information

Article

Optimizing the Deployment of an Aerial Base Station and the Phase-Shift of a Ground Reconfigurable Intelligent Surface for Wireless Communication Systems Using Deep Reinforcement Learning

Wendenda Nathanael Kabore [1], Rong-Terng Juang [2,*], Hsin-Piao Lin [2], Belayneh Abebe Tesfaw [3] and Getaneh Berie Tarekegn [4]

[1] Department of Electronic Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; nooptanio2007@gmail.com
[2] Institute of Aerospace and System Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; hplin@ntut.edu.tw
[3] Department of Electrical Engineering and Computer Science, National Taipei University of Technology, Taipei 10608, Taiwan; keskes2007@gmail.com
[4] Department of Electrical and Computer Engineering, National Yang-Ming Chiao Tung University, Hsinchu 30010, Taiwan; gechb21@gmail.com
* Correspondence: rtjuang@mail.ntut.edu.tw

Abstract: In wireless networks, drone base stations (DBSs) offer significant benefits in terms of Quality of Service (QoS) improvement due to their line-of-sight (LoS) transmission capabilities and adaptability. However, LoS links can suffer degradation in complex propagation environments, especially in urban areas with dense structures like buildings. As a promising technology to enhance the wireless communication networks, reconfigurable intelligent surfaces (RIS) have emerged in various Internet of Things (IoT) applications by adjusting the amplitude and phase of reflected signals, thereby improving signal strength and network efficiency. This study aims to propose a novel approach to enhance communication coverage and throughput for mobile ground users by intelligently leveraging signal reflection from DBSs using ground-based RIS. We employ Deep Reinforcement Learning (DRL) to optimize both the DBS location and RIS phase-shifts. Numerical results demonstrate significant improvements in system performance, including communication quality and network throughput, validating the effectiveness of the proposed approach.

Keywords: reconfigurable intelligent surface; drone base station deployment; phase-shift optimization; deep reinforcement learning

Citation: Kabore, W.N.; Juang, R.-T.; Lin, H.-P.; Tesfaw, B.A.; Tarekegn, G.B. Optimizing the Deployment of an Aerial Base Station and the Phase-Shift of a Ground Reconfigurable Intelligent Surface for Wireless Communication Systems Using Deep Reinforcement Learning. *Information* **2024**, *15*, 386. https://doi.org/10.3390/info15070386

Academic Editors: Wade Trappe, Zahir M. Hussain and Yu Zhang

Received: 26 April 2024
Revised: 28 June 2024
Accepted: 29 June 2024
Published: 1 July 2024

1. Introduction

The primary determinant of wireless communication performance lies in the radio propagation environment. Techniques such as channel coding, modulation, transmission/reception diversity, and beamforming have been employed to mitigate multipath effects between transmitters and receivers [1]. The burgeoning demands for high-quality wireless services in future communication systems, particularly in scenarios such as natural disasters, traffic offloading, emergency communications, etc., present significant challenges for current telecommunication sectors [2]. To address these challenges, it is becoming increasingly obvious that drones, or unmanned aerial vehicles (UAVs), are capable of performing a variety of aerial wireless communication operations [2]. Compared to conventional fixed base stations (FBSs), drone base stations (DBSs) offer enhanced communication coverage and system throughput attributed to features such as agility, high probability of line-of-sight (LoS) connections, and fully controllable mobility.

Concurrently, reconfigurable intelligent surfaces (RISs) have emerged as a promising technology, driven primarily by the forthcoming 6G applications and future Internet of Things (IoTs) networks [3]. Acting as anomalous mirrors, RISs manipulate wireless propagation by reflecting impinging radio waves towards arbitrary angles, applying phase-shifts, and modifying polarization [4]. Recent studies [3–8] have demonstrated the effectiveness of RISs in improving channel quality and communication performance, thereby enabling smart and controllable wireless propagation environments. Due to their energy efficiency and programmability, signal, interference, security, scattering, and security engineering may be achieved through RISs in future wireless networks [3].

While DBSs have the potential to offer extensive network coverage, the line-of-sight (LoS) link between DBSs and ground mobile users may be obstructed in complex environments, resulting in reflected, diffracted, and scattered replicas of the transmitted signal. Consequently, the integration of DBSs and RISs presents a viable solution for achieving high-quality network connectivity. Recent research has explored two scenarios: (1) DBSs carrying RISs and acting as passive relays in both uplink and downlink communications, and (2) RISs mounted on building surfaces to assist DBSs. This paper focuses on the latter scenario. However, deriving low-complexity algorithms using conventional optimization methods poses challenges given the system's complexity and dynamic environment. Some early works studied the UAV trajectory and RIS phase shift design using conventional optimization-based methods like successive convex approximation (SCA) [9–11].

Fortunately, advancements in machine learning, particularly reinforcement learning (RL), offer successful approaches to addressing complicated control tasks in wireless communication networks [12,13]. Due to the numerous variables involved, as well as unpredictable system behavior, it is difficult to perform sequential decision-making mathematically in real-time. Against this problem, recent works have demonstrated that deep reinforcement learning (DRL) can effectively control UAV trajectories and RIS phase shifts to maximize system rewards [14–17].

Based on deep reinforcement learning (DRL), our paper investigates RIS-assisted DBS wireless communications (RIS-DWC), where a drone flies at variable altitudes, communicating with multiple mobile ground users. The ground-based RIS is utilized to establish line-of-sight links, compensating for non-line-of-sight links between the DBSs and users. Consequently, as RIS elements are phase-shifted appropriately to enhance propagation environment, the proposed method is expected to improve the propagation environment. Taking a close look at this paper, the following highlights its novelty and main contributions:

(1) Formulation of an optimization problem for RIS-DWC that jointly optimizes RIS reflection coefficients and DBS trajectory to maximize coverage and overall achievable data rates of all users in a target area.

(2) Consideration of DBS trajectory in 3D space and passive phase-shift of RIS with a uniform linear array design [18]. DRL is employed to address computational challenges inherent in traditional methods, tackling the joint trajectory-phase-shift problem while considering DBS and ground users' mobility scenarios. The proposed method can be easily applied to various systems.

(3) Provision of numerical results demonstrating the benefits of the RIS-DWC system in order to maximize the communication coverage and achieving downlink data rates.

The remainder of this paper is arranged as follows: Section 2 discusses related works. Section 3 introduces the system model and formulates an optimized problem to optimize both service coverage and total network data rates. Section 4 develops the DRL-based RIS-DWC framework. The proposed joint method of phase-shift and trajectory optimization is verified by simulation results shown in Section 5. Finally, Section 6 presents concluding remarks and future works.

2. Related Works

Numerous studies have delved into modeling air-to-ground channels and deployment strategies for drone base stations (DBSs) to enhance cellular network performance [19–22].

Alzenad et al. [23] and Chen et al. [24] investigated the 3-D positioning of DBSs to optimize user coverage under various Quality of Service (QoS) constraints.

The incorporation of RIS in UAV-assisted communication systems has become a significant focus of research. In reference [9], the authors investigated a downlink transmission setup consisting of a rotary-wing UAV, a ground user, and an Intelligent Reflecting Surface (IRS). They introduced a SCA algorithm to optimize the trajectory of the UAV and the passive beamforming of the RIS. Similarly, reference [25] explored the potential of RIS in enhancing UAV-assisted cellular networks, highlighting significant performance gains. While existing algorithms rely on convex optimization theory, their computational complexity may increase with the number of reflecting elements.

