Deep Reinforcement Learning for RIS-Aided Multiuser Full-Duplex Secure Communications With Hardware Impairments

Zhangjie Peng, Zhibo Zhang, Lei Kong, Cunhua Pan, Member, IEEE, Li Li, and Jiangzhou Wang, Fellow, IEEE

Abstract—In this article, we investigate a reconfigurable intelligent surface (RIS)-aided multiuser full-duplex secure communication system with hardware impairments at transceivers and RIS, where multiple eavesdroppers overhear the two-way transmitted signals simultaneously, and a RIS is applied to enhance the secrecy performance. Aiming at maximizing the sum secrecy rate (SSR), a joint optimization problem of the transmit beamforming at the base station (BS) and the reflecting beamforming at the RIS is formulated under the transmit power constraint of the BS and the unit modulus constraint of the phase shifters. As the environment is time varying and the system is high dimensional, a nonconvex optimization problem is mathematically intractable. A deep reinforcement learning (DRL)-based algorithm is explored to obtain the satisfactory solution by repeatedly interacting with and learning from the dynamic environment. Extensive simulation results illustrate that the DRL-based secure beamforming algorithm is proved to be significantly effective in improving the SSR. It is also found that the performance of the DRL-based method can be greatly improved and the convergence speed of the neural network can be accelerated with appropriate neural network parameters.

Index Terms—Deep reinforcement learning (DRL), full duplex (FD), hardware impairment, reconfigurable intelligent surface (RIS), secure communication.

I. INTRODUCTION

RECONFIGURABLE intelligent surface (RIS), also known as intelligent reflecting surface (IRS), is one of the most promising and disruptive technologies in the sixth generation (6G) and beyond wireless communication [1]–[4]. The RIS is a uniform planar array made of numerous low-cost and nearly passive components. Every unit can reflect the incident electromagnetic wave independently to the desired direction by dynamically adjusting the phase shifts or amplitudes with the controller [5]. RIS can be expediently installed on ceilings or billboards to establish a virtual Line-of-Sight (LoS) link between the source and the destination when the direct link is blocked by obstacles [3], [6].

Due to the inherent broadcast nature of wireless communication, the transmitted information is easy to be eavesdropped. Thus, physical-layer security (PLS) is an indispensable part of the wireless communication, which attracts a lot of attention. Recently, RIS has been introduced into secure communication as an effective way with low-power consumption and anti-eavesdropping. With the assistance of RIS, the reflected signals received by the legitimate users will be constructively enhanced while the signals leaked to the eavesdroppers will be destructively weakened. Extensive research attention has been devoted to RIS-assisted PLS. The initial contributions were based on the simple system model with only one legitimate user in the presence of an eavesdropper [7]–[9]. While in [10] and [11], a legitimate user was wiretapped by multiple eavesdroppers. Different from these contributions considering a single legitimate user [7]–[11], a RIS-aided secure system, including multiple legitimate users and eavesdroppers, was considered in [12] and [13].

Most of the existing literature on RIS-assisted PLS [7]–[13] considered half-duplex (HD) communications, in which the signals cannot be transmitted and received at the same time and over the same frequency. However, it is less difficult for an eavesdropper to decode the desired signal from the eavesdropped signal in an HD communication system [14]. The development of full-duplex (FD) technology in wireless communication provides the possibility of enhancing PLS [14]. Meanwhile, FD communications enable the transceivers to exchange information simultaneously in the same frequency band, which can effectively improve spectrum efficiency with the development of residual self-interference (SI) cancelation...
technology. Recently, RIS-assisted FD systems were studied in [15] and [16], which highlighted the superiority of RIS-aided FD communications under proper residual SI cancellation. Furthermore, in RIS-aided FD secure communication systems, the bidirectional signals can be simultaneously received and transmitted in the same frequency band, which can theoretically lead to the overlap of multiple signals at the eavesdroppers to improve the secrecy performance [17]. Specifically, when the uplink/downlink signals are eavesdropped, the downlink/uplink signals reflected by the RIS will be reused as interference noise to degrade the eavesdropping performance of the eavesdroppers. To our knowledge, the potential of deploying RIS to secure FD communications has rarely been explored [17], [18]. The closed-form expression for the sum secrecy rate (SSR) of a RIS-aided multipair two-way communication system was derived in [17]. In [18], the SSR was maximized by optimizing the transmit power and the phase shifts. However, there is still a lack of research work on RIS-assisted multiuser FD secure communications.

It is worth noting that the above contributions on RIS-assisted secure communications [7]–[13], [17], [18] were based on the assumption that the hardware is perfect. However, transceiver hardware impairment (T-HWI) is generally inevitable considering the nonlinearities and phase noise, etc. [19], [20]. Even if the mitigation/compensation algorithm was proposed, T-HWI cannot be completely eliminated [19]. A robust beamforming designed algorithm was proposed in [21] to maximize the received signal-to-noise ratio (SNR) considering the impact of T-HWI. The robust transmission optimization of a RIS-assisted secure communication system with T-HWI was studied for the first time in [22]. At the same time, the high-precision configuration of the phase shifts at the RIS is not feasible [23], thus there is also the HWI at the RIS (RIS-HWI). The achievable rate expression of the RIS-assisted mmWave communication system was derived, taking into account the RIS-HWI [24]. The impact of RIS-HWI on the performance of a secure communication system was investigated in [25]. However, there were only a few studies that analyzed the RIS-aided communication systems with both the T-HWI and the RIS-HWI. Both analytical and experimental results in [26] showed that RIS-HWI, as well as T-HWI, reduced the achievable rate. The spectral and energy efficiency was analyzed in [27] considering a RIS-aided MISO communication system with both T-HWI and RIS-HWI.

All the above-mentioned contributions [7]–[27] were based on traditional mathematical optimization methods, such as alternating optimization (AO) and block coordinate descent (BCD) method to design the transmit beamforming and/or phase shifts, which are model-based design paradigms that require accurate mathematical models and expert knowledge. However, traditional model-based wireless techniques are hard to meet the demanding requirements of emerging applications in future 6G networks, including excessively complex communication scenarios with unknown channel model and communication scenarios that cannot be described mathematically, such as the nonlinearity due to the inevitable HWI [28]. On the contrary, deep learning (DL) is a powerful technique capable of learning complex interrelationships among variables, especially those that are difficult to accurately describe with mathematical models, which empowers us to design wireless communication systems without the knowledge of precise mathematical models [28]. In addition, dynamic 6G networks can lead to uncertainty in the wireless environment (for example, due to the changing of RIS configuration), which brings great difficulties to real-time sensing. Meanwhile, artificial intelligence (AI) techniques are more robust against the uncertainty in the systems, which can achieve accurate real-time sensing. Therefore, the application of AI to 6G networks can optimize the network structure and improve the communication performance [29]. Inspired by the advantages of model-free AI, extensive research attention has been devoted to apply DL to the phase shifts design [30], [31].

However, DL methods that have been applied in the existing contributions [30], [31] to solve the sophisticated problems are supervised learning. Although supervised learning has shown promising results in wireless problems, such as channel estimation, RIS phase shifts design, and so on, it requires a large number of prior labeled training and test data sets for offline training, which relies on either the AO algorithms or the exhaustive search. As known, the training data set is difficult to obtain in many circumstances. Furthermore, for a RIS-aided communication system, once the number of phase shifters changes, numerous training samples need to be reacquired and the DL network needs to be retrained from scratch, which greatly limits the application of the DL-based method. In the considered sophisticated optimization problem, the artificially labeled training data set is unable to be acquired. On the contrary, as a branch of machine learning (ML), reinforcement learning (RL) has been considered to be an effective approach to deal with complex control tasks, which possesses the characteristics of online learning and sample generation, and does not require the training labels with prior knowledge. Furthermore, deep RL (DRL), which combines the advantages of DL in function fitting and RL in policymaking, has been leveraged to optimize the phase shifts [32]–[34]. Recently, DRL-based methods have also been applied to solve the optimization problem for RIS-aided secure communications [35], [36]. On the other hand, due to the time-varying channels and the high-dimensional system, the proposed nonconvex optimization problem is mathematically intractable. It is challenging to derive the explicit mathematical optimal solution for this problem. Fortunately, no complicated mathematical formulations are necessary for DRL, which can significantly reduce the computational complexity and computing time. Meanwhile, DRL is able to learn the implicit knowledge about the radio channels through the trial-and-error interactions without knowing the channel model and mobility pattern. Therefore, DRL is particularly beneficial to deal with the problem in a high-dimensional wireless communication system where the radio channel varies over time. Motivated by the aforementioned facts, we explore a DRL-based method to address the proposed sophisticated optimization problem.

