# *A DRL-based Energy Efficiency Optimization of IRS-aided UAV WPT Networks*

Kimchheang Chhea, Sengly Muy, and Jung-Ryun Lee, *Senior Member, IEEE*

***Abstract*—Lower production costs have inspired studies on unmanned aerial vehicles (UAV) for wireless communication. However, limited transmission power and size of the UAV make it challenging to use advanced communication models while meeting the growing need for high data rates and energy efficiency (EE). In this paper, an energy transfer-enabled network with UAV aided by intelligent reflecting surface (IRS) is studied, where IRS is adopted to enhance the performance of ground user equipment (GUE). The problem of maximizing the average EE is studied by jointly controlling the UAV's flying route, IRS phase steer, UAV transmission power, and power splitting (PS) ratio of the energy transfer technology. The formulated problem is non-convex and thus challenging to be solved. To address this problem, we propose a deep reinforcement learning (DRL) approach. The modified reward function is implemented to enhance the efficiency of the DRL agent, which is formulated based on the expected signal-to-interference-plus-noise ratio (SINR) map. Simulation results demonstrate that the proposed DRL algorithm achieves lower energy consumption, higher data rate, and improved EE compared to the comparison algorithm.**

***Index Terms*—Intelligent reflecting surface, unmanned aerial vehicle, deep reinforcement learning, energy transfer.**

## I. Introduction

The latest technological progress of unmanned-aerial-vehicles (UAV) and the drop in production costs have inspired substantial studies on UAV applications in wireless communications. It is expected that UAV-enabled wireless communication can be utilized as a mobile base station (BS) in the upcoming wireless communication networks, providing ultra-reliable service and high data rate communications while supporting a massive number of users [1]. It is noted that UAV-aided data delivery is an *important* communication technology in Internet of Things (IoT) environments. This is because it can reduce the energy consumption of IoT devices by positioning the UAV near low-battery devices, thereby prolonging the network lifetime. *As opposed to the ground base station (BS) that is always supplied with reliable energy source, UAVs cannot obtain energy while flying. Additionally, the location of the UAV can impact the energy consumption of both the ground device and the UAV itself. Therefore, careful trajectory planning and transmission power control are required to satisfy the increasing demands for high data rate and energy efficiency (EE) [2].*

Recently, intelligent reflecting surface (IRS) has emerged as a highly promising and innovative communication paradigm within the area of wireless networks. The IRS is a type of metasurface that directs incoming communication signals to achieve extended coverage, improve physical layer security, and support massive network connections [3]. *Using IRS in UAV networks can extensively improve the communication range of the UAV without using a significant amount of energy. In addition, IRS enables passive beamforming, which mitigates high RF signal attenuation and establishes an effective transmission beam to ground devices. These capabilities to manage the wireless environment offer distinct advantages in addressing various challenges in wireless communications, such as improving spectrum efficiency and EE [4].*

On the other hand, IoT networks primarily consist of wireless nodes that are geographically dispersed or spatially spread out, such as sensor nodes and device-to-device communications [5]. One challenging issue for these networks is lowering energy usage and prolonging the lifespan of the network. Because battery replacement or regular recharging can be costly and inconvenient, *harvesting* energy from the surrounding environment is regarded as a sustainable way to offset the energy consumption of the devices [6]. Specifically, UAV-enabled wireless energy transfer holds great potential as it offers the flexibility to efficiently cover a specific area by dynamically adjusting source-to-destination distance. This adaptability allows for meeting the energy requirements of diverse nodes and enhancing energy harvesting efficiency. Furthermore, it is noted that the EE of an IoT network can be further enhanced by integration of wireless power transfer (WPT) technology into the IRS-aided UAV network platform.

When integrating UAV into the IoT network in wireless communication, one key aspect to consider is the optimization of UAV trajectory. Even though our work focuses mainly on the communications and optimization of the ground network/devices, the consideration of a trajectory optimization of the UAV in the networks allows for more robust and successful communications between the UAV and ground devices [7]. Our optimization approach introduces the SINR map concept which is an important measurement for communication quality and EE. *The SINR map is defined specifically for use in the DRL reward function.* Integrating the SINR map with learning-based algorithm such as DRL allows for smoother training, increased stability and better adaptation to complex environment, *which we will describe in more detail in Section IV.* This leads to better strategy in interference management and UAV route planning. In addition, the better adaptation to environment is advantageous for scalability which means the

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251105), and by the Human Resources Development (No.20214000000280) of the Korea Institute of Energy Technology Evaluation and Planning(KETEP) grant funded by the Korea government Ministry of Trade, Industry and Energy, and by the Chung-Ang University Young Scientist Scholarship in 0000(2019).

approach can allocate resources more efficiently among nodes.

In this work, we consider increasing the network's lifetime of the simultaneous wireless information and power transfer (SWIPT)-based UAV network, where an IRS is employed at edge of the network to improve the EE of the GUEs. The performance of the network's lifetime of SWIPT-based IRS-aided network is affected by various aspects, such as the geographical position of the UAV, the transmission power level of the UAV, the phases of the IRS units, and the energy harvesting/information reception ratio (power splitting ratio) in a SWIPT system. Therefore, our work aims to maximize the average EE of the GUEs by simultaneously optimizing the UAV flying route, transmission power, IRS phase steer, and power splitting (PS) ratio in a SWIPT-based IRS-aided UAV network. The key contributions of this paper are outlined as follows:

- We construct a system model of the IRS-aided UAV communication system equipped with SWIPT functionality. From the channel model for the UAV data delivery with an IRS and the energy model of the GUE for IRS-aided UAV data delivery, we develop an optimization model of maximizing the average EE of the GUEs with decision variables of the UAV flying route, transmission power, IRS phase steer, and energy harvesting ratio of SWIPT functionality.
- To solve this optimization problem with very high computational complexity, we propose a deep reinforcement learning (DRL) algorithm. The proposed DRL algorithm introduces the concept of the *SINR map (the average SINR of the UAV over the GUEs in the given network)*. From the *SINR map*, we construct the reward function using bivariate normal distribution in the proposed DRL algorithm.
- It is verified that the proposed DRL algorithm enhances the performance of the UAV in that it consumes less energy on average and maintains high data rate compared to the comparison schemes. Also, the results verify that using a UAV equipped with IRS functionality can significantly reduces the energy consumption of nodes in a network.

## II. Related Works

### A. IRS-enabled communications

Due to the essential qualities of IRS in tackling important challenges in wireless networks, many theoretical and experimental studies have been conducted concerning various scenarios with the focus on analysis and improvement of the system [8]. According to the study in [9], the authors investigated the use of IRS in mobile edge computing (MEC) systems. They formulated a latency-minimization problem and find solutions using the block coordinate descent and the alternative optimization approach. The performance of the MEC-aided IRS was improved by reducing the latency approximately 20% compared to the non-IRS system. In [10], the study on multiple IRS systems that incorporate location information is carried out, which highlighted the high location uncertainty of the IRS and BS. The authors introduced a power allocation scheme aimed at optimizing the system performance. In [11], the authors explored the use of an IRS in a mmWave multiple-input and multiple-output system. They aimed to analyze the data rate between a BS and a mobile user. To optimize performance, the authors proposed a method for determining the ideal IRS phase steer, relying on restricted input from the moving user for the optimal configuration. Alongside improving the data rate, it is possible to minimize both the error bound and the orientation error bound by introducing perfect channel state information.

