
C4\ &:\ 0\leq\lambda\leq 1.
\end{aligned} \tag{12}$$

$C1$ represents the required minimum SNR constraints on each UT to guarantee the quality of service (QoS) of wireless networks. $C2$ is the time constraint and $C3$ is the maximum power control constraint of AP. $C4$ is the constraint of the area ratio of the signal reflection units to all the meta-surfaces.

The optimization problem in (P1) is non-convex due to the non-convex constraints and coupling of multiple variables. Thus, (P1) cannot be efficiently solved via the standard convex optimization methods [8]. Therefore, we develop a DRL-based framework to cope with it in the next section. Besides, RIS always bears over 10×10 elements array. Hence, we regard λ as a continuous variable to simplify action space and accelerate the convergence of DRL.

IV. DDPG Based Framework

In this section, we present a DRL based meta-surface elements resource allocation framework that can maximize the total harvested energy of the UAV-RIS while guaranteeing the minimal SNR demand of its service coverage. A reinforcement learning (RL) framework comprises two components, an agent and an environment. The agent acts on the environment based on the environment's state. Then the agent updates its knowledge with the reward value returned from the environment to evaluate the last action. The interaction between the agent and the environment is a Markov Decision Process (MDP) [18].

DDPG is a model-free, off-policy, and actor-critic based RL approach [13]. Fig. 3 illustrates the architecture of the proposed DDPG-based EH method. From the figure, A deterministic policy network (DPN), $a = \pi(s \mid \delta^{\pi})$, is maintained as a actor to determine the actions in continuous space. The critic Q-network $Q(s, a \mid \delta^{Q})$ is used to criticize the performance of the actor. s represents the environment state in each step. The Q-network faces the challenge for Q update inclines to divergence and the single $Q(s, a \mid \delta^{Q})$ being updated is used in calculating the target value [13]. Therefore, a copy of the actor and critic nets, $\pi'(s \mid \delta^{\pi'})$ and $Q'\left(s, a \mid \delta^{Q'}\right)$, are created as the target nets to overcome the challenge by calculating the corresponding target values. The structure of each target net is the same as its corresponding net. This work use δ^{π} and δ^{Q} present the parameter of the DPN and the Q-network, respectively. Besides, $\pi(\cdot)$ maps the policy from state to action and $Q(\cdot)$ is the approximator that uses the state-action pairs to generate Q-value. The weights of these target networks are soft updated as follows:

$$\begin{aligned}
\delta^{Q'} &\leftarrow \psi_a\delta^{Q} + (1-\psi_a)\delta^{Q'},\\
\delta^{\pi'} &\leftarrow \psi_c\delta^{\pi} + (1-\psi_c)\delta^{\pi'},
\end{aligned} \tag{13}$$

where $\psi_a \ll 1$ and $\psi_c \ll 1$ are the learning rate for the soft update on training actor and critic networks, respectively. Furthermore, DDPG creates an exploration policy to cope with the challenge of learning in continuous action spaces via adding a noise sampled from the stochastic noise process $\mathcal{N}$,

$$\bar{\pi}(s_t) = \pi\left(s_t \mid \delta_t^{\pi}\right) + \mathcal{N}, \tag{14}$$

while $\mathcal{N}$ can be chosen to suit the environment.

In this work, the RL environment depends on the communication system assumption. The reward function, state space, and action space are given by:

- *State Space*: The state s_t at the t-th time step is constructed by the action from the $(t-1)$-th time step, the transmitted signal from the AP $\boldsymbol{G}$ at the t-th time step, the distance between AP and UTs d_k^{ru}, and baseband equivalent channels $\boldsymbol{Z} \in \mathbb{C}^{\mathcal{L}\times D}$.
- *Action Space*: At the t-th time step, action a_t of the proposed DRL based framework consists of two main components, the length of the EH phase $\tau(t)$ and the ratio of the signal reflection units to all the meta-surfaces units λ. Besides, both $\tau(t)$ and λ are defined in a continuously feasible region.
- *Reward function*: The positive reward represents the objective of the proposed framework, which efforts to maximize the total harvested energy of the UAV-RIS. At the t-th time step, the harvested energy $E(t)$ is