In this article, we present a RIS-assisted FD two-way wiretap communication system with multiple legitimate users as well as multiple eavesdroppers. Both T-HWI and RIS-HWI are
considered. The transmit beamforming at the base station (BS) 
and phase shifts at the RIS are jointly optimized to maximize 
the SSR. This complex optimization problem is nonconvex and 
the optimal solution is unknown. A novel DRL-based method 
is developed to obtain the satisfactory solution, in contrast 
to numerical optimization algorithms by utilizing complicated 
mathematical formulations. The main contributions of this 
article are summarized as follows.

1) To the best of our knowledge, this is the first attempt 
to study the performance of a RIS-assisted FD secure 
communication system with both T-HWI and RIS-
HWI. In order to improve the secrecy performance, an 
optimization problem for jointly designing the transmit 
and reflecting beamforming is formulated. In the 
complex and time-varying environment, it is a mathematically 
intractable nonconvex optimization problem, which is modeled as a Markov decision process (MDP).

2) A DRL-based secure beamforming algorithm is 
developed to jointly design the transmit and reflecting 
beamforming to maximize the SSR. Specifically, the 
SSR is used as the instant reward to train the DRL 
network. The training data are generated online through 
the trial-and-error interactions between the agent and 
the environment. The network parameters are adjusted 
accordingly, so as to design the transmit and reflecting 
beamforming simultaneously by gradually maximizing 
the SSR through repeated iterations. Since the action 
spase containing the transmit beamforming matrix and 
the phase shifts matrix is continuous, the DDPG tech-
nique is adopted.

3) Dispensing with explicit mathematical calculation, the 
proposed DRL-based algorithm has a standard framework, 
as well as low complexity in its implementation, 
which can be conveniently deployed to the communica-
tion system with different settings. On the other hand, 
compared with the DL-based algorithm that requires 
artificially labeled training data sets, the DRL-based 
algorithm learns the knowledge of the environment to 
adapt.

4) Extensive simulation results illustrate that the proposed 
DRL-based algorithm is significantly effective in 
improving the SSR. Specifically, the agent improves 
its action policy step by step according to the reward 
fed back from the environment, to obtain the satisfac-
tory transmit beamforming and phase shifts to improve 
the SSR. It is also found that the performance of the 
DRL-based method can be greatly improved and 
the convergence speed of the neural network can be 
accelerated with appropriate neural network parameters.

The remainder of this article is organized as follows. In 
Section II, we introduce the RIS-aided FD secure commu-
nication system with HWI and formulate the SSR maximization 
problem. In Section III, we present the proposed DRL-based 
algorithm for joint optimization of the transmit beamforming 
and the phase shifts. In Section IV, we provide exten-
sive simulation results to elaborate the performance of the 
developed DRL-based algorithm. In Section V, we draw a 
brief conclusion.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

Consider a RIS-assisted FD two-way wiretap communica-
tion system, as illustrated in Fig. 1. Both the BS and the 
legitimate users work in the FD mode. The BS transmits 
independent confidential information to K legitimate users, 
while each legitimate user sends a data stream to the BS. 
At the same time, L unauthorized eavesdroppers that are dis-
tributed around the BS and legitimate users are trying to 
wiretap the transmission information from the BS and the 
legitimate users independently. A RIS is applied to enhance the 
secrecy performance. Moreover, due to the existence of many 
obstacles, the signals are only transmitted via the RIS-assisted 
reflecting links. We denote the kth legitimate user as Bk and 
the ℓth eavesdropper as Ek, for k = 1, . . . , K and ℓ = 1, . . . , L. 
Each legitimate user has a single transmit antenna and a 
single receive antenna, while each eavesdropper is equipped with 
one antenna [37]. The BS is equipped with Nt transmit anten-
as and Nr receive antennas. The RIS consists of M passive 
reflecting elements. The phase shift of each reflecting element 
can be dynamically adjusted by a smart controller coordi-
nated with the BS. The phase shifts matrix Θ is denoted by 
Θ = diag(χ1eiθ1, . . . , χmeiθme, . . . , χMeiθM), where χm ∈ [0, 1] 
and θm ∈ [0, 2π] represent the amplitude and phase shift of the mth reflecting element, respectively.

Channel state information (CSI) is crucial for achieving the 
performance gain of the RIS-assisted secure communica-
tion. We suppose that the CSI of all channels is perfectly 
known at the BS before the data transmission. On the one 
hand, the acquisition of CSI for the legitimate users has been 
widely explored [38]–[40]. On the other hand, for the CSI 
of the eavesdroppers, if the eavesdroppers are active in the 
secure transmission system,1 then their CSI acquisition can 
be acquired by regarding them as the legitimate users. If the 

1For example, the eavesdroppers can be disguised as active users in the 
secure communication system trusted by the BS but untrusted by the legitimate 
users.
eavesdroppers are passive, their CSI can also be obtained by some technologies, such as detecting the local oscillator power accidentally leakage from the eavesdropper receivers’ RF front end [41], [42].

According to [43], we assume that the power of the signal reflected twice or more times by the RIS is imperceptible, and, therefore, is ignored. In addition, RIS is designed to maximize the power of the incident signal so that there is no energy loss during reflection.

B. Channel Model

According to the above-mentioned eavesdropper’s CSI acquisition method, we suppose that the BS perfectly knows the eavesdroppers’ CSI. The baseband equivalent channels BS → RIS, RIS → BS, RIS → kth legitimate user B_k, B_k → RIS, the downlink RIS → ℓth eavesdropper E_ℓ, and the uplink RIS → E_ℓ are represented as H_d ∈ ℂ^{M×N_t}, H_r ∈ ℂ^{M×N_r}, h_d,k ∈ ℂ^{M×1}, h_r,k ∈ ℂ^{M×1}, g_d,ℓ ∈ ℂ^{M×1}, and g_r,ℓ ∈ ℂ^{M×1}, respectively, as shown in Fig. 1.

It is assumed that the RIS is installed on the outer wall of a high building with only a few scatters. Therefore, we adopt a Rician fading channel model for all reflecting channels, which is composed of LoS components as well as NLoS components. The channels can be expressed as follows:

$$
h = \sqrt{\alpha_h} \left( \frac{\varepsilon_h}{\varepsilon_h + 1} \bar{h} + \frac{1}{\varepsilon_h + 1} \tilde{h} \right)$$

where h ∈ ℍ = {H_d, H_r, h_d,k, h_r,k, g_d,ℓ, g_r,ℓ}, \( \alpha_h \) denotes the large-scale fading coefficient, and \( \varepsilon_h \) denotes the Rician factor. \( \bar{h} \) is deterministic LoS channel component. \( \tilde{h} \) is the NLoS channel component whose elements are independent and identical distribution (i.i.d.) standard Gaussian random variables following the distribution of \( \mathcal{CN}(0, 1) \).

The BS and RIS are both equipped with uniform linear array (ULA) and the LoS components are modeled as follows:

$$\bar{h} = a_h(r(\theta_h,\text{AoA})a_h(H(\theta_h,\text{AoD}))$$

where \( a_h(r(\theta_h,\text{AoA}))-a_h(H(\theta_h,\text{AoD})) \) are the array responses of \( W_h \)-element/antenna ULA, which are, respectively, given as follows:

$$a_h(r(\theta_h,\text{AoA})) = \left[ 1, \ldots, e^{j2\pi \Delta \theta_h (W_h-1) \sin \theta_h,\text{AoA}} \right]^T$$

$$a_h(H(\theta_h,\text{AoD})) = \left[ 1, \ldots, e^{j2\pi \Delta \theta_h (W_h-1) \sin \theta_h,\text{AoD}} \right]^T$$

where \( W_h \) and \( W_h \) represent the numbers of antennas/elements at the receiver and the transmitter, respectively [16], \( \theta_h,\text{AoA} \) and \( \theta_h,\text{AoD} \) are the Angle of Arrival (AoA) and Angle of Departure (AoD), respectively, \( \lambda_h \) is the interelement spacing at the ULA, and \( \lambda_h \) is the carrier wavelength.