### B. UAV-assisted networks

The authors of [12] explored three-dimensional (3D) route mapping for a UAV connected to a cellular network to reduce the flight time while maintaining the target link quality. To address this issue, the authors derived the optimal UAV path by solving the equivalent shortest path problem. The results show the flight time was reduced while satisfying the communication quality constraints. In [13] the authors proposed a study considering gathering and dissemination of critical emergency information, transmission powers of the UAV and the GUE, and UAV acceleration. Their results demonstrated an improvement in the quality-of-experience factor and significant reduction in the flying duration of the UAV. In [14], the authors proposed a new scheme which can reduce the computational complexity of the UAV flying route calculation and communication overhead compared to the existing schemes, by optimizing only some selected way-points along the UAV's path. Their method substantially reduced the flying route design complexity and achieved an enhanced rate performance compared to the conventional path discretization method.

### C. IRS-aided UAV networks

Due to the aforementioned feasible characteristics of UAV and IRS, combining these technologies could further enhance the network performance. The study in [16] focused on the problem of minimum rate maximization of users in the UAV and IRS communication systems, and solved it by converting the original problem into successive convex optimization problems with rate constraint penalty. The authors in [17] investigated UAV with the assistance of IRS in wireless radio systems. The authors optimize for a weighted minimum bit error rate through the optimization of the UAV flying route, IRS phase steer, and IRS scheduling. The result is obtained by using relaxation-based method, and it shows the advantages of the IRS and the UAV flying route optimization in that there is a notable enhancement in the fairness of the system. The results also show that the UAV moves nearer to the IRS and hovers at the location longer when the total flying time is increased to benefit the reflecting signal to and from the IRS. In [19] the authors worked on maximizing the system sumrate in the system that utilizes IRS to aid UAV in orthogonal frequency division multiple access communications. A joint-iterative optimization algorithm is employed to optimize the allocation of resources, IRS phase steer, and flying route of the UAV. The results demonstrated the increase in the achievable sumrate when the maneuverability of the UAV and the IRS phase steer are optimized. The system's total rate can also be significantly influenced by the number of reflectors in an IRS when the phase steer is optimized. In [20], the authors conducted a study on the attainable data rate of a communication system enabled by UAV with the assistance of an IRS. The findings of the study indicated that the inclusion of the IRS resulted in significant improvements in the quality of communication in networks enabled by UAV.

It is noted that our study is distinguished from previous studies on the UAV-IRS communication system [15]–[21] in that it combines energy harvesting functionality of SWIPT with IRS-aided UAV communications to increase the EE of the network system, and apply the DRL algorithm to address the non-convexity in the proposed optimization problem with low computational complexity. The remaining sections of this paper are structured as follows. Section III presents the system model and outlines the problem formulation. Section IV provides a comprehensive explanation of the proposed DRL method. *Section V discusses the computational complexity analysis.* Simulation results are discussed in Section VI, and the paper concludes in Section VII.

## III. System Model and Problem Formulation

### A. System Model

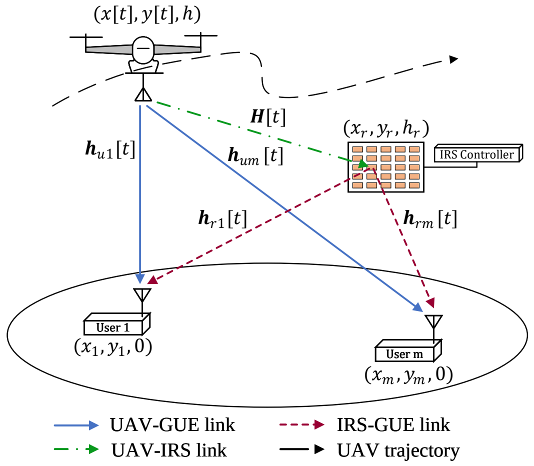

Fig. 1. System model of the IRS-aided UAV communication network with energy transfer

We consider a single-antenna aerial-BS UAV serving a set of $m = \{1, 2, ..., M\}$ users, providing downlink communications within a considered area, as shown in Fig. 1. A GUE receives information and energy at the same instant due to the embedded SWIPT technology. In our assumption, we utilize a 3D Cartesian coordinate such that the location of the GUE $m$ is fixed at $q_m = [x_m, y_m, 0]^T$. The initial flying location of the UAV is $q_{ui} = [x_0, y_0, h_0]^T$. We deploy one single IRS with multiple steerable elements which it is mounted fixed at $q_r = [x_r, y_r, h_r]^T$. The IRS consists of $N_r \times N_c$ passive reflecting element (PRE), which are consistently arranged as a uniform planar array (UPA), with $N_r$ and $N_c$ be the number of IRS unit in row and column, respectively. The UPA is structured such that each column contains PREs that are equidistant from each other, with a separation of $s_c$ meters. Similarly, the UPA is composed of PREs arranged in rows that are equidistant, with a spacing of $s_r$ meters. The UPA allows each PRE to independently re-scatter the incoming signal, and this process is characterized by a reflection coefficient comprising an amplitude $a$ ranging from 0 to 1 and a phase steer $\theta_{n_r,n_c} \in [0, 2\pi]$, i.e. $r_{n_r,n_c} = ae^{j(\theta_{n_r,n_c})}$, $\forall n_r \in \{1, 2, \dots, N_r\}$, and $\forall n_c \in \{1, 2, \dots, N_c\}$. In this paper we use fixed $a = 1$, and phase steer $\theta_{n_r,n_c}$ can be modified by the IRS decision maker.

### B. Channel Model for UAV data delivery with IRS

The UAV is dispatched to provide services to all GUEs. In contrast to terrestrial communication systems, where Rayleigh fading is commonly employed for small-scale fading, Rician fading is deemed more suitable for UAV-ground communications. This choice is justified by the typically prevalent Line-of-Sight (LoS) channel component and the occurrence of local scattering in UAV-ground communication scenarios. Thus, we utilize the channel model of Rician fading for both the UAV-GUE link and the IRS-GUE link. By considering the substantial path loss and reflection loss, we assume that signals reflected by the IRS two or more times have minimal power and are consequently disregarded. The channel between the UAV and the GUE $m$ can be described as

$$\mathbf{h}_{um}[t] = \sqrt{\frac{\beta_0}{d_{um}^{\alpha_{um}}[t]}} \left( \sqrt{\frac{\kappa}{\kappa + 1}} + \sqrt{\frac{1}{1 + \kappa}} \tilde{\mathbf{h}}_{um}[t] \right), \quad (1)$$

here, $\alpha_{um}$ is the path loss factor specifically associated with the link of the UAV to the GUE $m$. The Rician factor is denoted as $\kappa$, and $\tilde{\mathbf{h}}_{um}[t] \sim \mathcal{CN}(0, 1)$ represents the scattering element of GUE $m$ during time slot $t$, following a complex circularly symmetric Gaussian distribution with zero mean and unit variance. For ease of design, we presume that LoS propagation dominates the link of the UAV, IRS, and GUE. This is because the UAV operates at a high altitude, allowing the UAV-IRS link to typically occur in unobstructed field of view, reducing the impact of blockage. Consequently, we utilize an appropriate channel model of the UAV-IRS link to achieve the desired system performance and design [15]. At time $t$, the channel representing the communication of the UAV-IRS link is expressed as

$$\mathbf{H}[t] = \sqrt{\frac{\beta_0}{d_{ur}^2[t]}} \tilde{\mathbf{H}}[t], \quad (2)$$

where $\beta_0$ is gain at $d_0 = 1$m reference distance, and $\tilde{\mathbf{H}}[t]$ is the LoS channel of the UAV-GUE link, given in (7). We denote $\theta_{ur}[t]$, $\zeta_{ur}[t]$ and $z$ as the angle-of-arrivals at the IRS, and the height of the UAV, respectively, where

$$\sin\theta_{ur}[t] = \frac{z - h_r}{d_{ur}[n]}, \tag{3}$$

$$\sin\zeta_{ur}[t] = \frac{x_r - x[t]}{\sqrt{(x_r - x[t])^2 + (y_r - y[t])^2}}, \tag{4}$$

$$\cos\zeta_{ur}[t] = \frac{y[n] - y_r}{\sqrt{(x_r - x[t])^2 + (y_r - y[t])^2}}. \tag{5}$$

