applied sciences

Article

The Deep Deterministic Policy Gradient Algorithm Based on RIS Technology in a Coal Mine Tunnel

Shuqi Wang *, Fengjiao Wang, Wei Zhang and Yaqi Wang

School of Communication and Information Engineering, Xi'an University of Science and Technology,
Xi'an 710600, China; 15205848573@163.com (F.W.); zhangwei_xust@163.com (W.Z.);
24207040041@stu.xust.edu.cn (Y.W.)
* Correspondence: wangshuqi@xust.edu.cn

Abstract: Reconfigurable Intelligent Surface (RIS) technology relies on its reconfigurable electromagnetic properties and offers an efficient solution for enhancing signal quality in coal mine communications. RIS technology significantly enhances signal coverage and transmission quality in complex, confined environments. This paper proposes a channel propagation optimization scheme for coal mine RIS communication systems, using the Deep Deterministic Policy Gradient (DDPG) algorithm. By jointly optimizing base station power allocation and RIS phase shift, this paper comparatively analyzes RIS reflection performance under ideal and non-ideal conditions, focusing on its impact on system propagation rates. A comparison of system stability and convergence rates among the DDPG, A3C, and DQN algorithms reveals that, under the DDPG optimization scheme, the average link rate reaches 6.6 bps/Hz with ideal RIS reflection and 4.6 bps/Hz with non-ideal conditions when the base station transmit power is defined as 38 dBm. Furthermore, increasing the number of RIS units from 8 to 32 results in a system link rate improvement from 5 bps/Hz to 6.8 bps/Hz. The research results provide new design ideas for optimizing coal mine RIS communication systems and open up new solutions for the use of artificial intelligence in complex coal mine tunnel environments.

Keywords: reconfigurable intelligent surface; coal mine tunnel; deep deterministic policy gradient; link rate

check for updates

Citation: Wang, S.; Wang, F.; Zhang, W.; Wang, Y. The Deep Deterministic Policy Gradient Algorithm Based on RIS Technology in a Coal Mine Tunnel. *Appl. Sci.* **2024**, *14*, 12014. https://doi.org/10.3390/app142412014

Academic Editor: Tiago Miranda

Received: 22 November 2024
Revised: 17 December 2024
Accepted: 21 December 2024
Published: 22 December 2024

1. Introduction

Reconfigurable Intelligent Surfaces (RISs) is a cutting-edge technology in 6G communication systems [1,2]. It can optimize the wireless signal transmission path through a large number of adjustable metamaterial reflective elements, enhance signal strength and coverage, and is becoming a powerful means to solve coal mine communication problems [3–5].

In recent years, research on RIS has mainly focused on the ground RIS structure design and signal processing techniques [6–8], while research on RIS in coal mine environments is relatively limited. A prototype of RIS was proposed and developed, which investigated RIS channel models and conducted practical indoor signal coverage experiments [9]. Another research utilized Ray-tracing simulation tools to obtain channel data in offices, conference rooms, and laboratory scenarios, analyzing channel characteristics and performing channel modeling [10]. Research has also explored guided frequency power allocation in channel estimation for RIS-assisted communication systems and proposed corresponding channel estimation and passive beamforming design schemes [11]. Furthermore, the RIS-assisted system design method was introduced to optimize RIS reflection coefficients, maximizing the weighted sum rate of the ground multi-user systems [12].

With the rapid development of coal mine intelligence, by combining deep learning and Deep Reinforcement Learning (DRL) algorithms with the coal mine environment, various communication problems can be more accurately identified and solved [13–17]. Traditional Q-value-based DRL algorithms, such as the Deep Q-Network (DQN), face significant challenges when dealing with continuous action spaces. These algorithms are mainly designed for discrete action spaces and perform poorly when optimizing continuous variables such as RIS phase offset and base station power control. In dynamic and complex environments such as coal mine tunnels, they may encounter problems such as slow convergence, poor stability, and poor system performance [18,19]. To address these challenges, this paper proposes the use of the Deep Deterministic Policy Gradient (DDPG) algorithm. DDPG is particularly suitable for processing continuous action spaces and can effectively optimize the performance of RIS-assisted communication systems, especially in dynamic and non-stationary coal mine tunnel environments [20–22]. One research group applied the DDPG-based deep reinforcement learning algorithm to jointly optimize downlink beamforming and RIS configurations in MU-MISO systems, achieving improved system performance under perfect Channel State Information (CSI) conditions [23,24]. Another work further investigated system performance under hardware impairments and imperfect CSI conditions [24]. An innovative joint beamforming strategy was proposed based on the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm in deep reinforcement learning, minimizing the total transmit power of the base station through fine-tuning [25]. Additionally, the phase correlation amplitude was introduced, enabling beamforming optimization under this practical phase model [26]. Research that is based on the DDPG algorithm examined the performance of RIS-assisted communication under a Rayleigh fading model for direct channels and a Rician fading model for cascaded channels [27]. Other research proposed a beamforming algorithm based on CSI to optimize the RIS reflection coefficients using DRL techniques to maximize transmission efficiency [28]. Furthermore, a DDPG-based framework was designed for RIS-assisted non-orthogonal multiple access (NOMA) downlink, establishing a long-term stochastic optimization problem involving phase-shift optimization to maximize the total rate of mobile users in the NOMA downlink [29].

At present, there are still some deficiencies in the research of RIS technology in the field of coal mine communication. Existing research mainly focuses on performance analysis under ideal reflection conditions, usually assuming that RIS hardware can achieve precise phase control while ignoring the phase error problem caused by hardware accuracy limitations, which may have a significant impact on system performance in practical applications.

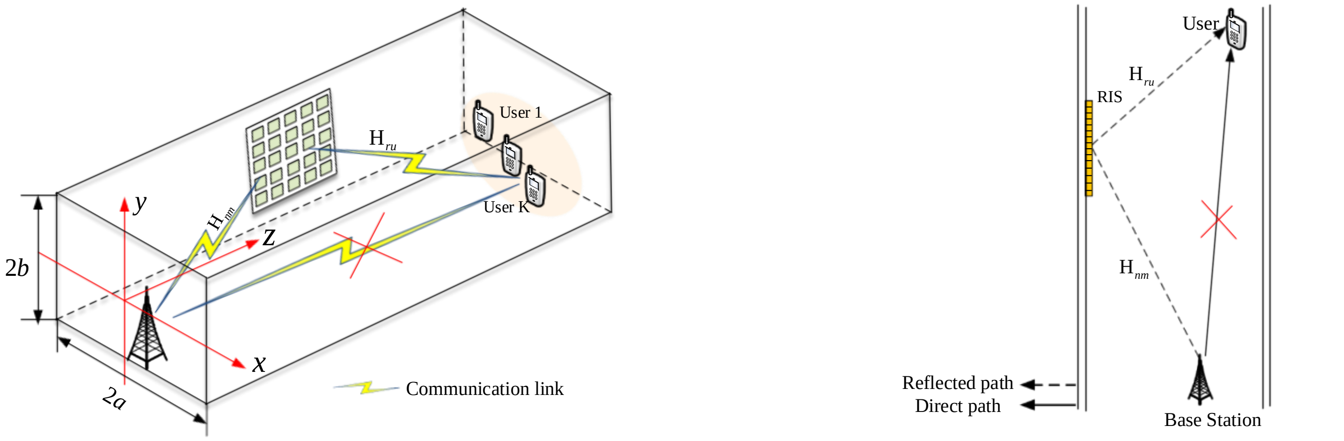
In view of these deficiencies, this paper uses the DDPG algorithm to conduct an in-depth study on the performance of RIS-assisted coal mine tunnel signal transmission. The DRL algorithm is applied to the coal mine RIS communication system for optimization design, and the DDPG algorithm is used to jointly optimize the base station beamforming matrix and the RIS phase shift matrix to maximize the system link rate. This article also compares the performance of DDPG with other DRL algorithms. Under the same conditions, the DDPG algorithm is superior to A3C and DQN in terms of convergence speed and stability. In addition, the article explores the impact of the system link rate under the two modes of RIS, ideal reflection and non-ideal reflection, as well as the impact of neural network parameter settings on algorithm performance.

