Smart Jamming for Secrecy: Deep Reinforcement Learning Enabled Secure Visible Light Communication

Sicong Liu, *Senior Member, IEEE,* Xianbin Liu, Xiaojiang Du, *Fellow, IEEE,* and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—As one of the indoor communication technologies, visible light communication (VLC) has drawn great attention for its advantages such as ultra-wide unlicensed spectrum, power saving and low complexity. The nature of the visible light propagation is an open channel, which is vulnerable to wiretapping. This paper investigates a secure VLC mechanism enabled by multiple light fixtures acting as friendly jammers. The goal of the friendly jammers is to diminish the capability of the eavesdropper to infer the undisclosed information, on the premise of causing minimal impact on the legitimate receiver. For this reason, an algorithm based on reinforcement learning is proposed to dynamically optimize the friendly jamming policy in realistic nonstationary environments. In order to resolve the difficult problem of the dimensional curse and to effectively represent the continuous state and action spaces, an algorithm based on deep reinforcement learning is devised, which utilizes deep convolutional neural networks to accelerate the convergence rate of the learning process. A differentiable neural dictionary is introduced to make full use of the experiences in similar anti-eavesdropping scenarios to improve the learning capability. Simulation results demonstrate that, the proposed schemes can achieve a higher secrecy rate and a lower bit error rate than some state-of-the-art schemes.**

***Index Terms*—Visible light communication, anti-eavesdropping, friendly jamming, deep reinforcement learning, multiple-input multiple-output.**

I. Introduction

Among the various technologies for indoor broadband communications, visible light communication (VLC) is one of the key technologies that have drawn great attention from both academia and industry [1]. VLC is operated in the ultra-wide unlicensed visible light spectrum, which naturally integrates information transfer with existing illumination infrastructure. Energy efficient and cost-effective light-emitting diodes (LEDs) that are widely applied for illumination are commonly adopted as a good choice of light signal source. Compared with traditional radio-frequency (RF) technologies [2], VLC enjoys many advantages such as license-free, high transmission rate, cost-effective implementation, and low energy consumption [3]–[5]. This makes VLC very suitable to be applied in various indoor and outdoor scenarios, such as high-capacity hotspots, ultra-dense networks, and high-rate underwater communications [6]–[8].

Due to the broadcast, unstable and open characteristics of visible light channels, the secrecy of the transmission between the transmitter and the legitimate user within the range of light exposure is threatened by eavesdropping, which is particularly prevalent in indoor scenarios such as offices, libraries, museums, and hotels [9]. It is also possible for the secrecy information to be wiretapped from outside the room through large windows [10]. Therefore, it is crucial to investigate effective anti-eavesdropping technologies while maintaining satisfactory transmission performance of the legitimate users.

With the ever enhancement of the deciphering ability of the eavesdroppers, the traditional secrecy protection scheme is more difficult to meet the security requirements, which is still a vital issue restricting the development of VLC techniques [11]–[14]. Recently, the technique of friendly jamming has attracted increasing attention as one of the physical layer security methods for secrecy protection [15]–[17]. A friendly jamming based cooperative communication scheme as proposed in [9] can degrade the signal quality received by an eavesdropper over the visible light wiretap channel, which is further extended to the multiple-input single-output (MISO) visible light channel in [18]. An artificial noise injection based technique can improve the secrecy performance of the free-space optical communication system [16]. An optical interference assisted secrecy enhancement method is introduced for a generalized space shift keying modulated VLC transmission over a Gaussian wiretap channel [17]. An efficient iterative algorithm based on transmit beamforming and friendly jamming techniques developed in [19] enhances the communication secrecy of MISO VLC systems with multiple eavesdroppers. A MISO VLC beamforming scheme was proposed in [20], which tried to find a strategy to reduce the receiving power of the eavesdropper through iterative learning process. However, the receiving power of the eavesdropper and the receiving power

This work is supported in part by National Natural Science Foundation of China (No. 62471414 and 62077040), in part by Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515030150), in part by the Natural Science Foundation of Fujian Province of China (No. 2023J01001), in part by the Science and Technology Key Project of Fujian Province, China (No. 2021HZ021004), and in part by the Science and Technology Planning Project of Fujian Province (No. 2022I0001). (*Corresponding Author: Sicong Liu*).

S. Liu and X. Liu are with the Department of Information and Communication Engineering, School of Informatics, Xiamen University, Xiamen 361005, China, and also with Shenzhen Research Institute of Xiamen University, Shenzhen 518057, China (E-mail: liusc@xmu.edu.cn).

X. Du is with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 USA (E-mail: dxj@ieee.org).

Mohsen Guizani is with Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates (E-mail: mguizani@ieee.org).

of the legitimate user are positively correlated since the signal is generated by the same transmitter, and thus the receiving power of legitimate users is constrained by eavesdroppers and the performance of the legitimate user is limited. This results in an inevitable tradeoff between the reception performance of the legitimate user and the secrecy protection ability of the system.

It is worth noting that, for common VLC systems, due to the non-symmetrical physical property of the visible light based downlink channel and the radio frequency or infrared based wireless uplink channel, and considering the incoherent nature of the IM/DD-based VLC transmission, many existing security mechanisms for classical wireless communications are not directly applicable to VLC systems [18]. Besides, realistic complex wiretapping environments tend to be temporally and spatially nonstationary due to dynamic changes, such as the blockage by clutters, human activities, and the mobility of the eavesdropper, making the wiretapping channel time-variant and complicated. Thus, it is very difficult to design a closed-form and continuously optimal policy of friendly jamming under the condition of realistic time-varying channels and complex dynamic environments. This makes it necessary to conceive an approach more adaptive to spatiotemporally nonstationary environments to improve the traditional friendly jamming schemes. Moreover, apart from deteriorating the ability of the wiretapper, some metrics concerning the legitimate user, such as the bit error rate (BER) that reflects the receive quality, can also be taken into consideration in the approach of friendly jamming, which might further improve the overall secrecy performance of the system.

To this end, we devise a smart friendly jamming approach in this paper inspired by the recently emerging technique of reinforcement learning (RL), which obtains optimal strategies by dynamically interacting with the environment modeled as a Markov decision process (MDP) [21]. As a subfield of artificial intelligence, the RL technology has been investigated in some telecommunication applications, including power control, anti-jamming, and relay selection [22]–[25], etc. By incorporating deep neural networks with RL, the deep reinforcement learning (DRL) technique is employed to find better strategies from more complex environments and accelerate the learning process [26]–[31]. Hence, in order to deal with the great challenge of secrecy protection in realistic spatiotemporally nonstationary environments, an RL-based friendly jamming (i.e., RL-FJ) scheme is proposed in this paper.

Specifically, multiple light fixtures acting as intelligent friendly jammers are controlled by the RL-FJ scheme to transmit friendly jamming signals to prevent the potential eavesdropper from inferring the private information, and meanwhile maintaining the quality of reception for legitimate users. Different from the scheme in [20], more degrees of freedom have been introduced by friendly jamming to search for the optimal anti-eavesdropping strategy in the solution space. The legitimate reception performance can be improved by increasing the receiving power of the legitimate user, which is not constrained by the eavesdropper. The proposed RL-based scheme searches for the best jamming policy that maximizes the jamming power on the eavesdropper, thereby improving the secrecy rate of the system. During the communication process, the proposed intelligent friendly jammers determine the friendly jamming policy, via Q-learning according to the state information such as the BER, secrecy rate and energy consumption. To determine the jamming policy is to determine the jamming beamformers for the multiple LED light fixtures, i.e., the friendly jammers. Thus, a multiple-input single-output (MISO) channel is formed between the multiple jammers and the legitimate user or eavesdropper.

Furthermore, it should be noted that Q-learning algorithms have only asymptotical optimality guarantee [32], i.e. converge to an optimal solution as the number of data samples tends to infinity, which might limit the learning performance. More importantly, in realistic complex VLC environments, the difficulty of effective representation of the continuous state and action spaces and the problem of dimensional curse, that is, the difficulties encountered by algorithms in model training due to the high dimensionality of data, still remain to be resolved. Meanwhile, the convergence rate of the learning process should be further accelerated in this complicated circumstance to ensure satisfactory quality-of-service. To this end, we propose a DRL-based friendly jamming (DRL-FJ) scheme that employs deep convolutional neural networks (CNNs) to extract the complex environmental features and effectively represent the continuous state and action spaces. Accompanied with the DRL-FJ scheme, a memory module called differentiable neural dictionary (DND) [33] is utilized to make full use of the experiences in similar anti-eavesdropping scenarios to further accelerate the learning process. To summarize, our contributions are listed as below.

- An RL-based scheme, i.e., RL-FJ, is proposed, which dynamically determines the optimal friendly jamming policy via Q-learning, to adaptively improve the performance of secure VLC transmission in realistic spatiotemporally nonstationary environments [1].
- A DRL-based scheme, i.e., DRL-FJ, is proposed, which utilizes deep CNNs to effectively represent the complex and continuous state and action spaces, and resolve the problem of the dimensional curse.
- A DND is introduced to make full use of the previous similar anti-eavesdropping experiences, which further accelerates the learning process.
- The performance of the proposed schemes are theoretically analyzed in terms of the receive quality of the legitimate user and the overall system utility, and the computational complexity of the proposed algorithms are derived.

The remainder of this paper is organized as follows. The

[1] Part of this work, i.e., part of the first contribution on Q-learning based algorithm, has been presented in IEEE International Conference on Communications (IEEE ICC) 2022 [34]. Compared with the conference version, this article has extensively extended the technical content, theoretical analysis and experimental results. A DRL-FJ algorithm is proposed to deal with the problem of high-dimensional curse and quantization error. A DND is introduced to further improve the learning performance and efficiency. Theoretical analysis of the performance of the proposed schemes as well as, the computational complexity are derived. More extensive and thorough simulations have been conducted and reported, and more benchmark schemes are compared.

visible light channel model and the VLC wiretapping model are described in Section II. In Section III and Section IV, the proposed RL-FJ and DRL-FJ schemes are introduced, respectively. The theoretical performance of the proposed algorithms are evaluated in Section V. In Section VI, the simulation results are reported with discussions, followed by the conclusions in Section VII.

II. System Model

A. Visible Light Channel Model

In this paper, a VLC system using pulse amplitude modulation (PAM) with a direct current (DC) bias is considered, where the transmitter is composed of multiple LED light fixtures driven by a fixed bias I_D and the receiver at the user terminal is equipped with a single photodiode (PD). The total electric current I_e driving the LED is the superposition of the DC bias I_D and the signal component x conveying the information to send. To avoid severe clipping distortion and maintain satisfactory linearity in optical conversion, and meanwhile to satisfy the illumination requirements, it is required that the total electric current I_e should not exceed a specific range [35]. This imposes a specific constraint on the amplitude of the information signal x, i.e., $|x| \leq \alpha I_\mathrm{D}$, where $\alpha \in [0, 1]$ can be regarded as the modulation index. The electric current I_e is then converted to the instantaneous transmit optical power P_T in the LED to radiate light via an electro-optical converter, i.e., $P_\mathrm{T} = \eta I_\mathrm{e}$, where η is the electro-optical conversion efficiency.

At the legitimate receiver of the user terminal, a PD receives the incident optical power represented by $P_\mathrm{R} = G P_\mathrm{T}$ with G denoting the path gain of the channel, and converts the optical power to the received electrical signal y via optic-electrical conversion with the PD responsivity of R and after the DC bias removal.