Algorithm 1: The proposed DDPG-based scheme

Input: $\boldsymbol{G}$, $\boldsymbol{Z} \in \mathbb{C}^{\mathcal{L}\times D}$, $\boldsymbol{\Theta}$, $\hat{h}_{i,j}^{H}$, $\boldsymbol{h}_{r,k}^{H}, \forall k \in \mathcal{K}$, the channel power gain $\left\{g_{i,j}\right\}_{i,j=1}^{M,N}$, the size of experience replay $N_{\mathcal{D}}$, the size of mini-batches N_B;
Initial: The actor $\pi\left(s \mid \delta^{\pi}\right)$ and critic $Q\left(s, a \mid \delta^{Q}\right)$ networks, the target actor $\pi'\left(s \mid \delta^{\pi'}\right)$ and target critic $Q'\left(s, a \mid \delta^{Q'}\right)$ networks, $\delta^{\pi'} = \delta^{\pi}, \delta^{Q'} = \delta^{Q}$; experience replay memory $\mathcal{D}$ with a capacity of $N_{\mathcal{D}}$;
Output: Optimal action $a = \{\tau(t), \lambda\}$, and the total harvested energy $\bar{E}$ of the UAV-RIS.
for *episode* $N_e = 1$ *to* N_{epoch} **do**
  Receive the current $\boldsymbol{G}$, $\boldsymbol{Z} \in \mathbb{C}^{\mathcal{L}\times D}$, $\boldsymbol{\Theta}$;
  Initialize a stochastic noise process $\mathcal{N}$;
  Collect $\boldsymbol{h}_{r,k}^{H}, \forall k \in \mathcal{K}$ for N_e-th episode to s_1;
  for $t = 1$ *to* T **do**
    Select $a_t = \pi\left(s_t \mid \delta_t^{\pi}\right) + \mathcal{N}$ according to current policy;
    Execute a_t to calculate its corresponding reward $\mathscr{R}_t$;
    Obtain the next state s_{t+1};
    Store the transition $(s_t, a_t, \mathscr{R}_t, s_{t+1})$ into $\mathcal{D}$;
    Randomly sample a mini-batch of N_B transitions $(s_j, a_j, \mathscr{R}_j, s_{j+1})$ from $\mathcal{D}$;
    Use (16) to set target Q-value;
    Update critic net via minimizing (17);
    Using the sampled policy gradient to update the actor policy: (18);
    According to (13), soft update the target critic and actor nets;
    Update $s_t = s_{t+1}$;
  end
end

taken account into as the reward value, which is defined in (3). The proposed framework also needs to account for users' minimum capacity requirement defined in the constraints $C1$. Hence, reward $\mathscr{R}_t$ can be described as the following:

$$\mathscr{R}_t = \begin{cases} E(t), & SNR_k(t) \geq SNR_{min}, \forall k \in \mathcal{K}. \\ 0, & SNR_k(t) < SNR_{min}, \forall k \in \mathcal{K}. \end{cases} \tag{15}$$

The cumulated reward is given by $\hat{\mathscr{R}} = \sum_{t=1}^{T} \mathscr{R}_t$.

Algorithm 1 illustrates the work principle of the proposed DDPG-based framework. The actor and critic networks and their corresponding target networks are generated at the initialization stage. An empty replay buffer $\mathcal{D}$ with the size of $N_{\mathcal{D}}$ is initialized for learning process. In each episode, the actor selects an action, a_t, according to the current policy and calculates the corresponding reward $\mathscr{R}_t$. The tuple $(s_t, a_t, \mathscr{R}_t, s_{t+1})$ is stored into $\mathcal{D}$. Then, a mini-batch of N_B transitions is sampled from the replay memory $\mathcal{D}$ to calculate