2. System Model

2.1. Coal Mine Channel Modeling and Reflection Characterization Analysis

Compared with the indoor space on the ground, the wireless channel in the mine exhibits unique propagation characteristics due to its narrow and closed structure and significant multipath effect. In this closed environment, signals undergo multiple reflections and attenuations between tunnel walls, making the propagation path more complex and uncertain. To thoroughly analyze these characteristics and optimize signal coverage and communication quality, this paper selects a rectangular tunnel with a simple structure and good geometric characteristics as the research model to facilitate the accurate construction of a channel propagation model. In this model, a Cartesian coordinate system is established at the center of the rectangular tunnel section, assuming that the tunnel width is $2a$ and the height is $2b$. The relative permittivity of the side walls, top, and bottom is ε_r. As shown in Figure 1, Figure 1a shows the structure of the rectangular straight tunnel model assisted by the coal mine RIS. Figure 1b shows the system model under the corresponding two-dimensional perspective. In addition, it is also worth noting that this paper only investigates the Non-Line of Sight (NLoS) propagation system characteristics and does not consider the effect of Line of Sight (LoS) on the system due to the complex long-distance propagation in the confined space.

The model consists of a Base Station (BS) with M antennas, a RIS with N intelligent reflection units, and K single-antenna users (User Equipment, UE). The RIS is deployed on tunnel walls or the surface of large equipment and devices within the environment. By dynamically adjusting the phase shift and amplitude of the RIS reflective element, the propagation direction of the reflected signal can be precisely controlled, thereby constructing an alternative reflection path to bypass obstacles in the tunnel. This enables RIS to effectively redirect signals to users that are originally in an NLoS state, overcoming the limitations of direct signal transmission. It is assumed that the BS-RIS and RIS-UE channels are frequency-flat fading, and all the wireless channels remain unchanged in each transmission block. $\mathbf{H}_{nm} \in \mathbb{C}^{N\times M}$ represents the channel from the BS to the RIS and $\mathbf{H}_{ru} \in \mathbb{C}^{N\times K}$ represents the channel from the RIS to the user.



(**a**) (**b**)

Figure 1. (**a**) RIS-assisted coal mine communication system; (**b**) RIS-assisted coal mine communication system from a two-dimensional perspective.

The cascaded channel of the system (BS-RIS-UE) can be expressed as follows [30]:

$$\mathbf{H} = \mathbf{H}_{nm}^{H}\boldsymbol{\Phi}\mathbf{H}_{nu} \tag{1}$$

$$\begin{aligned}\boldsymbol{\Phi} &= diag(\phi_1, \phi_2, \cdots, \phi_N) \\ &= \begin{bmatrix} \beta(\theta_1)e^{j\theta_1} & 0 & \cdots & 0 \\ 0 & \beta(\theta_2)e^{j\theta_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \beta(\theta_N)e^{j\theta_N} \end{bmatrix}_{N\times N}\end{aligned} \tag{2}$$

where $\phi_n = \beta(\theta_n)e^{j\theta_n}$,$\theta_n \in [0, 2\pi)$ is the phase shift change caused by the nth reflection element of the RIS, and $\beta(\theta_n)$ is an amplitude function based on the phase shift θ_n, which reflects the amplitude change of the signal reflected from the RIS. $\beta(\theta_n)$ and θ_n satisfy the following relationship:

$$\beta(\theta_n) = (1 - \beta_{\min})f(\theta_n) + \beta_{\min} \tag{3}$$

where $\beta_{\min} \in [0,1]$ represents the minimum amplitude of the reflection coefficient. $f(\theta_n)$ defined as $f_{\sin}(\theta_n) = \left(\frac{\sin(\theta_n-\mu)+1}{2}\right)^{\alpha}$; therefore,

$$\beta(\theta_n) = (1-\beta_{\min})\left(\frac{\sin(\theta_n-\mu)+1}{2}\right)^{\alpha} + \beta_{\min} \tag{4}$$

where μ represents the horizontal offset of the phase change, which determines the phase difference between the function and the angle θ_n, and α controls the steepness of the function curve; the larger α is, the steeper the curve is near μ, and vice versa, the flatter it is. $\beta_{\min}, \mu \geq 0, \alpha \geq 0$ depends on the hardware implementation constants of the RIS. However, in actual situations, due to hardware limitations, the horizontal offset μ of the phase change will produce errors, resulting in phase errors in the reflected signal, and perfect phase control cannot be achieved. In addition, the limited number of RIS reflective elements limits the accuracy of phase adjustment, making it impossible to achieve optimal beamforming, ultimately reducing system gain. If the signal is ideally reflected on the RIS, then $|\phi_n|^2 = 1$, i.e., $\beta_{\min} = 1$ or $\alpha = 0$.