C. Hardware Impairment Model

In practice, due to the nonideality of hardware, the transmitted and received signals are affected by HWI, which generally exists in practical communication systems, including RIS-aided systems with RIS-HWI [26]. In the considered system, there are two kinds of HWI with different mathematical models, including RIS-HWI and T-HWI.

First, RIS-HWI is caused by the inherent hardware imperfections of the reflecting elements or the imperfect channel estimations, which is also called phase noise [23]. The phase noise matrix is expressed as \( \Phi = \text{diag}(e^{j\Delta \theta_0}, \ldots, e^{j\Delta \theta_m}, \ldots, e^{j\Delta \theta_M}) \), where \( \Delta \theta_m \) is random phase noise uniformly distributed within \([-\pi/2, \pi/2]\) for \( m = 1, 2, \ldots, M \), according to [26].

Then, another kind of HWI, namely, T-HWI, is modeled as an independent Gaussian distortion noise, which will cause that the received signal does not match the expected signal or produce distortion on the received signal during signal processing [26]. The distortion power of the transmit antenna increases with the increase of the signal power allocated to the antenna [19].

D. Signal Transmission Model

At the BS side, let \( w_k ∈ ℂ_{N_t \times 1} \) be the beamforming vector for \( B_k \) and \( s_d,k \) be the confidential information sent to \( B_k \). Therefore, the signal transmitted from the BS is modeled as follows:

$$x_d = 3\bar{x}_d + \eta_d^S$$

$$\bar{x}_d = \sum_{k=1}^K w_k s_d,k \quad (5a)$$

$$s_d,k \quad (5b)$$

where \( s_d,k \) is modeled as i.i.d. random variable with zero mean and unit variance, i.e., \( s_d,k \sim \mathcal{CN}(0, 1) \) and \( E[|s_d,k|^2] = 1 \). According to [44], \( \eta_d^S \sim \mathcal{CN}(0, \Upsilon_d^S) \) is the independent Gaussian transmit distortion noise. The power of the distortion noise at each antenna is proportional to the transmit power at that antenna, which can be modeled as \( \eta_d^S \sim \mathcal{CN}(0, \Upsilon_d^S) \), where [21]

$$\Upsilon_d^S = \kappa_d^S \text{diag} \left[ \sum_{k=1}^K w_k w_k^H \right]$$

with \( \kappa_d^S ≥ 0 \) being a scale factor that describes the severity of HWI at the BS’s transmitter, \( \text{diag}[X] \) denotes a diagonal matrix whose diagonal entries are the diagonal elements of matrix \( X \). The transmit power at the BS is subject to the maximum power constraint

$$\text{Tr(WW}^H) ≤ P_{\text{max}}$$

(7)

where \( W \triangleq [w_1, w_2, \ldots, w_K] ∈ ℂ_{N_t \times K} \), and \( P_{\text{max}} \) is maximum transmit power of the BS.

Similarly, at the legitimate users side, the transmitted signal from \( B_k \) is as follows:

$$x_{u,k} = \sqrt{P_{\text{RIS}}} s_{u,k} \quad (8a)$$

$$s_{u,k} = s_{u,k}^{\text{RIS}} + n_{u,k} \quad (8b)$$
where $\tilde{s}_{u,k} \sim CN(0,1)$ represents the independent Gaussian uplink information symbol sent by $B_k$, $P_k$ is the transmit power of $B_k$, $\eta_{u,k}^B$ is an independent Gaussian transmit distortion noise, and $\eta_{u,k}^B \sim CN(0,\kappa_{u,k}^B \mathbb{E}(|\eta_{u,k}^B|^2))$, where $\kappa_{u,k} \geq 0$ represents a scale factor that describes the severity of HWI at the $k$th legitimate user’s transmitter.

For downlink, the received signal at $B_k$ is written as follows:

$$\begin{align*}
y_{B}^k & = \frac{h_{d,k}^H \Phi h_{d,k}}{\sqrt{P_k}} w_{d,i,k} + \sum_{i=1, i\neq k}^K \frac{h_{d,i,k}^H \Phi h_{d,i}}{\sqrt{P_k}} w_{d,i,k} + \sum_{i=1, i\neq k}^K \sqrt{P_k} h_{d,k}^H \Phi h_{d,i} (s_{u,i,k} + \eta_{u,i,k}^B) \\
& + \frac{h_{d,k}^H \Phi h_{d,k}}{\sqrt{P_k}} \eta_{u,k}^B + \sqrt{P_k} \eta_{d,k}^B \\
& + \frac{\sqrt{P_k} h_{d,k}^H \Phi h_{d,i} (s_{u,i,k} + \eta_{u,i,k}^B)}{\sqrt{P_k}} + \frac{\eta_{d,k}^B}{\sqrt{P_k}} + \eta_{d,k}^B \\
& + \sum_{i=1, i\neq k}^K \sqrt{P_k} h_{d,k}^H \Phi h_{d,i} (s_{u,i,k} + \eta_{u,i,k}^B) + \eta_{d,k}^B \\
& + \sum_{i=1, i\neq k}^K \frac{\sqrt{P_k} h_{d,k}^H \Phi h_{d,i} (s_{u,i,k} + \eta_{u,i,k}^B)}{\sqrt{P_k}} + \frac{\eta_{d,k}^B}{\sqrt{P_k}} + \eta_{d,k}^B \tag{9}
\end{align*}$$

where $h_{d,k}$ is the loop channel between the transmit antenna and receive antenna of the $k$th legitimate user $B_k$, $n_d^B$ is additive white Gaussian noise (AWGN) following the distribution of $CN(0, \sigma_d^2)$, and $\eta_{d,k}^B$ is an independent additional distortion noise term at the $k$th legitimate user’s receiver. Similar to [22], we denote the received signal as $\gamma_k^B = \frac{h_{d,k}^H \Phi h_{d,k}}{\sqrt{P_k}} + \eta_{d,k}^B$, then $\eta_{d,k}^B$ follows the distribution of $CN(0, \kappa_{u,k}^B \mathbb{E}(|\eta_{u,k}^B|^2))$, where $\kappa_{u,k}^B$ is a proportionality coefficient that describes the severity of HWI at the $k$th legitimate user’s receiver. The coefficients $\rho_L$ and $\rho_S$ with $0 \leq \rho_L, \rho_S \leq 1$ are the coefficients of loop interference (LI) and SI, respectively. $\rho_L$ and $\rho_S$ are introduced to characterize the LI and SI which cannot be completely eliminated by some interference elimination techniques, respectively. According to (16), LI term and noise term can be combined, the sum of which is denoted as $m_{d,k}$, whose average power is written as $\sigma_{d,k}^2 = \rho_L P_k |h_{d,k}|^2 + \sigma_d^2$. Accordingly, the signal-to-interference-plus-noise ratio (SINR) at $B_k$ is as follows:

$$\text{SINR}_B^k = \frac{|h_{d,k}^H \Phi h_{d,k} w_{d,k}|^2}{\Gamma_k^B(\mathbf{W}, \Theta) + |\eta_{d,k}^B|^2} \tag{10}$$

where $\Gamma_k^B(\mathbf{W}, \Theta)$ is given by (11), shown at the bottom of the page, where the coefficient $\rho$ is defined as

$$\rho = \begin{cases} 
\rho_S, & \text{if } i = k \\
1, & \text{otherwise}
\end{cases}$$