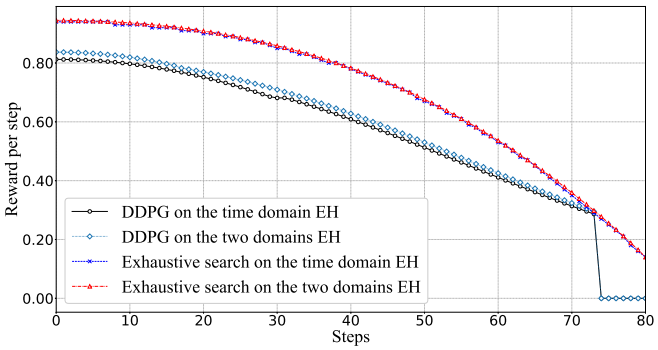

Fig. 4: Reward per step of the proposed DDPG-base method and benchmark.

the target Q-value $\mathscr{y}_j$,

$$\mathscr{y}_j = \begin{cases} \mathscr{R}_j, & j = N_B, \\ \mathscr{R}_j + \vartheta Q'\left(s_{j+1}, \pi'\left(s_{j+1} \mid \delta^{\pi'}\right) \mid \delta^{Q'}\right), & j < N_B. \end{cases} \tag{16}$$

$\vartheta \in [0, 1]$ is a discounting factor. The critic evaluation network is updated by minimizing the loss function:

$$L\left(\delta^{Q}\right) = \frac{1}{N_B} \sum_{j=1}^{N_B} \left(\mathscr{y}_j - Q\left(s_j, a_j \mid \delta^{Q}\right)\right)^2. \tag{17}$$

The actor evaluation network is updated using the policy gradient as following,

$$\Delta_{\delta^{\pi}} = \frac{1}{N_B} \sum_{j=1}^{N_B} \left(\nabla_a Q\left(s_j, \pi\left(s_j \mid \delta^{\pi}\right) \mid \delta^{Q}\right) \mid \nabla_{\delta^{\pi}} \pi\left(s_j \mid \delta^{\pi}\right)\right). \tag{18}$$

At the end of each time step, the target actor and target critic networks are updated through the soft update method in (13).

V. Simulation Results

We evaluate the proposed framework through simulation in python. Referring to [9], [6], [19], we set $\mathcal{A}$: 9.61; η: 0.7; $\mathcal{B}$: 0.16; κ: -30dB; φ: 20dB; P_{max}: 500W; $\bar{\alpha}$ and α to be 2.5 and 3, respectively. The number of RIS's meta-surface elements is 100, and the SNR_{min} is 12 dB. In this work, we use the exhaustive search approach as the benchmark. Besides, the complexity of the exhaustive search method is non-deterministic polynomial. Therefore, we set the number of users $K = 1$ in order to compare the performance of the proposed DDPG-based method with benchmarks. The multiple UTs scenario will be studied in the future work. The detail information for the simulation can be found in the source code: https://github.com/Haoran-Peng/UAV-RIS_EH_DDPG.

Fig. 4 illustrates the reward per step of the proposed DDPG-based scheme and benchmark methods for eighty-one interactions. The Euclidean distance between the UAV-RIS and the UT increased from twenty meters to sixty meters in eighty-one steps. The reward represents the ratio of the harvested energy to the received energy from impinging RF signals each

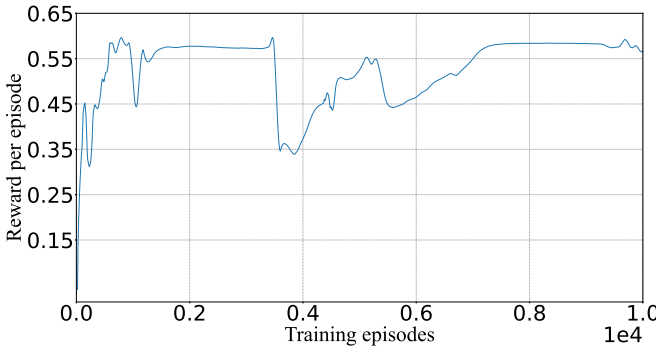

Fig. 5: Reward per training episode.

step. From the simulation results, the proposed DDPG-based method can harvest 57.03% and 55.32% of the energy from the received RF signal in the two-domains and time-domain schemes, respectively. The upper limit of the harvested energy obtained by searching all probabilistic actions is 65.1% and 64.7% in the two-domains and time-domain schemes, respectively. Nevertheless, the complexity of the exhaustive-search algorithm requires non-deterministic polynomial-time, which results in loss of practicality in a real-world application. Therefore, the proposed DDPG-based scheme achieves significant performance via trade-off energy efficiency and practicality to enhance the durability of the UAV-RIS. Compared to the time-domain optimization, the joint optimization of time and spatial is not significant, although it can achieve approximately 1% improvement in EH. The reason is that the experiment considers a single UT scenario while the optimization problem depends on the number of UTs to some extent. Hence, the multiple UTs scenario is worthwhile to study in future work.