Assuming the transmitted signal is $S(t)$, with zero mean and unit variance, $\mathbf{S} = [s_1, s_2, \cdots, s_K]^T \in \mathbb{C}^{K\times 1}$, then the received signal after RIS is expressed as follows [5]:

$$y = \sqrt{\mathbf{P}}\mathbf{HGS} + W = \sqrt{\mathbf{P}}\mathbf{H}_{nm}^{H}\mathbf{\Phi}\mathbf{H}_{nu}\mathbf{GS} + W \tag{5}$$

where $\mathbf{P}$ is the transmit power, $\mathbf{P} = diag[p_1, p_2, \cdots, p_K] \in \mathbb{R}^{K\times K}$, and $\mathbf{G}$ is the beamforming matrix applied at BS, $\mathbf{G} \in \mathbb{C}^{M\times K}$. W represents the total noise in the coal mine tunnel, which includes the background noise w_{bg}, i.e., continuous noise generated by mechanical operation and the impulse interference noise w_{imp}, i.e., some sudden noise (such as equipment failure). The total noise can be expressed as follows:

$$W = w_{\text{bg}} + w_{\text{imp}} \tag{6}$$

where the background noise is additive Gaussian white noise, expressed as $w_{\text{bg}} \sim \mathcal{N}(0, \sigma^2)$. The impulse interference noise is $w_{\text{imp}} = B \cdot G_a$, where B is a Bernoulli random process with mean 0 and variance 1, taking the value of 0 or 1, indicating whether there is impulse noise generated at a certain moment; G_a is a Gaussian random process related to B, which represents the amplitude of the impulse and obeys $\mathcal{N}\left(0, \sigma_{\text{imp}}^2\right)$. Therefore, the coal mine noise can be represented by an independent and identically distributed Bernoulli–Gaussian process [5], that is:

$$W = w_{\text{bg}} + B \cdot G_a \tag{7}$$

The received signal of the kth user can be expressed as follows:

$$y_k = \sqrt{\mathbf{P}}\boldsymbol{h}_{nm,k}^{H}\mathbf{\Phi}\mathbf{H}_{nu}\mathbf{GS} + W_k \tag{8}$$

Further, it can be written as follows:

$$y_k = \sqrt{P_k}\boldsymbol{h}_{nm,k}^{H}\mathbf{\Phi}\mathbf{H}_{nu}\boldsymbol{g}_k s_k + \sqrt{P_n}\sum_{n,n\neq k}^{K}\boldsymbol{h}_{nm,k}^{H}\mathbf{\Phi}\mathbf{H}_{nu}\boldsymbol{g}_n s_n + W_k \tag{9}$$

where the first term is the expected signal of the kth user, the second term is the interference caused by the signals of all other users $(n \neq k)$ to the kth user, i.e., Co-Channel Interference (CCI), and $\boldsymbol{g}_k$ is the kth column vector of the matrix G.

The SINR of the kth user is represented as follows:

$$\beta_k = \frac{\sqrt{P_k}\left|\boldsymbol{h}_{nm,k}^{H}\mathbf{\Phi}\mathbf{H}_{nu}\boldsymbol{g}_k\right|^2}{\sqrt{P_n}\sum\limits_{n.n\neq k}^{K}\left|\boldsymbol{h}_{nm,k}^{\boldsymbol{H}}\mathbf{\Phi}H_{nu}\boldsymbol{g}_n\right|^2+\sigma^2+p\cdot\sigma_{\text{imp}}^2} \tag{10}$$

where p is the probability that $\mathbf{B}$ equals to 1, assuming that each signal has the same transmission power, i.e., P_t.

Therefore, the total link rate (in units of bps/Hz) in the system can be expressed as follows:

$$C = \sum_{k=1}^{K}\log_2(1+\beta_k) \tag{11}$$

To more accurately describe the characteristics of the coal mine RIS channel, this paper analyzes, in detail, the relationship between the number of reflections of the signal between the tunnel walls and the path length based on the "second law of tent" in geometric optics. As shown in Figure 2, by dividing the propagation path into horizontal and vertical mapping planes, the propagation law of the signal under multiple reflections is clarified. In the figure, the blue dotted line represents the horizontal mapping plane, the green dotted line represents the vertical mapping plane, and the yellow curve represents the propagation path of the signal from the base station to the user after being reflected by the RIS.

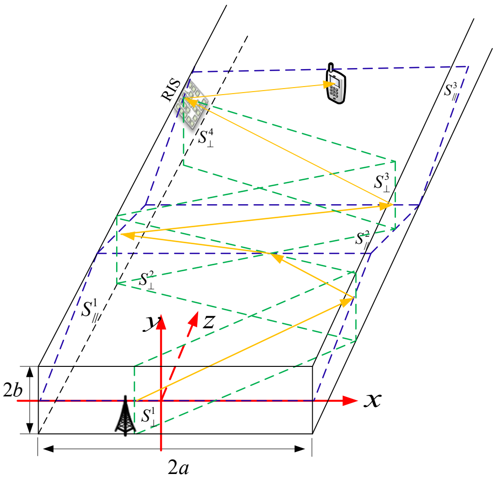


Figure 2. Schematic diagram of the second law of tent.

Assuming that the number of reflections of the signal in the coal mine tunnel is L, the total number of effective paths to the RIS formed by multiple reflections of the signal through the coal mine tunnel walls is as follows:

$$R = 2L^2 + 2L + 1 \tag{12}$$

According to Fresnel's law, the reflection coefficients of vertically polarized waves and horizontally polarized waves can be expressed as follows:

$$R_{\perp} = \frac{\cos\theta_i - \sqrt{\varepsilon_r - \sin^2\theta_i}}{\cos\theta_i + \sqrt{\varepsilon_r - \sin^2\theta_i}} \tag{13}$$

$$R_{\|} = \frac{\varepsilon_r \cos\theta_i - \sqrt{\varepsilon_r - \sin^2\theta_i}}{\varepsilon_r \cos\theta_i + \sqrt{\varepsilon_r - \sin^2\theta_i}} \tag{14}$$

where $\boldsymbol{\varepsilon_r}$ is the relative dielectric constant of the coal mine tunnel wall. Assuming that the roughness distribution of the coal mine tunnel wall follows the Gaussian distribution with a mean of 0 and a variance of h, then $\boldsymbol{\rho_r}$ represents the roughness loss factor. The roughness loss coefficient caused by the rough surface of the coal mine wall is as follows:

$$\rho_\perp = \rho_r R_\perp, \rho \| = \rho_r R_{\|} \tag{15}$$

where $\rho_r = \exp[-8(\frac{\pi h \cos\theta_i}{\lambda})^2]$. Multiple reflections of the signal on the two side walls and the top and bottom plates of the tunnel will cumulatively affect the total reflection coefficient. Assuming that the ray is reflected m times at the two side walls and n times at the top and bottom plates, the reflection coefficient at this point can be expressed as follows:

$$\omega_p = R_\perp^m R_{\|}^n \rho_r^{m+n} \tag{16}$$

It can be expressed uniformly:

$$w_p = \prod_{k=1}^{m+n} R_k \rho_r^k \tag{17}$$

where R_k denotes the polarized reflection coefficient corresponding to each reflection.

2.2. Systematic Performance Optimization Problem

The power of the transmitted signal of a multi-antenna BS is constrained by the maximum transmit power:

$$E\left\{Tr(\mathbf{G}\mathbf{G}^H)\right\} \le P_{\max} \tag{18}$$

where $P_{\max}$ is the maximum value of the transmitted power at the BS, E represents the statistical expectation value, and $Tr(\cdot)$ represents the trace of the matrix.