A model-free reinforcement learning approach [26] showed the promise in developing autonomous drone mobility control strategies. The Q-learning algorithm introduced fundamental concepts like environment, state, action, reward, and Q-value in DBS to optimize trajectory policies through interaction with the environment. However, Q-learning is constrained by discrete state and action spaces because of the finite size of the Q-table. To overcome this limitation, the authors of reference [27] suggested the deep Q-network (DQN) algorithm, which integrates Reinforcement Learning (RL) with deep neural networks (DNNs). DQN's ability to handle infinite state spaces makes it a powerful tool, although its action space remains discrete. Training a DQN differs from traditional neural network training in supervised learning, with weights updated online based on previous experiences. Despite its effectiveness, DQN may overestimate action values under certain conditions. To mitigate this issue, reference [28] introduced the double deep Q network (DDQN), which reduces overestimation by decomposing the max operation in the target. Subsequent works extended this approach, such as [29], which proposed a DDQN with a multi-step learning algorithm. In [30], researchers investigated RIS-assisted UAV secure communication systems in order to improve the average worst-case secrecy rate. Reference [31] suggested a decaying deep Q-network-based algorithm to both improve the UAV's trajectory, resource allocation, and energy consumption, and the RIS's passive beamforming. Similarly, reference [32] leveraged DDQN and Deep Deterministic Policy Gradient (DDPG) algorithms to optimize the UAV's three-dimensional (3D) trajectory and adjust the phase-shifts of the RIS, aiming to maximize UAV energy efficiency.

Recent works, by considering the state-of-the-art advances, have proposed the DRL model for UAV navigation [33] and radio resource management [34] in cellular networks.

Despite these advancements, most existing works overlook target area coverage and ground user movement. This paper addresses this gap by applying a novel DRL technique in an RIS-aided DBS system, accounting for ground user movement and coverage considerations.

3. System Model and Problem Formulation

3.1. System Model

Figure 1 depicts the scenario of the RIS-assisted DBS wireless communication system. In this setup, a single DBS functions as an aerial base station, catering to a group of ground users $\boldsymbol{U} = \{1, 2, \ldots, U\}$ across a designated target area T_a. It is assumed that the DBS possesses location awareness via GPS, while each user is mobile and randomly distributed within the target area. Alongside the DBS, a ground-based RIS is deployed to augment communication quality. Given the practical uncertainty of direct links between the DBS and users, the RIS facilitates direct connections between them. For simplicity, it is assumed that both the DBS and ground users are equipped with a single antenna. Furthermore, the DBS's 3D position in each time slot is defined by $\mathrm{L} = \{x^D, y^D, z^D\}$, where latitude and longitude coordinates are denoted as (x^D, y^D), and z^D represents the flight altitude. To guarantee robust wireless communication and achieve high data rates, the DBS adjusts its altitude based on user movement, i.e., $z^D \in \left(z^{min}, z^{max}\right)$. Each user's coordinates are represented as (x^U, y^U).

The RIS maintains a fixed location specified by (x^R, y^R, z^R), and it consists of an array of M reflective elements arranged in a uniform linear pattern, denoted as $\boldsymbol{M} = \{1, \ldots m, \ldots M\}$.

The RIS controller dynamically adjusts the phase-shift of each reflecting element within the range of [0, 2π), denoted as $\theta m \in [0, 2\pi)$, $\forall m \in M$. To aid reader comprehension, Table 1 provides a summary of the mathematical symbols utilized in this article.

Figure 1. The scenario of the RIS-assisted DBS-enabled wireless communications system.

Table 1. Main notations.

Notation	Detailed Definition
x^D, y^D, z^D	Location of the DBS
x^U, y^U	Location of the ground users
x^R, y^R, z^R	Location of the RIS
z^{min}, z^{max}	Minimal and maximal altitudes of the DBS
d^{min}, d^{max}	Minimal and maximal flying distances of the DBS
T_a	Target area
Θ_t	Phase-shift matrix of the RIS in timeslot t
$h_{u,t}^{DU}$	Channel gain of user–DBS link
h_t^{DR}	Channel gain of DBS-RIS link
h_u^{RU}	Channel gain of RIS–ground user link
$h_{u,t}^{DRU}$	Channel gain of ground user–RIS–DBS link
$d_{u,t}^{DU}$	Distance between the DBS and ground users
d_t^{DR}	Distance between the DBS and RIS
d_t^{RU}	Distance between the RIS and ground users
$p_{u,t}$	Blockage probability
P, σ^2	Transmission power, noise power
α	Path loss at the reference distance of 1 m
β	Path loss exponent for RIS–user link
$R_{u,t}$	Data rate of DBS–RIS–user link
C	The coverage score

3.1.1. Channel Model

In Figure 1, the DBS typically operates at high altitudes, while the RIS is commonly positioned on building surfaces. If the direct link between the DBS and users (DBS–ground user link) is obstructed, scattering occurs via the RIS. This paper considers the downlink channel to comprise a direct link (i.e., the DBS–user path) and multiple reflected links through the RIS (i.e., the DBS–RIS–user path).

Assuming the RIS always maintains a direct link with both the DBS and users, it ensures compensation for the DBS–user link through the DBS-RIS and RIS–user links. Although there is a possibility of the DBS-RIS link being closely blocked by ground obstacles, this pessimistic scenario will be considered in a future work.

Specifically, the direct link channel between the DBS and the u-th user at time slot t is denoted as $h_{u,t}^{DU}$ and can be expressed as

$$h_{u,t}^{DU} = \frac{\alpha}{\left(d_{u,t}^{DU}\right)^2} \tag{1}$$

where $d_{u,t}^{DU}$ is the distance from the DBS to the ground user. The channel of the indirect link between the DBS and the u-th user is denoted as $h_{u,t}^{DRU}$. The channel gain for the link between the DBS and the RIS is express as

$$h_t^{DR} = \sqrt{\frac{\alpha}{\left(d_t^{DR}\right)^2}}\left[1, e^{-j\frac{2\pi}{\lambda}d\varphi_t^{DR}}, \ldots, e^{-j\frac{2\pi}{\lambda}(M-1)d\varphi_t^{DR}}\right]^T \tag{2}$$

where α is defined as the path loss at the reference distances of 1 m, 2 is defined as the path loss exponent, $\varphi_t^{DR} = \frac{x^R - x_t^D}{d_t^{DR}}$ represents the cosine value of the angle of arrival (AoA), and d_t^{DR} is the distance between the DBS and RIS. The channel gain of the RIS–user link is denoted as

$$h_{u,t}^{RU} = \sqrt{\frac{\alpha}{\left(d_t^{RU}\right)^\beta}}\left[1, e^{-j\frac{2\pi}{\lambda}d\varphi_t^{RU}}, \ldots, e^{-j\frac{2\pi}{\lambda}(M-1)d\varphi_t^{RU}}\right]^T \tag{3}$$

where α is defined as the path loss at the reference distance of 1 m, β is the path loss exponent, $\varphi_t^{RU} = \frac{x^R - x_t^U}{d_{u,t}^{RU}}$ represents the cosine value of the angle of departure (AoD), and d_t^{RU} is the distance from the RIS to the user. The overall channel gain of the DBS–RIS–user link can be expressed as

$$h_{u,t}^{DRU} = \left(h_{u,t}^{RU}\right)^T \cdot \theta_t \cdot h_t^{DR} \tag{4}$$

where θ_t represents the diagonal phase matrix of the RIS in each timeslot t.