Fig. 5 presents the convergence behavior of the proposed DDPG-based method for harvesting energy on both time and space domains. The cumulative rewards for each training episode are increasing as the training iteration continues. From the figure, the cumulative rewards per episode approximately reach 60% in 1,000 training iterations. The training performance converges since 7,000 training episodes after some fluctuations caused by the exploration. Simulation results demonstrate the effectiveness of the proposed DDPG-based approach for collecting energy simultaneous on time and space domains of RIS.

VI. Conclusions and Future Work

In this work, we investigated the energy-harvesting policy for unmanned aerial vehicle (UAV) mounted RIS (UAV-RIS). We proposed a novel long-endurance scheme via allocating resources of passive reflect-arrays to harvest energy on both time and space dimensions. The scheme aims to maximize the total harvested energy within the finite time horizon while satisfying the required communication quality of service (QoS) constraints. Therefore, a novel deep deterministic policy gradient (DDPG)-based approach was used to solve the formulated non-convex optimization problem in this work. The proposed DDPG-based approach can harvest 57.03% of the energy from received radio-frequency (RF) signals with an acceptable time complexity. The optimized long-lasting scheme for the multiple user terminals (UTs) scenario is worthwhile studied in the future.

References

[1] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. Cogn. Commun. Netw*, vol. 6, no. 3, pp. 990–1002, May 2020.

[2] E. Basar, M. Di Renzo, J. De 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.

[3] 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.

[4] A. Balakrishnan, S. De, and L.-C. Wang, "Traffic skewness-aware performance analysis of dual-powered green cellular networks," in *IEEE Glob. Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020.

[5] H. Peng, A.-H. Tsai, L.-C. Wang, and Z. Han, "LEOPARD: Parallel optimal deep echo state network prediction improves service coverage for UAV-assisted outdoor hotspots," *IEEE Trans. Cogn. Commun. Netw.*, Sep. 2021, doi: 10.1109/TCCN.2021.3115765.

[6] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jul. 2020.

[7] M. Lei, X. Zhang, B. Yu, S. Fowler, and B. Yu, "Throughput maximization for UAV-assisted wireless powered D2D communication networks with a hybrid time division duplex/frequency division duplex scheme," *Wireless Netw.*, vol. 27, pp. 2147–2157, Feb. 2021.

[8] Z. Yang, W. Xu, C. Huang, J. Shi, and M. Shikh-Bahaei, "Beamforming design for multiuser transmission through reconfigurable intelligent surface," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 589–601, Oct. 2021.

[9] H. Mei, K. Yang, J. Shen, and Q. Liu, "Joint trajectory-task-cache optimization with phase-shift design of RIS-assisted UAV for MEC," *IEEE Wireless Commun. Lett.*, Apr. 2021.

[10] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted SWIPT under QoS constraints," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1735–1748, Jul. 2020.

[11] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.

[12] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[13] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, Sep. 2015.

[14] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Dec. 2013.

[15] H. Wang, J. Wang, G. Ding, L. Wang, T. A. Tsiftsis, and P. K. Sharma, "Resource allocation for energy harvesting-powered D2D communication underlaying UAV-assisted networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 14–24, May. 2018.

[16] Y. Tang, G. Ma, H. Xie, J. Xu, and X. Han, "Joint transmit and reflective beamforming design for IRS-assisted multiuser MISO SWIPT systems," in *IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020.

[17] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *IEEE Glob. Commun. Conf.*, Austin, TX, Dec. 2014.

[18] R. Bellman, "A markovian decision process," *J. Appl. Math. Mech.*, vol. 6, no. 5, pp. 679–684, May 1957.

[19] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.