To maximize the total system link rate by optimizing $\mathbf{G}$ and $\boldsymbol{\Phi}$, the corresponding optimization problem is defined as follows:

$$\begin{array}{l} \max\limits_{G,\Phi} \sum\limits_{k=1}^{K} \log_2(1+\beta_k) \\ s.t.\ Tr(\mathbf{G}\mathbf{G}^H) \le P_{\max} \\ |\phi_n|^2 = 1 \\ 0 \le \theta_n < 2\pi, n = \{1, 2, \ldots, N\} \end{array} \tag{19}$$

The optimization problem in Equation (19) is non-convex due to the non-convexity of the objective function and the complexity of the constraints. If the traditional method is used to solve this problem, a large number of iterative calculations are required, resulting in high computational complexity. Therefore, this paper adopts a DRL-based solution to solve the optimization problem in Equation (19) to obtain a feasible $\mathbf{G}$ and $\boldsymbol{\Phi}$.

3. Deep Reinforcement Learning Based Optimization

3.1. Overview of the DDPG Algorithm

To cope with the limitations of traditional reinforcement learning methods in continuous action spaces, the DDPG algorithm combines the expressive capabilities of deep learning with the optimization strategies of policy gradient methods, achieving policy optimization through the actor–critic structure. The framework of the DDPG algorithm is shown in Figure 3. The actor and critic training networks are used to learn to select and evaluate actions, respectively. The actor target network provides stable target strategies for the actor training network, while the critic target network provides stable Q-value for the critic training network and is updated synchronously regularly.

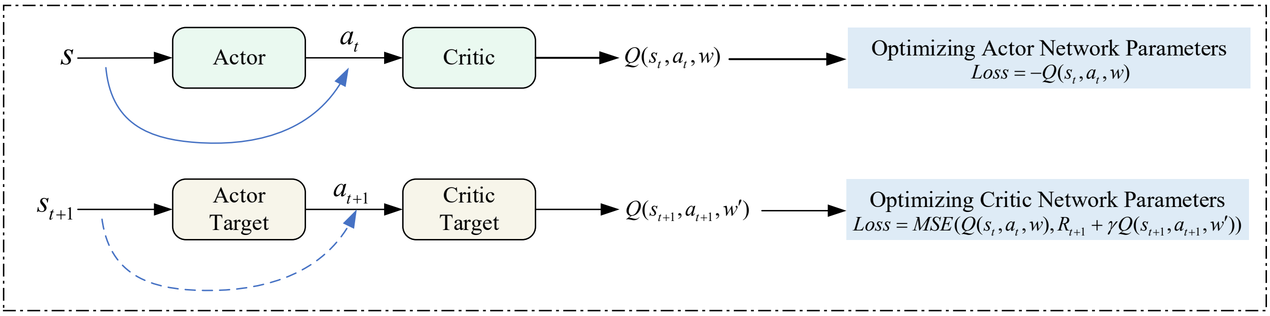

s
Actor
a_t
Critic
$Q(s_t, a_t, w)$
Optimizing Actor Network Parameters
$Loss = -Q(s_t, a_t, w)$
s_{t+1}
Actor Target
a_{t+1}
Critic Target
$Q(s_{t+1}, a_{t+1}, w')$
Optimizing Critic Network Parameters
$Loss = MSE(Q(s_t, a_t, w), R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}, w'))$

Figure 3. DDPG algorithm block diagram.

The action value function Q_π evaluates the expected cumulative reward obtained by taking action a_t in state s_t. The Bellman expectation equation for the Q-value under policy π is expressed as follows:

$$Q_\pi(s_t, a_t) = \mathbb{E}[R_{t+1} + \gamma \cdot \sum_{a'} \pi(a'|s_{t+1}) Q_\pi(s_{t+1}, a')|S_t = s_t, A_t = a_t] \tag{20}$$

where R_{t+1} is the reward obtained after taking action a_t in the state s_t, s_{t+1} is the next state, a' is the possible action to be taken in the next action, and $\pi(a'|s_{t+1})$ is the probability of the policy to take the action a' in the next state.

The optimal policy is determined by maximizing the Q-value:

$$\pi^* = \underset{\pi}{\text{argmax}} Q_\pi(s_t, a_t) \tag{21}$$

To directly evaluate the current state and action without considering the influence of policy π, the optimal action value function $Q^*(s_t, a_t)$ is introduced. For the optimal Q-value function, the Bellman optimality equation is expressed as follows:

$$Q^*(s_t, a_t) = \mathbb{E}[R_{t+1} + \gamma \cdot \max_{a'} Q^*(s_{t+1}, a')|S_t = s_t, A_t = a_t] \tag{22}$$

The $Q^*(s_t, a_t)$ function is updated as follows:

$$Q^*(s_t, a_t) \leftarrow (1 - \alpha) Q^*(s_t, a_t) + \alpha \cdot (R_{t+1} + \gamma \cdot \max_{a'} Q(s_{t+1}, a')) \tag{23}$$

where α is the learning rate of the $Q^*(s_t, a_t)$ function update.

3.2. Neural Network Structure and Training

The structure of the critic network and action network includes an input layer, two hidden layers, a batch normalization layer (Batch Normalization, BN), and an output layer. All layers of deep neural networks use the tanh activation function, and all network optimizers use Adam. The learning rate of Adam will be adaptively adjusted according to the gradient changes of each parameter, i.e., $\beta_{\mathrm{a}}^{(t+1)} = \lambda_{\mathrm{a}} \beta_{\mathrm{a}}^{(t)}$ and $\beta_{\mathrm{c}}^{(t+1)} = \lambda_{\mathrm{c}} \beta_{\mathrm{c}}^{(t)}$, where λ_{a} and λ_{c} are the decay rate of the critic network and action network, respectively.

The parameters of the critic training network can be updated according to gradient descent:

$$w_\mathrm{c} = w_\mathrm{c} - \beta_\mathrm{c} \cdot \nabla_{w_\mathrm{c}} l(w_\mathrm{c}) \tag{24}$$

$$y_t = R_{t+1} + \gamma \cdot \max_{a'} q\left(s_{t+1}, a'; w_t\right) \tag{25}$$

$$l(w_\mathrm{c}) = \frac{1}{W} \sum_{k=1}^{W} \left[y_t - q(s_t, a_t; w_\mathrm{c})\right]^2 \tag{26}$$

where W is a mini-batch of size sampled from the experience replay buffer, y_t is the Q-value generated by the transfer tuple $(s_t, a_t, R_{t+1}, s_{t+1})$, β_c is the learning rate of the critic training network update, and $\nabla_{w_\mathrm{c}} l(w_\mathrm{c})$ represents the gradient of the loss function with respect to w_c.

The actor training network parameters are updated using the following equation:

$$w_\mathrm{a} = w_\mathrm{a} - \beta_\mathrm{a} \cdot \nabla_\mathrm{a} q(s_t, a_t; w_\mathrm{c}) \nabla_\mathrm{a} \pi(s_t; w_\mathrm{a}) \tag{27}$$

where β_a is the learning of the actor training network update, $a_t = \pi(s_t, w_\mathrm{a})$.