The LEDs are commonly assumed to have a Lambertian radiation pattern [36], [37], where the path gain of the visible light channel G is given by

$$G = \begin{cases} \frac{n_0^2 A_\mathrm{P} (\log\cos\phi_{1/2} - \log 2)}{2\pi d^2 \sin^2(\varphi_\mathrm{F}) \log\cos\phi_{1/2}} \cos^{\frac{-\log 2}{\log\cos\phi_{1/2}}}(\phi)\cos(\varphi), & |\varphi| \leq \varphi_\mathrm{F}, \\ 0, & |\varphi| > \varphi_\mathrm{F}, \end{cases} \tag{1}$$

where the parameters are defined as follows: n_0 is the optical concentrator refractive index; A_P is the PD detector area; ϕ is the angle of irradiance relative to the LED optical axis; $\phi_{1/2}$ is the LED half irradiation intensity semi-angle; φ is the angle of incidence relative to the PD optical axis; φ_F is the PD field-of-view (FoV); d is the distance from the LED to the PD.

B. VLC Wiretapping Model

A typical indoor VLC framework against eavesdropping assisted by friendly jamming is shown in Fig. 1. An LED light fixture acting as the transmitter (Alice) is sending private information to the receiver of the legitimate user (Bob) equipped with a PD via the VLC link. An eavesdropper (Eve) in this environment is attempting to wiretap the private information from the VLC signal that reaches the PD of its receiver. The friendly jammers are composed of N_J LED light fixtures that are transmitting jamming signals with a specifically designed beamforming pattern in order to prevent Eve from effective wiretapping.

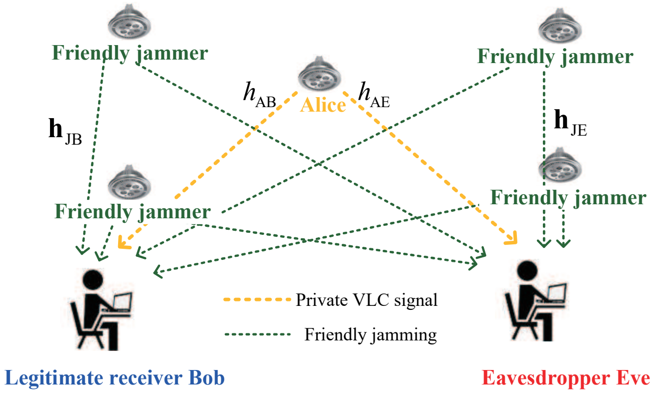


Fig. 1. A typical indoor VLC framework against eavesdropping, including an LED light fixture acting as the transmitter (Alice), the legitimate receiver (Bob), an eavesdropper (Eve), and the friendly jammers composed of several LED light fixtures.

Based on the visible light channel model, the received electrical signals of Bob and Eve are respectively given by

$$y_\mathrm{B} = h_\mathrm{AB} x + \mathbf{h}_\mathrm{JB}^\mathrm{T} \mathbf{j} + n_\mathrm{B}, \tag{2a}$$

$$y_\mathrm{E} = h_\mathrm{AE} x + \mathbf{h}_\mathrm{JE}^\mathrm{T} \mathbf{j} + n_\mathrm{E}, \tag{2b}$$

where h_AB and h_AE represent the channel gains from Alice to Bob and Eve, respectively, with $h_\mathrm{AB} = R\eta G_\mathrm{AB}$ and $h_\mathrm{AE} = R\eta G_\mathrm{AE}$, where G_AB and G_AE represent the path gain from Alice to Bob and Eve respectively; $x \in \mathbb{R}$ represents the transmit information signal with zero mean; The vectors $\mathbf{h}_\mathrm{JB}$, $\mathbf{h}_\mathrm{JE} \in \mathbb{R}^{N_\mathrm{J}}$ denote the channel gain vectors from the N_J friendly jammers to Bob and Eve, respectively, where $\mathbf{h}_\mathrm{JB} = R\eta[G_\mathrm{1B}, G_\mathrm{2B}, \cdots, G_{N_\mathrm{J}\mathrm{B}}]^\mathrm{T}$ and $\mathbf{h}_\mathrm{JE} = R\eta[G_\mathrm{1E}, G_\mathrm{2E}, \cdots, G_{N_\mathrm{J}\mathrm{E}}]^\mathrm{T}$, with $G_{i\mathrm{B}}$ and $G_{i\mathrm{E}}$, $i = 1, 2, \cdots, N_\mathrm{J}$, denoting the gains of the propagation links from the i-th friendly jammer to the legitimate user and the wiretapper, respectively; The vector $\mathbf{j} \in \mathbb{R}^{N_\mathrm{J}}$ represents the jamming signal transmitted by the N_J friendly jammers; The background noise at Bob and Eve is represented by n_B and n_E, which can be modelled by zero-mean additive white Gaussian noise (AWGN) with the variance of σ_B^2 and σ_E^2, respectively.

As previously mentioned, subject to the constraint of the dynamic range of the LED driving current due to nonlinearity distortion and the requirements of illumination purpose, the amplitude of both the information signal and the jamming signals should satisfy a certain constraint. Specifically, it is satisfied that $|x| \leq \alpha I_\mathrm{D}$ and $|\mathbf{j}| \preceq \mathbb{1}\alpha I_\mathrm{D}$, where $\mathbb{1}$ is an all-one vector, and the operator $\preceq$ denotes elementwise inequality between two vectors, i.e., the absolute value of each element in the jamming signal $\mathbf{j}$ is no greater than αI_D.

The jamming signal $\mathbf{j}$ can be equivalently rewritten in a format of MIMO beamforming for the purpose of simplifying the representation of the jamming strategy to be determined by the proposed learning scheme. Specifically, a jamming beamforming vector, i.e., the jamming beamformer $\mathbf{w} = [w_1, w_2, \cdots, w_{N_\mathrm{J}}]^\mathrm{T}$ subject to $|\mathbf{w}| \preceq \mathbb{1}$, can be allocated to the N_J friendly jammers to determine the jamming signals

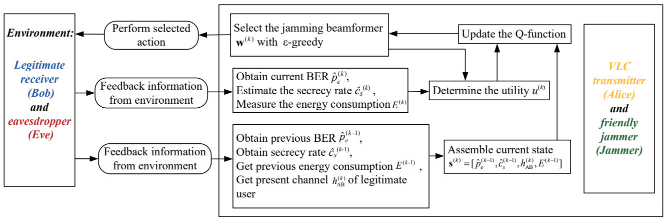

Environment:
Legitimate receiver (Bob) and eavesdropper (Eve)
Perform selected action
Feedback information from environment
Feedback information from environment
Select the jamming beamformer $\mathbf{w}^{(k)}$ with ε-greedy
Update the Q-function
Obtain current BER $\hat{p}_e^{(k)}$, Estimate the secrecy rate $\hat{c}_s^{(k)}$, Measure the energy consumption $E^{(k)}$
Determine the utility $u^{(k)}$
Obtain previous BER $\hat{p}_e^{(k-1)}$, Obtain secrecy rate $\hat{c}_s^{(k-1)}$, Get previous energy consumption $E^{(k-1)}$, Get present channel $h_{\mathrm{AB}}^{(k)}$ of legitimate user
Assemble current state $\mathbf{s}^{(k)}=[\hat{p}_e^{(k-1)},\hat{c}_s^{(k-1)},h_{\mathrm{AB}}^{(k)},E^{(k-1)}]$
VLC transmitter (Alice) and friendly jammer (Jammer)

Fig. 2. Proposed RL-based friendly jamming beamforming scheme.

to transmit. Then, the jamming signals can be rewritten as $\mathbf{j} = \mathbf{w}j$, where the amplitude variable j is zero-mean and subject to $|j| \leq \alpha I_{\mathrm{D}}$. Thus, the received electrical signals in equation (2) is simplified as

$$y_{\mathrm{B}} = h_{\mathrm{AB}}x + \mathbf{h}_{\mathrm{JB}}^{\mathrm{T}}\mathbf{w}j + n_{\mathrm{B}}, \tag{3a}$$

$$y_{\mathrm{E}} = h_{\mathrm{AE}}x + \mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}j + n_{\mathrm{E}}. \tag{3b}$$

III. RL-Based Friendly Jamming Beamforming Scheme Against Eavesdropping

Assuming that the amplitude of the transmit and received information signals, i.e., x and j, is uniformly distributed within $[-\alpha I_{\mathrm{D}}, \alpha I_{\mathrm{D}}]$, the achievable secrecy rate c_s of the VLC system in the wiretap channel [9] is given by

$$\begin{aligned} c_s =& \frac{1}{2}\log\left(1+\frac{2h_{\mathrm{AB}}^2\alpha^2 I_{\mathrm{D}}^2}{\pi e\left(\sigma_{\mathrm{B}}^2 + 2\left|\mathbf{h}_{\mathrm{JB}}^{\mathrm{T}}\mathbf{w}\right|^2\alpha^2 I_{\mathrm{D}}^2\right)}\right) \\ & - \min\left(\log\frac{h_{\mathrm{AE}}}{|\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}|} + \frac{\left|\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}\right|}{h_{\mathrm{AE}}}\log\sqrt{e}, \frac{h_{\mathrm{AE}}}{|\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}|}\log\sqrt{e}\right). \end{aligned} \tag{4}$$

The objective of the VLC system should be to maximize the secrecy rate c_s as given in (4), i.e., $\arg\max\limits_{\mathbf{w}} c_s$. This is, however, an intractable problem that is difficult to find a closed-form solution [9]. As a tractable alternative solution, a suboptimal jamming beamformer $\mathbf{w}_0$ that tries to maximize the secrecy rate in (4) based on zero-forcing can be obtained by

$$\begin{aligned} \mathbf{w}_0 = & \arg\max_{\mathbf{w}} c_s, \\ & \text{s.t. } \mathbf{h}_{\mathrm{JB}}^{\mathrm{T}}\mathbf{w} = 0,\ |\mathbf{w}| \preceq \mathbb{1}. \end{aligned} \tag{5}$$

The solution to the above problem is to force the jamming power applied by the friendly jammer on Bob to be zero, i.e. $\mathbf{h}_{\mathrm{JB}}^{\mathrm{T}}\mathbf{w} = 0$. It should be noted that this zero-forcing solution can be suboptimal. It is intuitively understandable that the jamming beamformer $\mathbf{w}$ is confined in the null-space of $\mathbf{h}_{\mathrm{JB}}^{\mathrm{T}}$ in the zero-forcing solution, so all possible solutions of $\mathbf{w}$ can not be traversed. Thus, only a local optima within the zero-forcing subspace can be found, while the globally optimal solution of the jamming beamformer policy that maximizes c_s cannot be achieved. For example, it is possible that in a certain case the jamming power on Bob is a small value but not strictly zero, but the jamming power on Eve might be larger than the case where Bob receives zero jamming, which results in a higher secrecy rate than the zero-forcing case.

Besides, the optimization problem in (5) only considers the ability to degrade the wiretapping performance of the eavesdropper, but the receive quality of the legitimate user, which is reflected by some metrics such as the BER, is not guaranteed, which might have a severe impact on the quality-of-service of Bob and limit the overall system utility. Hence, it is necessary to design an efficient scheme to tackle this problem.

In order to take both the secrecy rate and the receive quality of the legitimate user into account in realistic spatiotemporally nonstationary environments, an RL-based friendly jamming (RL-FJ) beamforming method is devised, which aims to improve the overall system utility through an online iterative learning process, and fully explore the solution space of the problem in (4) to dynamically update the optimal policy of friendly jamming. The system utility adopted in the learning process can be carefully devised to thoroughly consider the secrecy rate performance, the receive quality of the legitimate user, and in addition, the energy consumption, which will be described later this section.

Specifically, during the VLC transmission process composed of a series of time slots indexed by k, the intelligent friendly jammers dynamically determine the optimal jamming beamformer by interacting with the spatiotemporally nonstationary environment. Consequently, an MDP is formulated to select the optimal policy, in which the state, action, and reward are elaborated as follows.