Then, we consider the uplink, the received signal at the BS can be written as follows:

$$\begin{align*}
y_{B}^k & = \frac{H_u^H \Phi h_{u,k} \sqrt{F_k s_{u,k} + H_u^H \Phi h_{u,k} \sqrt{F_k \eta_{u,k}^B}}}{\sqrt{P_k}} + \sum_{i=1, i\neq k}^K \frac{H_u^H \Phi h_{u,i} \sqrt{F_k (s_{u,i} + \eta_{u,i}^B)}}{\sqrt{P_k}} \\
& + \frac{H_u^H \Phi h_{u,k}}{\sqrt{P_k}} \eta_{u,k}^S + \sum_{i=1}^K \frac{H_u^H \Phi h_{u,i} \sqrt{F_k (s_{u,i} + \eta_{u,i}^B)}}{\sqrt{P_k}} + \frac{\eta_{u,k}^S}{\sqrt{P_k}} + n_u^S \tag{12}
\end{align*}$$

where $H_u$ is the loop channel between the transmit antennas and receive antennas of the BS, and $n_u^S$ is an AWGN vector, whose elements are independent and follow the same distribution of $CN(0, \sigma_u^2)$. $\eta_{u,k}^S$ is an independent additional distortion noise term at the BS’s receiver. Similar to the downlink, we express the received signal as $y_B^k = y_B^k + \eta_{u,k}^S$, then $\eta_{u,k}^S$ follows the distribution of $CN(0, \kappa_{u,k}^S \mathbb{E}(|\eta_{u,k}^S|^2))$, where $\kappa_{u,k}^S$ is a proportionality coefficient that describes the severity of HWI at the BS’s receiver. According to [45] and [46], we assume that the BS is equipped with the SI and LI cancelation capabilities, thus LI and SI can be effectively canceled. The residual noise generated in the interference cancelation process can be modeled as i.i.d. AWGN. Therefore, the total noise can be denoted as $m_u \Delta [m_1, m_2, \ldots, m_N]^T$, where $m_i \sim CN(0, \delta_i^2)$, $i = 1, 2, \ldots, N_r$.

Note that the decoding matrix at the BS, which is denoted by $F_k \Delta [f_{u,1}, \ldots, f_{u,k}, \ldots, f_{u,K}] \subset \mathbb{C}^{N_r \times K}$, where $f_{u,k} \subset \mathbb{C}^{N_r \times 1}$ denotes the combining vector at the BS for $B_k$. Therefore, the recovered signal at the BS is obtained by

$$\begin{align*}
y_{u,k}^f & = f_{u,k}^H H_u^H \Phi h_{u,k} \sqrt{F_k s_{u,k} + f_{u,k}^H \eta_{u,k}^S} + f_{u,k}^H \sum_{i=1, i\neq k}^K H_u^H \Phi h_{u,i} \sqrt{F_k (s_{u,i} + \eta_{u,i}^B)} \\
& + f_{u,k}^H H_u^H \Phi h_{u,k} \sqrt{P_k \eta_{u,k}^B} \tag{13}
\end{align*}$$

Accordingly, the SINR of the recovered signal at the BS is obtained by

$$\begin{align*}
\text{SINR}_S^k = \frac{P_k |f_{u,k}^H H_u^H \Phi h_{u,k}|^2}{\Gamma_k^S(\mathbf{W}, \Theta) + |f_{u,k}^H \eta_{u,k}^S|^2 + |f_{u,k}^H m_u|^2} \tag{14}
\end{align*}$$

$$\begin{align*}
\Gamma_k^B(\mathbf{W}, \Theta) = \sum_{i=1, i\neq k}^K |h_{d,i,k}^H \Phi h_{d,i}|^2 + \kappa_{d,k}^B \sum_{i=1, i\neq k}^K |h_{d,i,k}^H \Phi h_{d,i}|^2 + \sum_{i=1}^K (1 + \kappa_{d,i}^B) \rho_k |h_{d,i,k}^H \Phi h_{d,i}|^2 \\
+ \sum_{i=1}^K (1 + \kappa_{d,i}^B) \rho_k |h_{d,i,k}^H \Phi h_{d,i}|^2 \tag{11}
\end{align*}$$
where $\Gamma^S(E, \Theta)$ is given by (15), shown at the bottom of the page.

Since the full knowledge about the eavesdroppers cannot be well acquired by the BS, we make the worst case assumptions that the signal processing capabilities of the eavesdroppers are significantly strong and the hardware of the eavesdroppers are well acquired by the BS, we make the worst case assumptions that the signal processing capabilities of the eavesdroppers are significantly strong and the hardware of the eavesdroppers are.

Correspondingly, the signal eavesdropped by the $\ell$th eavesdropper $E_{\ell}$ can be written as follows:

$$y_{E,\ell}^k = g_{u,\ell}^H \Phi h_{u,k} \sqrt{P_k} s_{u,k} + \sum_{i=1,i \neq k}^K g_{u,i}^H \Phi h_{u,i} \sqrt{P_i} s_{u,i} + \sum_{i=1,i \neq k}^K g_{d,i}^H \Phi h_{d,i} w_{d,k}$$

where $P_{u,\ell}$ is the unit modulus constraint. In the traditional mathematical optimization to maximize $\Gamma^S(E, \Theta)$ under the transmit power constraint of the BS and the unit modulus constraint of the phase shifters. The optimization problem is formulated as follows:

$$\max_{W, \Theta} \Gamma^S(E, \Theta)$$

s.t. $\text{Tr}(WW^H) \leq P_{\text{max}}$}

(25)

It can be seen that Problem (25) is a nonconvex problem because of the nonconvexity of the objective function and the unit modulus constraint. In the traditional mathematical optimization algorithm, the exhaustive algorithm can be used.
to obtain the optimal solution, but it is difficult to be practically implemented due to the large amount of optimization variables in our problem. We may also explore an AO algorithm to develop the suboptimal solution of the optimization problem, like BCD algorithm in [16], but this also requires complex mathematical derivations, and the optimal solution is not guaranteed. Moreover, the AO algorithm cannot be directly applied to solve the considered problem due to its complicated SINR expression. Furthermore, the artificially labeled data set is unable to be acquired for DL technology in dealing with this problem. The aforementioned facts motivate us to adopt the DRL-based method, which is different from other traditional mathematical methods as well as DL-based algorithms. Therefore, we utilize the advanced DRL-based algorithm to solve the joint optimization problem without complex mathematical derivations to achieve a satisfactory transmit beamforming matrix \( W \) and phase shifts matrix \( \Theta \).

III. DRL-BASED JOINT OPTIMIZATION OF TRANSMIT BEAMFORMING AND PHASE SHIFTS

The optimization problem formulated in (25) is mathematically intractable as it is a nonlinear and nonconvex problem. Furthermore, in a practical RIS-assisted FD secure communication system with both T-HWI and RIS-HWI, the hardware quality is unknown, the capabilities of the legitimate users and the channel quality are time varying. The traditional optimization algorithms (such as AO and BCD) can address the single time-slot optimization problem, while they ignore the historical information and long-term benefits of the system and may achieve suboptimal solutions or performances similar to the greedy search. Therefore, it is usually not feasible to apply the traditional optimization techniques to obtain the satisfactory secure beamforming in an uncertain dynamic environment [35].

Model-free RL is a dynamic programming technique which learns the knowledge of the uncertain dynamic environment to deal with decision-making problems. Therefore, the Problem (25) is modeled as an RL problem in this article, and a DRL-based secure beamforming algorithm is developed to jointly design the transmit and reflecting beamforming to maximize the SSR.

A. Optimization Problem Transformation Based on RL

In DRL, the RIS-aided multiuser FD secure communication system is regarded as an environment, and the smart controller coordinating with the BS is assumed to be an agent, which can collect CSI, such as \( H_{\text{u}} \) and all other channels. The agent interacts with the environment in discrete time slots. In time slot \( t \), the agent gets the current state \( s_t \) of the environment and chooses an action \( a_t \) based on the policy \( \pi(s_t, a_t) \). After receiving the action \( a_t \), the environment will update the state to \( s_{t+1} \) and feed back a reward \( r_t \), which represents the performance of the action \( a_t \) under current state \( s_t \), and then the agent chooses action \( a_{t+1} \) for the new state \( s_{t+1} \). The agent exploits feedback to learn the policy to maximize cumulative rewards. The interaction process between the agent and the environment can be modeled as an MDP, which consists of a tuple denoted by \( (S, A, P, r, \gamma) \). The key elements of the proposed DRL-based algorithm are defined in the following.