At time slot $t$, the channel model from the IRS to the GUE $m$ is described as

$$\mathbf{h}_{rm}[t] = \sqrt{\frac{\beta_0}{d_{rm}^{\alpha_{rm}}[t]}}\left(\sqrt{\frac{\kappa}{\kappa+1}}\mathbf{h}_{rm}^{LoS}[t] + \sqrt{\frac{1}{1+\kappa}}\tilde{\mathbf{h}}_{rm}[t]\right), \tag{6}$$

where $\tilde{\mathbf{h}}_{rm}[t] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r N_c})$, $\mathbf{I}_{N_r N_c}$ is the covariance matrix, and $\mathbf{h}_{rm}^{LoS}[t]$ is given by (8). The reflector coefficient of the IRS at time slot $t$ is described by

$$\mathbf{\Theta}[t] = \mathrm{diag}(\theta[t]) \in \mathbb{C}^{N_r N_c \times N_r N_c}, \tag{9}$$

where $\theta[t] = [e^{j\theta_{1,1}}[t], ..., e^{j\theta_{n_r,n_c}}[t], ..., e^{j\theta_{N_r,N_c}[t]}]^T \in \mathbb{C}^{N_r N_c \times 1}$. The composite of the UAV-GUE channel can be formulated as

$$\mathbf{h}_m^H[t] \triangleq \mathbf{h}_{rm}^H[t]\mathbf{\Theta}[t]\mathbf{H}[t] + \mathbf{h}_{um}^H[t]. \tag{10}$$

Because of the scattering elements present in the Rician fading channels described in equations (1) and (6), the overall channel becomes probabilistic in nature. To ensure effective control over the IRS phase for coherent signal composition at the GUE and to facilitate UAV route planning, it is necessary to approximate and provide real-time channel information between the PRE, UAV, and the individual GUE. This information is essential to be shared with the central decision maker of both the IRS and the UAV during the entire flight operation.

### C. Energy Model for IRS-aided UAV data delivery

In our work, we consider the SWIPT-equipped GUE using PS technique. With the PS ratio $\rho_m$, the GUE can exploit energy and information signal at the same instant with the received SINR at the GUE $m$, which can be expressed as

$$SINR_m[t] = \frac{\rho_m p_m[t]|\mathbf{h}_m^H[t]|^2}{\rho_m \sum_{m' \neq m} p_{m'}[t]|\mathbf{h}_{m'}^H[t]|^2 + \sigma_m^2}, \tag{11}$$

where $p_m$ and $\sigma_m$ are the received power and noise, respectively, at the GUE $m$. The energy dissipation of the GUE $m$ can be described as

$$\begin{aligned} ED_m[t] =& P_c + p_m - (1-\rho_m)\big(p_m[t]|\mathbf{h}_m^H[t]|^2 \\ &+ \sum_{m' \neq m} p_{m'}[t]|\mathbf{h}_{m'}^H[t]|^2\big), \end{aligned} \tag{12}$$

where $P_c$ is the circuit power consumption and $(1-\rho_m)p_m[t]|\mathbf{h}_m^H[t]|^2$ is the energy signal received at the GUE $m$. Therefore, we can express the data rate received at the GUE $m$ as

$$R_m[t] = \log(1 + SINR_m[t]). \tag{13}$$

### D. Problem Formulation

From (11) and (12), we define the EE at the GUE $m$ as

$$EE_m(q_u, \rho_m, p_m, \mathbf{\Theta})[t] = \frac{R_m(q_u, \rho_m, p_m, \mathbf{\Theta})[t]}{ED_m(q_u, \rho_m, p_m, \mathbf{\Theta})[t]}. \tag{14}$$

Accordingly, we formulate a problem that simultaneously optimizes the route planning of the UAV, PS ratio, transmission power, and IRS phase steer to maximize the average EE which can be described as

$$\max_{q_u, \rho_m, p_m, \mathbf{\Theta}} \frac{1}{M}\sum_{m=1}^{M} EE_m(q_u, \rho_m, p_m, \mathbf{\Theta})[t] \tag{15}$$

$$\text{s.t.} \quad C1: p_m[t] \leq p_{max}, \tag{16}$$

$$C2: 0 \leq \Theta[t] \leq 2\pi, \tag{17}$$

$$C3: 0 < \rho_m[t] \leq 1, \tag{18}$$

$$C4: \|q_u[t+1] - q_u[t]\| \leq D_{max}, \ \forall t, \tag{19}$$

where constraints $C1, C2, C3$ indicate that the power cannot exceed the maximum power, phase steer is within the range of $[0, 2\pi]$, and the PS ratio is in range of $[0, 1]$, respectively. In addition, constraint $C4$ defines the displacement of the UAV from one location at time slot $t$ to the next location at time slot $t+1$. None of the constraints is convex with respect to the controlling parameters, leading to the non-convex problem. *While conventional approaches could solve the optimization problem (15), the learning-based approach might be a more appropriate choice. In our work, we handle high-dimensional control parameters such as the route planning of the UAV, PS ratio, transmission power allocation, and IRS phase steer in dynamic environment. This justifies the use of DRL with reduced computational complexity.*

$$\begin{aligned} \tilde{\mathbf{H}}[t] =& \Big[1, e^{-j2\pi d \frac{\sin\theta_{ur}[t]\cos\zeta^{ur}[t]}{\lambda}}, ..., e^{-j2\pi d(N_r-1)\frac{\sin\theta_{ur}[t]\cos\zeta^{ur}[t]}{\lambda}}\Big]^T \\ &\times \Big[1, e^{-j2\pi d \frac{\sin\theta_{ur}[t]\sin\zeta^{ur}[t]}{\lambda}}, ..., e^{-j2\pi d(N_c-1)\frac{\sin\theta_{ur}[t]\sin\zeta^{ur}[t]}{\lambda}}\Big]^T \end{aligned} \tag{7}$$

$$\begin{aligned} \mathbf{h}_{rm}^{LoS} =& \Big[1, e^{-j2\pi d \frac{\sin\theta_{rm}[t]\cos\zeta^{rm}[t]}{\lambda}}, ..., e^{-j2\pi d(N_r-1)\frac{\sin\theta_{rm}[t]\cos\zeta^{rm}[t]}{\lambda}}\Big]^T \\ &\times \Big[1, e^{-j2\pi d \frac{\sin\theta_{rm}[t]\sin\zeta^{rm}[t]}{\lambda}}, ..., e^{-j2\pi d(N_c-1)\frac{\sin\theta_{rm}[t]\sin\zeta^{rm}[t]}{\lambda}}\Big]^T \end{aligned} \tag{8}$$