State: As depicted in Fig. 2, the friendly jammers receive the feedback information from the communication environment, including the legitimate channel gain $h_{\mathrm{AB}}^{(k)}$ and the previous BER $\hat{p}_e^{(k-1)}$. The achievable rate $\hat{c}_s^{(k-1)}$ at previous time slot is calculated according to (4), in which the channel state information (CSI) of h_{AB} and $\mathbf{h}_{\mathrm{JB}}$ can be obtained from the feedback information from Bob, while the CSI of $\mathbf{h}_{\mathrm{JE}}$ and h_{AE} can be estimated by roughly predicting Eve's possible location with the aid of some prior information, such as the geometric information of the indoor environment, the furniture layout, the obstruction, etc. The energy consumption $E^{(k-1)}$ of the transmitter and the friendly jammers at the previous slot is also obtained. Thus, the current state $\mathbf{s}^{(k)}$ of the system can be expressed as

$$\mathbf{s}^{(k)} = \left[\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, h_{\mathrm{AB}}^{(k)}, E^{(k-1)}\right] \in \Lambda, \tag{6}$$

where Λ denotes the state space consisting of all the possible states.

Action: The jamming beamformer $\mathbf{w}$ is selected as the action, which determines the jamming power transmitted by the N_{J} intelligent friendly jammers as in (3). The jamming beamformer $\mathbf{w}^{(k)}$ of time slot k is selected from the action space $\mathbf{W}$, i.e. $\mathbf{w}^{(k)} = [w_1^{(k)}, \ldots, w_{N_{\mathrm{J}}}^{(k)}]^{\mathrm{T}} \in \mathbf{W}$ with $|\mathbf{w}^{(k)}| \preceq \mathbb{1}$.

Reward: The reward to be optimized in the MDP is regarded as the utility function of the system. In order to ensure the receive quality of the legitimate user, the BER of Bob should be taken into consideration in the design of the utility function. After obtaining the feedback information from the environment, the BER of Bob $\hat{p}_e^{(k)}$ and energy cost $E^{(k)}$ are obtained, and secrecy rate $\hat{c}_s^{(k)}$ is estimated. Thus, the utility function $u^{(k)}$ is defined as

$$u^{(k)} = \hat{c}_s^{(k)} - \delta_1\hat{p}_e^{(k)} - \delta_2 E^{(k)}, \tag{7}$$

where the coefficients of δ_1 and δ_2 play the role of balancing the contribution of the BER, energy consumption and secrecy rate to the overall utility, which can provide an appropriate tradeoff among them. The second penalty term in the utility function is introduced to evaluate the energy consumed by the friendly jammers and the transmitter, which drives the agent to search for an optimal solution that considers both benefits and costs. In the dynamic interaction with the time-varying channel, maximizing the utility given in (7) is regarded as an optimization objective by the RL-FJ scheme, i.e. finding an optimal jamming strategy that maximizes $u^{(k)}$, so not only the secrecy rate but also the receive quality of Bob and the energy consumption are considered thoroughly in the scheme, which leads to a better overall system performance.

To deal with the dynamic MDP in the complex environment, RL-based algorithms such as Q-learning can be adopted thanks to the capability of learning sophisticated strategy out of dynamic environments. In this regard, the friendly jammers act as the smart RL agent that determines the optimal jamming beamformer according to the current state and Q-function to maximize the utility function. The Q-function therein denoted by $Q(\mathbf{s}, \mathbf{w})$ represents the expectation of the cumulative discount reward of the friendly jammers performing action $\mathbf{w}$ in the current state $\mathbf{s}$. The actions taken by the friendly jammers will have an impact on the next state, thus further influencing the future actions and rewards. To facilitate the online iterative Q-learning processing, each element of the jamming beamformer is quantized to $2L_x + 1$ equally spaced discrete values, i.e., $w_i^{(k)} \in \{\frac{l}{L_x} | l = -L_x, \cdots, L_x, l \in \mathbb{N}\}$, where L_x can be chosen to realize a proper tradeoff between learning accuracy and computational complexity. To avoid being stuck in local optima and achieve a compromise between exploitation and exploration for the RL scheme, the jamming beamformer $\mathbf{w}^{(k)}$ is selected based on the ε-greedy method as given by

$$\Pr\left(\mathbf{w}^{(k)} = \tilde{\mathbf{w}}\right) = \begin{cases} 1-\varepsilon, & \tilde{\mathbf{w}} = \arg\max_{\mathbf{w}' \in \mathbf{W}} Q\left(\mathbf{s}^{(k)}, \mathbf{w}'\right) \\ \frac{\varepsilon}{|\mathbf{W}|-1}, & \text{o.w.} \end{cases} \tag{8}$$

where $\tilde{\mathbf{w}}$ is the jamming beamformer that the friendly jammers tend to choose in state $\mathbf{s}^{(k)}$ with a large probability $1-\varepsilon$, and ε is a very small value representing the low probability of new exploration in the action space. In this way, the friendly jammers would most likely exploit the Q-function to determine the action, while also possibly explore another random jamming beamformer with a small probability to effectively prevent falling into a local optimum.

Algorithm 1 RL-based friendly jamming (RL-FJ) algorithm

```
Initialize: $\lambda$, $\beta$, $\mathbf{Q} = \mathbf{0}$ and $\mathbf{V} = \mathbf{0}$
for $k = 1, 2, ...$ do
    Obtain previous BER of Bob $\hat{p}_e^{(k-1)}$ via feedback
    Estimate the secrecy rate at $(k-1)$-th time slot $\hat{c}_s^{(k-1)}$
    Obtain the channel gain of legitimate user at $k$-th time slot $h_{\text{AB}}^{(k)}$
    Measure the energy consumption at $(k-1)$-th time slot $E^{(k-1)}$
    Formulate the system state $\mathbf{s}^{(k)} = \left[\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, h_{\text{AB}}^{(k)}, E^{(k-1)}\right]$
    Select the jamming beamformer $\mathbf{w}^{(k)}$ based on $\varepsilon$-greedy given in (8)
    Transmit friendly jamming signals using the selected action, i.e., the jamming beamformer, $\mathbf{w}^{(k)}$
    Obtain the BER at $k$-th time slot $\hat{p}_e^{(k)}$
    Calculate the secrecy rate at $k$-th time slot $\hat{c}_s^{(k)}$
    Measure the energy consumption at $k$-th time slot $E^{(k)}$
    Determine the utility $u^{(k)}$ via (7)
    Update the Q-function via (9)
end for
```

The pseudo-code of the proposed RL-FJ scheme is summarized in **Algorithm 1**. Specifically, the friendly jammers observe the current state $\mathbf{s}^{(k)}$ and employ the optimal jamming beamformer $\mathbf{w}^{(k)}$ selected based on the ε-greedy policy defined in (8). Then the reward $u^{(k)}$ determined by (7) is obtained, and the environment turns to the next state $\mathbf{s}^{(k+1)}$. The Q-function is updated based on the iterative Bellman equation as given by

$$\begin{aligned} Q\left(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}\right) \leftarrow & (1-\lambda) Q\left(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}\right) \\ & + \lambda\left(u^{(k)} + \beta \max_{\mathbf{w}' \in \mathbf{W}} Q\left(\mathbf{s}^{(k+1)}, \mathbf{w}'\right)\right), \end{aligned} \tag{9}$$

where $\lambda \in [0, 1]$ is the learning rate indicating the extent to which the new information overrides the previous information, and the discount factor β valued between zero and one reflects the contribution by the long-term rewards. To be more specific, the core idea of the Q-learning algorithm is an iterative update of the weighted average of the previous information and the new information of the reward: $Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)})$ represents the previous Q-values; $u^{(k)}$ represents the immediate reward obtained in the transition from the current state to the next state; A value function $V\left(\mathbf{s}^{(k+1)}\right) = \max_{\mathbf{w}' \in \mathbf{W}} Q\left(\mathbf{s}^{(k+1)}, \mathbf{w}'\right)$ is defined to represent the maximum future reward that can be obtained based on all feasible actions in state $\mathbf{s}^{(k+1)}$; Thus, the new information is a weighted average of the immediate reward and the discounted estimated long-term reward.

IV. DRL-Based Friendly Jamming Beamforming Scheme Against Eavesdropping

In this section, a DRL-FJ beamforming scheme is further proposed, which employs deep CNNs to extract the complex environmental features and effectively represent the continuous state and action spaces, and resolve the problem of the dimensional curse therein. Accompanied with the DRL-FJ scheme, a memory module called differentiable neural dictionary (DND) is utilized to store the previous similar anti-eavesdropping experiences to be used in the future learning process, which further accelerates the convergence in non-stationary environments. Specifically, the friendly jammers select the jamming beamformer $\mathbf{w} \in \mathbf{W}$ according to the

Algorithm 2 DRL-based friendly jamming (DRL-FJ) algorithm

```
Initialize: $\lambda$, $\beta$, $\boldsymbol{\theta}$, $T$, and $M$
for $k = 1, 2, ...$ do
    Obtain previous BER of Bob $\hat{p}_e^{(k-1)}$ via feedback
    Estimate the secrecy rate at $(k-1)$-th time slot $\hat{c}_s^{(k-1)}$
    Obtain the channel gain of legitimate user at $k$-th time slot $h_{\text{AB}}^{(k)}$
    Measure the energy consumption at $(k-1)$-th time slot $E^{(k-1)}$
    Formulate the system state $\mathbf{s}^{(k)} = [\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, h_{\text{AB}}^{(k)}, E^{(k-1)}]$
    if $k \leq M$ then
        Select a jamming beamformer $\mathbf{w}^{(k)}$ randomly
    else
        Assemble state-action sequence $\boldsymbol{\varphi}^{(k)} = \left\{\mathbf{s}^{(k-M)}, \mathbf{w}^{(k-M)}, \cdots, \mathbf{w}^{(k-1)}, \mathbf{s}^{(k)}\right\}$
        Feed $\boldsymbol{\varphi}^{(k)}$ into the CNN as input
        Obtain CNN output and use it as look-up key $\hat{\boldsymbol{h}}$ in the DND $D_{\mathbf{w}}$ for each action $\mathbf{w} \in \mathbf{W}$
        Generate weight $\omega_j$ via (10)
        Estimate the corresponding Q-function $Q(\mathbf{s}^{(k)}, \mathbf{w})$ for each action $\mathbf{w}$ via (12)
        Collect all estimated Q-values $Q(\mathbf{s}^{(k)}, \mathbf{w})$, $\mathbf{w} \in \mathbf{W}$ to update the overall Q-function
        Select jamming beamformer $\mathbf{w}^{(k)}$ via (8)
        Append a new entry $(\hat{\boldsymbol{h}}, Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}))$ to the DND memory $D_{\mathbf{w}^{(k)}}$ for selected action $\mathbf{w}^{(k)}$
    end if
    Perform friendly jamming using the selected action $\mathbf{w}^{(k)}$
    Obtain the BER at $k$-th time slot $\hat{p}_e^{(k)}$
    Calculate the secrecy rate at $k$-th time slot $\hat{c}_s^{(k)}$
    Measure the energy consumption at $k$-th time slot $E^{(k)}$
    Determine the system utility $u^{(k)}$ via (7)
    Append jamming experience $\left\{\boldsymbol{\varphi}^{(k)}, \mathbf{w}^{(k)}, u^{(k)}, \boldsymbol{\varphi}^{(k+1)}\right\}$ to $\mathcal{B}$
    for $t = 1, 2, ..., T$ do
        Choose an experience $\mathbf{e}^{(t)} \in \mathcal{B}$ randomly
    end for
    Set up an experience mini-batch $\mathcal{T} = \left\{\mathbf{e}^{(t)}\right\}_{1 \leq t \leq T}$
    Update the weights $\boldsymbol{\theta}^{(k)}$ of the CNN via (13)
end for
```

Q-function determined by the mapping entries stored in a DND memory corresponding to the action $\mathbf{w}$, which is denoted as $D_{\mathbf{w}} = (\boldsymbol{K}_{\mathbf{w}}, \boldsymbol{V}_{\mathbf{w}})$. Each mapping entry in the DND $D_{\mathbf{w}}$ contains a look-up key array $\boldsymbol{K}_{\mathbf{w}}$ and a Q-value array $\boldsymbol{V}_{\mathbf{w}}$, in which the look-up keys $\hat{\boldsymbol{h}} \in \mathbb{R}^{|\mathbf{W}|}$ for query and the corresponding estimated Q-values are stored, respectively. The look-up key $\hat{\boldsymbol{h}}$ used to query the estimated Q-value is obtained from the output of the preceding CNN, which can be regarded as extracted high-level features corresponding to current state $\mathbf{s}^{(k)}$.