1) **State Space**: \( s_t \in S \) represents the state observed from the environment at time slot \( t \). The state \( s_t \) at time slot \( t \) is composed of the rate part at the \((t-1)\)th time slot, the cascaded channel part at the \(t\)th time slot, the phase noise part at the \(t\)th time slot, the action at the \((t-1)\)th time slot, the transmit power of the BS, and the received power of the legitimate users at the \((t-1)\)th time slot [33]. The rate part consists of the sum rate at the legitimate users, the sum rate at the BS, the sum rate at the eavesdroppers, and the SSR. The cascaded channel part contains the BS-RIS- legitimate users channel \( G_{\text{ld}} = h_{\text{d}}^H \Phi H_{\text{u}} W \in \mathbb{C}^{K \times K} \), the BS-RIS-eavesdroppers channel \( G_{\text{md}} = g_{\text{d}}^H \Phi h_{\text{u}} W \in \mathbb{C}^{L \times K} \), the legitimate users-RIS channel \( G_{\text{lu}} = H_{\text{lu}}^H \Phi H_{\text{u}} \in \mathbb{C}^{K \times K} \), and the legitimate users-RIS-eavesdroppers channel \( G_{\text{mu}} = g_{\text{u}}^H \Phi h_{\text{d}} \in \mathbb{C}^{L \times K} \). The phase noise part consists of the random phase noise \( \Delta_{\text{ph}} \). In order to reduce the dimension of the state space and the computational complexity, we only take the cascaded channels as the inputs instead of all the channels. Note that neural network is only capable of taking real number as input instead of complex number, so if there is complex number in state \( s_t \), which needs to be divided into real part and imaginary part as independent input items. In total, there are \( 2K^2 + 4LK + 2N_rK \) entries of state \( s_t \) made up of the cascaded channel part. The transmit power for \( B_k \) is given by \( \| w_k \|^2 = |R_i \{ w_k H_{\text{u}} w_k \} |^2 + |R_i \{ w_k^H w_k \} |^2 \). The received power for \( B_k \) is given by \( |G_{\text{ld},k}|^2 = |R_i \{ G_{\text{ld},k} \} |^2 + |R_i \{ G_{\text{md},k} \} |^2 \). The number of entries of the action at the \((t-1)\)th time slot is \( 2M+2N_rK \). To sum up, the dimension of the state space is \( D_s = 4+2K^2+4L+2N_rK + 4K + 3M + 2N_rK \).

2) **Action Space**: Action space includes the transmit beamforming matrix \( W \) and the phase shifts matrix \( \Theta \). The diagonal elements of phase shifts matrix is regarded as entries of the action. Specifically, the real part and imaginary part, \( \Theta = \Re(\Theta) + \Im(\Theta) \) and \( W = \Re(W) + \Im(W) \) make up the action. Then, the dimension of the action space is \( D_a = 2M + 2N_rK \). Therefore, the action vector \( a_t \) is defined as follows:

\[
a_t = \left[ \begin{array}{c} w_k^{(t)} \\ \Theta^{(t)}_{m,m} \end{array} \right]_{k \in K, m \in M}.
\]

The action \( a_t \) will be reformed into a transmit beamforming matrix \( W^{(t)} \) and a RIS phase shifts matrix \( \Theta^{(t)} \).

3) **State Transition Function**: \( P \) is the transition probability matrix, where \( P(s_t, a_t, s_{t+1}) \in [0, 1] \) represents the probability that state \( s_t \) changes to the new state \( s_{t+1} \) when the agent chooses an action \( a_t \).

4) **Reward Function**: The objective of the optimization problem is to maximize the SSR. Thus, the reward function is defined as follows:

\[
r_t = C^{(t)}. \tag{27}
\]

5) **Discount Factor**: \( \gamma \in [0, 1] \) is applied to discount future rewards and represents the uncertainty of future rewards.
6) Policy: The policy \[ \pi(s_t, a_t) \] denotes probability distribution of choosing an action \[ a_t \] over the state \[ s_t \], which satisfies \[ \sum_{a \in A} \pi(s_t, a_t) = 1 \].

B. DDPG-Based Joint Transmit and Reflecting Beamforming Optimization

In DRL, whether transition probability is known or not is a big difference between model-based and model-free algorithms. The proposed multiuser FD PLS problem is built as an MDP model, once the transition probability set is given, the problem can be effectively solved by dynamic programming techniques [49]. However, this is not the case, as the transition probability set is difficult to obtain in most cases. Therefore, some standard RL algorithms, such as Q-learning, are considered to be instrumental in solving the model-free problem, in which the transition probability is not required. However, Q-learning is unable to deal with the high-dimensional input problem and the continuous state–action space. To put it in practical terms, the channel matrix and beamforming matrix in the considered RIS-aided FD secure communication system, which are difficult to enumerate or even to discretize.

In order to overcome the shortcomings of tabular RL algorithms as well as to tackle the continuous action and state space, we proposed the DDPG-based algorithm, which is based on the actor–critic framework and shown in Fig. 2, to jointly design the BS transmit beamforming and the phase shifts for the RIS-aided FD secure communication system. There are four neural networks in the proposed DDPG-based secure beamforming algorithm, named actor network, target actor network, critic network, and target critic network, the function of each network is listed as follows [50].

1) Actor Network: It is also known as policy network, which takes state \( s_t \) as input, outputs an continuous action \( a_t \), and updates network parameter \( \theta_\mu \). The actor network is trained by (37), which aims to maximize the state-value function. The update on the actor network is expressed as follows:

\[
\theta_\mu = \theta_\mu - \alpha_\mu \nabla Q(\theta_\mu|s_t, a_t) \nabla \pi(\theta_\mu|s_t) \tag{28}
\]

where \( \alpha_\mu \) is the learning rate for the actor network, \( \nabla (\cdot) \) denotes the gradient, and \( \theta_\mu \) is the parameter of the target critic network.

2) Target Actor Network: It is parameterized by \( \theta_\mu \), and \( \theta_\mu \) is periodically soft updated by

\[
\theta_\mu \leftarrow \beta_\mu \theta_\mu + (1 - \beta_\mu) \theta_\mu \tag{29}
\]

where \( \beta_\mu \) is the learning rate for the target actor network.

3) Critic Network: It is also called Q-network, parameterized by \( \theta_c \), takes state \( s \) and action \( a \) as input, and outputs the Q-function defined in (35). The loss function \( L(\theta) \) is a difference between the output Q-value of the target neural network and the predicted neural network, therefore, the loss function is defined by

\[
L(\theta_c) = (y - Q(\theta_c|s_t, a_t))^2 \tag{30}
\]

where \( y \) is the target value, which is estimated by

\[
y = r_{t+1} + \gamma \max_{a_{t+1}} Q(\theta_c|s_{t+1}, a_{t+1}). \tag{31}
\]

The gradient update of the loss function (30) with respect to \( \theta_c \), is updated by

\[
\theta_c = \theta_c - \alpha_c \nabla L(\theta_c) \tag{32}
\]

where \( \alpha_c \) is the learning rate for the critic network.

4) Target Critic Network: It is parameterized by \( \theta_c \), and outputs \( Q_\tau(\theta_c|s_{t+1}, a_{t+1}) \), \( \theta_c \), is soft updated by

\[
\theta_c \leftarrow \beta_c \theta_c + (1 - \beta_c) \theta_c \tag{33}
\]

where \( \beta_c \) is the learning rate for the target critic network.