## IV. Proposed Algorithm

The non-convex nature of problem in (15) arises from the interference term and the dynamic positioning of the UAV in relation to both ground users and IRS. To address this problem, this paper employs the *Deep Q-learning (DQL) framework (a branch of DRL) as the core framework in this study due to its exceptional ability to handle complex, non-convex optimization challenges common in dynamic and incomplete wireless network environments [22]. Specifically, our work focuses on maximizing EE within a UAV/IRS/WPT scenario, where the problem involves dynamic and high-dimensional parameters such as UAV route planning, IRS phase steer, transmission power allocation, and PS ratio. As the UAV navigates a largely unknown and constantly changing environment to serve GUEs, DQL's deep neural network-based function approximation [23] proves essential in managing the high-dimensional state and action spaces. Additionally, the SINR map concept is used in the reward function to enhance the training of the DQL.*

### A. Markov Decision Process Formulation

Generally, the concept of RL aims to learn optimal decision-making policy in an environment using trial-and-error strategy. Prior to turning to the RL algorithm, the formulated problem is regarded as a Markov Decision Process. The central decision maker, also known as RL agent, makes sequential decisions which affects the observed state (the route planning of the UAV, PS ratio, transmission power, IRS phase steer, channel gain of $m$ GUE and current timeslot in our work) at the next time slot. Thus, the IRS-aided UAV data delivery problem can be formulated as a Markov Decision Process.

### B. Proposed DRL based IRS-aided UAV data delivery

In this subsection, we introduce a DRL-based algorithm to optimize the route planning of the UAV, PS ratio, transmission power, and IRS phase steer. We will introduce RL, and describe the proposed DRL and its design details afterward.

*It is noted that among many RL techniques, Q-learning is suitable for our study because it is well-suited for discrete action spaces and balances exploration and exploitation, making it adequate for optimizing various discretized control parameters including the UAV route planning, IRS phase steer, UAV transmission power, and PS ratio of SWIPT. Additionally, DQL is chosen over traditional Q-learning due to its ability to effectively manage high-dimensional state and action spaces using deep neural networks, which is essential for our complex optimization problem.* In Q-learning, the agent gets an optimal policy through interactions with the environment and earns rewards for its actions. The goal is to learn a policy that maximizes the discounted sum of rewards. Let denote $Q^{\pi}(s,a)$ as the action-value function, also known as the Q-value function. It details the expected discounted total rewards obtained by the agent when performing action $a$ in state $s_t$, given the policy $\pi$. The objective of the agent is to have the ability to select the best action value $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$ and choose the best policy. The approximated Q-values are stored in the Q-table, which associates states with corresponding actions.

At each time slot $t$, The Q-value is computed based on the present state and the action chosen in the previous step. The recorded value is kept in a Q-function, which plays a crucial role in determining the policy $\Pi$. The intent of the DRL model is to empower the agent with the ability to make the choices optimally, ensuring the maximization of cumulative rewards in the long run.

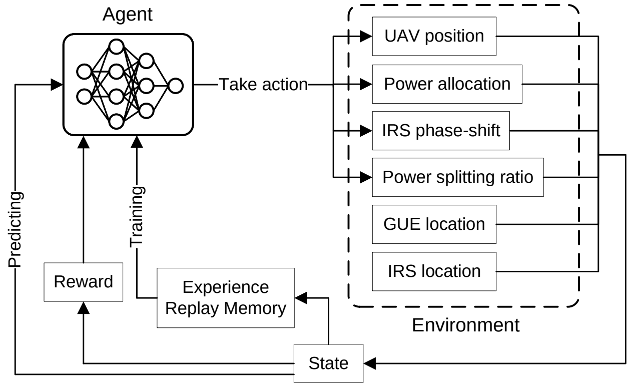

Fig. 2. Proposed DRL model

During each time slot $t$, the Q-value and Q-function undergo continuous updates to incorporate new information and improve the agent's decision-making capabilities by considering the present state, previously executed actions, and the obtained reward. This update process is performed using the equation as follows

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \eta[r_t + \gamma \max_{a} Q_t(s_{t+1}, a)] \quad (20)$$

where $\eta$ and $\gamma$ are the step size and discount rate, respectively. In equation (20), the reward $r_t$ is acquired from $r := S \times A \rightarrow R$, where $\mathbb{E}\{r_t|(s,a,s^{'}) = (s_t, a_t, s_{t+1})\} = R^{s^{'}}_{s,a}$. From iterative updating of the equation, the optimal value function can be acquired as follows

$$Q^*(s,a) = \mathbb{E}_{s^{'}}[R + \gamma \max_{a^{'}} Q^*(s^{'} a^{'})|s,a]. \quad (21)$$

In the DRL based model, we make the assumption that a single decision maker, acting as an agent, governs both the IRS and the UAV. The agent receives information on state $s_t$ from the state space $\mathbf{S}$ during time slot $t$, which includes the positions of the UAV and all GUEs, and the IRS phase steer. The agent acquires the current state and make choice according to the decision policy $\Pi$, then selects one action $a_t$ from a set of possible action $\mathbf{A}$ that includes the direction of movement for the UAV, PS ratio, transmission power, and IRS phase steer. Following the agent's action, it receives a reward or penalty $r_t$ determined by the average EE of the UAV. We explain the detailed definition of the states, actions, and reward function in our work, as follows.

*1) States:* The state space of the proposed DRL model is defined by

$$s^t = \{q_u[t], \rho[t], p[t], \mathbf{\Theta}[t], \mathbf{h}^H_m[t], t\} \quad (22)$$

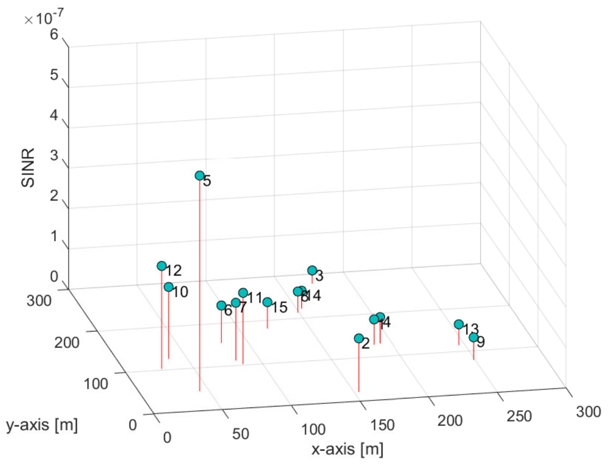

Fig. 3. SINR levels of the UAV (located at $[0, 0, z]^T$) to given 15 GUEs

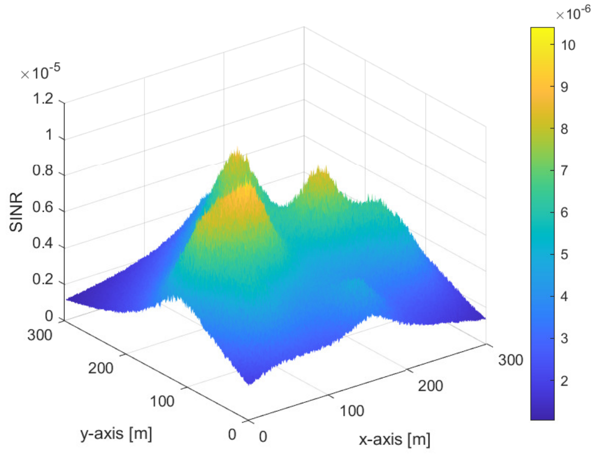

Fig. 4. Average SINR values of the UAV over given 15 GUEs

where $q_u[t] = [x[t], y[t], h]^T$ is the UAV two-dimensional (2D) coordinate, $\rho[t] \in [0, 1]$ is the PS ratio, $p[t]$ is the transmission power, $\boldsymbol{\Theta}[t] \in \mathbb{C}^{N_rN_c \times N_rN_c}$ is the IRS phase steer at time slot $t$.