3.1.2. Phase-Shift Designs and Channel Rate

The RIS manipulates the signal from the DBS to the ground users by adjusting the phase-shifts of numerous scattering reflective elements. These reflected signals can be coherently combined to amplify the received signal or suppress interferences.

The reflection coefficient matrix associated with effective phase-shifts at the RIS can be represented as

$$\Theta_t = \{\theta_{1,t}, \theta_{2,t}, \theta_{m,t}, \ldots, \theta_{M,t}\}, \quad with \quad \theta_{m,t} = \eta_{m,t}e^{j\phi_{m,t}} \tag{5}$$

where $\theta_{m,t}$ represents the diagonal phase-shift of the reflecting elements, $\eta_{m,t}$ is the amplitude, and $\phi_{m,t} \in [0, 2\pi)$ defines the phase-shift. A set of discrete values is used for phase-shift $\phi_{m,t}$; because of hardware limitations [20], $\phi_{m,t} \in \{0, \Delta\varnothing, \ldots, \Delta\varnothing(L-1)\}$. $\Delta\varnothing = 2\pi/L$, where $L = 2^b$ indicates the number of phase-shift levels, and b is the number of bits. There are 2^b different angles for tuning phase-shifts on RIS elements.

The air-to-ground (A2G) channel model considers two propagation groups: line-of-sight (LoS) propagation and non-line-of-sight (NLoS) propagation. Unlike LoS, NLoS signals experience much stronger reflections and diffractions. To assess the likelihood of the direct DBS-to-ground user link being blocked, we utilize the A2G channel model in urban environments from [19], where the blocking probability is defined by

$$p_{u,t} = 1 - \frac{1}{1 + a\ \exp\left[-b(\omega - a)\right]} \tag{6}$$

where $\omega = arctan\left(z_t^{DBS}/d_{u,t}^{DU}\right)$ is the elevation angle from the DBS to the ground users. Additionally, a and b are constants whose values depend on the environment, such as rural or urban areas. By applying a threshold value, $p_{u,t}$ is designed as an expected value to determine if the user link is blocked or not.

From (1), (4), and (6), the SNR of the ground user u can be defined by

$$\gamma_{u,t}^{snr} = \frac{P\left|(1-p_{u,t})\, h_{u,t}^{DU} + p_{u,t} h_{u,t}^{DRU}\right|^2}{B\sigma^2} \tag{7}$$

where P is the transmitted power from the DBS to users, B represents the bandwidth, and σ^2 is the noise variance. The total achievable rate of user u in timeslot t can be rewritten as

$$R_{u,t} = B\log_2\left(1+\gamma_{u,t}^{snr}\right) \tag{8}$$

3.2. Problem Formulation

The aim of this research is to optimize the communication coverage and attainable data rates for all ground users. Let $\tau_{u,t} \in \{0,1\}$, indicating the connection between the DBS and the u-th user, either directly or via the RIS. When $\tau_{u,t} = 1$, it signifies that the user is served by the DBS either directly or via the RIS; otherwise, the user is not served by the DBS.

$$\tau_{u,t} \in \{0,1\},\ \forall u \in U,\ \forall u \in T_a,\ \forall t \tag{9}$$

Accordingly, the coverage scores of the RIS-assisted DBS wireless communication system can be expressed as

$$C_t = \left(\frac{\Pi_t}{U}\right) \times 100\%,\ \ \Pi_t = \sum_{u=1}^{U} \tau_{u,t},\ \tau_{u,t} \in 1,\ \forall t \tag{10}$$

where Π_t is the number of users connected to the deployed DBS (the total ground users with a reception SNR above a threshold) at each time slot t.

The proposed method aims to improve the throughput of each user by optimizing the DBS mobility and RIS phase-shifts. It follows a two-stage approach: the communication stage and the mobility control stage. In the communication stage, the DBS remains stationary while the RISs reflect the signal from the DBS to the ground users by adjusting the phase-shift of each element. However, if the data rate and area coverage fall below a threshold, most users' downlink channels become blocked. In such cases, the current communication stage ends, and the DBS is repositioned to establish line-of-sight (LoS) links, thereby offering improved services to all ground users. Thus, the problem can be mathematically formulated as

$$\begin{gathered} \binom{Argmax}{L,\Theta} \sum_{t=1}^{T}\sum_{u=1}^{U} R_{u,t}\cdot C_t \\ R_{u,t} \geq D_u \forall u \in U \\ 0 \leq x_t^D \leq x^{max} \\ 0 \leq y_t^D \leq y^{max} \\ z^{min} \leq z_t^D \leq z^{max} \\ 0 \leq v_u \leq V_{max} \text{ for all user } u \end{gathered} \tag{11}$$

where $R_{u,t} \geq D_u \forall$u $\in U$ denotes the required minimum rates constraints on each ground user; $0 \leq x_t^D \leq x^{max}, 0 \leq y_t^D \leq y^{max}, z^{min} \leq z_t^D \leq z^{max}$ indicate the DBS flying constrains; and $0 \leq v_u \leq V_{max}$ represents the user's speed limit.

4. Stage of RIS-DWC Approach

4.1. General Reinforcement Learning Framework

RL [35] is a dynamic learning approach capable of automatically adapting to changing environments, making it suitable for solving decision-making problems such as maximizing the throughput of a user in real time, as well as the resource allocation in aerial base station

networks [36]. RL learns from its interaction with the environment, generating data as it goes along. The proposed RIS-DWC process can be represented as a Markov decision process (MDP) and it can be tackled using DRL algorithms. MDP comprises four key components: the system state space **s**, the system action space **a**, the reward function **r**, and the state transition probabilities **p**.

The RL agent, at each timeslot, observes the system state s_t and takes action a_t according to a policy π. The agent receives a reward r_t and transitions to a new system state s_{t+1}. The goal of MDP is to find an optimal policy $\pi*$ that maximizes long-term reward. In RL, a policy maps observed states to actions to maximize accumulated reward. Additionally, the value function predicts future rewards, evaluating the desirability of actions in a given state. Q-learning, a renowned algorithm of RL, employs a Q-table or look-up table to store Q-values corresponding to each state-action pair (s, a) [35]. The agent, for a given state under policy $\pi*$, aims to enhance a long-term cumulative reward by selecting the optimal action. The discount factor $\gamma \in [0, 1]$ weighs future rewards: a low γ prioritizes immediate rewards, while a high γ focuses on long-term rewards. The agent, by consulting the look-up table, can make decisions based on any observed state to formulate an effective decision-making strategy. Nevertheless, traditional Q-learning encounters challenges, especially in large-scale environments, due to issues with dimensionality [36]. In such environments, discovering the optimal policy through Q-learning proves challenging because the RL agent may encounter difficulty exploring a vast number of state–action pairs, rendering the storage requirements for the Q-table impractical. To address these challenges, deep neural networks (DNNs) augment the conventional Q-learning approach with enhanced capabilities. The Q-learning algorithm encounters challenges in determining the optimal policy when the action and state spaces expand significantly. To address this issue, the DQN algorithm was introduced, which integrates the traditional Q-learning approach with a DNN [37]. Unlike Q-learning, DQN replaces the table with a function approximator, typically a deep neural network, offering both linear and nonlinear function approximators [37]. In each timeslot, actions are selected based on the ε-greedy policy, and transition tuples are primarily stored in a replay memory to enhance stability. However, a significant drawback of DQN is the ambiguity in action selection or evaluation, often leading to an overestimation of action values and unstable training. To mitigate this issue, Hasselt et al. proposed the Double DQN (DDQN) architecture, which decomposes the max function estimators into separate action selection and evaluation components. A look-up table in DDQN is updated based on agent experiences when the Q-values of state–action pairs are selected, utilizing the Bellman equation [36]. Here, the learning rate $\alpha \in (0, 1]$ determines the extent of parameter updates, while $r(s, a)$ denotes the received reward for a particular action a. DDQN was applied to optimize system actions, resulting in accelerated learning and improved RL performance [37]. Further details on the DRL-based RIS-DWC mechanism are provided in Section 4.2.