In the process of interaction between agent and environment, the goal of the agent is to search the optimal policy \( \pi^* \) to maximize the long-term expected discounted reward, and the cumulative discounted reward function is defined as follows:

\[
R = \sum_{\tau=0}^{\infty} \gamma^\tau r_{\tau+1} \tag{34}
\]

where \( \tau \) is the number of time slots. When given a certain policy \( \pi \) and a state–action pair \( (s_t, a_t) \), the cumulative discounted reward can be approximately calculated by

\[
Q_\tau(s_t, a_t) = \mathbb{E}_Q[R_t|s_t = s_t, a_t = a_t] \tag{35}
\]

which is called the action-value function and is also known as the Q-function. It can be updated by the Bellman expectation equation [35]. The optimal Q-function in (35) can be solved by the Bellman optimality equation as follows:

\[
Q^*(s_t, a_t) = r_t + \gamma \max_{a_{t+1} \in A} Q^*(s_{t+1}, a_{t+1}) \tag{36}
\]

the corresponding optimal action \( a^*_t \) can be obtained by

\[
a^*_t = \arg \max_{a \in A} Q^*(s, a). \tag{37}
\]
Algorithm 1 DDPG-Based Joint Optimization of Transmit Beamforming and Phase Shifts

**Input:** All the channels, $H_d$, $H_a$, $h_{d,k}$, $h_{a,k}$, $g_{d,\ell}$, $g_{a,\ell}$; the coefficients of HWI, $k_{d,k}^B$, $k_{a,k}^B$, $k_{d,k}^S$, $k_{a,k}^S$.

**Output:** The optimal phase shifts matrix $\Theta^*$, the optimal transmit beamforming matrix $W^*$, the maximized SSR $C^*$ under current channel state.

**Initialize:** actor network parameter $\theta_a$, target actor network parameter $\theta_{a\mu}$, critic network parameter $\theta_c$, target critic network parameter $\theta_{c\mu}$, learning rate $\alpha_a$, $\alpha_c$, $\beta_a$ and $\beta_c$, mini-batch size $H$, experience replay memory $M$ with size $D$.

1. for episode $= 1, 2, \ldots, N^{epi}$ do
2. Initialize the exploration noise $N'$;
3. Initialize the phase shifts matrix $\Theta_0$ and the beamforming matrix $W_0$;
4. Obtain the initial state $s_0$;
5. for time slot $t = 1, 2, \ldots, T$ do
6. Obtain action $a_t$ from the actor network;
7. Add exploration noise to $a_t$ as $\tilde{a}_t = a_t + N'$;
8. Calculate the instant reward $r_t$ using (27);
9. Observe the new state $s_{t+1}$;
10. Store transition $\{s_t, a_t, r_t, s_{t+1}\}$ into $M$;
11. Sample a $H$ mini-batch transitions from $M$;
12. Set target function $y$ according to (31);
13. Minimize the loss function $L(\theta_c)$ by equation (30);
14. Update the actor network $\theta_{a\mu}$ by (28) and the critic network $\theta_c$ by (32);
15. Update the target actor network $\theta_{a\mu\mu}$ by (29) and the target critic network $\theta_{c\mu}$ by (33);
16. Update the state $s_t = s_{t+1}$;
17. Reduce the exploration noise $N'$;
18. end for
19. end for

The details of the proposed DRL-based method for joint optimization of the transmit beamforming and the phase shifts are shown in Algorithm 1. The algorithm runs over $N^{epi}$ episodes and each episode iterates $T$ steps. At the beginning of each episode, the parameters of the four networks are initialized, including $\theta_a$ in the actor network, $\theta_{a\mu}$ in the target actor network, $\theta_c$ in the critic network, and $\theta_{c\mu}$ in the target critic network, all of which are uniformly distributed. In addition, the experience replay memory $M$ with size $D$ will be emptied. For the sake of encouraging the agent to fully explore the environment, exploration noise $N'$ is added to the output of the actor network. It implies that the action to be chosen for state $s$ is given by $\tilde{a}_t = a_t + N'$, where $a_t$ is the output of the actor network at time slot $t$, and the exploration noise $N'$ can be chosen appropriately to suit the environment and will decrease with the increase of the number of iterations. We adopt the identity matrix to initialize the action $a_0$, including the transmit beamforming matrix $W_0$ and the phase shifts matrix $\Theta_0$. In addition, all the channels and the exploration noise $N'$ will be initialized at the beginning of each episode.

All the proposed neural network structures are based on a fully connected layer. The actor network structure is composed of five layers, including an input layer, an output layer, and three hidden layers. The critic network shares almost the same network structure as the actor network. The input and output dimensions of the actor network are equal to state and action dimensions, respectively, and the input dimension of the critic network is equal to the sum of state and action dimensions. Due to the negative inputs, the activation function $\tanh$ is utilized for the network. The optimizer applied for the network is Adam, which is with adaptive learning rate. More details are shown in Table II. Furthermore, the correlation between input data will adversely affect the learning of neural networks. In order to reduce the correlation among the data in the state, the state needs to go through the whitening process before being input into the actor network and critic network [33], which is conducive to the learning of the neural network.

Note that the transmit beamforming matrix $W$ and the phase shifts matrix $\Theta$ need to meet the power constraint defined in (7) and the unit modulus constraints, respectively. In order to satisfy these two constraints, we add a batch normalization layer after the output layer of actor network, where $\text{Tr}(WW^H) \leq P_{\text{max}}$ and the phase shifts satisfy $|\chi_m\theta_m^\ast| = 1$, the later constraint means the transmission signal will change the direction without power loss when going through RIS. Specifically, to satisfy the BS maximum power constraint, the following projection operator is applied:

$$\prod\{W\} = \left\{ \begin{array}{ll} W, & \text{Tr}(WW^H) \leq P_{\text{max}} \\ \frac{W}{\sqrt{\text{Tr}(WW^H)}}, & \text{otherwise} \end{array} \right.$$

(38)

After the normalization operation, we can obtain and reformulate the normalized transmit beamforming $W_{\text{nor}} = \prod\{W\}$, and then $\text{Tr}(W_{\text{nor}}W_{\text{nor}}^H)^\ast \leq P_{\text{max}}$, which always satisfies the BS maximum power constraint in Problem (25).

Our objective is to take advantage of DRL to find satisfactory $W$ and $\Theta$ by maximizing the SSR under given a particular CSI, according to [33], rather than offline training and online deployment, such as those in [35]. In the presented DRL-based algorithm, each CSI is utilized to construct the state, then the DDPG algorithm is run to find the optimal solutions. In this article, the optimal transmit beamforming matrix $W^*$ and the phase shifts matrix $\Theta^*$ are obtained by the action with the largest instant reward.

IV. SIMULATION RESULTS

In this section, extensive simulation results are presented to evaluate the performance of the DRL-based algorithm in the RIS-aided multiuser FD secure communication system with multiple eavesdroppers.

A. Simulation Setup

Fig. 3 depicts a 2-D plane of the proposed RIS-assisted FD secure communication system, the default unit in the figure is meter (m). As shown in the figure, $K$ legitimate users and $L$ eavesdroppers are randomly distributed in the light blue square in the right half of the figure, with the size of $100\ m \times 100\ m$. The BS and RIS are located at $(0, 0)$ and $(20, 100)$, respectively.

Unless stated otherwise, the following parameters are employed in the simulations. The numbers of antennas at the
Fig. 3. Simulation setup for the RIS-aided multiuser FD two-way communication system with multieavesdropper.

Fig. 4. Average reward and instant reward performances versus time steps at $P_{\text{max}} = 20$ dBm and $P_{\text{max}} = 10$ dBm, respectively.

The other parameters are given in Table I.