The framework of the proposed DRL-FJ beamforming scheme is illustrated in Fig. 3, whose pseudo-code is summarized in **Algorithm 2**. Specifically, the friendly jammers observe the current VLC system state $\mathbf{s}^{(k)}$ as given in (6). A series of previous states and actions, along with the current state, are stacked to formulate a state-action sequence $\boldsymbol{\varphi}^{(k)} = \left\{\mathbf{s}^{(k-M)}, \mathbf{w}^{(k-M)}, \mathbf{s}^{(k-M-1)}, \mathbf{w}^{(k-M-1)}, \cdots, \mathbf{w}^{(k-1)}, \mathbf{s}^{(k)}\right\}$, including the state-action pairs of the previous M time slots and the current VLC system state $\mathbf{s}^{(k)}$. Using the state-action sequence, the temporal correlation between different system states can be strengthened and exploited over the learning process. The state-action sequence $\boldsymbol{\varphi}^{(k)}$ is reshaped into an $m_0 \times m_0$ matrix as the input to the CNN with two convolutional (Conv) layers followed by two fully connected (FC) layers. The l-th Conv layer has f_l filters of size $m_l \times m_l$ and stride of s_l, followed by a rectified linear unit (ReLU) with f_l feature maps as the output for $l = 1, 2$. The feature maps in the second Conv layer are fed into the first FC layer with n_1 neurons. The second FC layer outputs a length-n_2 vector $\hat{\boldsymbol{h}}$, which is used as the look-up key in the subsequent DNDs. For the convenience of notations, the parameters of the CNNs are assembled as a hyper-parameter vector $\mathbf{F} = [f_1, f_2, m_1, m_2, s_1, s_2, n_1, n_2]$.

For each feasible action in the action space $\mathbf{w} \in \mathbf{W}$, a corresponding DND memory $D_{\mathbf{w}} = (\boldsymbol{K}_{\mathbf{w}}, \boldsymbol{V}_{\mathbf{w}})$ is maintained for query purpose. Specifically, the friendly jammers use the look-up key $\hat{\boldsymbol{h}}$, which is the output of the preceding CNN module, to query the corresponding estimated Q-values stored in the DND $D_{\mathbf{w}}$ for each action $\mathbf{w}$. During the query in the DND $D_{\mathbf{w}}$, a weight ω_j for the j-th entry of the Q-value array $\boldsymbol{V}_{\mathbf{w}}$ is generated as given by

$$\omega_j = \frac{\operatorname{Ker}(\hat{\boldsymbol{h}}, \boldsymbol{h}_j)}{\sum\limits_{p=1}^{\zeta} \operatorname{Ker}(\hat{\boldsymbol{h}}, \boldsymbol{h}_p)}, \tag{10}$$

where $1 \leq j \leq \zeta$, $\zeta \leq k$ means that the total number of mapping entries accumulated in the previous k time slots does not exceed k; $\boldsymbol{h}_j$ is the j-th entry stored in the look-up key array $\boldsymbol{K}_{\mathbf{w}}$, and $\operatorname{Ker}(\hat{\boldsymbol{h}}, \boldsymbol{h}_j)$ represents a Gaussian kernel function between the look-up keys of $\hat{\boldsymbol{h}}$ and $\boldsymbol{h}_j$, which returns the vector distance between the two keys as given by

$$\operatorname{Ker}(\hat{\boldsymbol{h}}, \boldsymbol{h}_j) = \exp\left(-\frac{1}{2}\left\|\hat{\boldsymbol{h}} - \boldsymbol{h}_j\right\|_2^2\right). \tag{11}$$

After performing the look-up operation on the DND $D_{\mathbf{w}}$ for a specific action $\mathbf{w}$, an estimate of the corresponding Q-function $Q(\mathbf{s}^{(k)}, \mathbf{w})$ for the given action $\mathbf{w}$ at the current state $\mathbf{s}^{(k)}$ is returned, which is given by

$$Q(\mathbf{s}^{(k)}, \mathbf{w}) = \sum_{j=1}^{\zeta} \omega_j v_j, \tag{12}$$

where v_j is the estimated Q-value stored in the j-th entry of the Q-value array $\boldsymbol{V}_{\mathbf{w}}$. It is noted that the Q-function obtained in (12) is a weighted sum of the Q-values stored in the array $\boldsymbol{V}_{\mathbf{w}}$ of the DND $D_{\mathbf{w}}$, whose weights are determined by the normalization kernel between the look-up key $\hat{\boldsymbol{h}}$ to be queried and each look-up key $\boldsymbol{h}_j$ in the array $\boldsymbol{K}_{\mathbf{w}}$.

This estimation process is repeated for each action $\mathbf{w} \in \mathbf{W}$, and in this way, the estimated Q-values for all the actions are

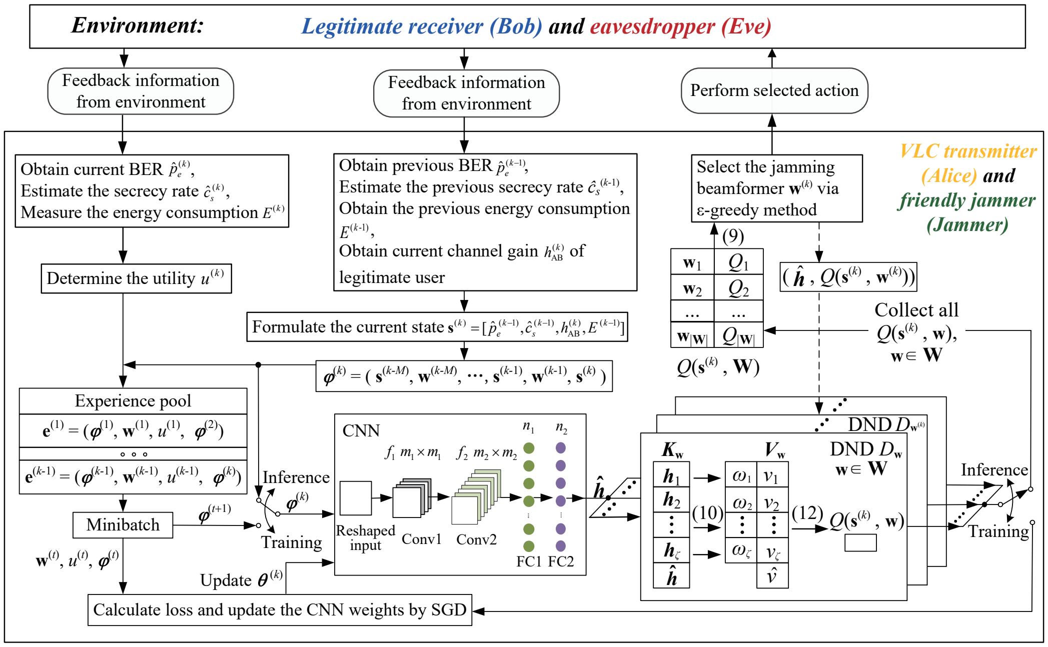

Environment: Legitimate receiver (Bob) and eavesdropper (Eve)
Feedback information from environment
Feedback information from environment
Perform selected action
VLC transmitter (Alice) and friendly jammer (Jammer)
Obtain current BER $\hat{p}_e^{(k)}$, Estimate the secrecy rate $\hat{c}_s^{(k)}$, Measure the energy consumption $E^{(k)}$
Obtain previous BER $\hat{p}_e^{(k-1)}$, Estimate the previous secrecy rate $\hat{c}_s^{(k-1)}$, Obtain the previous energy consumption $E^{(k-1)}$, Obtain current channel gain $h_{AB}^{(k)}$ of legitimate user
Select the jamming beamformer $\mathbf{w}^{(k)}$ via ε-greedy method
(9)
$\mathbf{w}_1$ Q_1
$\mathbf{w}_2$ Q_2
… …
$\mathbf{w}_{|\mathbf{W}|}$ $Q_{|\mathbf{W}|}$
$Q(\mathbf{s}^{(k)}, \mathbf{W})$
$(\hat{\boldsymbol{h}}, Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}))$
Collect all $Q(\mathbf{s}^{(k)}, \mathbf{w})$, $\mathbf{w} \in \mathbf{W}$
Determine the utility $u^{(k)}$
Formulate the current state $\mathbf{s}^{(k)} = [\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, h_{AB}^{(k)}, E^{(k-1)}]$
$\boldsymbol{\varphi}^{(k)} = (\mathbf{s}^{(k-M)}, \mathbf{w}^{(k-M)}, \cdots, \mathbf{s}^{(k-1)}, \mathbf{w}^{(k-1)}, \mathbf{s}^{(k)})$
Experience pool
$\mathbf{e}^{(1)} = (\boldsymbol{\varphi}^{(1)}, \mathbf{w}^{(1)}, u^{(1)}, \boldsymbol{\varphi}^{(2)})$
$\mathbf{e}^{(k-1)} = (\boldsymbol{\varphi}^{(k-1)}, \mathbf{w}^{(k-1)}, u^{(k-1)}, \boldsymbol{\varphi}^{(k)})$
Minibatch
$\mathbf{w}^{(t)}, u^{(t)}, \boldsymbol{\varphi}^{(t)}$
$\boldsymbol{\varphi}^{(t+1)}$
Inference
$\boldsymbol{\varphi}^{(k)}$
Training
CNN
f_1 $m_1 \times m_1$
f_2 $m_2 \times m_2$
n_1
n_2
Reshaped input
Conv1
Conv2
FC1 FC2
$\hat{\boldsymbol{h}}$
Update $\boldsymbol{\theta}^{(k)}$
Calculate loss and update the CNN weights by SGD
DND $D_{\mathbf{w}^{(k)}}$
DND $D_{\mathbf{w}}$ $\mathbf{w} \in \mathbf{W}$
$\boldsymbol{K}_{\mathbf{w}}$
$\boldsymbol{V}_{\mathbf{w}}$
$\boldsymbol{h}_1$ ω_1 v_1
$\boldsymbol{h}_2$ ω_2 v_2
$\boldsymbol{h}_\zeta$ ω_ζ v_ζ
$\hat{\boldsymbol{h}}$ $\hat{v}$
(10)
(12)
$Q(\mathbf{s}^{(k)}, \mathbf{w})$
Inference
Training

Fig. 3. DRL-based friendly jamming beamforming scheme for the friendly jammers to dynamically and rapidly determine the optimal jamming beamformer adaptively in spatiotemporally nonstationary complex environments.

collected to update the overall Q-function, i.e., $Q(\mathbf{s}^{(k)}, \mathbf{w}), \mathbf{w} \in \mathbf{W}$. Using the updated Q-function, the friendly jammers can determine the jamming beamformer $\mathbf{w}^{(k)}$ for the current time slot based on the ε-greedy method, and then the selected friendly jamming beamforming action is performed.

Afterwards, a new mapping entry $(\hat{\boldsymbol{h}}, Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}))$ is recorded and appended to the end of the DND memory $D_{\mathbf{w}^{(k)}}$ for the selected action $\mathbf{w}^{(k)}$, where $Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)})$ is the updated Q-value for $\mathbf{w}^{(k)}$ in the updated Q-function. In the case where the look-up key $\hat{\boldsymbol{h}}$ already exists in the memory $D_{\mathbf{w}^{(k)}}$, its corresponding Q-value is updated to $Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)})$.