To maintain the stability of learning, "soft update" is used to smoothly update the parameters of the target network:

$$\begin{gathered} w_{\mathrm{a}'} \leftarrow \tau w_\mathrm{a} + (1-\tau) w_{\mathrm{a}'},\ 0 < \tau \ll 1 \\ w_{\mathrm{c}'} \leftarrow \tau w_\mathrm{c} + (1-\tau) w_{\mathrm{c}'},\ 0 < \tau \ll 1 \end{gathered} \tag{28}$$

where τ represents the learning rate of the target network.

3.3. DDPG Algorithm Elements and Processes

To enhance the stability of training, DDPG incorporates an experience replay mechanism. The experience replay buffer stores samples generated from interactions between the agent and the environment, improving training efficiency by reducing data correlation through random sampling. Each experience usually includes the current state, the action taken, the reward received, the next state, and a termination flag indicating whether the interaction has ended. The specific definitions of state s_t, action a_t, and reward r_t are as follows:

(1) State s_t: Represents the environment, comprising the transmit power P_t at time t, the channel matrix $\mathbf{H}_{nm}$ from the BS to the RIS, and the signal matrix $\mathbf{H}_{ru}$ from the RIS to the user; the size of the state space is as follows:

$$N_i = 2K + 2K^2 + 2MK + 2N + 2MN + 2NK \tag{29}$$

(2) Action a_t: Comprises the beamforming matrix $\mathbf{G}$ and phase shift matrix $\mathbf{\Phi}$ at time t. The size of the action space is as follows:

$$N_o = 2MK + 2N \tag{30}$$

(3) Reward r_t: The reward at the time t corresponds to the value of the objective function defined in Equation (19).

$$r_t = \log_2(1 + \beta_k) \tag{31}$$

The flow of the DDPG algorithm proposed in this paper is as follows:

Algorithm 1 DDPG Algorithm

```
Initialization: transmit beam forming matrix $G$, phase shift matrix $\Phi$, experience replay buffer $E$, parameters of critic training network $w_{\mathrm{c}}$, parameters of actor training network $w_{\mathrm{a}}$, parameters of the target network $w_{\mathrm{c}'}$, $w_{\mathrm{a}'}$
Inputs: channel matrix from BS to RIS $\mathbf{H}_{nm}$, channel matrix from RIS to user $\mathbf{H}_{nu}$
Outputs: Q-value function, optimal action $a = \{G, \Phi\}$

Get the initial state $s_0$ from the environment;
Calculate the action $a_t$ at each moment in turn, $a_t = \pi(s_t, w_{\mathrm{a}})$;
Based on the action $a_t$, interact with the environment to get the next state $s_{t+1}$ and instant reward $R_{t+1}$;
Store the transfer tuple $(s_t, a_t, R_{t+1}, s_{t+1})$ in the experience replay buffer $E$;
Randomly draw a mini-batch of size $W$ from the experience replay buffer $E$;
Calculate the target Q-value according to Equation (25);
Update the parameters $w_{\mathrm{c}}$ and $w_{\mathrm{a}}$ of the critic and actor networks according to Equations (24) and (27) and the target network parameters $w_{\mathrm{c}\,\prime}$, $w_{\mathrm{a}\,\prime}$ according to Equation (28);
Update the state to the next state $s_t = s_{t+1}$.
```

4. Results and Discussions

The performance of the RIS-assisted coal mine communication system based on the DDPG algorithm was evaluated, and it was divided into two scenarios: ideal reflection and non-ideal reflection of RIS. In both scenarios, system performance was influenced by factors such as the learning rate lr, fading rate dr, base station power, and the number of RIS units. The key simulation parameters are shown in Table 1.

Table 1. Simulation parameters.

Parameter	Value
Height of the tunnel $2b$	6 m
Width of the tunnel $2a$	5 m
Buffer size for experience replay E	10^6
The number of experiences in the mini-batch W	16
The number of episodes	$5 \cdot 10^3$
The number of steps in each episode	$2 \cdot 10^4$
Discounted rate for future reward γ	0.999
Learning rate of action/critic networks $\beta_{\mathrm{a}}/\beta_{\mathrm{c}}$	10^{-3}
Targeted action/critic network learning rate $\lambda_{\mathrm{a}}/\lambda_{\mathrm{c}}$	10^{-3}
Action/critic of network decaying rate τ	10^{-6}

In the simulation, the randomly generated channel matrices $\mathbf{H}_{nm}$ and $\mathbf{H}_{nu}$ obey Rayleigh distribution, with the total effective number of scattering paths being $R = 128$. The gain of the receiving and transmitting antennas was $5dBm$, the roughness variance of the coal mine tunnel wall was $h = 0.01$, and the relative permittivity was $\varepsilon_r = 5$. In this paper, we used the average reward as a measure of the performance of the system, which was defined as follows:

$$R_{avg} = \frac{\sum_{t=1}^{T} reward(t)}{t}, t = 1, 2, \ldots, T \tag{32}$$

4.1. Performance Comparison of Different DRL Algorithms

This paper selects three typical DRL algorithms—DDPG, DQN, and A3 C—to explore their effects on link rate optimization in RIS-assisted coal mine tunnel communication systems. The simulation conditions are defined as $P_t = 38dBm, M = K = 8, N = 16$, and RIS ideal reflection.

Figure 4 shows that the link rate optimization performance of the three algorithms is DDPG > A3C > DQN. Specifically, DDPG achieves the highest average link rate (about 6.5 bps/Hz) with superior convergence speed and stability, followed by A3C with a link rate of about 5.5 bps/Hz and DQN with the worst performance with a link rate of about 3.4 bps/Hz. The DDPG algorithm is superior to DQN and A3C, mainly because its "actor–critic" structure is suitable for continuous action space and can more accurately optimize the phase and gain of the RIS reflection unit, thereby increasing the link rate. At the same time, DDPG introduces the target network and experience replay mechanism to enhance the stability of convergence, making communication quality improvement even more significant in complex coal mine environments.

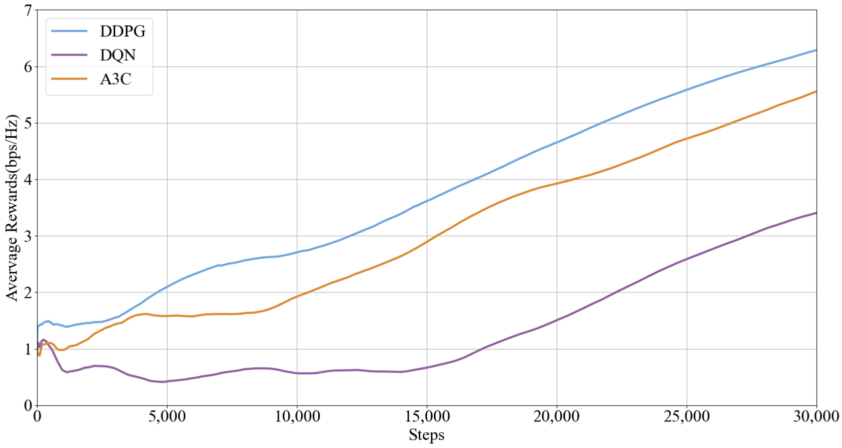


Figure 4. Comparison of average rewards under different algorithms.