**TABLE I**

| Parameters | Description | Settings |
|------------|-------------|----------|
| $P_{\text{max}}$ | transmit power at BS | 10 dBm ~ 40 dBm |
| $P_k$ | transmit power at users | 100 mW |
| $\alpha$ | path loss exponent | 2.0 |
| $\varepsilon_h$ | Rician factor | 10 |
| $BW$ | channel bandwidth | 10 MHz |
| $\delta^2$ | noise power density | $-174$ dBm/Hz |
| $\rho_S$ | SI coefficient | 1 |

**TABLE II**

| Network | Layer | Neurons | Activation |
|---------|-------|---------|------------|
| critic network | Input | $D_s + D_a$ | / |
| | Hidden 1 | 128 | ReLU |
| | Hidden 2 | 128 | ReLU |
| | Output | 1 | ReLU |
| actor network | Input | $D_s$ | / |
| | Hidden 1 | 128 | ReLU |
| | Hidden 2 | 128 | ReLU |
| | Hidden 3 | 128 | ReLU |
| | Output | $D_a$ | tanh |

**B. Convergence**

Before showing the simulation results, the convergence performance of the presented DDPG algorithm is verified. Consider the nonconvexity of Problem (25), different initial points may lead to different local optimal solutions of the developed algorithm. Fig. 5 investigates the impact of different initializations on the performance of the proposed algorithm. The initialization strategy of Algorithm 1 is that the identity matrix is adopted to initialize the action $a_0$, including the transmit beamforming matrix $W_0$ and the phase shifts matrix $\Theta_0$. “Random initialization” denotes that the initial point of Algorithm 1 is randomly selected under the constraints in Problem (25). As can be seen from the figure that although the initial points of each curve are different, the convergence performance of the proposed algorithm, instant reward and average reward are considered, among which the average reward is defined as follows [51]:

$$r_{\text{Average}}^{t+1} = \sigma r_t^{\text{Average}} + (1 - \sigma)r_t^{\text{Instant}}$$

where $\sigma$ is the smoothing factor, and $t$ is the time step. In this article, the reward function represents the optimization objective, and our objective is to maximize the SSR. Thus, the reward function is defined as the SSR. The convergence behavior of the DDPG algorithm under different BS transmit power is shown in Fig. 4. It is obvious from the figure that both the instant rewards and the average rewards will converge with the increase of time step $t$. This result indicates that the DRL-based algorithm explores the environment through interactions between the agent and environment and then learns from the exploration experience to converge at about $t = 8000$ steps and obtains a satisfactory solution.

BS are set to $N_t = N_r = 4$, the number of the RIS reflecting elements is set to $M = 8$, the number of the legitimate users is set to $K = 2$, and the number of the eavesdroppers is set to $L = 2$. The large-scale path loss in (1) is defined by $PL = PL_0 - 10 \alpha \log_{10}(d/d_0)$, where $PL_0 = -30$ dB is the path loss at $d_0 = 1$ m, and $d$ is the distance between the transmitter and the receiver. All channels follow the Rician distribution. Let $\delta_{u,k}^2 = 1.1\delta_u^2$, $\sigma_{d,k}^2 = 1.1\sigma_d^2$, the coefficients of HWI $\kappa_u^B = \kappa_u^S = \kappa_d^B = \kappa_d^S = \kappa = 0.01$, and the random phase errors are uniformly distributed within $[-\pi/2, \pi/2]$. The other parameters are given in Table I.

In the proposed DRL algorithm, the discount factor is set to $\gamma = 0.9$, the learning rate is set to $\alpha_{\mu} = 0.0005$, $\alpha_c = 0.001$, the batch size of the minibatch is set to $H = 128$, the experience replay capacity is set to $D = 100,000$, and the number of steps in each episode is set to $T = 20,000$. Other neural network parameter settings, such as activation function, are given in Table II.

**TABLE II**

| Network | Layer | Neurons | Activation |
|---------|-------|---------|------------|
| critic network | Input | $D_s + D_a$ | / |
| | Hidden 1 | 128 | ReLU |
| | Hidden 2 | 128 | ReLU |
| | Output | 1 | ReLU |
| actor network | Input | $D_s$ | / |
| | Hidden 1 | 128 | ReLU |
| | Hidden 2 | 128 | ReLU |
| | Hidden 3 | 128 | ReLU |
| | Output | $D_a$ | tanh |
points are almost the same. First, it shows that the initialization strategy adopted by the proposed algorithm is a good choice. In addition, it also proved that the proposed algorithm is robust to the initial points. In other words, the proposed DRL-based algorithm can converge to a satisfactory solution that is almost the same even though the initial points are different.

**C. Comparisons With Benchmarks**

Fig. 6 illustrates the SSR versus $P_{\text{max}}$ under several different schemes. We consider six schemes, “Proposed DRL method, FD” and “Proposed DRL method, HD” denote Algorithm 1 with FD and HD, i.e., $P_k = 0$ mW. For the sake of comparison, classical TD3 [52] is applied to the joint design of transmit beamforming and RIS phase shifts as the performance benchmark, which are denoted by “Classical TD3, FD” and “Classical TD3, HD.” Moreover, “Fixed phase RIS, FD” and “Fixed phase RIS, HD” denote that the phase shift for each reflecting unit is fixed while the transmit beamforming is optimized by Algorithm 1, which serves as another performance benchmark. From the figure, first, the proposed DRL scheme achieves much higher SSR than both classical TD3 and fixed phase RIS schemes. Furthermore, with the increase of BS maximum transmit power $P_{\text{max}}$, the SSRs of all schemes increase as expected, either in the FD mode or HD mode.

Fig. 7 shows the SSR performance versus $\kappa$ under different schemes. It is obvious from the figure that the performance of various schemes deteriorates with the increase of $\kappa$, owing to the fact that the adverse effect of hardware impairments is more prominent. In addition, it can be found that the SSR in the FD mode decreases faster than the SSR in the HD mode with the increase of $\kappa$. This is because the distortion noise power in the FD mode increases much more than the distortion noise power in the HD mode with the increase of $\kappa$. In other words, for a relatively small $\kappa$, a greater performance gain will be achieved by the FD mode in contrast with the HD mode.

**D. Impact of the System Settings**

In order to further verify the effectiveness of our proposed DRL-based method, we evaluate the performance under various system settings. We consider three cases, i.e., \{NT = 2, NR = 2, M = 8\}, \{NT = 4, NR = 4, M = 8\}, and \{NT = 4, NR = 4, M = 16\}, which are shown in Fig. 8. From this figure, we see that the SSR improves with the increase of the BS transmit power $P_{\text{max}}$. As more transmit power is allocated to BS, higher SSR can be obtained. This observation is consistent with that of conventional secure systems. In addition, larger size of RIS and more antennas at the BS lead to higher SSR.

Furthermore, we also compare the cumulative distribution function (CDF) of SSRs under different system settings, which are shown in Fig. 9. It is obvious from the figure that the CDF curves demonstrate the observations from Fig. 8, where the SSR improves with the increase of the BS transmit power $P_{\text{max}}$, the number of antennas at the BS, and the number of RIS reflecting elements $M$. The performance gain of the proposed DRL-based algorithms is stable in the light of the CDF curves, which means that the presented algorithm is robust to the system settings and can perform well with high probability. On the other hand, the CDF curves demonstrate the convergence of the proposed algorithm shown in Fig. 4. For example, focus
Fig. 8. SSR as a function of $P_{\text{max}}$ under different system settings.

Fig. 9. CDF of SSR for various system settings.

on the curve with $\{N_t = 2, N_r = 2, P_{\text{max}} = 10 \text{ dBm}, M = 8\}$, there is a 40% probability that the abscissa of this curve is less than 4.5, and 60% is stable at about 4.5. This proves that the agent in the proposed DRL-based method explores in the environment in the first 8000 steps, then learns from the exploration experience to converge, and obtains a satisfactory solution in 8000–20 000 steps.

E. Impact of the Number of Reflecting Elements at RIS

Fig. 10 depicts the impact of the number of the reflecting elements $M$ on the SSR performance with $K = 2, L = 2$, $\kappa_r = \kappa^R_{\text{e}_{\text{ck}}}$, $\kappa^S_{\text{e}_{\text{ck}}}$ = $\{0.01, 0.05\}$, and $\kappa_t = \kappa^R_{\text{e}_{\text{ck}}}$ = $\kappa^S_{\text{e}_{\text{ck}}}$ = $\{0.01, 0.05\}$ when $P_{\text{max}} = 10 \text{ dBm}$. On the one hand, it can be seen that increasing the number of RIS’s reflecting elements will increase the SSR of the system. However, the performance gain decreases while the level of the HWI increases. On the other hand, the SSR in the system with $\{\kappa_r = 0.01, \kappa_t = 0.05\}$ is much higher than the SSR in the system with $\{\kappa_r = 0.05, \kappa_t = 0.01\}$. This observation means that the HWI on receivers has a greater negative impact on the SSR than that of the HWI on the transmitters. This is because the proportion of the receive distortion noise power in the total received noise is much larger than that in the transmit distortion noise power.