After performing the friendly jamming action, the friendly jammers obtain the feedback information from the environment. The BER $\hat{p}_e^{(k)}$ is obtained, the energy consumption $E^{(k)}$ is measured, and the secrecy rate $\hat{c}_s^{(k)}$ is estimated. Then the current utility $u^{(k)}$ is obtained, and the system turns to the next $\mathbf{s}^{(k+1)}$.

To memorize the jamming experiences for exploitation, the friendly jammers formulate the current jamming experience $\mathbf{e}^{(k)} = \{\boldsymbol{\varphi}^{(k)}, \mathbf{w}^{(k)}, u^{(k)}, \boldsymbol{\varphi}^{(k+1)}\}$, and save it in the experience pool $\mathcal{B} = \{\mathbf{e}^{(k)}\}$. A jamming experience consists of the current state-action sequence $\boldsymbol{\varphi}^{(k)}$, the next state-action sequence $\boldsymbol{\varphi}^{(k+1)}$, the jamming beamformer $\mathbf{w}^{(k)}$, and the utility $u^{(k)}$. Then, a technology named experience relay can be performed, in which a proportion of the experiences denoted as $\mathcal{T}$ are randomly selected from the experience pool $\mathcal{B}$, to update the weights $\boldsymbol{\theta}^{(k)}$ of the CNN via stochastic gradient descent (SGD) [38]. Note that it is necessary to disrupt the order of the sampling sequence over time to improve the generalization of the model. Thus, the friendly jammers randomly select T jamming experiences from $\mathcal{B}$ to formulate an experience mini-batch sequence $\mathcal{T} = \{\mathbf{e}^{(t)}\}_{1 \le t \le T}$, where $\mathbf{e}^{(t)}$ represents the t-th selected jamming experience. Then, using the SGD method, the weights $\boldsymbol{\theta}^{(k)}$ of the CNN are updated by minimizing the loss function over $\mathcal{T}$, i.e., the squared error between the output of the network and the target Q-value, as given by

$$\boldsymbol{\theta}^{(k)} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{T}} \left[\left(u - Q\left(\boldsymbol{\varphi}, \mathbf{w}; \boldsymbol{\theta}^{(k)}\right) + \beta \max_{\mathbf{w}'} Q\left(\boldsymbol{\varphi}', \mathbf{w}'; \boldsymbol{\theta}^{(k-1)}\right) \right)^2 \right], \tag{13}$$

where $\mathbb{E}$ is the expectation operator; the experience $(\boldsymbol{\varphi}, \mathbf{w}, u, \boldsymbol{\varphi}') \in \mathcal{T}$; β is the discount factor defined in (9); $\mathbf{w}'$ is the next jamming beamformer selected by the friendly jammers; $\boldsymbol{\varphi}'$ is the next state-action sequence in the jamming experience from $\mathcal{T}$.

V. Performance Evaluations

The performance of the proposed friendly jamming beamforming scheme for secure VLC is theoretically analyzed in terms of the overall VLC system utility and the receive quality of the legitimate user. The receive quality is evaluated via Bob's BER. The theoretically optimal overall utility, which is concerned with secrecy rate, energy consumption, and BER, is derived. Apart from that, the computational complexity of the proposed RL-FJ and DRL-FJ schemes are evaluated.

For simplicity, it is assumed that 4PAM is used as the modulation scheme, while the proposed approach works for other higher-order modulation schemes. Thus, the BER of the legitimate user is given by

$$p_e = \frac{3}{4} \operatorname{erfc}\left(\sqrt{\frac{2P_{\mathrm{T}} \left(\mathbf{w}^{\mathrm{T}} \mathbf{h}_{\mathrm{JB}} + h_{\mathrm{AB}}\right) \left(h_{\mathrm{AB}} + \mathbf{h}_{\mathrm{JB}}^{\mathrm{T}} \mathbf{w}\right)}{5{\sigma_{\mathrm{B}}}^2}} \right), \tag{14}$$

where $\operatorname{erfc}(x)$ represents the complementary error function as shown below

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-z^2} dz. \tag{15}$$

According to the literature [39], the energy consumption E is given by

$$E = \varrho_s \sum_{i=1}^{N_{\mathrm{J}}} w_i, \tag{16}$$

where ϱ_s denotes the unit jamming cost and w_i, $1 \leq i \leq N_{\mathrm{J}}$ is the friendly jamming beamformer for the N_{J} friendly jammers.

For a given current state and friendly jamming policy, the future state observed by the friendly jammers, which is composed of BER, secrecy rate, channel gain and energy consumption, is independent of the previous states. Thus, the decision process of the friendly jamming policy in the replicated interactions between the friendly jammers and the spatiotemporally nonstationary environment can be regarded as an MDP. The theoretical performance of the proposed friendly jamming beamforming algorithms is derived as given by the following corollary.

Corollary 1. *The proposed RL-FJ scheme as in* **Algorithm 1** *and DRL-FJ scheme as in* **Algorithm 2** *will reach the optimal friendly jamming beamformer* $\mathbf{w}^*$ *as given by*

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} u, \tag{17}$$

where the optimal utility u^* *of the VLC system is achieved, which is given by*

$$\begin{aligned} u^* =& \frac{1}{2} \log \left(1 + \frac{2 h_{\mathrm{AB}}^2 \alpha^2 I_{\mathrm{D}}^2}{\pi e \sigma_{\mathrm{B}}^2} \right) \\ & - \min \left(\log \frac{h_{\mathrm{AE}}}{\left| \mathbf{h}_{\mathrm{JE}}^{\mathrm{T}} \mathbf{w}^* \right|} + \frac{\left| \mathbf{h}_{\mathrm{JE}}^{\mathrm{T}} \mathbf{w}^* \right|}{h_{\mathrm{AE}}} \log \sqrt{e}, \frac{h_{\mathrm{AE}}}{\left| \mathbf{h}_{\mathrm{JE}}^{\mathrm{T}} \mathbf{w}^* \right|} \log \sqrt{e} \right) \\ & - \frac{3}{4} \delta_1 \operatorname{erfc} \left(\sqrt{\frac{2 \eta I_{\mathrm{e}} h_{\mathrm{AB}}^2}{5 \sigma_{\mathrm{B}}^2}} \right) - \delta_2 \varrho_s \sum_{i=1}^{N_{\mathrm{J}}} w_i, \end{aligned} \tag{18}$$

if the channel state information (CSI) of the legitimate user is known at the friendly jammers, and

$$\mathbf{h}_{\mathrm{JB}}^{\mathrm{T}} \mathbf{w} = 0, \ |\mathbf{w}| \preceq \mathbb{1}; \tag{19}$$

In particular, if

$$\frac{\left| \mathbf{h}_{\mathrm{JE}}^{\mathrm{T}} \mathbf{w} \right|}{h_{\mathrm{AE}}} \leq 1, \tag{20}$$

holds, the closed-form expression of the achievable secrecy rate c_s *and the optimal utility* u^* *can be further derived as*

$$c_s = \frac{1}{2} \log \left(1 + \frac{2 h_{\mathrm{AB}}^2 \alpha^2 I_{\mathrm{D}}^2}{\pi e \sigma_{\mathrm{B}}^2} \right) - (\log(\log \sqrt{e}) + 1), \tag{21}$$

and

$$\begin{aligned} u^* =& \frac{1}{2} \log \left(1 + \frac{2 h_{\mathrm{AB}}^2 \alpha^2 I_{\mathrm{D}}^2}{\pi e \sigma_{\mathrm{B}}^2} \right) - (\log(\log \sqrt{e}) + 1) \\ & - \frac{3}{4} \delta_1 \operatorname{erfc} \left(\sqrt{\frac{2 \eta I_{\mathrm{e}} h_{\mathrm{AB}}^2}{5 \sigma_{\mathrm{B}}^2}} \right) - \delta_2 \varrho_s \sum_{i=1}^{N_{\mathrm{J}}} w_i. \end{aligned} \tag{22}$$

Proof: See Appendix A. ■

Remark I: If the channel state information of the legitimate user is known at the friendly jammers, the optimal jamming beamformer as given by $\mathbf{w}^* = \arg\max_{\mathbf{w}} u$ that maximizes the system utility is selected, which takes the energy consumption of the transmitter and the friendly jammers, the system secrecy rate, and the receive quality of Bob into account and the best tradeoff between these factors is achieved. The constraint in (20) means that the jamming power imposed on the eavesdropper is no greater than the channel gain from the transmitter to the eavesdropper. In this case, according to (35) in the proof in Appendix A, the friendly jammers will keep optimizing the jamming power imposed on the eavesdropper and after sufficient interactions with the environment, it will converge to

$$\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}} \mathbf{w}^* = \frac{h_{\mathrm{AE}}}{\log \sqrt{e}}. \tag{23}$$

Next, consider the computational complexity of the proposed scheme. Let $\mathcal{O}(\Gamma_1)$ and $\mathcal{O}(\Gamma_2)$ denote the computational complexity of **Algorithm 1** and **Algorithm 2**, respectively. The total number of steps to finally reach the optimal jamming beamformer largely determines $\mathcal{O}(\Gamma_1)$ [40]. For simplicity, the number of steps to reach the optimal jamming beamformer in an interaction between the friendly jammers and the spatiotemporally nonstationary environment is denoted as K, and the number of interactions is denoted as Z_{F}. The complexity of the CNN module is the main source of $\mathcal{O}(\Gamma_2)$. Let f_0 represent the number of input channels of the CNN, and let c_l represent the spatial size of the output feature maps of the l-th Conv layer. The size of the feature maps that are output from the two convolutional layers is $c_1 = (m_0 - m_1)/s_1 + 1$ and $c_2 = (m_0 - m_1)/(s_1 s_2) - (m_2 - 1)/s_2 + 1$ [41], respectively. Then, the computational complexity of the proposed algorithms is derived in the following corollary.

Corollary 2. *The computational complexity of the proposed RL-FJ algorithm is given by*

$$\mathcal{O}(\Gamma_1) = \mathcal{O}(K Z_{\mathrm{F}}), \tag{24}$$

if

$$K Z_{\mathrm{F}} \geq \operatorname{poly}\{[|\Lambda|, |\mathbf{W}|, K]\}, \tag{25}$$

where $\operatorname{poly}\{\mathbf{v}\}$ *function can find the characteristic polynomial of equations or square matrices with the vector* $\mathbf{v}$ *as the solution; The computational complexity of the proposed DRL-FJ algorithm is given by*

$$\mathcal{O}(\Gamma_2) = \mathcal{O}\left(f_1 m_2^2 f_2 \left(\frac{m_0 - m_1}{s_1 s_2} - \frac{m_2 - 1}{s_2} + 1 \right)^2 \right), \tag{26}$$

if

$$f_0 = 1. \tag{27}$$

Proof: See Appendix B. ■

Remark II: The computational complexity of the proposed RL-FJ algorithm increases with totally number of steps to finally reach the optimal jamming beamformer. In order to reduce the computational complexity, transfer learning tech-

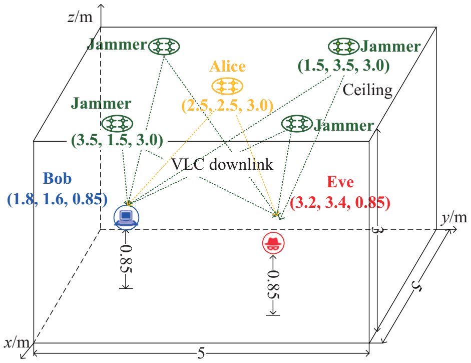


Fig. 4. Experimental simulation scenario setup for evaluating the secure VLC system against eavesdropping, including four down-facing LED light fixtures deployed on the top acting as the friendly jammers, an LED light fixture as the transmitter (Alice), a legitimate user (Bob) and a wiretapper (Eve) both equipped with VLC receivers located in random positions.