*2) Actions:* The action space of the proposed DRL model is designed as

$$a^t = \{\Delta q_u[t], \Delta\rho[t], \Delta p[t], \Delta\boldsymbol{\Theta}[t]\}, \tag{23}$$

where $\Delta q_u[t] \in \{(-\delta x, 0), (\delta x, 0), (0, -\delta y), (0, \delta y), (-\delta xy, 0), (\delta xy, 0), (-\delta yx, 0), (\delta yx, 0), (0, 0)\}$ means the moving directions of the UAV, $\Delta\rho[t] \in \{0, ..., 1\}$, $\Delta p[t] \in \{0, ..., p_{max}\}$, $\Delta\boldsymbol{\Theta}[t] \in \{0, ..., 2\pi\}$.

*3) Reward:* Once an action is taken, the environment gives feedback as a reward or penalty to the agent, which indicates how good the taken actions are concerning the designed objective. Since the DRL in our work controls many parameters, we facilitate the design by introducing a reward function based on *the SINR map which consists of necessary information about the communication quality between the UAV and GUEs in the given area. To create the SINR map,* we first consider a UAV located at the specific location, and calculate the SINR levels between the UAV and the GUEs under the given network topology by using (11). Fig. 3 shows the SINR distribution assuming that the UAV is located at $[0, 0, z]^T$ and 15 GUEs are deployed over the given area. These SINR values are averaged over the number of GUEs, which denotes the expected SINR value of the UAV under the given network topology. We repeat the above procedure varying the location of the UAV and obtain the *SINR map* which indicates the average SINR of the UAV over all GUEs under the given network topology. Fig. 4 shows the map of the average SINR of the UAV under the same environment in Fig. 3. Based on this basis, we will calculate the expected EE that will be used in the reward function in the following.

Using the concept of the *SINR map*, we build a reward function, which indicates the average EE of the UAV under given network topology. The reward function is designed as

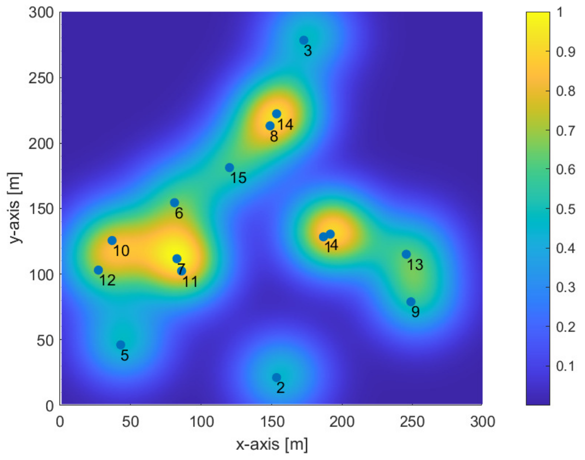

Fig. 5. Reward values of the UAV in 2D

$$r(\tilde{q}_u, q_u, \rho_m, p_m, \boldsymbol{\Theta})[t] = \frac{1}{M}\sum_{m=1}^{M} A_m f_{\tilde{q}_m}(\tilde{q}_u) EE_m(q_u, \rho_m, p_m, \boldsymbol{\Theta})[t], \tag{24}$$

where $f$ is the multivariate (bivariate in our work) normal distribution, $\tilde{q}_m$ and $\tilde{q}_u$ denote the projections of $q_m$ and $q_u$ over $x$-$y$ plane, respectively, and $A_m$ is the normalization weight. It is noted that the function $f$ is multiplied by $EE_m$ to obtain a continuous and differentiable reward function while preserving the characteristics of the SINR distribution of multi-modal Gaussian distribution. $f$ is given by

$$f_{\vec{\psi}}(\vec{a}) = \frac{\exp\left(-\frac{1}{2}(\vec{a}-\vec{\psi})^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\vec{a}-\vec{\psi})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Delta}|}}, \tag{25}$$

where $\vec{a}$ is a column vector, $\vec{\mu}$ is the mean vector, $k$ is the dimension of the function, $\boldsymbol{\Delta}$ is the covariance matrix and $|\boldsymbol{\Delta}| \equiv \det \boldsymbol{\Delta}$ is the determinant of the $\boldsymbol{\Delta}$. The procedure of obtaining the proposed reward function is summarized into two steps as follows.

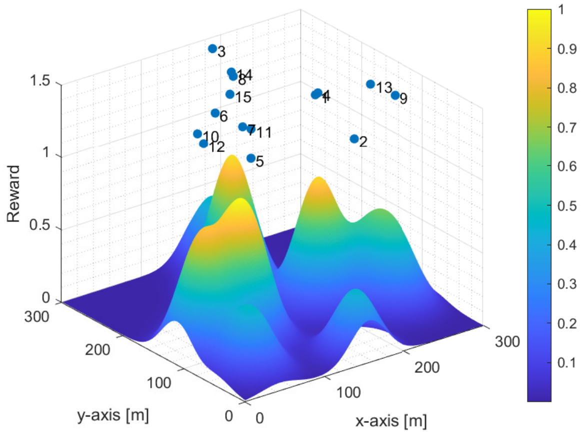

Fig. 6. Reward values of the UAV in 3D

- Step 1: We assume a finite number of quantized grid on the considered area. Next, we sample the SINRs from the given GUEs location when the UAV is located at a fixed position. We repeat this procedure until we obtain all the SINR samples changing the UAV's locations. After that, we can construct the SINR map by averaging the collected SINR samples.
- Step 2: However, the SINR map shows considerable fluctuation due to the nature of the SINR, making the map inefficient when implementing with DRL. To address this issue, we applied the multivariate function to the SINR map, and obtained the continuous smooth reward function which can be implemented with DRL. *First, we calculate the average EE obtained via the SINR map using (14). $f$ is the bivariate normal distribution, having peaks at all of the GUE locations. To obtain the reward function, $f$ is multiplied by the average EE with the normalization weight $A_m$. Because $f$ is inherently smooth, we can construct the smooth and continuous reward function $r$. We can verify that the reward in Fig. 6 is continuous which is different from the SINR value in Fig. 4 (discrete and fluctuated).*

*It is noted that the SINR map is unique to each specific set of generated GUE locations, meaning that any change in the GUE locations necessitates retraining the DRL model to reflect the new scenario accurately. This requirement stems from the fact that the model's learned parameters are highly dependent on the spatial distribution of the GUEs, and thus, a different topology of GUE locations alters the underlying problem dynamics. Therefore, for each unique set of GUE locations, one SINR map is computed, and a corresponding reward is determined based on this map.*

Figs. 5 and 6 show the average EE of the UAV in 2D and 3D, respectively, under the same environment in Fig. 4. It is noticeable that the EE in Fig. 6 is a surface, and the product of $EE_m$ and the bivariate function gives the EE a smoother contour shape as opposed to the average SINR value in Fig. 4. In addition, the EE of the UAV is designed in such a way that the highest EE region provides the highest reward, which induces the UAV to fly over or hover around this region with high probability.

Consequently, the penalty of the proposed DRL algorithm is defined as

$$a_t^{'} = \begin{cases} r_t & \text{satisfies} \quad C1, C2, C3, C4, \\ 0 & \text{otherwise}. \end{cases} \tag{26}$$

*4) DNN design:* In this subsection, we discuss in detail the design of the DNN in our proposed algorithm. The DNN is designed with fully connected layer and zero bias. The DNN architecture comprises three distinct layers: an initial layer for input, intermediate layers or hidden layers for processing, and a final layer for generating output. For this model, the input layer consists of the UAV coordinates, GUE coordinates, transmission power, PS ratio, IRS phase steer, and composite channel gain. The neural networks with all the samples are trained using minibatch and the average output of the network is computed. The utilization of a DNN allows us to approximate the Q-value, and thus, the output of the neural network directly corresponds to the Q-value. Within our neural network architecture, we incorporate $L$ hidden layers, with the number of nodes in each hidden layer matching that of the layer at the output. The rectified linear unit (ReLU) activation function is utilized to activate every layer of the DNN. To calculate the Q-value, denoted as $Q(s, a, \phi)$, we consider the neural network weights represented by $\phi$. These weights are updated using the backpropagation algorithm, which involves computing the partial derivative of the loss function concerning to the network weights. By iteratively adjusting the weights through backpropagation, we aim to optimize the Q-value estimation process. The objective of the DNN is to reduce the value of the loss function, which can be expressed as follows

$$L(\phi) = \mathbb{E}[(y - Q(s_t, a_t, \phi))^2], \tag{27}$$

where $y = r_t + \gamma \max_{a \in A} Q_{prev}(s_t, a_t, \phi)$. During the training phase of the DNN, the parameter $\phi$ is updated using a technique called experience replay. This involves randomly selecting a minibatch, denoted as $\hat{D}$, from the experience replay memory $D$. The selected minibatch is then utilized as the input data for updating the parameter $\phi$ of the neural network. This approach allows for efficient and effective utilization of past experiences to improve the training process of the DNN.

Error gradient is obtained by chain rule, which is given by

$$\nabla_\phi L \approx \frac{1}{|\hat{D}|} \sum 2(y - Q(s_t, a_t, \phi)) \nabla_\phi Q(s_t, a_t, \phi). \tag{28}$$

At each iteration, the agent modifies its decision-making strategy based on the current estimate of the Q-value. The agent employs an $\epsilon$-greedy policy to choose an action from the action space. This policy is defined as follows

$$a_t^{'} = \begin{cases} \text{argmax}_{a \in A} Q(s_t, a_t, \phi) & \text{with probability } 1 - \epsilon \\ \text{random action} & \text{with probability } \epsilon. \end{cases} \tag{29}$$

## V. Computational Complexity Analysis

*In this section, we analyze and evaluate the computational complexity of the algorithms used in this paper. For the*

*REINFORCE algorithm, the computational complexity can be calculated as $O(\frac{1}{\epsilon_r{}^4})$ [24], where $\epsilon_r$ is the error threshold. The successive hover and fly (SHF) algorithm's complexity is influenced by the number of variables and the computation of the objective function. For a convex subproblem with $\kappa = q_e + \rho_e + p_e + \theta_e$ variables, the ellipsoid method can be applied using iterative techniques like gradient descent with barrier, which has a computational complexity denoted by $O\big(\kappa^2 \times \log\left(\frac{1}{\epsilon_e}\right)\big)$. Furthermore, the exhaustive 2-Dimensional search of the SHF algorithm involves $O(\xi^2)$ possible candidates, with $\xi$ representing the quantized levels along the x and y axes. Incorporating the ellipsoid update [25] and addressing (15) as discussed in [26], the complexity $O(\kappa^2)$ and $O(\kappa \times \xi)$ are considered in each iteration. Consequently, the overall computational complexity of the SHF algorithm is represented by*

$$
\begin{aligned}
&O(\xi^2) \times O\left(\kappa^2 \log\left(\frac{1}{\epsilon_e}\right)\right) \times \left(O(\kappa^2) + O(\kappa\xi)\right) \\
&\qquad = O\left(\kappa^3 \times \xi^2 \times \log\left(\frac{1}{\epsilon_e}\right) \times \max\{\kappa, \xi\}\right). \qquad (30)
\end{aligned}
$$

*The computational complexity of DRL is mainly calculated based on the DNN. The DNN used in the proposed DRL consists of 4 layers. We set the number node for input and output layer as equal to the number of states and the number of actions, respectively. With the number of control variables $N_{con}$, the number of inputs can be defined as $(N_{con}+2) \times M$. With fully connected layers and zero biased, we assume the algorithm converges at $E$ episodes with $I$ iterations. The computational complexity for feedforward network can be computed as $O(L \times N_{act} \times M)$ where $L$ and $N_{act}$ are the number of layers and actions, respectively, and $N_{act} = q_e \times \rho_e \times p_e \times \theta_e$. Therefore, the computational complexity of the proposed DRL can be calculated as*

$$
O(E \times I \times L \times N_{act} \times (N_{con}+2) \times N_{IRS} \times M^2), \qquad (31)
$$

*where $N_{IRS}$ is the number of IRS element. By setting the value of $\epsilon_r = \epsilon_e = 10^{-4}, E = 100, I = 100, L = 4, q_e = 10, \rho_e = 10, p_e = 10, \theta_e = 10, \xi = 10^5, M = 15, N_{IRS} = 80$, we can compute the computational complexity and the number of operation, which are summarized as in Table I.*

TABLE I
COMPUTATIONAL COMPLEXITY COMPARISON.

| Algorithm | Computational Complexity | Operations |
|---|---|---|
| DRL | $O(EILN_{act}(N_{con}+2)N_{IRS}M^2)$ | $3.49 \times 10^{12}$ |
| REINFORCE | $O(\frac{1}{\epsilon_r{}^4})$ | $10^{16}$ |
| SHF | $O\left(\kappa^3\xi^2 \log\left(\frac{1}{\epsilon_e}\right) \times \max\{\kappa, \xi\}\right)$ | $5.46 \times 10^{20}$ |

## VI. Results and Discussion

In this section, we assess the performance of the proposed DRL with *SINR map*-based reward through comprehensive evaluations of various simulations. For comparison, we select the SHF scheme [27] with random IRS phase steer, DRL without IRS and REINFORCE [28]. REINFORCE is derived as a Monte-Carlo policy gradient learning algorithm, which trains the agent to generate a stochastic policy. Due to the challenge of balancing exploration and exploitation during training, REINFORCE often converges to suboptimal solution. The Table II provides a summary of the parameters employed in the simulation.

TABLE II
SIMULATION PARAMETERS.

| Parameter | Value |
|---|---|
| Coverage area | 300m×300m |
| Number of the GUEs | 15 |
| Number of reflecting units | {10,20,...,80} |
| Velocity of the UAV | 5m/s |
| UAV flying height | 100m |
| IRS's height | 30m |
| transmission power | {10,12,14,...,28} [dBm] |
| Energy transfer efficiency | 50% |
| Path loss exponent (NLoS) | 3.6 |
| Path loss exponent (LoS) | 2.2 |
| Rician factor | 2 |
| Discount factor | 0.8 |

In this simulation, we set the number of IRS element to 80 units. In Fig. 7 we compare the average EE under various $p_{max}$ values. This figure shows that EE performance of the proposed algorithm is higher than that of the SHF. Additionally, it shows the performance improvement of the IRS in the IRS-aided UAV communication scenario. We simulate SHF with various PS ratio $\rho = 1, \rho = 0.5, \rho << 1$, where $\rho = 1$ means that all received signals are used for information decoding. It is noticed that the SHF with $\rho << 1$ yields a very low average EE compared to that of the proposed DRL. The REINFORCE obtains slightly lower EE than SHF with $\rho = 1$ and significantly lower EE than the proposed DRL.