DQN is a multilayer neural network system that computes action values Q(s, -; θ) for a given state vector, where θ represents the network parameters. When the state space has n dimensions and the action space has m dimensions, the neural model operates between $\mathbb{R}^n$ and $\mathbb{R}^m$. DQN is commonly employed for Q-value prediction, wherein the Q-value signifies the state–action value, indicating the expected benefit derived from the agent executing a specific action while considering its current state. A crucial aspect of the DQN algorithm involves the incorporation of a target network and experience replays. The target network, akin to the online model, possesses identical settings, with the exception that these settings are copied every τ steps from the online model. Consequently, the target network, characterized by parameters θ', mirrors the online network but undergoes parameter updates every τ steps from the online network, remaining fixed during all other steps. Therefore, $\theta' = \theta$ and is constant throughout other iterations. The equation for the target network is given by

$$Y_{target}^{DQN} = \mathrm{r}(s_t, a_t) + \gamma max_a \mathrm{Q}(s_{t+1}, a_t; \theta') \tag{12a}$$

DDQN represents an enhancement over the original DQN framework. The structural design of DDQN closely resembles that of DQN, with both models employing standard Q-learning and DQN maximum operators to select actions and evaluate stocks based on similar value metrics. However, this shared methodology can lead to the selection of overestimated values, resulting in overly optimistic value estimations. To address this issue, DDQN decouples action selection from action evaluation. In DDQN, two value functions are learned simultaneously through alternating updates, resulting in two sets of weights. Greedy policies are identified using one set of weights for each update, while the other set is utilized to evaluate the algorithm's value. By separating action selection from action evaluation within the target, DDQN mitigates overestimation issues. The architecture of DDQN's target network provides an efficient solution without additional interconnections, enhancing the accuracy of greedy policy measurements. To alleviate overestimation problems associated with Q-values, DDQN introduces a new target network, denoted as

$$Y_{target}^{DDQN} = \mathrm{r}(s_t, a_t) + \gamma \mathrm{Q}'\left[s_{t+1}, argmax_a \mathrm{Q}(s_{t+1}, a_{t+1}, \theta); \theta'\right] \tag{12b}$$

Similar to DQN, DDQN employs a periodic copy of the online network for target updates, maintaining the core principles of the DQN algorithm while reaping the benefits of dual Q-learning. Computational demands are kept to a minimum, with DDQN utilizing parameters from current Q-networks for each selection, unlike DQN where target parameters are unavailable during target value calculation. By replacing the selection with DDQN, overestimation is reduced, bringing Q-values closer to reality. However, it is worth noting that the Q-value calculated after such a replacement must be less than or equal to the original Q-value, effectively decreasing overestimation and improving accuracy.

4.2. DRL-Based RIS-DWC Mobility Control Strategy

The proposed RIS-DWC method aims to enhance the wireless communication services based on ground users' locations. The DDQN framework comprises DNN and RL modules. The DNN approximates Q-values, eliminating the need for a look-up table, which are then input into the RL module for decision-making based on the DNN outputs. We use DRL over traditional RL for our work, because compared with traditional RL, DRL can converge to an optimal solution faster and is more robust against non-optimal parameter settings [38]. Our proposed system uses the DBS as the DRL agent, and its system state, action spaces, and reward function are defined as follows:

System Action Space: The DBS updates its location based on the ground user's mobility, and the actions represent the flying direction, in order to enhance communication service. After the DBS takes an action, the RIS controller adjusts phase-shifts of RIS elements to improve communication quality between the DBS and ground users.

System State Space: The system includes both direct links and reflective channels. The DBS serves as a DRL agent, with system states defined as the DBS coordinates, U (the current number of users available in T_a), and $\gamma_{u,t}^{snr}$ (the SNR value) as the system states. Additionally, we define a threshold state value γ_{def} to assess the quality of the user communication channel.

Reward function: The reward function aims to optimize the total achievable rate and coverage. During training, the DRL agent receives rewards from the function in each timeslot based on the received reward, updating its policy π in order to enhance our proposed strategy network. The reward function $\epsilon_{s(t)}$ is defined as

$$\epsilon_{s(t)} = \gamma_{u,t}^{snr} \times C_t \tag{13}$$

where $r(t) = 1$ if $\epsilon_{s(t+1)} > \epsilon_{s(t)}$, $r(t) = 0$ if $\epsilon_{s(t+1)} = \epsilon_{s(t)}$, and $r(t) = -1$ if $\epsilon_{s(t+1)} < \epsilon_{s(t)}$.

Algorithm for Phase-Shift optimization: Each timeslot involves determining the appropriate phase-shift based on the positions of the DBS and currently connected ground users. To address this challenge, we employ an alternating optimization robust algorithm previously utilized in [39] to handle the RIS discrete phase-shifts. Algorithm 1 outlines

the pseudocode for this process. Initially, the RIS's phase-shift elements are randomly assigned from available angle values. Subsequently, we iteratively optimize each reflecting element while holding the others constant. For a given element m, its phase shift is adjusted to maximize the achievable rate $R_{u,t}$ by evaluating all possible phase-shift values. This procedure continues in each timeslot until the condition $R_{u,t} \geq D_u$ is met or the maximum number of iterations is reached. Finally, the phase-shifts of all RIS elements for each timeslot are determined.

Algorithm 1: RIS-assisted DBS Trajectory and Phase Shift Design Strategy

Input: System action space, size of mini-batch, discount factor, number of timeslots, number of episodes, set of actions, learning rate β, exploration coefficient ∈.
Output: The optimal DBS trajectory and the ground users' coverage
Initialize main network Q with random weights θ;
Initialize target network Q′ with random weights θ′ = θ;
Initialize experience replay memory;
/* Exploration Stage */
For episode = 1 to E **do**
  Set t =*1*, initialize the system to s_1 and update the DBS location;
  While $0 \leq t \leq T$ **do**
    Obtain the state s_t;
    The DBS select the action a_t using the ε-greedy approach; with the probability ϵ, or select action by a_t;
    Execute a_t;
    Obtain the RIS phase-shifts based on Algorithm 2;
    The DBS observes the next system state s_{t+1} and receives the reward based on the Equation (13);
      if DBS out of the target area, **then** cancel the action and apply s_{t+1} accordingly;
      end if
    Store (s_t, a_t, r_t, s_{t+1}) in the experience relay memory;
  /* Training Stage */
    Select a random mini-batch from the experience replay memory;
    Calculate the target Q-value using (15);
    The DRL agent updates the weights θ by minimizing the loss function using (16);
    Update weights of the main network θ;
    Update weights of the target network θ′;
    t ← t+1;
  end while
 episode ← episode+1
end for