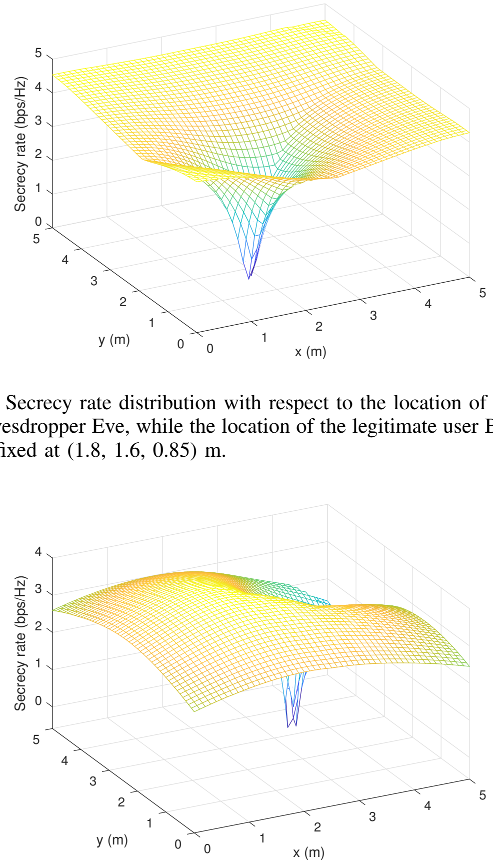
niques such as hot-booting can be utilized to obtain the initial Q-table for the RL-based algorithm, so that the overhead of initial random exploration can be greatly reduced. The size of the CNN input, along with the number, size, and stride of the filters in each Conv layer, determine the computational complexity of the proposed DRL-FJ algorithm. With the increase of the number of filters and the size of the CNN input, the learning capacity of the CNN module might improve, but a higher computational complexity is introduced. Thus, it is necessary to properly configure the CNN parameters to make a good tradeoff between the anti-eavesdropping capability and the computational complexity. In realistic applications, for user terminals with limited computing resource, the RL-FJ algorithm with lower computational complexity can be chosen to achieve moderate system performance; On the other hand, the DRL-FJ algorithm is a better choice for user terminals equipped with sufficient computational resource if the task requires high-quality secrecy communication in complex and nonstationary environments.

VI. Simulation Results and Discussions

Extensive simulation experiments are carried out to evaluate the performance and validate the effectiveness of the proposed RL-FJ and DRL-FJ beamforming schemes for secure indoor VLC. A diagram of the experimental simulation environment setup for the secure VLC system is illustrated in Fig. 4, and the main transmission parameters set according to [18] are listed in Table I. The size of the room is $5 \times 5 \times 3$ m^3 with five downward LED light fixtures on the ceiling, each quipped with four LEDs. The transmitter (Alice) sending private VLC signals is a light fixture located in the center of the ceiling, while the other four light fixtures deployed around act as the intelligent friendly jammers. The specific locations of them in the indoor three-dimensional coordinate system are marked in Fig. 4.

TABLE I
Parameter Configuration for VLC Transmission

Parameters	Value
LED average power	1 W
Modulation index α	10%
Half irradiation intensity semi-angle $\phi_{1/2}$	60°
FoV φ_{F} of the receiver PD	60°
Optical concentrator refractive index n_0	1.5
Background noise power σ_{B}^2 of Bob	-98.79 dBm
PD detector area A_{P}	1 cm^2
PD responsivity R	0.54 A/W



(a) Secrecy rate distribution with respect to the location of the eavesdropper Eve, while the location of the legitimate user Bob is fixed at (1.8, 1.6, 0.85) m.

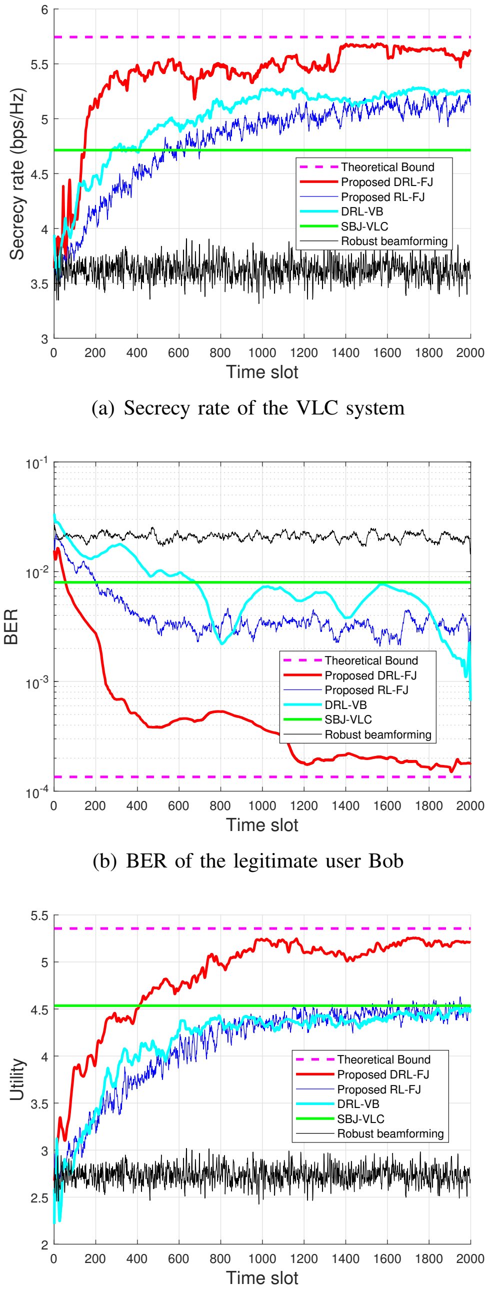
(b) Secrecy rate distribution with respect to the location of the legitimate user Bob, while the wiretapper is located in (3.2, 3.4, 0.85) m.

Fig. 5. Spatial distribution of secrecy rate with respect to locations of the legitimate user and wiretapper.

Some parameters related to the learning schemes are set as follows: The coefficients δ_1 and δ_2 in the utility function in (7) are set as 3.5 and 0.4, respectively; The learning rate λ and the discount factor β in the Q-function update in (9) are both set as 0.5; the size of the sliding window M for the state-action sequence $\boldsymbol{\varphi}^{(k)}$ is set to 11; The size of the experience mini-batch T is set as 4; We determine the hyper-parameter vector of the CNN using random search, which involves randomly sampling combinations of hyper-parameters to find the optimal hyper-parameter that strike a good balance between the secrecy performance and the computational complexity. Thus, the hyper-parameter vector of the CNN parameters is set as $\mathbf{F} = [20, 40, 3, 2, 1, 1, 180, 180]$. As a benchmark, the state-of-the-art robust beamforming scheme [42] and simultaneous beamforming and jamming for VLC (SBJ-VLC) scheme [43] are evaluated in the same simulation setup for comparison.

We first investigate the spatial distribution of the achievable secrecy rate by changing the geometric locations of Bob or Eve. As shown in Fig. 5(a) where Bob's location is fixed, as Eve moves closer towards Bob, the secrecy rate decreases rapidly and finally reaches the global minimum at the location of Bob, i.e., (1.8, 1.6, 0.85) m. In Fig. 5(b) where Eve's location is fixed, it is observed that the secrecy rate reaches the minimum value when Bob arrives at the location of Eve. The reason therein is that, when Eve and Bob are getting closer, it is difficult for the friendly jammers to find a jamming beamformer to separately impose different levels of jamming power on Eve and Bob. In fact, if the receive quality of Bob should be guaranteed, the jamming power imposed on the location of Bob (note that it is located close to Eve) should be sufficiently small, which makes it also easy for Eve to wiretap the private signal, thus resulting in significant degradation in secrecy rate. On the other hand, when the distance between Bob and Eve reaches a certain level, it is much likely that the intelligent friendly jammers can gradually learn and converge to the optimal jamming beamformer that greatly diminishes the capability of Eve to infer the private information, so the secrecy rate can be maintained at a satisfactory level in a substantial portion of the room using the proposed scheme.

The performance of secrecy rate, BER, and system utility of the proposed intelligent friendly jamming beamforming schemes over a series of time slots in the interactive learning process is shown in Fig. 6, and the benchmark schemes of robust beamforming [42] and SBJ-VLC [43] are also reported for comparison. The results demonstrate that, the proposed intelligent friendly jamming beamforming schemes can achieve a higher secrecy rate, higher utility and a lower BER over the interaction with the environment after a certain time slots. Specifically, as reported in Fig. 6(a), the proposed RL-FJ beamforming scheme improves the secrecy rate from 3.64 to 5.14 with an increase of 41.2%, while the DRL-FJ beamforming scheme can further improve it to 5.68 at the 1400-th time slot with an increase of 56.1%. As reported in Fig. 6(b), the RL-FJ beamforming scheme reduces the BER from 2.2×10^{-2} to 3.2×10^{-3} with a decrease of 85.5%, while the DRL-FJ beamforming scheme can reduce the BER to 2.0×10^{-4} at the 1200-th time slot with a decreased of 99.1%. Meanwhile, as reported in Fig. 6(c), the RL-FJ beamforming scheme improves the utility from 2.68 to 4.48 with an increase of 67.2%, while the DRL-FJ beamforming scheme converges faster and improves the utility to 5.24 with an increase of 95.5% in the middle of the process.



(a) Secrecy rate of the VLC system

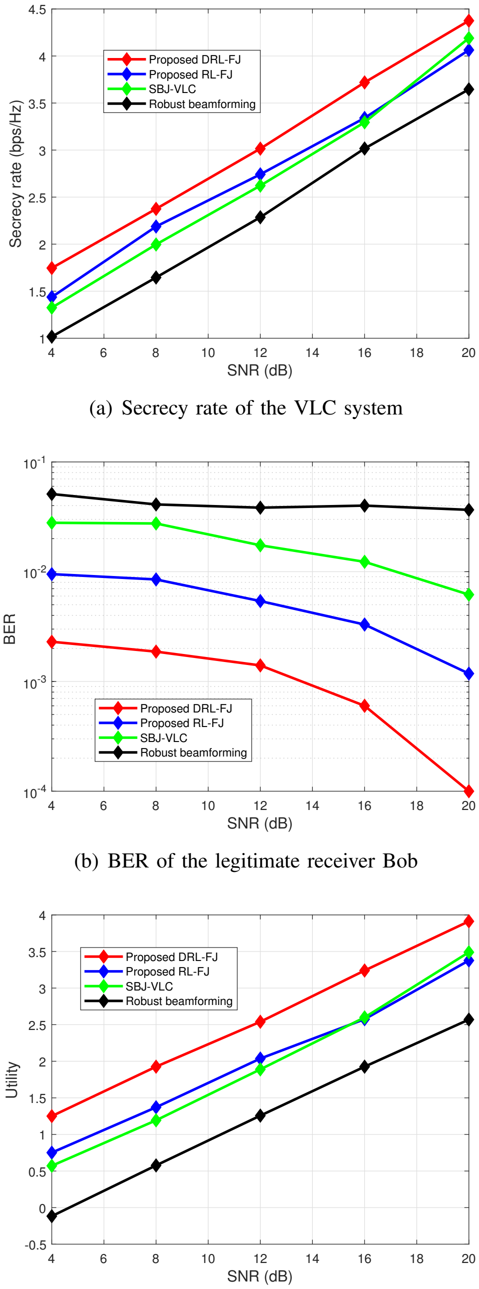
(b) BER of the legitimate user Bob

(c) Utility of the VLC system

Fig. 6. Performance of proposed intelligent friendly jamming beamforming schemes for secure VLC system against eavesdropping over a series of time slots in the interactive learning process.

We also note that the proposed intelligent friendly jamming beamforming schemes outperform the robust beamforming and the SBJ-VLC schemes in Fig. 6. Specifically, in Fig. 6(a), close to the end of the process, the RL-FJ beamforming scheme is 45.4% and 10.9% higher than the robust beamforming and SBJ-VLC schemes in terms of secrecy rate, respectively. In addition, it is worth noting that the DRL-FJ beamforming scheme is approaching the theoretical bound of the secrecy rate, which is further improved by 9.8% compared with the RL-FJ beamforming scheme. This verifies that the deep CNN module conceived in the DRL-FJ beamforming scheme can effectively represent the high-dimensional continuous state and action spaces, and the introduction of the DND memory can make full use of the experiences in previous similar anti-eavesdropping scenarios.