4.2. Effect of Neural Network Parameters on Average Reward

4.2.1. Effect of Learning Rate on Average Reward

The effect of different learning rates on the average reward as a function of the number of steps is shown in Figure 5. When *lr* is 0.01, the average reward R_{avg} is 2 bps/Hz. For lower values (e.g., 0.0001 and 0.00001), the final reward value is 3 bps/Hz, and when *lr* is 0.001, R_{avg} reaches 6.6 bps/Hz. It helps the model to achieve the best average reward.

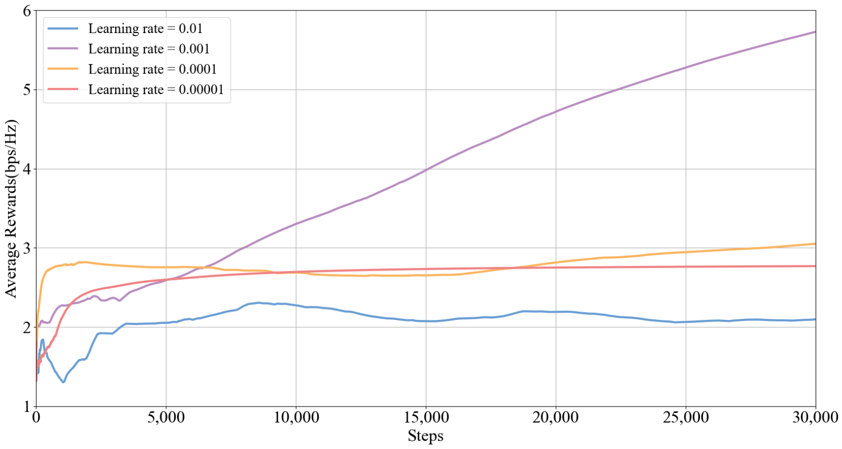


Figure 5. The effect of learning rate on average reward.

4.2.2. Effect of Decaying Rate on Average Reward

The effect of different decaying rates on the average reward as the number of steps changes is shown in Figure 6. As can be seen from the figure, with the accumulation of training steps, R_{avg} shows an upward trend at all decay rates. Specifically, when $dr = 0.000001$, the performance is the best in the entire training process, with the most significant upward trend in R_{avg}, and finally reaches 6.6 bps/Hz. Then, when dr is 0.001 and 0.00001, the reward level is slightly lower than $dr = 0.000001$, which is about 5.7 bps/Hz. Therefore, choosing an appropriate decay rate is crucial to improving the average reward of the model.

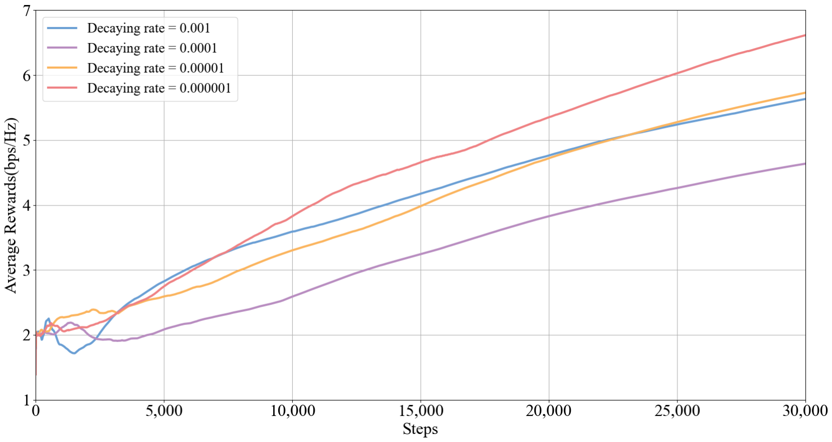


Figure 6. Effect of decay rate on average reward.

4.3. Effect of Base Station Transmit Power on Average Reward

4.3.1. Comparative Analysis of Average Reward at Different Power

The effect of different base station transmit powers on the average reward as the number of steps changes is shown in Figure 7. The simulation conditions are set to $M = N = K = 8$. The average reward under different powers is shown in Table 2. As the transmission power increases from 36 dBm to 38 dBm, the link rate of the communication system is improved regardless of ideal or non-ideal conditions. This shows that increasing the transmit power can significantly enhance the communication capability of the system. Under the ideal reflection condition with 38 dBm power, the rate increases from 2 bps/Hz to 6.7 bps/Hz with an increase of 4.7 bps/Hz, while under the non-ideal condition, the rate increases from 2 bps/Hz to 4.7 bps/Hz with an increase of only 2.7 bps/Hz. In short, RIS technology can significantly improve the performance of communication systems under appropriate transmit power, especially under ideal reflection conditions.

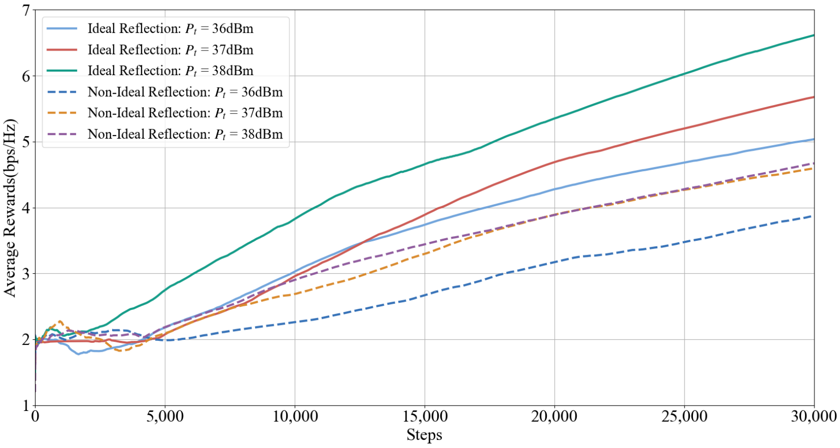


Figure 7. Effect of power on average reward.

Table 2. Average reward changes at different powers.

State	Power (dBm)	Initial Reward (bps/Hz)	Final Reward (bps/Hz)	Difference in Value (bps/Hz)
RIS Ideal Reflection	36	2	5	3
	37	2	5.8	3.8
	38	2	6.7	4.7
RIS Non-Ideal Reflection	36	2	3.9	1.9
	37	2	4.6	2.6
	38	2	4.7	2.7

4.3.2. Comparative Analysis of Instant and Average Rewards

The effect of reward changes with the number of steps under different transmit powers is shown in Figure 8. In terms of instant reward, when the power is 36 dBm, R_{avg} is 8.2 bps/Hz, while when the power is 38 dBm, R_{avg} reaches 10.4 bps/Hz and its fluctuation is also larger, indicating that the system's reward varies more drastically at higher power. The average reward curve is smoother. At a transmit power of 36 dBm, R_{avg} reaches 5 bps/Hz, and at 38 dBm, R_{avg} increases to 6.5 bps/Hz, indicating that increasing the transmission power helps to improve the system performance.

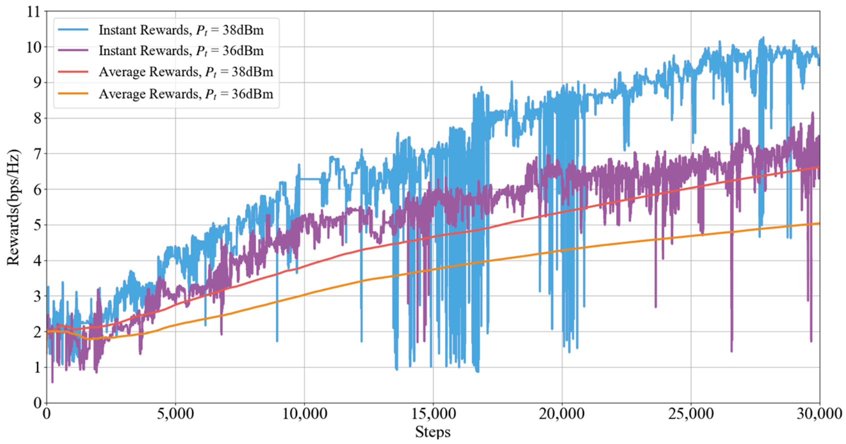


Figure 8. Comparative analysis of instant rewards and average rewards.

4.4. Effect of the Number of RIS Units on Average Rewards

The effect of different numbers of RIS units on the average reward as the number of step changes is shown in Figure 9. The simulation conditions are set to $p_t = 38\text{dBm}, \mu = \pi, \alpha = 1$ with a non-ideal reflection of $\beta_{\text{min}} = 0.5$.

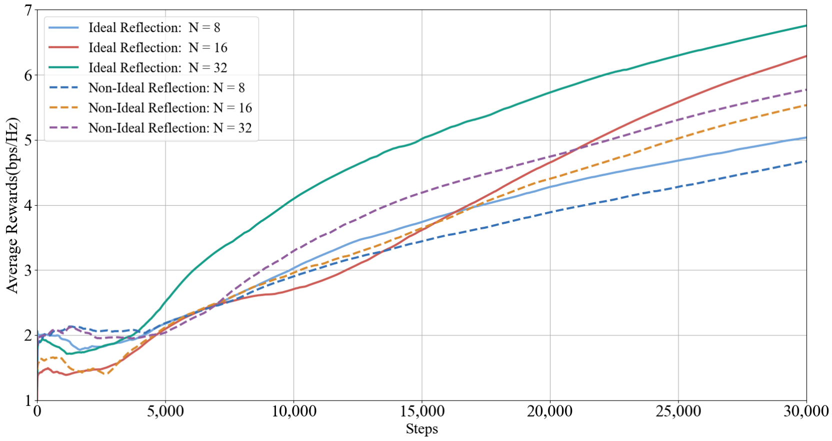


Figure 9. Effect of RIS unit quantity on average reward.

The average rewards under different numbers of RIS units are shown in Table 3. Under ideal and non-ideal reflection conditions, the average rate of the system increases with the increase in the number of RIS units. When the number of RIS units is 32, the link rate is significantly increased under ideal conditions, exhibiting a maximum increase of 4.9 bps/Hz; under non-ideal conditions, the link rate increased from 2 bps/Hz to 5.8 bps/Hz. In addition, the performance gap between ideal and non-ideal conditions increases with the number of RIS units. For example, the rate difference is 0.3 bps/Hz at 8 RIS units, and the difference widens to 1 bps/Hz at 32 RIS units, which is due to the decrease in signal reflection efficiency under non-ideal conditions.

Table 3. Average reward changes under different numbers of RIS units.

State	Number of RIS units N	Initial Reward (bps/Hz)	Final Reward (bps/Hz)	Difference in Value (bps/Hz)
RIS Ideal Reflection	8	2	5	3
	16	1.5	6.2	4.7
	32	1.9	6.8	4.9
RIS Non-Ideal Reflection	8	2	4.7	2.7
	16	1.7	5.5	3.8
	32	1.9	5.8	3.9

Increasing the number of RIS units can significantly increase the link rate, but the performance gain gradually decreases. Specifically, under non-ideal reflection conditions, the link rate increases by about 1.1 bps/Hz when the number of RIS units increases from 8 to 16; it only increases by 0.1 bps/Hz when the number increases from 16 to 32. In addition, an increase in the number of RIS units will lead to an increase in hardware and maintenance costs. Hardware costs increase linearly, and increased system complexity will lead to increased maintenance costs. Therefore, when optimizing the system, performance gains and costs should be balanced to ensure the best balance between overall performance and economic benefits.

5. Conclusions

In this paper, the performance optimization of RIS-assisted wireless communication systems in coal mine tunnels was investigated through simulation and algorithmic analysis. The conclusions are as follows:

(1) Aiming at the complex environment of limited space in the coal mine tunnel, RIS technology is introduced into the coal mine wireless communication system, and combined with the DDPG optimization method, an effective signal enhancement solution is provided;

(2) By jointly optimizing the base station power and RIS phase shift, the link rate is significantly enhanced under both ideal and non-ideal reflection conditions. Simulation results show that the link rate optimization effect of DDPG is better than that of A3C and DQN under the same conditions. At a transmit power of 38 dBm, the DDPG algorithm significantly optimizes the system performance, especially in the ideal reflection condition; the average link rate of 2 bps/Hz is improved to 6.6 bps/Hz. In addition, the specific impact of the number of RIS units on the system performance is explored, and it is found that increasing the number of RIS units can further improve the system performance;

(3) Future research will further explore the application and optimization of the DDPG algorithm and other algorithms in more complex coal mine tunnel systems, especially in coal mine tunnel environments with bends, different frequency bands, and multiple modulation technologies.

Author Contributions: Conceptualization, S.W. and F.W.; methodology, S.W. and F.W.; software, S.W. and F.W.; validation, S.W., F.W. and W.Z.; formal analysis, W.Z. and Y.W.; investigation, S.W., F.W., W.Z. and Y.W.; resources, S.W. and F.W.; data curation, S.W., F.W. and W.Z.; writing—original draft preparation, F.W.; writing—review and editing, S.W.; visualization, F.W., W.Z. and Y.W.; supervision, S.W.; project administration, S.W.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the National Natural Science Foundation of China (NSFC52174197) and the UWB Radar Life Information Feature Extraction and Quantitative Identification for the Mine Drill Hole Rescue project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

Acknowledgments: I would like to express special thanks to Shuqi Wang for his teaching and guidance, which has benefited me greatly.