Training Procedure: The training procedure of the proposed framework is depicted in Figure 2 and follows the steps outlined in Algorithm 1. We set a simulation environment in our work that models the real environment of DBS, RIS configurations, and wireless channel characteristics using Equations (1) to (6). We defined the DBS, RIS, and mobile ground users' parameters. The agent (i.e., DBS) interacts with the environment to collect user channel data. At each time slot, the agent observes environmental information, such as the DBS's location, the ground users' location, and the users' received SNR values. Then, based on the received state information, the agent selects the proper action to make a decision. The environment responds with rewards and new states. These interactions at each time are stored as experiences (s_t, a_t, r_t, s_{t+1}) in the experience replay buffer memory. This experience dataset is continually updated and used for training the proposed UAV trajectory and RIS phase shift scheme. The complexity of a DRL algorithm is determined by the number of training time-slots per episode and the total number of episodes. Initially

(lines 1–3), the algorithm initializes the system's action space, learning rate β, discount factor γ, and the initial position of the DBS. At the outset, the locations and distribution of users are known. Subsequently, the training stage begins from line 4, where the DBS, from the environment, acquires observations using Equation (13), and the DBS is controlled by the DRL agent. Based on the obtained observations, the DBS selects an appropriate action using policy π. To strike a balance between exploiting past experiences and exploring the target environment, to attain better rewards (line 8), we employ an epsilon ε-greedy approach in the action selection mechanism. During the initial learning phase, the DRL agent may lack confidence in the estimated value of the Q-function as it might not have encountered certain state and action pairs within the target area. Hence, the DRL agent must explore the environment to a certain extent to prevent getting stuck in suboptimal policies.

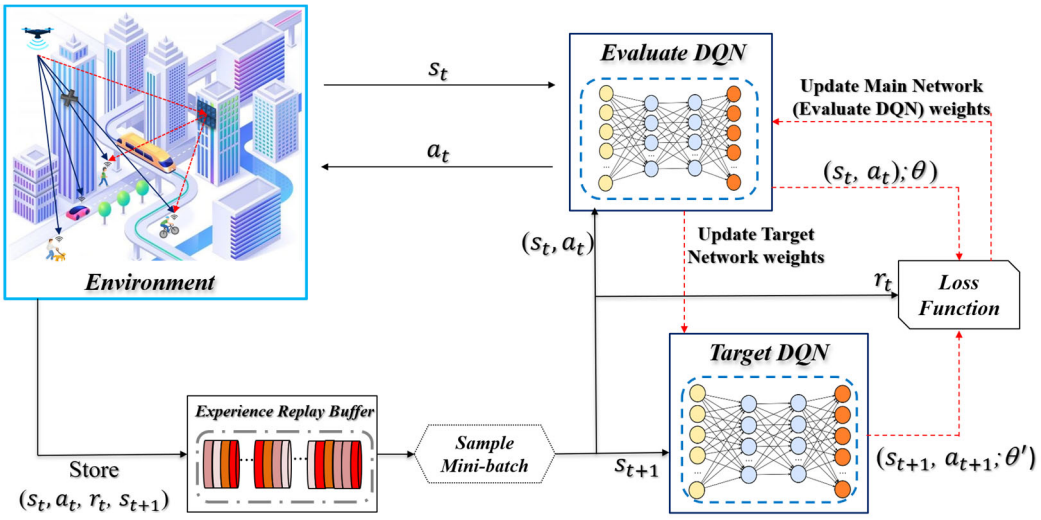

Figure 2. Structure of DRL-based on DBS trajectory and phase shift design strategy.

Consequently, the action selection probability of a DRL agent is defined by the ε parameter, which signifies the likelihood of selecting exploration over exploitation. When ε equals 1, the agent dedicates itself entirely to exploration, neglecting exploitation. The agent's decision between exploration and exploitation hinges on a randomly generated number between 0 and 1. Subsequently, through the Q-learning process, a policy is adopted to determine the best action. A greedy policy is used in most cases and expressed as

$$\pi(s_t) = argmax_{a_t} Q(s_t, a_t) \tag{14}$$

where the optimal DBS trajectory acquired through the DDQN algorithm corresponds to selecting action a_t that possesses the highest Q-value related to a given state s_t.

$$Q(s_t, a_t) = \mathbb{E}\left[\sum_{t+1=t}^{T} \gamma \mathrm{r}(s_{t+1}, a_{t+1}); s_t, a_t\right] \tag{15}$$

The DRL agent selects an action with the highest Q-value from the current system state. Otherwise, from the action space, the next action is randomly chosen. This decision is important to allow the DBS to explore and discover new states to develop an optimal control strategy. Moreover, the RIS-DWC strategy can explore and learn the environment using a high exploration coefficient at the beginning of training. Then, the DBS executes trajectory control according to the selected action a_t. If the DBS flies outside the border, cancel this movement and give a penalty. Once the decision is taken by the DBS, the RIS controller adjusts the phase-shifts of all reflection elements based on our proposed Algorithm 2 (line 10). Additionally, the DBS will remain at its current location if the next location is obtained beyond the target area (lines 11–13). After executing a_t, we obtained

the reward r_t and the next system state s_{t+1} (line 11). After obtaining a new generated sample s_t, a_t, r_t, and s_{t+1} in (line 14) the DRL agent are stored into an experience replay memory F. Once sufficient experiences have been collected in the replay memory F, we sample a mini-batch of H samples that is extracted randomly from the replay memory to train the DDQN network (line 15). To make the network more stable and for a better learning convergence, the main network and the target network in our proposed DDQN network are both used.

Algorithm 2: Algorithm for Phase-Shift optimization

Input: Number of the timeslots, number of episodes, set of actions, learning rate, size of mini-batch, state.
Output: The optimal phase-shifts
$\Theta_t\ (0) = \phi_{m,t} \in \{0,\ \Delta\varnothing,\ \ldots,\ \Delta\varnothing(L-1)\}$;
Initialize the maximum number of iterations *Imax*;
Obtain the DBSs, ground users' position, and the initial RIS phase-shifts; then, calculate the achievable rate $R_{u,t}(0)$;
For $i = 1, 2, \ldots, Imax$ **do**
  For $m = 1, 2, \ldots, M$ **do**
    Fix the phase-shift of the *m-th* element, ∀m = m;
    Set $p_{m,t}$ when $R_{u,t}(i)$ is maximum;
  end for
  If $R_{u,t}(i) > R_{th}$ **then**
    Find $\Theta_t = \Theta_t(i)$;
    break;
  end if
end for

To achieve the optimal RIS-DWC system, DNN updates the weights θ for each tuple within the sampled minibatch. This update is performed to minimize the prediction error using a gradient descent method. Then, the target value Y_{target}^{DDQN} is used to update the weights θ of the original network by minimizing the loss J(θ) in lines 18–19. We define the loss function as the following formula:

$$J(\theta) = \mathbb{E}\left[Y_{target}^{DDQN} - Q(s_t, a_t; \theta)\right] \tag{16}$$

where the target value Y_{target}^{DDQN} expressed is calculated following the double Q-learning approach taking into account the Q-values of the original network and the target network. For the DDQN, we employ a DNN with weight vectors θ to estimate the Q-function $Q(s_t, a_t; \theta)$ expressed in Equation (15). Subsequently, the weights of the target network are updated (lines 18–19), but at a lower frequency than the original network, to mitigate the issue of overestimation.

5. Performance Analysis

This section presents some numerical results to assess the performance of the proposed method. It begins by detailing the simulation parameters, followed by the presentation of results and discussions.