As for the receive quality of Bob, as shown in Fig. 6(b), the proposed RL-FJ beamforming scheme at the 1400-th time slot is reduced by 84.1% and 58.9% compared with the robust beamforming and SBJ-VLC schemes in terms of BER, respectively. The BER of the proposed DRL-FJ beamforming scheme is further reduced by 93.2% compared to the proposed RL-FJ beamforming scheme, which is approaching the theoretical bound after 1800 time slots. Considering the overall



(a) Secrecy rate of the VLC system

(b) BER of the legitimate receiver Bob

(c) Utility of the VLC system

Fig. 7. Performance of proposed intelligent friendly jamming beamforming schemes with respect to SNR.

performance reflected by system utility, as shown in Fig. 6(c), the utility of the RL-FJ beamforming scheme at the 1700-th slot is improved by 64.9% and 1.2% compared with the robust beamforming and SBJ-VLC schemes, respectively. The DRL-FJ beamforming scheme further improves the utility by 16.7% over RL-FJ beamforming, and gradually approaches the theoretical bound. The simulation results verifies that the proposed intelligent friendly jamming beamforming schemes, especially the DRL-FJ beamforming scheme, can approach the optimal solution through rapid interactive learning.

It can also be seen from Fig. 6 that at 1800-th time slot, the secrecy rate and the system utility of the DRL-FJ scheme are higher by 7.0% and 17.1% respectively than that of the DRL-based MISO VLC beamforming (DRL-VB) scheme in [20]. Moreover, the BER of the DRL-FJ scheme is lower by 96.3% compared to the DRL-VB scheme. This verifies that the introduction of intelligent friendly jammers brings additional degrees of freedom in the strategy of the RL algorithm, which can find an optimal solution that maximizes the transmitting power of Alice to Bob and meanwhile minimizes the signal-to-noise ratio (SNR) of the eavesdropper Eve, leading to a better performance of the receive quality of the legitimate user Bob, and higher secrecy rate of the VLC system in the presence of Eve, and further improves the performance of the system.

Furthermore, the proposed method is evaluated in different values of SNR, which is reported in Fig. 7. According to the visible light channel propagation characteristics [7], the SNR level varies significantly with the spatial location. A typical SNR range is between 4 dB and 20 dB, which is considered in the simulations. The performance shown in Fig. 7 is calculated from the average of the first 2000 time slots. It is demonstrated from Fig. 7(a) and Fig. 7(c) that, the secrecy rate and system utility of the RL-FJ and DRL-FJ beamforming schemes increase with the SNR. As the SNR increases from 4 dB to 20 dB, the secrecy rate and the system utility of the DRL-FJ beamforming scheme are increased by 151% and 213%, respectively. It can also be concluded from Fig. 7(b) that the BER decreases with the SNR. It is observed that at the target BER of 1.0×10^{-3}, the DRL-FJ beamforming scheme can achieve an SNR gain of approximately 6 dB over the RL-FJ beamforming scheme, and enjoys an even larger SNR gain over the robust beamforming scheme and the SBJ-VLC scheme.

Moreover, it can be concluded from Fig. 7 that, at the SNR of 12 dB, the RL-FJ beamforming scheme improves the secrecy rate by 20.0% and 4.6%, and improves the system utility by 62.1% and 7.8%, compared with the robust beamforming and SBJ-VLC schemes, respectively. This validates the anti-eavesdropping capability of the proposed RL-based framework of friendly jamming for secure VLC. Besides, the utility and secrecy rate of the DRL-FJ beamforming scheme are 24.7% and 10.0% higher than the RL-FJ beamforming scheme, respectively, which verifies the effectiveness of the conceived DRL-based architecture.

VII. Conclusion

With the emergence and rapid adoption of VLC technology, the problem of secure privacy-preserving yet highly energy-efficient and high-rate transmission in complex and open visible light channels remains to be resolved in a more effective approach. In light of this great challenge, an RL-based intelligent anti-eavesdropping framework has been formulated in this paper, which dynamically optimize the friendly jamming policy to achieve the optimal system utility in realistic spatiotemporally nonstationary environments. Meanwhile, a DRL-FJ beamforming scheme is devised to resolve the difficult problem of the dimensional curse and effectively represent the continuous state and action spaces. A DND memory module is introduced to store the previous experiences in similar anti-eavesdropping scenarios to accelerate the convergence of learning. The simulation results have verified the superiority of the conceived method in secrecy rate, bit error rate, and system utility compared with some of the existing benchmark schemes. It is promising for the proposed scheme to be applied in various indoor scenarios where an intelligent and adaptive solution of secure, efficient, and high-rate transmission link should be established.

Appendix A
Proof of Corollary 1

According to equations (4), (7), (14) and (16), if (19) holds, the system utility can be rewritten as

$$\begin{aligned} u =&\frac{1}{2}\log\left(1+\frac{2h_{\mathrm{AB}}^2\alpha^2 I_{\mathrm{D}}^2}{\pi e\sigma_{\mathrm{B}}^2}\right) \\ &-\min\left(\log\frac{h_{\mathrm{AE}}}{\left|\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}\right|}+\frac{\left|\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}\right|}{h_{\mathrm{AE}}}\log\sqrt{e},\frac{h_{\mathrm{AE}}}{\left|\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}\right|}\log\sqrt{e}\right) \\ &-\frac{3}{4}\delta_1\,\mathrm{erfc}\left(\sqrt{\frac{2\eta I_{\mathrm{e}}h_{\mathrm{AB}}^2}{5\sigma_{\mathrm{B}}^2}}\right)-\delta_2\varrho_s\sum_{i=1}^{N_{\mathrm{J}}}w_i. \end{aligned} \quad (28)$$

Without loss of generality, the jamming power $\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}$ imposed on the eavesdropper is assumed to be non-negative. In fact, if $\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}<0$, one can replace $\mathbf{w}$ with $-\mathbf{w}$ without changing the system secrecy rate result or violating the amplitude constraints. For simplicity of notations, let H denote $\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}$. Then, we have

$$\frac{\partial u}{\partial H}=\begin{cases}\frac{1}{H}-\frac{\log\sqrt{e}}{h_{\mathrm{AE}}} & \frac{H}{h_{\mathrm{AE}}}\leq 1,\\ \frac{h_{\mathrm{AE}}\log\sqrt{e}}{H^2} & \frac{H}{h_{\mathrm{AE}}}>1.\end{cases} \quad (29)$$

Since $H\geq 0$, thus $\partial u/\partial H\geq 0$, which suggests that the system utility increases monotonically with H. Hence, the optimal system utility can be achieved when the maximum jamming power is applied to Eve, and thus the optimal jamming beamformer is derived by $\mathbf{w}^*=\arg\max_{\mathbf{w}} u=\arg\max_{\mathbf{w}}\mathbf{h}_{\mathrm{JE}}^{\mathrm{T}}\mathbf{w}$, which leads to the optimal system utility as given by (18).

Moreover, if (20) holds, we have

$$c_s=\frac{1}{2}\log\left(1+\frac{2h_{\mathrm{AB}}^2\alpha^2 I_{\mathrm{D}}^2}{\pi e\sigma_{\mathrm{B}}^2}\right)-\left(\log\frac{h_{\mathrm{AE}}}{H}+\frac{H}{h_{\mathrm{AE}}}\log\sqrt{e}\right), \quad (30)$$

$$\begin{aligned} u^* =&\frac{1}{2}\log\left(1+\frac{2h_{\mathrm{AB}}^2\alpha^2 I_{\mathrm{D}}^2}{\pi e\sigma_{\mathrm{B}}^2}\right)-\left(\log\frac{h_{\mathrm{AE}}}{H}+\frac{H}{h_{\mathrm{AE}}}\log\sqrt{e}\right) \\ &-\frac{3}{4}\delta_1\,\mathrm{erfc}\left(\sqrt{\frac{2\eta I_{\mathrm{e}}h_{\mathrm{AB}}^2}{5\sigma_{\mathrm{B}}^2}}\right)-\delta_2\varrho_s\sum_{i=1}^{N_{\mathrm{J}}}w_i, \end{aligned} \quad (31)$$

$$\frac{\partial u}{\partial H}=\frac{1}{H}-\frac{\log\sqrt{e}}{h_{\mathrm{AE}}}, \quad (32)$$

and

$$\frac{\partial^2 u}{\partial H^2}=-\frac{1}{H^2}<0. \quad (33)$$

From (32), we have

$$\left.\frac{\partial u}{\partial H}\right|_{H=H^*}=0, \quad (34)$$

where

$$H^*=\frac{h_{\mathrm{AE}}}{\log\sqrt{e}}. \quad (35)$$

Since $|\mathbf{w}|\preceq\mathbb{1}$, H will not be infinite. Let $\mathcal{P}_{\max}$ represent the maximum value of H, i.e., $0\leq H\leq\mathcal{P}_{\max}$. Then, if (33) holds, we have

$$u(H^*)\geq u(H). \quad (36)$$

According to the literature [21] and [44], the RL-based algorithms on episodic MDP can converge to the strategy H^* after a sufficiently long time. Thus, **Algorithm 1** and **Algorithm 2** can achieve H^* in (35). By integrating (35) into (30) and (31), we have (21) and (22).

Appendix B
Proof of Corollary 2

It takes KZ_{F} operations to finally reach the optimal jamming beamformer as stated in **Algorithm 1**. Let $|\Lambda|$ represent the size of the state space, and $|\mathbf{W}|$ represent the size of the action space. If $KZ_{\mathrm{F}}\geq\mathrm{poly}\{[|\Lambda|,|\mathbf{W}|,K]\}$, where $\mathrm{poly}\{[|\Lambda|,|\mathbf{W}|,K]\}$ function can find the characteristic polynomial of equations or square matrices with the vector $[|\Lambda|,|\mathbf{W}|,K]$ as the solution, then the computational complexity of the RL-FJ algorithm is given by $\mathcal{O}(\Gamma_1)=\mathcal{O}(KZ_{\mathrm{F}})$ [40]. That is, if (25) holds, we have (24).

The computational complexity $\mathcal{O}(\Gamma_2)$ of the DRL-FJ scheme in **Algorithm 2** is mainly contributed by the two Conv layers of the CNN, because the computational complexity of the quadratic operation in the two Conv layers is much larger than that of the linear operation in the two FC layers. Then, if (27) holds and according to [45], we have

$$\begin{aligned} \mathcal{O}(\Gamma_2) =&\mathcal{O}\left(\sum_{l=1}^{2}f_{l-1}m_l^2 f_l c_l^2\right) \\ =&\mathcal{O}\left(f_0 m_1^2 f_1\left(\frac{m_0-m_1}{s_1}+1\right)^2\right. \\ &\left.+f_1 m_2^2 f_2\left(\frac{m_0-m_1}{s_1 s_2}-\frac{m_2-1}{s_2}+1\right)^2\right) \\ =&\mathcal{O}\left(m_1^2 f_1\left(\frac{m_0-m_1}{s_1}+1\right)^2\right. \\ &\left.+f_1 m_2^2 f_2\left(\frac{m_0-m_1}{s_1 s_2}-\frac{m_2-1}{s_2}+1\right)^2\right). \end{aligned} \quad (37)$$

According to the CNN architecture in [38], we have

$$m_1^2 f_1\left(\frac{m_0-m_1}{s_1}+1\right)^2\ll f_1 m_2^2 f_2\left(\frac{m_0-m_1}{s_1 s_2}-\frac{m_2-1}{s_2}+1\right)^2. \quad (38)$$

Thus, from (37) and (38), we have (26).