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

References

1. You, X.H.; Wang, C.X.; Huang, J.; Gao, X.Q.; Zhang, Z.C.; Wang, M.; Huang, Y.M. Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts. *Sci. China Inf. Sci.* **2021**, *64*, 1–74. [CrossRef]
2. Gao, Z.L.; Sun, S.H.; Li, L. Overview of intelligent surface for new-generation mobile communication. *Telecommun. Sci.* **2022**, *38*, 20–35.
3. Li, S.Y.; Yang, R.X.; Yang, L.; Shen, T.Q.; Li, F.F.; Hu, Q.S. Survey of the non-line-of-sight wireless coverage technology by reconfigurable intelligent surfaces in underground coal mines. *J. China Univ. Min. Technol.* **2024**, *53*, 613622.
4. Li, S.Y.; Zhang, P.; Min, M.H.; Li, Z.W.; Zhang, M.D.; Xiao, J.Y. Discussion on intelligent reflecting surface technology and its application in wireless blind spot coverage in coal mines. *J. Mine Autom.* **2023**, *49*, 112–119.
5. Wang, S.Q.; Zhang, W. RIS-Assisted Wireless Channel Characteristic in Coal Mine Tunnel Based on 6G Mobile Communication System. *Prog. Electromagn. Res. C* **2024**, *141*, 13–23. [CrossRef]
6. Xiao, J. *Research on Deep Learning-Based Channel Estimation for Reconfigurable Intelligent Surface-Aided Communication Systems*; Hunan Institute of Science and Technology: Hunan, China, 2023.
7. Wu, Q.; Zhang, R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network. *IEEE Commun. Mag.* **2019**, *58*, 106–112. [CrossRef]
8. Jung, M.; Saad, W.; Debbah, M.; Hong, C.S. On the optimality of reconfigurable intelligent surfaces (RISs): Passive beamforming, modulation, and resource allocation. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 4347–4363. [CrossRef]
9. Xiong, R.J.; Zhang, J.N.; Wang, F.H.; Wang, Z.Y.; Ren, Y.; Liu, J.S.; Lu, J.L.; Wan, K.; Mi, T.B.; Qiu, C.M. Designing reconfigurable intelligent surfaces for wireless communication: A review. *J. Huazhong Univ. of Sci. and Technol. (Nat. Sci. Ed.)* **2023**, *51*, 1–32.
10. Li, X.L. *Ray Tracing Based Channel Modeling for RIS Empowered Indoor Wireless Communication Scenarios*; Shandong University: Shandong, China, 2023.
11. An, J.C. *Channel Estimation and Passive Beamforming for Reconfigurable Intelligent Surface-Assisted Communications*; University of Electronic Science and Technology of China: Chengdu, China, 2022.
12. Niu, H.H.; Lin, Z.; Wang, Y.; Wang, L.; Zhao, Q.C. Weighted sum rate optimization for intelligent reflecting surface-aided wireless network. *J. Natl. Univ. Def. Technol.* **2023**, *45*, 56–63.
13. Mismar, F.B.; Evans, B.L.; Alkhateeb, A. Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination. *IEEE Trans. Commun.* **2019**, *68*, 1581–1592. [CrossRef]
14. Abdalla, A.S.; Marojevic, V. DDPG learning for aerial RIS-assisted MU-MISO communications. In Proceedings of the 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Kyoto, Japan, 12–15 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 701–706.
15. Li, W.S.; Wang, L.; Wei, C. Application and Design of LSTM in Coal Mine Gas Prediction and Early Warning System. *J. Xi'an Univ. Sci. Technol.* **2018**, *38*, 1027–1035.
16. Li, H.; Gao, L.S.; Liu, L.; Shao, K. Application of deep learning in microseismic detection of hydraulic fracturing in coal mine. *J. Xi'an Univ. Sci. Technol.* **2023**, *43*, 686–696.
17. Zhu, Y.; Shi, E.; Liu, Z.; Zhang, J.; Ai, B. Multi-agent Reinforcement Learning-based Joint Precoding and Phase Shift Optimization for RIS-aided Cell-Free Massive MIMO Systems. *IEEE Trans. Veh. Technol.* **2024**, *73*, 14015–14020. [CrossRef]
18. Thanh, P.D.; Giang, H.T.H.; Hong, I.P. Anti-jamming RIS communications using DQN-based algorithm. *IEEE Access* **2022**, *10*, 28422–28433. [CrossRef]

19. Li, Z.J.; Wu, Y.J.; Jin, S.; Zhong, X.H. DQN based downlink-beamforming for millimeter wave MISO system in mobile environment. *J. South-Cent. Univ. Natl. (Nat. Sci. Ed.)* **2021**, *40*, 278–285.
20. Tan, H. Reinforcement learning with deep deterministic policy gradient. In Proceedings of the 2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Xi'an, China, 28–30 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 82–85.
21. Chen, L.; Liang, C.; Zhang, J.Y.; Liu, Y.T. A multi-agent reinforcement learning algorithm based on improved DDPG in Actor-Critic framework. *Control. Decis.* **2021**, *36*, 75–82.
22. Xu, Y.H.; Yang, C.C.; Hua, M.; Zhou, W. Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications. *IEEE Access* **2020**, *8*, 18797–18807. [CrossRef]
23. Huang, C.; Mo, R.; Yuen, C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 1839–1850. [CrossRef]
24. Saglam, B.; Gurgunoglu, D.; Kozat, S.S. Deep reinforcement learning based joint downlink beamforming and RIS configuration in RIS-aided MU-MISO systems under hardware impairments and imperfect CSI. In Proceedings of the 2023 IEEE International Conference on Communications Workshops (ICC Workshops), Rome, Italy, 28 May–1 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 66–72.
25. Ya, H.Y.; Wan, H.B.; Qin, T.F. Research on beamforming algorithms incoporating reconfigurable intelligent surface and deep reinforcement learning. *J. Chin. Comput. Syst.* **2024**, *45*, 1311–1317.
26. 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]
27. Feng, K.; Wang, Q.; Li, X.; Wen, C.K. Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 745–749. [CrossRef]
28. Xu, W.Y. *Deep Learning-Based Channel Estimation and Beamforming for RIS-Assisted Communication Systems*; University of Electronic Science and Technology of China: Chengdu, China, 2023.
29. Yang, Z.; Liu, Y.; Chen, Y.; Zhou, J.T. Deep reinforcement learning for RIS-aided non-orthogonal multiple access downlink networks. In Proceedings of the GLOBECOM 2020-2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
30. Yao, J.C.; Xu, W.; Huang, Y.M.; Xiao, H.H.; Lu, Z.H. Techniques for Reconfigurable Intelligent Surface-Aided 6G Communication Network: An Overview. *J. Signal Process.* **2022**, *38*, 1555–1567.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.