The average data rate versus the maximum transmission power is illustrated in Fig. 8. We can verify that the data rate of the proposed algorithm outperforms the SHF with $\rho = 1$. Moreover, the proposed algorithm achieves higher data rate than the SHF with $\rho = 0.5$ and exhibits high data rate compared to the SHF with $\rho << 1$.

We further illustrate the average energy used by the GUE in Fig. 9. Overall, the trend of the average energy consumption of the GUE becomes saturated around $p_{max} = 22$dBm. We can see that the SHF with $\rho = 1$ consumes the highest energy. The proposed algorithm consumes notably lower energy than the SHF with $\rho = 1$, and slightly lower energy than both REINFORCE and the DRL without IRS. We can notice that the SHF with $\rho = 0.5, \rho << 1$ uses low energy for their communications. This is because the received signal is mostly used for energy harvesting, which results in very low data rate and the degrades in the EE. It is noted that even though the proposed algorithm consumes slightly more energy than that of the SHF with $\rho = 0.5, \rho << 1$, it achieves higher EE.

In the following figures, we analyze the performances varying the number of reflecting units. In Fig. 10 the average EE increases as the number of reflecting unit goes up. We can clearly see the IRS performance gain of the proposed algorithm compared to the DRL without IRS method.

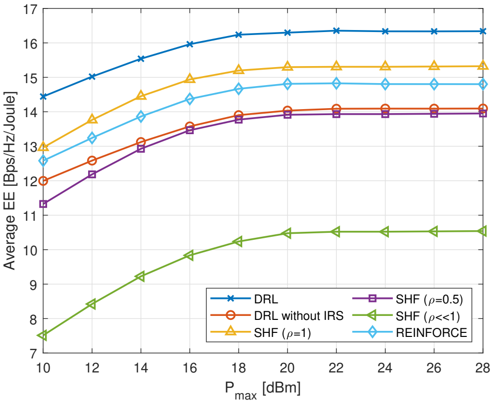

Fig. 7. Average EE vs. $p_{max}$

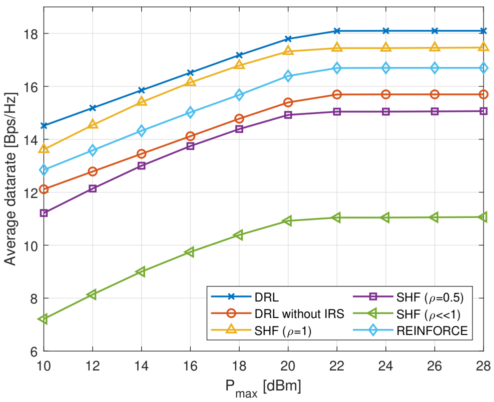

Fig. 8. Average data rate vs. $p_{max}$

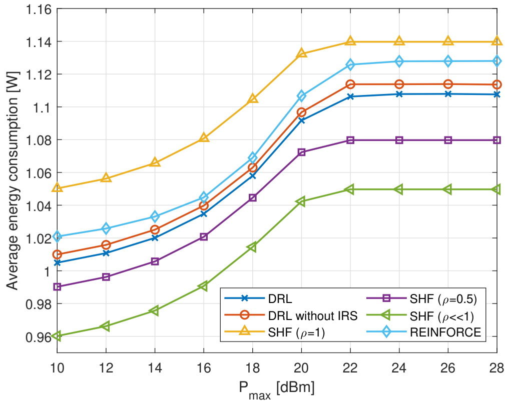

Fig. 9. Average energy consumption of the GUE vs. $p_{max}$

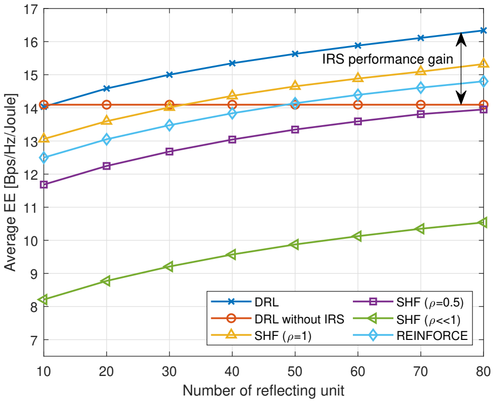

Fig. 10. Average EE vs. the number of reflecting unit

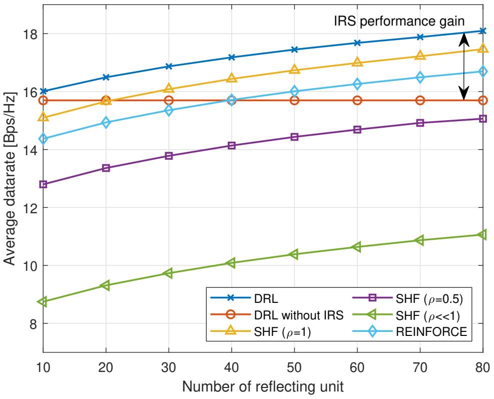

Fig. 11. Average data rate vs. the number of reflecting unit

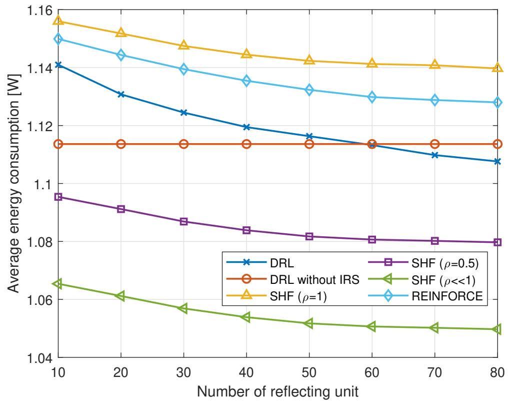

Fig. 12. Average energy consumption of the GUE vs. the number of reflecting unit

Fig. 11 shows the data rate achieved as we increase the number of reflecting units. It is verified that the proposed method achieves a higher data rate compared to that of the SHF with $\rho = 1$. Although REINFORCE achieves lower energy consumption than SHF with $\rho = 1$, its datarate is also lower than the SHF, which results in lower EE. Fig. 12 depicts the performance considering the average energy consumption versus the number of reflecting units. From the result in Fig. 12, we can verify that the proposed algorithm takes advantage of the reflecting units of the IRS to reduce the energy consumption. Furthermore, as the number of reflecting units increases to a certain amount, the performance gradually goes to convergent state, highlighting the limitation of IRS. Additionally, it is noteworthy highlighting that the achieved performance of the proposed algorithm with 10 units of IRS is approximately comparable to that of the DRL without any IRS. Therefore, it is important to employ more than 10 IRS units to reap the advantage of the IRS.

## VII. Conclusion

In this paper, we investigate IRS-aided UAV data delivery with energy transfer using a DRL-based method. We formulate the optimization problem as a maximization problem for the average EE. To address this problem, we employ the DRL approach and proposed the reward function based on the EE of the GUEs to optimize flying route of the UAV, the IRS phase steer, the transmission power of the UAV and PS ratio at the same instant. The simulation results demonstrate the potential gain of the proposed algorithm with and without IRS.