F. Impact of the Path-Loss Exponent

In the aforementioned simulations, the path-loss exponent is set as 2.0 due to the assumption that the RIS is deployed in an unobstructed location so that the link of BS-RIS-Users is established. However, in some practical communication scenarios, this ideal place may not exist. Therefore, the impact of the large-scale fading channel on the performance of the secure system is investigated. Fig. 11 shows the SSR versus the path-loss exponent. The SSR obtained by the proposed DRL algorithm decreases with the increase of $\alpha$ as expected. This is because larger scale fading will lead to a weaker signal reflected from the RIS, which weakens the benefits of RIS. Fortunately, this provides practical engineering guidance that RIS should be carefully deployed in a place with less obstacles in the legitimate link and more obstacles in the eavesdropping link to improve the system’s secrecy performance.

Furthermore, we also compare the CDF of SSR for various $\alpha, P_{\text{max}}$ as well as $M$, i.e., $\alpha = \{2.0, 2.8\}, P_{\text{max}} = \{10, 20\} \text{ dBm}$, and $M = \{8, 16\}$, which are shown in Fig. 12. It is obvious from the figure that the CDF curves are consistent with the results in Fig. 11. In addition, the performance gain obtained by the proposed DRL method is stable according to the CDF curves.

G. Impact of Hyperparameters of DRL

Network parameters play an important role in DRL, which determine the convergence speed and efficiency of learning. By choosing appropriate network parameters, the DRL can achieve the desired results. Here, we take the learning rate and the discount factor as examples to illustrate that selecting network parameter appropriately is crucial to DRL. Fig. 13 exhibits the average rewards versus time steps under different learning rates, i.e., $\{0.1, 0.01, 0.001, 0.0001\}$. It is seen from the figure that learning rates with different levels
Fig. 11. SSR versus pass loss exponent $\alpha$ with $P_{\text{max}} = \{10, 20\}$ dBm and $M = \{8, 16\}$.

Fig. 12. CDF of SSR for various $\alpha$, $P_{\text{max}}$, as well as $M$.

Fig. 13. Average rewards performance versus time steps under different learning rates, i.e., $\{0.1, 0.01, 0.001, 0.0001, 0.00001\}$.

Fig. 14. Average rewards performance versus time steps under different discount factors, i.e., $\{0.3, 0.5, 0.7, 0.9\}$.

Choosing appropriate parameters can improve the performance of the DRL-based algorithm and accelerate the convergence speed of the neural network.

V. CONCLUSION

In this article, we studied a RIS-assisted FD secure communication system in the presence of HWI. An SSR maximization problem was formulated by jointly optimizing both the transmit beamforming at the BS and the phase shifts at the RIS. To tackle this mathematically intractable NP-hard problem, a DRL-based algorithm was proposed to obtain the satisfactory solution. Moreover, the proposed DRL method had a standard framework and low complexity in implementation, without explicit mathematical calculation. Therefore, it was conveniently deployed to communication systems with different settings. Extensive simulation results confirmed the effectiveness of the developed DRL-based algorithm in improving the SSR. It was also found that appropriate neural network parameters can improve the performance of the DRL-based algorithm and accelerate the convergence speed of the neural network.
[46] H. Limori, G. T. F. de Abreu, and G. C. Alexandropoulos, “MIMO beamforming schemes for hybrid SIC FD radios with imperfect hardware and CSI,” IEEE Trans. Wireless Commun., vol. 18, no. 10, pp. 4816–4830, Oct. 2019.

[47] J. Zhu, D. W. K. Ng, N. Wang, R. Schober, and V. K. Bhargava, “Analysis and design of secure massive MIMO systems in the presence of hardware impairments,” IEEE Trans. Wireless Commun., vol. 16, no. 3, pp. 2001–2016, Mar. 2017.

[48] V.-D. Nguyen, H. V. Nguyen, O. A. Dobre, and O.-S. Shin, “A new design paradigm for full-duplex multiser user systems,” IEEE J. Sel. Areas Commun., vol. 36, no. 7, pp. 1480–1495, Jul. 2018.

[49] T. Chen and W. Su, “Local energy trading behavior modeling with deep reinforcement learning,” IEEE Access, vol. 6, pp. 62806–62814, 2018.

[50] Z. Yang, Y. Liu, Y. Chen, and J. T. Zhou, “Deep reinforcement learning for RIS-aided non-orthogonal multiple access downlink networks,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Taipei, Taiwan, Nov. 2019, pp. 567–571, Mar. 2022.

[51] H. Ren, C. Pan, L. Wang, W. Liu, Z. Kou, and K. Wang, “Long-term CSI-based design for RIS-aided multiuser MISO systems exploiting deep reinforcement learning,” IEEE Commun. Lett., vol. 26, no. 3, pp. 567–571, Mar. 2022.

[52] S. Fujimoto, H. van Hoof, and D. Meier, “Addressing function approximation error in actor-critic methods,” 2018, arXiv:1802.09477.

Zhaobin Zhang received the B.S. degree from the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China. His research interests include reconfigurable intelligent surface, cooperative communications, information theory, physical-layer security, and machine learning for wireless communications.

Lei Kong received the Ph.D. degree from the School of Information Science and Engineering, National Mobile Communications Research Laboratory, Southeast University, Nanjing, China, in 2016. From 2016 to 2021, he worked as a System Engineer with Nokia, Hangzhou, China. He is currently a Postdoctoral Researcher with New H3C Corporation, Hangzhou, China, and Southeast University. His research interests span on the URLLC, deterministic networks, full-duplex communication, network energy saving, and RIS.

Jiangzhou Wang (Fellow, IEEE) is a Professor with the University of Kent, Canterbury, U.K. He has published over 400 papers and four books. His research interest is in the area of mobile communications.

From 2015 to 2016, he was a Research Associate with the University of Kent, Canterbury, U.K. He held a postdoctoral position with the Queen Mary University of London, London, U.K., from 2016 to 2019, and was a Lecturer from 2019 to 2021. From 2021, he is a Full Professor with Southeast University. He has published over 120 IEEE journal papers. His research interests mainly include reconfigurable intelligent surfaces (RIS), intelligent reflection surface, ultrareliable low-latency communication (URLLC), machine learning, UAV, Internet of Things, and mobile-edge computing.

Prof. Pan received the IEEE ComSoc Leonard G. Abraham Prize in 2022. He is currently an Editor of IEEE WIRELESS COMMUNICATION LETTERS, IEEE COMMUNICATIONS LETTERS, and IEEE ACCESS. He serves as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS on the special issue on xURLLC in 6G: Next Generation Ultra-Reliable and Low-Latency Communications. He also serves as a Leading Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING on the special issue on Advanced Signal Processing for Reconfigurable Intelligent Surface-Aided 6G Networks, IEEE VEHICULAR TECHNOLOGY MAGAZINE on the special issue on Backscatter and Reconfigurable Intelligent Surface Empowered Wireless Communications in 6G, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY on the special issue of Reconfigurable Intelligent Surface Empowered Wireless Communications in 6G and Beyond, and IEEE ACCESS on the special issue on Reconfigurable Intelligent Surface-Aided Communications for 6G and Beyond. He is a Workshop Organizer in IEEE ICCC 2021 on the topic of Reconfigurable Intelligent Surfaces for Next Generation Wireless Communications (RIS for 6G Networks), and IEEE Globecom 2021 on the topic of Reconfigurable Intelligent Surfaces for future wireless communications. He is currently the Workshop and Symposium Officer for Reconfigurable Intelligent Surfaces Emerging Technology Initiative. He is the Workshop Chair for IEEE WCNC 2024, and the TPC Co-Chair for IEEE ICCT 2022. He serves as a TPC Member for numerous conferences, such as ICC and GLOBECOM, and the Student Travel Grant Chair for ICC 2019.

Li Li received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1985, the M.S. degree in electrical engineering from China Research Institute of Radiowave Propagation, Qingdao, China, in 1988, and the Ph.D degree in science from Peking University, Beijing, China, in 1997.

Her research interest is currently with the adaptive digital signal processing, spectrum allocation and interference alignment algorithms in heterogeneous cognitive radio network, and network modeling of ultradense wireless mobile communication.