5.1. Implementation Details

This paper employs a model for user movement simulation, considering two types of user mobility: slow and fast walking speeds. The DBS is deployed initially at the coordinate position [0, 0, 100] to serve ground users. The permissible altitude range for the DBS is constrained between 30 and 100 m, denoted as $30 \leq z^D \leq 100$. In our simulation, we assume that U represents 6 ground users, assigned randomly to a target area, and moving across it based on the Random Walk Model. Users move on the ground at a maximum speed of 0.5 m each step. This study's primary objective is to devise an RIS-assisted DBS

mobile control system aimed at maximizing communication coverage while maintaining an acceptable level of throughput for mobile users, utilizing DRL techniques. The coordinates of the RIS are fixed at [50, 50, 50].

Our proposed double deep neural network architecture consists of two layers, each comprising 30 neurons. The parameters of each DNN are initialized randomly from a zero-mean normal distribution. To mitigate overfitting during the training process, after each pooling operation, a dropout layer is included, ensuring the model's generalization capability. Rectified Linear Unit (ReLU) activation functions are employed for the first two hidden layers of the DNN, while RMSProp is utilized as the optimizer, adjusting the learning rate ($L_r = 1 \times 10^{-3}$) for each epoch with a minibatch size of 512.

Furthermore, the AdamOptimizer is utilized to train the DNNs comprising the Q-network. We adopt the air-to-ground communication scenario as discussed in [19]. Our experiment was conducted on a PC equipped with an NVIDIA GeForce GTX 1080, with Ubuntu as the operating system and Python 3.7 as the programming language. About the deep learning framework, Keras 2.2.5 and TensorFlow 1.14. libraries are employed for implementing the DNNs in the DDQN algorithm [40]. Due to the lack of equipment, we simulated our environment closely to a real-world scenario. For further details, Table 2 provides additional simulation information.

Table 2. Simulation parameters.

Parameters	Value
DBS transmit power	5 mW
Bandwidth	2 MHz
Noise power	−173 dBm
Path loss at 1 m α	−20 dB
Path loss β	2.5
Blocking parameters a, b [19]	9.61, 0.16
Carrier frequency fc	2 Ghz
Number of RIS elements	100
Target area size	1000 m × 1000 m
Episodes and time slots	500, 2000

5.2. Results and Discussion

In this part, we validate our proposed method (DBS-RIS with PS) through simulation and compare it with the following numerical results based on two different systems:

(i) The DBS trajectory using RIS in the absence of the optimal passive phase-shift (DBS-RIS no PS): In this scheme, the DBS position is optimized while the RIS has a random phase shifts configuration.

(ii) The DBS trajectory without RIS (DBS no RIS): In this setting, the DBS either transmits or relays signals to ground users.

Both of these benchmark systems optimize the trajectory and coverage between the DBS and the mobile ground users using the DDQN algorithm method.

In this work, the DBS in all different systems endeavors to find the best trajectories to serve the ground users effectively. Due to its proximity to the center, the area offers the best conditions for minimizing communication errors and improving communication coverage. The DBS is scheduled to move closer to users with a poor SNR, thereby directly enhancing their signal quality using direct or indirect links. This may involve algorithms that calculate the optimal position of the DBS to maximize the average or minimum data rate among all users.

Figure 3 presents the 3D and 2D trajectory planes of different systems for visualization and comparison purposes.

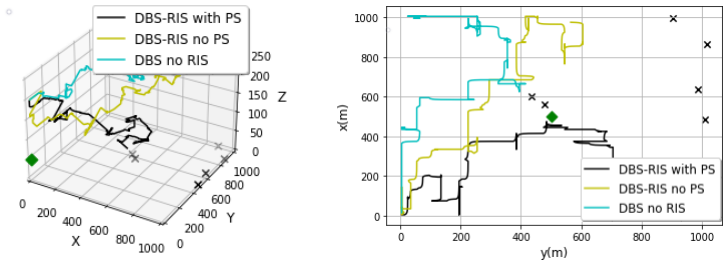

Figure 3. The 2D and 3D DBS trajectories over 3 different systems.

In the proposed case (DBS-RIS with PS), it is evident that the DBS descends to a lower altitude and attempts to approach the RIS. This strategy aims to enhance the SNR by minimizing the communication distance, as outlined in Equation (7). By flying closer to the RIS, the DBS optimizes its positioning to maximize signal quality and improve communication performance.

Figure 4 illustrates the coverage scores obtained by our proposed DRL method and two benchmark systems over time. Our experimental findings indicate that, as the learning iterations progress, the coverage score also increases.

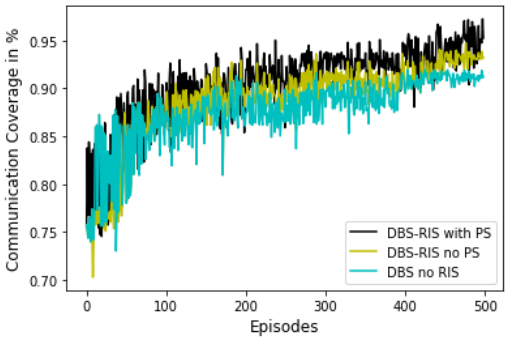

Figure 4. Comparison of the overall communication coverage scores over different systems at different learning iterations.

Specifically, with an increase in the number of timeslots, our proposed solution (DBS-RIS with PS) achieves an impressive 95.58% coefficient of coverage. However, the DRL agent associated with DBS-RIS (no PS) and DBS (no RIS) may struggle to maneuver the DBS to optimum positions during a limited number of training intervals. Moreover, at the outset, the DBS may not cover all states, requiring time to explore and reach unexplored states by executing actions based on the current state to gain a comprehensive understanding of the system environment. For comparison, the performances of DBS-RIS (no PS) and DBS (no RIS) are also depicted. These benchmark systems exhibit a lower user coverage compared to DBS-RIS (with PS). Additionally, due to learning instability, the estimation of expected rewards in these two approaches leads to a high-variance problem.

We compared the performance of the proposed system by changing its network hyperparameters as shown in Table 3. We used RMSProp, ReLU, 0.0001, 512, and 0.9 for the optimizer, activation function, learning rate, minibatch size, and discount factor, resulting in a better communication coverage.

Table 3. Proposed method performance under different DRL network hyperparameters.

	Optimizer		**Activation Function**		**Learning Rate**		**Minibatch Size**		**Discount Factor**	
	Adam	**RMSProp**	Sigmoid	**ReLU**	0.001	**0.0001**	256	**512**	0.89	**0.9**
Coverage score	0.942	**0.955**	0.942	**0.955**	0.942	**0.955**	0.942	**0.955**	0.942	**0.955**

Next, Figure 5 highlights the importance of deploying RIS to augment coverage in DBS-assisted wireless communication systems. The blockage probability $p_{u,t}$ as defined in Equation (7) characterizes the received signal of ground users from both direct ($p_{u,t} = 0$) and indirect links ($p_{u,t} = 1$). Various blockage probability values, representing different environmental densities ranging from low to high, are established. An analysis of the figure reveals that as the environment density increases, the coverage provided by DBS (no RIS) decreases, while the coverage facilitated by DBS-RIS (no PS) increases. Consequently, the integration of RIS technology enhances DBS coverage, with further enhancements observed by incorporating phase shift capabilities.

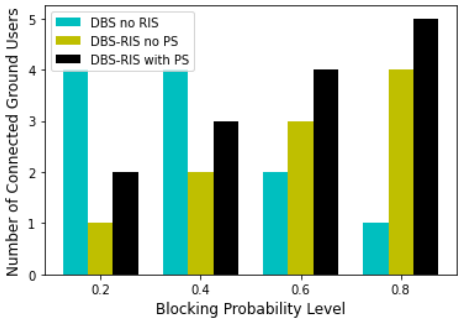

Figure 5. Comparison of the ground users' coverage by DBS over different systems.

Figure 6 illustrates the Cumulative Distribution Function (CDF) of the data rate performance across three compared systems with P = 5 mW. CDF serves as a tool to depict outcomes across various episodes. It aids in validating algorithm convergence and visually communicates the point at which satisfactory performance is achieved across episodes. This allows for the selection of an optimal algorithm endpoint. The proposed method iteratively learns from mistakes in the system environment by taking actions aimed at enhancing the data rate of each user. However, due to factors such as user mobility and distribution, as well as the phase shift algorithm in the target area, the data rate performance may degrade in some timeslots. This degradation causes the DBS position to become suboptimal in balancing coverage and communication quality. As a result, the proposed method continuously adapts the DBS location based on the received reward until the data of each ground user surpasses the threshold value (defined for the SNR). Notably, the performance of (DBS-RIS with PS) in the simulated scenario is significantly high. This can be attributed to the algorithm's capability to adjust phase shifts and DBS location, thereby increasing the difficulty of signal alignment and subsequently enhancing ground user throughput.

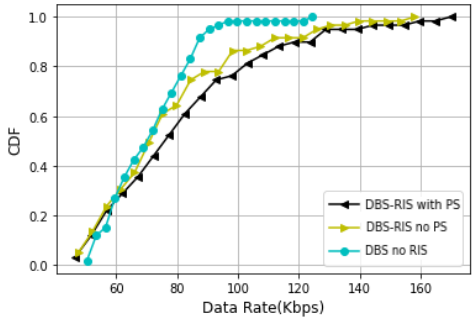

Figure 6. Performance of data rate over different systems with P = 5 mW.

Figure 7 illustrates the Cumulative Distribution Function (CDF) of the data rate performance across three compared systems with P = 1 mW. Based on a comparison of Figures 6 and 7, we can conclude that a higher transmit power from the DBS can result in stronger received signals for the ground users. This can improve the signal-to-noise ratio (SNR) at the users' devices, leading to better data rates.

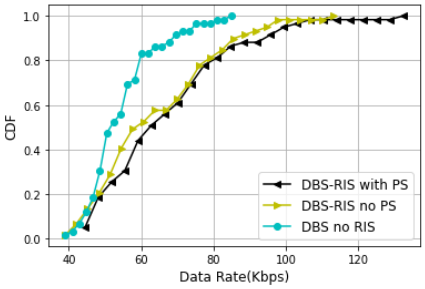

Figure 7. Performance of data rate over different systems with P = 1 mW.

Figure 8 presents a comparison of the average data rate for ground users across various systems, considering two to six ground users. It is observed that as the number of ground users increases, the average user data rate tends to decrease for all three algorithms. Furthermore, this decrease becomes more pronounced with an increasing number of ground users. In terms of algorithm efficacy, the proposed DBS-RIS (with PS) algorithm demonstrates superior performance compared to DBS-RIS (no PS), with both significantly outperforming the DBS (no RIS) approach.

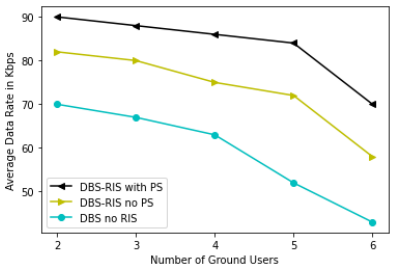

Figure 8. Performance of average data rate over different systems.

We can conclude that DBS-RIS (with PS) achieves a much higher average throughput performance compared to DBS-RIS (no PS) and DBS (no RIS). It attains higher communication rates and better communication quality than DBS-RIS (no PS) and DBS (no RIS). This behavior is mainly influenced by the mobile ground users' distribution.

5.3. Computational Time and Complexity

The RIS-assisted DBS in the wireless communication system model requires less computation time and complexity to assess the quality of service and the DBS coverage at each time slot. The computational complexity of the DRL depends on the size of the neural network layers and the number of parameters. The DDQN's architecture and training scheme help reduce complexity and computation time compared to traditional Q-learning algorithms, making it a more efficient and effective choice for training deep reinforcement learning agents. By updating the target network's parameters less frequently, DDQN reduces the number of network updates required per time slot, resulting in faster convergence and lower computational overhead. DDQN training typically involves iteratively updating the Q-network parameters using gradient descent, which can be computationally intensive. During each training iteration, the forward pass involves passing the current state through the neural network to obtain the Q-values for all possible actions. This technique can expedite the training process by enabling the simultaneous computation of multiple training samples or parallel updates of network parameters. This simplicity can lead to shorter training times and also reduce memory consumption. Hence, we implemented and optimized our proposed method using the DDQN model by optimizing hyperparameters, reducing unnecessary computations, and ensuring efficient memory usage. Our proposed system's computation time per episode(s) is 8.21×10^{-1}.

6. Conclusions and Open Work

Our work introduces an autonomous 3D-DBS deployment strategy aimed at enhancing communication services for ground users through RIS technology. A learning-based approach is proposed to tackle this challenge by optimizing the communication coverage while ensuring high data rates for ground users. The proposed DRL-based RIS-assisted DBS wireless communication control strategy aims to control the DBS movement based on ground users' locations to achieve the objective function. Simulation results demonstrate the promising performance of the proposed system in terms of both overall network throughput and communication coverage.

As open work, the authors intend to advance this approach by developing a satellite-assisted DBS wireless communications system. This advancement aims to further enhance overall user throughput, mitigate the channel fading effects, and improve the communication coverage using artificial intelligence models.

Author Contributions: W.N.K. contributed to the experiment, data analysis, programming, and manuscript preparation. The problem was suggested by H.-P.L., who provided guidance and instructions for the work. R.-T.J. offered suggestions and feedback during discussions. G.B.T. contributed to the experiment and collaborated with Nathanael on programming tasks. B.A.T. collected relevant references and offered comments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Council, Taiwan (R.O.C.) (Grant No: 112-2221-E-027-123).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: This research is partly funded by the National Science and Technology Council, Taiwan (R.O.C.).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khawaja, W.; Guvenc, I.; Matolak, D.W.; Fiebig, U.C.; Schneckenburger, N. A survey of air-to-ground propagation channel modeling for unmanned aerial vehicles. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2361–2391. [CrossRef]
2. Cicek, C.T.; Gultekin, H.; Tavli, B.; Yanikomeroglu, H. UAV Base Station Location Optimization for Next Generation Wireless Networks: Overview and Future Research Directions. In Proceedings of the 2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS), Muscat, Oman, 5–7 February 2019; pp. 1–6.
3. Di Renzo, M.; Zappone, A.; Debbah, M.; Alouini, M.S.; Yuen, C.; De Rosny, J.; Tretyakov, S. Smart radio environments empowered by reconfigurable intelligent surfaces: How it works state of research the road, ahead. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 2450–2525. [CrossRef]
4. Renzo, M.D.; Debbah, M.; Phan-Huy, D.T.; Zappone, A.; Alouini, M.S.; Yuen, C.; Sciancalepore, V.; Alexandropoulos, G.C.; Hoydis, J.; Gacanin, H.; et al. Smart radio environments empowered by reconfigurable AI metasurfaces: An idea whose time has come. *Commun. Netw.* **2019**, *2019*, 129.
5. Xu, F.; Hussain, T.; Ahmed, M.; Ali, K.; Mirza, M.A.; Khan, W.U.; Ihsan, A.; Han, Z. The state of ai-empowered backscatter communications: A comprehensive survey. *IEEE Internet Things J.* **2023**, *10*, 21763–21786. [CrossRef]
6. Jiao, H.; Liu, H.; Wang, Z. Reconfigurable Intelligent Surfaces aided Wireless Communication: Key Technologies and Challenges. In Proceedings of the 2022 International Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, 30 May–3 June 2022.
7. Siddiqi, M.Z.; Mir, T. Reconfigurable intelligent surface-aided wireless communications: An overview. *Intell. Converg. Netw.* **2022**, *3*, 33–63. [CrossRef]
8. Tesfaw, B.A.; Juang, R.T.; Tai, L.C.; Lin, H.P.; Tarekegn, G.B.; Nathanael, K.W. Deep Learning-Based Link Quality Estimation for RIS-Assisted UAV-Enabled Wireless Communications System. *Sensors* **2023**, *23*, 8041. [CrossRef] [PubMed]
9. Li, S.; Duo, B.; Yuan, X.; Liang, Y.-C.; Di Renzo, M. Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 716–720. [CrossRef]
10. Zhao, J.; Yu, L.; Cai, K.; Zhu, Y.; Han, Z. RIS-aided ground-aerial NOMA communications: A distributionally robust DRL approach. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 1287–1301. [CrossRef]

11. Zhang, J.; Tang, J.; Feng, W.; Zhang, X.Y.; So, D.K.C.; Wong, K.-K.; Chambers, J.A. Throughput Maximization for RIS-assisted UAV-enabled WPCN. *IEEE Accesss* **2024**, *12*, 3352085. [CrossRef]
12. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.C.; Kim, D.I. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3133–3174. [CrossRef]
13. Xiong, Z.; Zhang, Y.; Niyato, D.; Deng, R.; Wang, P.; Wang, L.C. Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges. *IEEE Veh. Technol. Mag.* **2019**, *14*, 44–52. [CrossRef]
14. Ji, P.; Jia, J.; Chen, J.; Guo, L.; Du, A.; Wang, X. Reinforcement learning based joint trajectory design and resource allocation for RIS-aided UAV multicast networks. *Comput. Netw.* **2023**, *227*, 109697. [CrossRef]
15. Fan, X.; Liu, M.; Chen, Y.; Sun, S.; Li, Z.; Guo, X. Ris-assisted uav for fresh data collection in 3d urban environments: A deep reinforcement learning approach. *IEEE Trans. Veh. Technol.* **2022**, *72*, 632–647. [CrossRef]
16. Zhang, H.; Huang, M.; Zhou, H.; Wang, X.; Wang, N.; Long, K. Capacity maximization in RIS-UAV networks: A DDQN-based trajectory and phase shift optimization approach. *IEEE Trans. Wirel. Commun.* **2022**, *22*, 2583–2591. [CrossRef]
17. Tarekegn, G.B.; Juang, R.T.; Lin, H.P.; Munaye, Y.Y.; Wang, L.C.; Bitew, M.A. Deep-Reinforcement-Learning-Based Drone Base Station Deployment for Wireless Communication Services. *IEEE Internet Things J.* **2022**, *9*, 21899–21915. [CrossRef]
18. Ranjha, A.; Kaddoum, G. URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning. *IEEE Internet Things J.* **2020**, *8*, 4618–4627. [CrossRef]
19. Al-Hourani, A.; Kandeepan, S.; Lardner, S. Optimal LAP altitude for maximum coverage. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 569–572. [CrossRef]
20. Wu, Q.; Zhang, R. Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts. *IEEE Trans. Commun.* **2019**, *68*, 1838–1851. [CrossRef]
21. Wang, J.L.; Li, Y.R.; Adege, A.B.; Wang, L.C.; Jeng, S.S.; Chen, J.Y. Machine learning based rapid 3D channel modeling for UAV communication. In Proceedings of the 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 11–14 January 2019; pp. 1–5.
22. Tarekegn, G.B.; Juang, R.T.; Lin, H.P.; Munaye, Y.Y.; Wang, L.C.; Jeng, S.S. Channel Quality Estimation in 3D Drone Base Stations for Future Wireless Network. In Proceedings of the 2021 30th Wireless and Optical Communication Conference (WOCC), Taipei, Taiwan, 7–8 October 2021.
23. Alzenad, M.; El-Keyi, A.; Yanikomeroglu, H. 3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different QoS requirements. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 38–41. [CrossRef]
24. Chen, Y.; Li, N.; Wang, C.; Xie, W.; Xv, J. A 3D placement of unmanned aerial vehicle base station based on multi-population genetic algorithm for maximizing users with different QoS requirements. In Proceedings of the 2018 IEEE 18th International Conference on Communication Technology (ICCT), Chongqing, China, 8–11 October 2018; pp. 967–972.
25. Ma, D.; Ding, M.; Hassan, M. Enhancing cellular communications for UAVs via intelligent reflective surface. *arXiv* **2019**, arXiv:1911.07631.
26. Diamanti, M.; Charatsaris, P.; Tsiropoulou, E.E.; Papavassiliou, S. The prospect of reconfigurable intelligent surfaces in integrated access and backhaul networks. *IEEE Trans. Green Commun. Netw.* **2021**, *6*, 859–872. [CrossRef]
27. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
28. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double q-learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 16, pp. 2094–2100.
29. Deng, H.; Yin, S.; Deng, X.; Li, S. Value-based algorithms optimization with discounted multiple-step learning method in deep reinforcement learning. In Proceedings of the 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Yanuca Island, Fiji, 14–16 December 2020; pp. 979–984.
30. Li, S.; Duo, B.; Di Renzo, M.; Tao, M.; Yuan, X. Robust secure UAV communications with the aid of reconfigurable intelligent surfaces. *IEEE Trans. Wireless Commun.* **2021**, *20*, 6402–6417. [CrossRef]
31. Liu, X.; Liu, Y.; Chen, Y. Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2042–2055. [CrossRef]
32. Mei, H.; Yang, K.; Liu, Q.; Wang, K. 3D-trajectory and phase-shift design for RIS-assisted UAV systems using deep reinforcement learning. *IEEE Trans. Veh. Technol.* **2022**, *71*, 3020–3029. [CrossRef]
33. Li, Y.; Aghvami, A.H.; Dong, D. Path planning for cellular-connected UAV: A DRL solution with quantum-inspired experience replay. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 7897–7912. [CrossRef]
34. Li, Y.; Aghvami, A.H. Radio resource management for cellular-connected uav: A learning approach. *IEEE Trans. Commun.* **2023**, *71*, 2784–2800. [CrossRef]
35. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018; pp. 1–775.
36. Mohi Ud Din, N.; Assad, A.; Ul Sabha, S.; Rasool, M. Optimizing deep reinforcement learning in data-scarce domains: A cross-domain evaluation of double DQN and dueling DQN. *Int. J. Syst. Assur. Eng. Manag.* **2024**, 1–12. [CrossRef]
37. Alharin, A.; Doan, T.N.; Sartipi, M. Reinforcement learning interpretation methods: A survey. *IEEE Access* **2020**, *8*, 171058–171077. [CrossRef]

38. Yu, Y.; Wang, T.; Liew, S.C. Deep-reinforcement learning multiple access for heterogeneous wireless networks. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1277–1290. [CrossRef]
39. Abeywickrama, S.; Zhang, R.; Wu, Q.; Yuen, C. Intelligent reflecting surface: Practical phase shift model and beamforming optimization. *IEEE Trans. Commun.* **2020**, *68*, 5849–5863. [CrossRef]
40. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for Large-Scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.