For future work, we will investigate energy-efficient UAV 3D route planning in a multiple-UAV environment using multi-agent DRL, where a detailed energy consumption model of the UAV is considered. *Additionally, we will focus on implementing the transformer model to address the need for retraining the network when there are changes in network configuration, such as changes in users' locations and the number of users or UAVs.*

## References

[1] A. Fotouhi et al., "Survey on UAV Cellular Communications: Practical Aspects, Standardization Advancements, Regulation, and Security Challenges," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3417-3442, Fourthquarter 2019, doi: 10.1109/COMST.2019.2906228.

[2] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," Proc. IEEE, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.

[3] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," IEEE Commun. Mag., vol. 58, no. 1, pp. 106–112, Jan. 2020.

[4] Y. Liu et al., "Reconfigurable Intelligent Surfaces: Principles and Opportunities," in IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1546-1577, thirdquarter 2021, doi: 10.1109/COMST.2021.3077737.

[5] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang and L. Wang, "Power Control in UAV-Supported Ultra Dense Networks: Communications, Caching, and Energy Transfer," in IEEE Communications Magazine, vol. 56, no. 6, pp. 28-34, June 2018, doi: 10.1109/MCOM.2018.1700431.

[6] N. Zhao, S. Zhang, F. R. Yu, Y. Chen, A. Nallanathan and V. C. M. Leung, "Exploiting Interference for Energy Harvesting: A Survey, Research Issues, and Challenges," in IEEE Access, vol. 5, pp. 10403-10421, 2017, doi: 10.1109/ACCESS.2017.2705638.

[7] N. Cheng et al., "AI for UAV-assisted IOT Applications: A comprehensive review," IEEE Internet of Things Journal, vol. 10, no. 16, pp. 14438–14461, Aug. 2023. doi:10.1109/jiot.2023.3268316

[8] Q. Wu, S. Zhang, B. Zheng, C. You and R. Zhang, "Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial," in IEEE Transactions on Communications, vol. 69, no. 5, pp. 3313-3351, May 2021, doi: 10.1109/TCOMM.2021.3051897.

[9] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan and L. Hanzo, "Latency Minimization for Intelligent Reflecting Surface Aided Mobile Edge Computing," in IEEE Journal on Selected Areas in Communications, vol. 38, no. 11, pp. 2666-2682, Nov. 2020, doi: 10.1109/JSAC.2020.3007035.

[10] X. Hu, C. Zhong, Y. Zhang, X. Chen and Z. Zhang, "Location Information Aided Multiple Intelligent Reflecting Surface Systems," in IEEE Transactions on Communications, vol. 68, no. 12, pp. 7948-7962, Dec. 2020, doi: 10.1109/TCOMM.2020.3020577.

[11] J. He, H. Wymeersch, T. Sanguanpuak, O. Silvén, and M. Juntti, "Adaptive beamforming design for mmWave RIS-aided joint localization and communication," in Proc. IEEE WCNC, Apr. 2020, pp. 1–6.

[12] S. Zhang and R. Zhang, "Radio Map-Based 3D Path Planning for Cellular-Connected UAV," in IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1975-1989, March 2021, doi: 10.1109/TWC.2020.3037916.

[13] Z. Huang, C. Chen and M. Pan, "Multiobjective UAV Path Planning for Emergency Information Collection and Transmission," in IEEE Internet of Things Journal, vol. 7, no. 8, pp. 6993-7009, Aug. 2020, doi: 10.1109/JIOT.2020.2979521.

[14] Y. Guo, C. You, C. Yin and R. Zhang, "UAV flying route and Communication Co-design: Flexible Path Discretization and Path Compression," in IEEE Journal on Selected Areas in Communications, doi: 10.1109/JSAC.2021.3088690.

[15] Z. Wei et al., "Sum-Rate Maximization for IRS-aided UAV OFDMA Communication Systems," in IEEE Transactions on Wireless Communications, vol. 20, no. 4, pp. 2530-2550, April 2021, doi: 10.1109/TWC.2020.3042977.

[16] Y. Pan, K. Wang, C. Pan, H. Zhu and J. Wang, "UAV-Assisted and Intelligent Reflecting Surfaces-Supported Terahertz Communications," in IEEE Wireless Communications Letters, vol. 10, no. 6, pp. 1256-1260, June 2021, doi: 10.1109/LWC.2021.3063365.

[17] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li and A. L. Swindlehurst, "UAV-Assisted Intelligent Reflecting Surface Symbiotic Radio System," in IEEE Transactions on Wireless Communications, vol. 20, no. 9, pp. 5769-5785, Sept. 2021, doi: 10.1109/TWC.2021.3070014.

[18] Y. Cai, Z. Wei, S. Hu, D. W. K. Ng and J. Yuan, "Resource Allocation for Power-Efficient IRS-aided UAV Communications," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 2020, pp. 1-7, doi: 10.1109/ICCWorkshops49005.2020.9145224.

[19] Z. Mohamed and S. Aïssa, "Leveraging UAVs with Intelligent Reflecting Surfaces for Energy-Efficient Communications with Cell-Edge Users," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 2020, pp. 1-6, doi: 10.1109/ICCWorkshops49005.2020.9145273.

[20] S. Li, B. Duo, X. Yuan, Y. -C. Liang and M. Di Renzo, "Reconfigurable Intelligent Surface Assisted UAV Communication: Joint flying route Design and Passive Beamforming," in IEEE Wireless Communications Letters, vol. 9, no. 5, pp. 716-720, May 2020, doi: 10.1109/LWC.2020.2966705.

[21] Z. Li, W. Chen, H. Cao, H. Tang, K. Wang and J. Li, "Joint Communication and flying route Design for Intelligent Reflecting Surface Empowered UAV SWIPT Networks," in IEEE Transactions on Vehicular Technology, 2022, doi: 10.1109/TVT.2022.3196039.

[22] *M. S. Frikha, S. M. Gammar, A. Lahmadi, and L. Andrey, "Reinforcement and deep reinforcement learning for wireless internet of things: A survey," Computer Communications, vol. 178, pp. 98–113, Oct. 2021. doi:10.1016/j.comcom.2021.07.014*

[23] *K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26–38, Nov. 2017. doi:10.1109/msp.2017.2743240*

[24] *J. Zhang, J. Kim, B. O'Donoghue, and S. Boyd, "Sample efficient reinforcement learning with reinforce," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 10887–10895, May 2021. doi:10.1609/aaai.v35i12.17300*

[25] *S. Bubeck, Convex optimization: Algorithms and complexity, 2015. doi:10.1561/9781601988614*

[26] *R. Hunger, "Floating point operations in matrix-vector calculus," Munich Univ. of Technology, Inst. for Circuit Theory and Signal Processing, Munich, Germany, Tech. Rep., 2005.*

[27] J. Xu, Y. Zeng and R. Zhang, "UAV-Enabled Wireless Power Transfer: flying route Design and Energy Optimization," in IEEE Transactions on Wireless Communications, vol. 17, no. 8, pp. 5092-5106, Aug. 2018, doi: 10.1109/TWC.2018.2838134.

[28] P. S. Thomas and E. Brunskill, "Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines," CoRR, vol. abs/1706.06643, pp. 1–2, Jun. 2017. [Online]. Available: http://arxiv.org/abs/1706.06643