References

[1] P. Zhang, L. Li, K. Niu, Y. Li, G. Lu, and Z. Wang, "An intelligent wireless transmission toward 6G," *Intelligent and Converged Networks*, vol. 2, no. 3, pp. 244–257, Sep. 2021.

[2] M. A. Arfaoui, M. D. Soltani, I. Tavakkolnia, A. Ghrayeb, M. Safari, C. M. Assi, and H. Haas, "Physical layer security for visible light communication systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1887–1908, 3rd Quart., 2020.

[3] H. Haas, L. Yin, Y. Wang, and C. Chen, "What is LiFi?," *J. Lightw. Technol.*, vol. 34, no. 6, pp. 1533–1544, Mar. 2016.

[4] L. Yin and H. Haas, "Physical-layer security in multiuser visible light communication networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 1, pp. 162–174, Jan. 2018.

[5] J. Chen and T. Shu, "Statistical modeling and analysis on the confidentiality of indoor VLC systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4744–4757, Jul. 2020.

[6] F. Xing, S. He, V. C. M. Leung, and H. Yin, "Energy efficiency optimization for rate-splitting multiple access-based indoor visible light communication networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1706–1720, May 2022.

[7] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2047–2077, 4th Quart., 2015.

[8] X. Ma, F. Yang, S. Liu, and J. Song, "Channel estimation for wideband underwater visible light communication: A compressive sensing perspective," *Opt. Express*, vol. 26, no. 1, pp. 311–321, Aug. 2018.

[9] A. Mostafa and L. Lampe, "Securing visible light communications via friendly jamming," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, pp. 524–529, Austin, TX, USA, Dec. 2014.

[10] Y. Xu, J.-M. Frahm, and F. Monrose, "Watching the watchers: Automatically inferring TV content from outdoor light effusions," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, pp. 418–428, New York, NY, USA, 2014.

[11] A. G. Fragkiadakis, E. Z. Tragos, and I. G. Askoxylakis, "A survey on security threats and detection techniques in cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 428–445, 1st Quart., 2013.

[12] A. Mukherjee, S. A. A. Fakoorian, J. Huang, and A. L. Swindlehurst, "Principles of physical layer security in multiuser wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1550–1573, 3rd Quart., 2014.

[13] H. V. Poor, "Information and inference in the wireless physical layer," *IEEE Wireless Commun.*, vol. 19, no. 1, pp. 40–47, Feb. 2012.

[14] A. Chorti, S. M. Perlaza, Z. Han, and H. V. Poor, "On the resilience of wireless multiuser networks to passive and active eavesdroppers," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1850–1863, Sep. 2013.

[15] H. Zaid, Z. Rezki, A. Chaaban, and M. S. Alouini, "Improved achievable secrecy rate of visible light communication with cooperative jamming," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, pp. 1165–1169, Orlando, FL, USA, Dec. 2015.

[16] A. Sikri, A. Mathur, M. Bhatnagar, G. Kaddoum, P. Saxena, and J. Nebhen, "Artificial noise injection-based secrecy improvement for FSO systems," *IEEE Photon. J.*, vol. 13, no. 2, pp. 1–12, Apr. 2021.

[17] F. Wang, C. Liu, Q. Wang, J. Zhang, R. Zhang, L.-L. Yang, and L. Hanzo, "Optical jamming enhances the secrecy performance of the generalized space-shift-keying-aided visible-light downlink," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4087–4102, Sep. 2018.

[18] A. Mostafa and L. Lampe, "Physical-layer security for MISO visible light communication channels," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 9, pp. 1806–1818, Sep. 2015.

[19] H. Shen, Y. Deng, W. Xu, and C. Zhao, "Secrecy-oriented transmitter optimization for visible light communication systems," *IEEE Photon. J.*, vol. 8, no. 5, pp. 1–14, Oct. 2016.

[20] L. Xiao, G. Sheng, S. Liu, H. Dai, M. Peng, and J. Song, "Deep reinforcement learning-enabled secure visible light communication against eavesdropping," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6994–7005, Oct. 2019.

[21] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[22] A. Uprety and D. B. Rawat, "Reinforcement learning for IoT security: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8693–8706, Jun. 2021.

[23] G. Faraci, C. Grasso, and G. Schembra, "Design of a 5G network slice extension with MEC UAVs managed with reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2356–2371, Oct. 2020.

[24] S. Liu, C. Zheng, Y. Huang, and T. Q. S. Quek, "Distributed reinforcement learning for privacy-preserving dynamic edge caching," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 749–760, Mar. 2022.

[25] N. Jiang, Y. Deng, A. Nallanathan, and J. A. Chambers, "Reinforcement learning for real-time optimization in NB-IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1424–1440, Jun. 2019.

[26] D. A. Saifaldeen, B. S. Ciftler, M. M. Abdallah, and K. A. Qaraqe, "DRL-based IRS-assisted secure visible light communications," *IEEE Photon. J.*, vol. 14, no. 6, pp. 1–9, Dec. 2022.

[27] H. Yang, A. Alphones, W.-D. Zhong, C. Chen, and X. Xie, "Learning-based energy-efficient resource management by heterogeneous RF/VLC for ultra-reliable low-latency industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5565–5576, Aug. 2020.

[28] B. S. Ciftler, A. Alwarafy, and M. Abdallah, "Distributed DRL-based downlink power allocation for hybrid RF/VLC networks," *IEEE Photon. J.*, vol. 14, no. 3, pp. 1–10, Jun. 2022.

[29] M. R. Maleki, M. R. Mili, M. R. Javan, N. Mokari, and E. A. Jorswieck, "Multi-agent reinforcement learning trajectory design and two-stage resource management in CoMP UAV VLC networks," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7464–7476, Nov. 2022.

[30] A. A. hammadi, L. Bariah, S. Muhaidat, M. Al-Qutayri, P. C. Sofotasios, and M. Debbah, "Deep Q-learning-based resource allocation in NOMA visible light communications," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 2284–2297, Nov. 2022.

[31] F. H. Panahi, F. H. Panahi, and T. Ohtsuki, "Intelligent cellular offloading with VLC-enabled unmanned aerial vehicles," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 17718–17733, May 2023.

[32] S. Amuru, C. Tekin, M. V. der Schaar, and R. M. Buehrer, "Jamming bandits—A novel learning method for optimal jamming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2792–2808, Apr. 2016.

[33] A. Pritzel, B. Uria, S. Srinivasan, A. Puigdomènech, O. Vinyals, D. Hassabis, D. Wierstra, and C. Blundell, "Neural episodic control," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, pp. 2827–2836, Sydney, NSW, Australia, Aug. 2017.

[34] X. Liu and S. Liu, "Jamming for secrecy: Reinforcement learning based anti-eavesdropping visible light communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 2053–2058, Seoul, Korea, Republic of, May 2022.

[35] F. Yang, Y. Sun, and J. Gao, "Adaptive LACO-OFDM with variable layer for visible light communication," *IEEE Photon. J.*, vol. 9, no. 6, pp. 1–8, Dec. 2017.

[36] L. Zeng, D. C. O'Brien, H. L. Minh, G. E. Faulkner, K. Lee, D. Jung, Y. Oh, and E. T. Won, "High data rate multiple input multiple output (MIMO) optical wireless communications using white LED lighting," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 9, pp. 1654–1662, Dec. 2009.

[37] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using LED lights," *IEEE Trans. Consum. Electron.*, vol. 50, no. 1, pp. 100–107, Feb. 2004.

[38] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[39] L. Xiao, H. Zhang, Y. Xiao, X. Wan, S. Liu, L.-C. Wang, and H. V. Poor, "Reinforcement learning-based downlink interference control for ultra-dense small cells," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 423–434, Jan. 2020.

[40] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 4868–4878, Montréal, QC, Canada, Dec. 2018.

[41] C. C. T. Mendes, V. Frémont, and D. F. Wolf, "Exploiting fully convolutional neural networks for fast road detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 3174–3179, Stockholm, Sweden, May 2016.

[42] H. Ma, A. Mostafa, L. Lampe, and S. Hranilovic, "Coordinated beamforming for downlink visible light communication networks," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3571–3582, Aug. 2018.

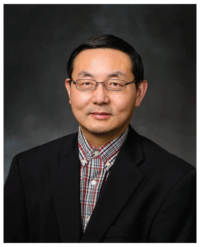
[43] S. Cho, G. Chen, and J. P. Coon, "Enhancement of physical layer security with simultaneous beamforming and jamming for visible light communication systems," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2633–2648, Oct. 2019.

[44] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Comput.*, vol. 6, no. 6, pp. 1185–1201, Nov. 1994.

[45] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Reconit. (CVPR)*, pp. 5353–5360, Boston, MA, USA, Jun. 2015.

Sicong Liu (Senior Member, IEEE) is an Associate Professor in the Department of Information and Communication Engineering, School of Informatics, Xiamen University, China. Dr. Liu received his B.S.E. and the PhD degree both in electronic engineering from Tsinghua University, Beijing, China in 2012 and 2017 with the highest honor. He was with the Department of Electronics and Communications Engineering, City University of Hong Kong, as a visiting scholar from 2010 to 2011. His current research interests are compressed sensing, AI-assisted communications, integrated sensing and communications, and visible light communications. He is on the editorial board of several academic journals. He has served as the TPC chair of several IEEE/ACM international conferences. He is a Senior Member of IEEE and the China Institute of Communications.

Xianbin Liu received the B.S. degree in electronic information engineering from Fujian Normal University, Fuzhou, China in 2021. He is currently pursuing the M.S. degree with the Department of Information and Communication Engineering, Xiamen University, Xiamen, China. His research interests include visible light communications and reinforcement learning.



Xiaojiang (James) Du (Fellow, IEEE) is the Anson Wood Burchard Endowed-Chair Professor in the Department of Electrical and Computer Engineering at Stevens Institute of Technology. He was a professor at Temple University between August 2009 and August 2021. Dr. Du received his B.S. from Tsinghua University, Beijing, China in 1996. He received his M.S. and Ph.D. degree in Electrical Engineering from the University of Maryland, College Park in 2002 and 2003, respectively. His research interests are security, wireless networks, and systems. He has authored over 500 journal and conference papers in these areas, including the top security conferences IEEE S&P, USENIX Security, and NDSS. Dr. Du has been awarded more than 8 million US Dollars research grants from the US National Science Foundation (NSF), Army Research Office, Air Force Research Lab, the State of Pennsylvania, and Amazon. He won the best paper award at IEEE ICC 2020, IEEE GLOBECOM 2014 and the best poster runner-up award at the ACM MobiHoc 2014. He serves on the editorial boards of three IEEE journals. Dr. Du is an IEEE Fellow, an ACM Distinguished Member, and an ACM Life Member.

Mohsen Guizani (Fellow, IEEE) received the BS (with distinction), MS and PhD degrees in Electrical and Computer engineering from Syracuse University, Syracuse, NY, USA in 1985, 1987 and 1990, respectively. He is currently a Professor of Machine Learning at the Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Previously, he worked in different institutions in the USA. His research interests include applied machine learning and artificial intelligence, smart city, Internet of Things (IoT), intelligent autonomous systems, and cybersecurity. He became an IEEE Fellow in 2009 and was listed as a *Clarivate Analytics Highly Cited Researcher in Computer Science in 2019, 2020, 2021 and 2022*. Dr. Guizani has won several research awards including the "2015 IEEE Communications Society Best Survey Paper Award", the Best ComSoc Journal Paper Award in 2021 as well 5 Best Paper Awards from ICC and Globecom Conferences. He is the author of 11 books, more than 1000 publications and several US patents. He is also the recipient of the 2017 IEEE Communications Society Wireless Technical Committee (WTC) Recognition Award, the 2018 AdHoc Technical Committee Recognition Award, and the 2019 IEEE Communications and Information Security Technical Recognition (CISTC) Award. He served as the Editor-in-Chief of IEEE Network and is currently serving on the Editorial Boards of many IEEE Transactions and Magazines. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer.