Deep Reinforcement Learning-Enabled Secure Visible Light Communication Against Eavesdropping

Liang Xiao, *Senior Member, IEEE*, Geyi Sheng, Sicong Liu, *Member, IEEE*, Huaiyu Dai, *Fellow, IEEE*, Mugen Peng, *Senior Member, IEEE*, and Jian Song, *Fellow, IEEE*

***Abstract*— The inherent broadcast characteristics of the visible light communication (VLC) channel makes VLC downlinks susceptible to unauthorized terminals in many actual VLC scenarios, such as offices and shopping centers. This paper considers a multiple-input-single-output (MISO) VLC scenario with multiple light fixtures acting as the transmitter, a VLC receiver as the legitimate user, and an eavesdropper attempting to intercept the undisclosed information. To improve the confidentiality of VLC links, a physical-layer anti-eavesdropping framework is proposed to obscure the unauthorized eavesdroppers and diminishes their capability of inferring the information through smart beamforming over the MISO VLC wiretap channel. To cope with the intractable problem of finding the theoretically optimal solution of the secrecy rate and utility for the MISO VLC wiretapping channel, a reinforcement learning (RL)-based VLC beamforming control scheme is proposed to achieve the optimal beamforming policy against the eavesdropper. Furthermore, a deep RL-based VLC beamforming control scheme is proposed to handle the curse of dimensionality for both observation space and action space and avoid the quantization error of the RL-based algorithm. Simulation results show that the proposed learning-based VLC beamforming control schemes can significantly decrease the bit error rate of the legitimate receiver and increase the secrecy rate and utility of the anti-eavesdropping MISO VLC system, compared with the benchmark strategy.**

***Index Terms*— Eavesdropping, visible light communication, secrecy rate, beamforming, deep reinforcement learning.**

Manuscript received January 23, 2019; revised June 1, 2019; accepted July 11, 2019. Date of publication July 22, 2019; date of current version October 16, 2019. This work was supported in part by the Natural Science Foundation of China (Grant No. 61671396, 61871339, 61731012 and 61831002), in part by the Natural Science Foundation of Fujian Province of China (Grant No. 2019J05001 and 2019J01843), in part by Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (Grant No. 2018D08), in part by Fundamental Research Funds for the Central Universities of China (Grant No. 20720190029), in part by Natural Science Foundation of Guangdong Province (Grant No. 2015A030312006), in part by US National Science Foundation (Grant No. EARS-1444009 and CNS-1824518), and in part by the State Major Science and Technology Special Project under 2017ZX03001025-006. The associate editor coordinating the review of this paper and approving it for publication was W. Xu. *(Corresponding Author: Sicong Liu.)*

L. Xiao is with the Department of Communication Engineering, Xiamen University, Xiamen 361005, China, with the Key Laboratory of Digital Fujian on IoT Communication, Architecture and Security Technology, Xiamen University, Xiamen 361005, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China (e-mail: lxiao@xmu.edu.cn).

G. Sheng and S. Liu are with the Department of Communication Engineering, Xiamen University, Xiamen 361005, China, and also with the Key Laboratory of Digital Fujian on IoT Communication, Architecture and Security Technology, Xiamen University, Xiamen 361005, China (e-mail: liusc@xmu.edu.cn).

H. Dai is with the Department of Electrical and Computer Engineering, NC State University, Raleigh, NC 27607 USA (e-mail: huaiyu_dai@ncsu.edu).

M. Peng is with the Key Laboratory of Universal Wireless Communication (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: pmg@bupt.edu.cn).

J. Song is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, with the Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China, and also with the Key Laboratory of Digital TV System of Guangdong Province and Shenzhen City, Research Institute of Tsinghua University in Shenzhen, Shenzhen 518057, China (e-mail: jsong@tsinghua.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2019.2930247

I. Introduction

VISIBLE light communication (VLC) has been recently acknowledged as a promising technology for the next-generation wireless communications, which uses light-emitting diodes (LEDs) to transmit signals to receivers such as photodiodes (PDs) [1], [2]. In the face of the limited spectrum of the conventional radio-frequency bands, VLC can meet the rapidly increasing requirement for high rate transmission and ubiquitous coverage, especially for indoor communication environments [3], [4]. Benefiting from the ultra-wide license-free light spectrum, the robustness against interference and low implementation cost [5], [6], VLC has been widely applied in many different kinds of areas including indoor localization, vehicular networks and mobile health-monitoring [7], [8], etc. Because of the inherent broadcast characteristics of the line-of-sight (LoS) VLC transmission and coverage properties, many potential security loopholes have emerged that jeopardize the secure transmission of the legitimate VLC users and network administrators. It is much likely for the transmitted information to be wiretapped by malicious attackers within the range of the light, especially in public areas such as offices, train stations, coffee shops and libraries [9]. Therefore, it is essential for VLC systems to devise a secure and efficient mechanism against eavesdropping, as is investigated and emphasized in [10].

Conventional security approaches protect legitimate users from eavesdropping via various high-layer techniques such as authentication and encryption, etc. However, with the rapid development of the computational capability of illegitimate terminals, the methods mentioned above may be ineffective

because the encryption methods can be decrypted by the potential advanced eavesdroppers. To close the vulnerable backdoor of these classical encryption techniques to potential attackers, physical (PHY)-layer security technologies have recently gained considerable attention, which exploit the PHY-layer channel characteristics and communication and signal processing techniques to hide the secrecy information from unauthorized terminals and secure the transmission over the wireless or VLC wiretap channel [11]–[14]. Among the related studies, the PHY-layer security framework pioneered in [15] proposes a fundamental information-theoretic security metric termed as secrecy capacity over the noisy wiretap environment, where an wiretapper will have access to the signal of interest with some degradation. Friendly jamming signals are utilized in [9], [16] to achieve an effective beamforming scheme to suppress the reception quality of the eavesdropper and further improve the secrecy rate of the VLC system. The achievable secrecy rate is maximized in [10] through zero-forcing beamforming over the VLC wiretap channel with multiple LED transmitters.

Nevertheless, it might be difficult to obtain the ideal wiretapping channel state information (CSI) when the location of the eavesdropper is unknown in practice, which is necessary for implementing zero-forcing beamforming. To improve state-of-the-art methods in practical VLC systems, a smart beamforming approach based on reinforcement learning (RL) is proposed in this work for the multiple-input single-output (MISO) VLC wiretap channel. Multiple light sources are adopted as the transmitter for smart beamforming policies obtained through the RL process. The purpose of the proposed RL-based smart beamforming control scheme is to reduce the information received by the eavesdropper and improve the desired signal level at the legitimate receiver in the MISO VLC wiretap channel.

In this paper, the proposed smart transmitter determines its beamforming policy based on the previous security performance such as the secrecy rate of the VLC system and the bit error rate (BER) of the legitimate receiver. Since the repeated beamforming policy control in a dynamic anti-eavesdropping communication process is modeled by Markov decision process (MDP), the methods like Q-learning based on RL [17]–[20] will provide an effective beamforming selection for the MISO VLC system. However, the Q-learning technique could only effectively handle discrete and low-dimensional action spaces. Neural network based methods, such as recurrent neural network have been used to achieve the channel state information (CSI) compression and improve the accuracy of quantized CSI feedback [21]. To satisfy the security requirement of practical VLC systems against wiretapping and keep track of the information of the complicated and high-dimensional structure of the beamforming policy domain, deep reinforcement learning (DRL) based algorithm is introduced, i.e., deterministic policy gradient (DDPG) [22]. The proposed DRL-based algorithm for MISO VLC smart beamforming control fully exploits the actor-critic approach [23] and deep Q network [24] to make it easier for learning, and to achieve a better overall anti-eavesdropping performance.

Consequently, an RL-based smart beamforming framework and an RL-based MISO VLC beamforming (RL-VB) algorithm are proposed for MISO VLC systems to protect the legitimate receiver from eavesdropping via trial and error. It is shown that the proposed RL-based scheme is able to achieve the optimal beamforming policy after a sufficient number of time slots of learning iterations. Furthermore, a DRL-based smart beamforming framework along with a DRL-based MISO VLC beamforming (DRL-VB) algorithm is proposed to cope with the continuous action space and high-dimensional state space to support efficient and effective anti-eavesdropping with higher security requirement. It is verified by the simulations that the DRL beamforming control method further increases the utility and the secrecy rate of the MISO VLC system compared with the RL-based scheme, and significantly decrease the BER of the legitimate receiver compared with the state-of-the-art benchmark.

To summarize, the main contributions of the work are as follows:

- A PHY-layer anti-eavesdropping framework is formulated over the MISO VLC wiretap channel, in which the transmitter exploits smart beamforming to decrease the eavesdropped signal level and improve the desired signal level of the legitimate receiver.
- An RL-based smart beamforming scheme is proposed to achieve the optimal beamforming, where the transmitter chooses the beamforming policy dynamically in an MDP.
- A DRL-based smart beamforming method is further proposed to increase the learning rate and performance by fully exploiting the information of the complicated and high-dimensional structure of the beamforming policy domain and jointly utilizing the actor-critic approach and deep neural networks.

The remainder of the paper is summarized in the following: Section II shows the related work and Section III presents the VLC model and eavesdropping model. Section IV and V show the RL-based and DRL-based smart beamforming control schemes, respectively. The performance bound of the proposed beamforming control scheme is analyzed in Section VI. Section VII analyzes the simulation results and Section VIII summarizes the work with several remarks.

II. Related Work

The lower and upper bounds of a free-space optical SISO channel capacity with a peak intensity constraint or an average intensity constraint are investigated in [25]. As an extension, the lower and upper bounds of the capacity with a total average intensity constraint is concluded in [26] for the constant parallel VLC channel assuming perfect CSI at the transmitter. The perfect secrecy capacity of the MIMO broadcast channel is developed in [27] to guarantee that the eavesdropper receives no signal. Considering some commonly adopted metrics of secrecy performance, a precoding scheme is proposed in [28] for multi-user MISO VLC broadcast channels.

As a key method for secrecy protection, the beamforming technique has drawn particular attention [10]. A robust transmit beamformer is proposed for the maximization of the achievable secrecy rate of a MISO VLC system subject to amplitude constraints, when part of the CSI or location

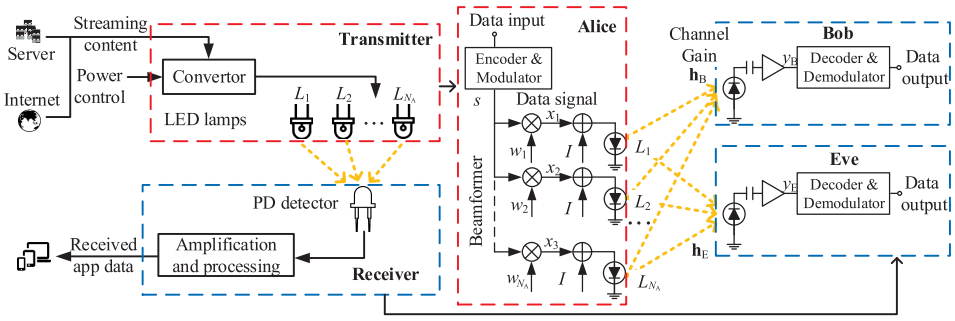


Fig. 1. Transmit beamforming over a MISO VLC wiretap channel, consisting of one transmitter (Alice), one legitimate VLC receiver (Bob), and one eavesdropper (Eve).

of the eavesdropper is known [29]. A beamforming method using artificial noise is proposed in [30] to decrease the eavesdropper's SNR over the MISO VLC wiretap channel. To enhance the communication secrecy, a PHY-layer scheme combining transmit beamforming and friendly jamming is designed for a MISO VLC system in the presence of multiple eavesdroppers [31]. The work in [32] provides the analysis of the mutual information and the achievable secrecy rate for the generalized space shift keying VLC system with a jamming-aided secrecy enhancement scheme. To solve the beamforming precoding problem, a robust MMSE beamforming scheme is proposed to minimize the cost function indicated by mean squared error in the VLC beamforming system, which can be used to suppress both intersymbol interference and crosstalk interference in wired channels [33], [34].

There have been some studies on RL-based methods for optimal strategy decision in a MDP [35], [36]. For instance, a deep RL-based power allocation strategy for unmanned aerial vehicles is investigated in [19] to obtain the optimal power allocation against intelligent attacks in a dynamic game. A hotbooting unmanned aerial vehicle relay algorithm based on policy hill climbing is investigated in [20] to facilitate the vehicular ad hoc network to prevent from malicious interference.

III. System Model

A. Channel Model

A PAM modulated VLC system with a DC bias is investigated with an LED transmitter driven by a current bias I to obtain positive values for the pulses. The zero-mean current data signal $x \in \mathbb{R}$ is superimposed on I to modulate the optical power P_{T} going out of the LED. For the sake of the current-light conversion linearity and clipping distortion avoidance, the total current $I+x$ should be limited in a specific range [37]. Thus the amplitude of the current data signal x satisfies the constraint denoted by $|x| \leq \alpha I$, where $\alpha \in [0, 1]$ is the modulation index. Then the transmitter implements electro-optical conversion and the instantaneous optical power is derived by $P_{\mathrm{T}} = \eta(I + x)$, where η is the electro-optical conversion efficiency of the LED.

The optical power collected by a PD at the receiver is given by $P_{\mathrm{R}} = GP_{\mathrm{T}}$, where G represents the path gain. The PD with the responsivity of R converts the received optical power P_{R} to the corresponding current. After removing the DC bias, the signal is amplified with the gain of T to generate the received signal denoted by y.

Assuming the LEDs are of the Lambertian emission pattern, the emission angle with respect to the optical axis of the transmitter is denoted by ϕ, the LED half luminous intensity semi-angle is denoted by $\phi_{1/2}$, the incidence angle of the light at the receiver is denoted by φ, the receiver field-of-view (FoV) is denoted by φ_{F} , the optical concentrator refractive index is denoted by n_0 , and the photodetector area is denoted by A_{P}. According to [38], [39], the path gain G between the transmitter and the receiver is given by

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

where d is the distance between the LED at the transmitter and the PD at the receiver.

B. Attack Model

As shown in Fig. 1, an indoor MISO VLC scenario is considered which consists of one transmitter (Alice) equipped with N_{A} down-facing light fixtures on the ceiling, one legitimate receiver (Bob) communicating with Alice and one eavesdropper (Eve) attempting to intercept the secret information prepared for Bob. Each light fixture of the transmitter is composed of a group of LEDs connected in serial and modulated by the same current signal. The MISO VLC channel is considered in this paper, because the cost and complexity of putting several PDs to the receiver as multiple receive antennas in VLC scenarios are high. Using multiple LEDs as multiple transmit antennas and a single PD as the receive antenna can greatly save the cost and complexity. Applying MISO VLC transmission also allows multiuser transmission, as investigated in literature [33], [34]. Besides, if multiple PDs are applied at the VLC receiver, the diversity of the VLC MIMO channel is insufficient due to the LoS propagation property, so the spatial diversity or multiplexing gain is not worth the costs it takes.

Alice applies N_{A} light fixtures to transmit the modulated data symbol denoted by s to Bob, and meanwhile

TABLE I
LIST OF NOTATIONS

Symbol	Description
I	Fixed bias
α	Modulation index
x	Current of a transmitted data symbol
$\mathbf{x}$	Transmitted data signal vector via beamforming
$\mathbf{w}$	Beamformer vector
s	Transmitted data symbol
$y_{\mathrm{B}}, y_{\mathrm{E}}$	Voltage signal received by Bob or Eve
$\mathbf{h}_{\mathrm{B}}, \mathbf{h}_{\mathrm{E}}$	VLC channel gains from Alice to Bob or Eve
$n_{\mathrm{B}}, n_{\mathrm{E}}$	Zero-mean AWGNs at the receiver of Bob or Eve
c_s	Achievable secrecy rate
$\mathbf{w}_0$	Zero-forcing beamforming vector
$u^{(k)}$	Utility of the VLC system
$p_e^{(k)}$	BER of the legitimate receiver
$c_s^{(k)}$	Secrecy rate of the current state

hopes to keep the information hidden from Eve without using high-layer secret-key encryption. Then the transmitted signal vector denoted by $\mathbf{x} = [x_1, x_2, \cdots, x_{N_{\mathrm{A}}}]^{\mathrm{T}}$ can be expressed as $\mathbf{x} = \mathbf{w}s$, where $\mathbf{w} = [w_1, w_2, \cdots, w_{N_{\mathrm{A}}}]^{\mathrm{T}}$ is the beamforming vector. Then the channel gains from Alice to Bob and Eve are denoted by $\mathbf{h}_{\mathrm{B}}$ and $\mathbf{h}_{\mathrm{E}}$, where $\mathbf{h}_{\mathrm{B}} = RT\eta[G_{1\mathrm{B}}, G_{2\mathrm{B}}, \cdots, G_{N_{\mathrm{A}}\mathrm{B}}]^{\mathrm{T}}$ and $\mathbf{h}_{\mathrm{E}} = RT\eta[G_{1\mathrm{E}}, G_{2\mathrm{E}}, \cdots, G_{N_{\mathrm{A}}\mathrm{E}}]^{\mathrm{T}}$ with $G_{i\mathrm{B}}$ and $G_{i\mathrm{E}}$ being the path gains from the i-th light fixture to Bob and Eve, respectively. The signals received by Bob and Eve denoted by y_{B} and y_{E}, respectively, in the MISO wiretap channel are given by

$$y_{\mathrm{B}} = \mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}s + n_{\mathrm{B}}, \tag{2a}$$

$$y_{\mathrm{E}} = \mathbf{h}_{\mathrm{E}}^{\mathrm{T}}\mathbf{w}s + n_{\mathrm{E}}, \tag{2b}$$

where n_{B} and n_{E} are the corresponding zero-mean additive white Gaussian noise (AWGN) vectors with variance of σ^2. Because the LED dynamic range is constrained, the transmitted data signal should meet the constraint for amplitude, i.e. $|\mathbf{x}| \preceq \alpha I \mathbb{1}$, where $\mathbb{1}$ is an all-one vector and the curled inequality symbol $\preceq$ denotes entry-wise inequality, i.e., the amplitude of each entry of $\mathbf{x}$ is not greater than αI.

C. Secrecy Rate and Zero-Forcing

For the MISO VLC wiretap channel subject to the average power constraint $|\mathbf{x}| \preceq \alpha I \mathbb{1}$, the achievable secrecy rate of the beamforming system denoted by c_s can be obtained as follows [10]

$$c_s = \frac{1}{2}\log\frac{6\alpha^2 I^2 \mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{B}}\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w} + 3\pi e\sigma^2}{\pi e\alpha^2 I^2 \mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{E}}\mathbf{h}_{\mathrm{E}}^{\mathrm{T}}\mathbf{w} + 3\pi e\sigma^2}, \tag{3}$$

The optimal beamforming vector $\mathbf{w}^*$ maximizing the secrecy rate in (3) is theoretically derived by the solution of problem as given by

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{1}{2}\log\frac{6\alpha^2 I^2 \mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{B}}\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w} + 3\pi e\sigma^2}{\pi e\alpha^2 I^2 \mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{E}}\mathbf{h}_{\mathrm{E}}^{\mathrm{T}}\mathbf{w} + 3\pi e\sigma^2} \quad \text{s.t. } |\mathbf{w}| \preceq \mathbb{1}. \tag{4}$$

The problem in (4) has been shown to be non-convex and NP-hard [10]. However, there is a chance to derive a suboptimal solution, when Eve's geometric location is known to the transmitter and thus the CSI at transmitter can be obtained. Then, zero-forcing beamforming can be exploited to obtain a suboptimal secrecy rate by imposing zero beamforming gain to Eve while maximizing the beamforming gain of Bob. Hence, the zero-forcing beamformer $\mathbf{w}_0$ is obtained by

$$\mathbf{w}_0 = \arg\max_{\mathbf{w}} \mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w} \quad \text{s.t. } \mathbf{h}_{\mathrm{E}}^{\mathrm{T}}\mathbf{w} = 0, \quad |\mathbf{w}| \preceq \mathbb{1}. \tag{5}$$

Consequently, the secrecy rate via zero-forcing beamforming can be obtained in closed-form by

$$c_s^* = \frac{1}{2}\log\left(1 + \frac{2\alpha^2 I^2 \mathbf{w}_0^{\mathrm{T}}\mathbf{h}_{\mathrm{B}}\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}_0}{\pi e\sigma^2}\right). \tag{6}$$

Table I provides a list of the frequently used symbols.

Nevertheless, the zero-forcing method is aimed at finding a beamforming vector such that the channel gain of Eve is nulled, so the degree of freedom of the beamforming vector is inevitably reduced. Since the solution space of the beamforming vector $\mathbf{w}$ significantly shrinks to only the null space of $\mathbf{h}_{\mathrm{E}}$, it is much unlikely for the zero-forcing method to approach the optimal secrecy rate as given by the solution of the original NP-hard problem in equation (4). Meanwhile, the BER of the legitimate receiver is not taken into account by the zero-forcing method. Thus, the system utility as defined by equation (8), which is influenced by the secrecy rate and the BER of Bob, might be severely limited. Therefore, there is a desperate need to explore a more effective method to solve the problem and converge to the theoretical global optimum.

IV. RL-BASED MISO VLC BEAMFORMING CONTROL SCHEME AGAINST EAVESDROPPING

In order to balance the influence of the secrecy rate of the system and the BER of the legitimate receiver on the system utility, an RL-based MISO VLC beamforming control scheme is proposed, which is aimed at further improving the overall performance of the VLC anti-eavesdropping system and find an approach that can converge to the theoretically optimal solution of the non-convex problem given by (4). In the dynamic VLC communication process, the transmitter Alice selects its beamforming policy to transmit the data signal to the legitimate receiver Bob based on the learnt Q-values of the current system state. The current state is composed of the previous BER and the current channel state information of Bob, and the system secrecy rate. The action, i.e. the beamforming policy, taken by the transmitter will influence the next system state, thus having an impact on the future rewards and future actions in the dynamic learning process. As a matter of fact, the next system state is dependent only on the state

Algorithm 1 RL-Based MISO VLC Beamforming Control Scheme Against Eavesdropping (RL-VB)

1 Initialize discount factor β and learning rate λ
2 $Q(\mathbf{s}, \mathbf{w}) = 0, \forall \mathbf{s} \in \Lambda, \mathbf{w} \in W$
3 **for** $k = 1, 2, 3, \ldots$ **do**
4 Measure the BER of the previous time slot $\hat{p}_e^{(k-1)}$
5 Estimate the previous secrecy rate $\hat{c}_s^{(k-1)}$
6 Obtain the current legitimate channel gain $\mathbf{h}_{\mathrm{B}}^{(k)}$
7 Formulate current state $\mathbf{s}^{(k)} = \left[\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, \mathbf{h}_{\mathrm{B}}^{(k)}\right]$
8 Choose beamforming policy $\mathbf{w}^{(k)}$ via ε-greedy method
9 Apply the selected beamforming policy $\mathbf{w}^{(k)}$
10 Measure the current BER $\hat{p}_e^{(k)}$
11 Estimate the current secrecy rate $\hat{c}_s^{(k)}$
12 Calculate the utility $u^{(k)}$
13 Update $Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}) \leftarrow$
$(1-\lambda)Q(\mathbf{s}^{(k)}, \mathbf{w}^{(k)}) + \lambda(u^{(k)} + \beta \max_{\mathbf{w}} Q(\mathbf{s}^{(k+1)}, \mathbf{w}))$
14 **end**

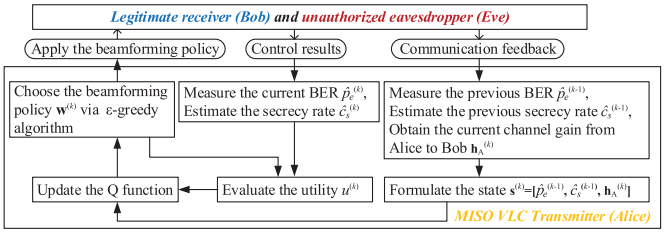


Fig. 2. RL-based secure MISO VLC beamforming control scheme against eavesdropping.

and action of the current time slot, and has nothing to do with the previous states or actions. Therefore, the VLC communication process can be modeled by an MDP. To achieve an optimal beamforming policy via trial and error, an RL-VB algorithm is proposed with the pseudo-code summarized in **Algorithm 1**.

More specifically, Alice applies the RL technique of Q-learning to achieve an optimal beamforming policy according to the quality function, i.e., Q-function, and the current state. As illustrated in Fig. 2, Alice receives the feedback information from the legitimate user Bob to estimate the previous BER of Bob denoted by $\hat{p}_e^{(k-1)}$. The prior geometric information and the statistical model information of the VLC transmission environment can be utilized to acquire Eve's approximate location and the corresponding channel state information, so the previous secrecy rate denoted by $\hat{c}_s^{(k-1)}$ can be estimated. Specifically, due to the LoS property of the VLC propagation channel, the possible location of Eve can be roughly estimated, such as on top of another table in the LoS range within the VLC transmitters, and thus the coarse channel of Eve is also obtained. The accuracy of the roughly estimated location might not be very high, but a relatively coarse estimation is sufficient for the evaluation of the utility and secrecy rate exploited in the dynamic learning process, which is demonstrated by the simulations in Section VII.

Then the transmitter Alice can observe the communication state at time slot k denoted by $\mathbf{s}^{(k)}$ composed of the BER of the legitimate receiver $\hat{p}_e^{(k-1)}$ in the previous time slot, the predicted secrecy rate $\hat{c}_s^{(k-1)}$, and the current legitimate channel gain from Alice to Bob $\mathbf{h}_{\mathrm{B}}^{(k)}$. Therefore, the current state is formulated by $\mathbf{s}^{(k)} = [\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, \mathbf{h}_{\mathrm{B}}^{(k)}] \in \Lambda$, where Λ is the state vector space containing all the possible states.

Based on the communication state, the transmitter chooses the beamforming policy $\mathbf{w}^{(k)} = [w_1^{(k)}, w_2^{(k)}, \ldots, w_{N_{\mathrm{A}}}^{(k)}]^{\mathrm{T}}$ with $|w_i| \leq 1$, $1 \leq i \leq N_{\mathrm{A}}$ and $\mathbf{w}^{(k)} \in \mathbf{W}$, where $\mathbf{W}$ is the action vector space composed of all the feasible beamforming policies. For simplicity, during the RL-based learning process in practice, each entry of the beamforming vector $w_i^{(k)}$ is quantized into $2L_x+1$ discrete values with equal space in between, i.e., $w_i^{(k)} \in \{l/L_x | -L_x \leq l \leq L_x\}$, where L_x can be properly set as a compromise between the beamforming learning accuracy and the computational complexity. The transmitter uses ε-greedy algorithm to choose the VLC beamforming policy which aims at keeping the balance between exploration and exploitation. More specifically, the VLC beamforming policy $\mathbf{w}^{(k)}$ will be chosen with a high probability of $1-\varepsilon$ to maximize the Q-value, while other beamforming policy are selected randomly with a low probability as ε to avoid local optimum, i.e.,

$$\Pr\left(\mathbf{w}^{(k)} = \hat{\mathbf{w}}\right) = \begin{cases} 1-\varepsilon, & \hat{\mathbf{w}} = \arg\max\limits_{\mathbf{w}'} Q\left(\mathbf{s}^{(k)}, \mathbf{w}'\right) \\ \frac{\varepsilon}{|\mathbf{W}|-1}, & \text{o.w.} \end{cases} \tag{7}$$

The selected beamforming policy is then applied to the MISO VLC system. According to the control results and communication feedback, Alice observes the current BER of the legitimate receiver Bob $\hat{p}_e^{(k)}$, obtains the predicted secrecy rate $\hat{c}_s^{(k)}$, and determines the utility $u^{(k)}$ at time slot k for the MISO VLC system. The utility is thus given by

$$u^{(k)} = \hat{c}_s^{(k)} - \delta \hat{p}_e^{(k)}, \tag{8}$$

where δ is a coefficient which balances the contribution of secrecy rate and BER to the utility. The coefficient δ can play a role of the tradeoff between the reception quality of the legitimate user and the overall secrecy metric of the VLC wiretap channel.

In the learning process, the proposed RL-VB algorithm obtains a Q-value for each beamforming policy, denoted by $Q(\mathbf{s}, \mathbf{w})$. Alice observes the next state $\mathbf{s}^{(k+1)}$ to update the Q-function at time slot k using iterative Bellman equation given by Line 13 in **Algorithm 1**. In the Bellman iterative equation, the learning rate $\lambda \in [0, 1]$ indicates the weight of the current Q-values. A learning rate of 0 will prevent the agent from learning anything, while a learning rate of 1 will make the transmitter consider only the latest Q-values. The discount factor $\beta \in [0, 1]$ is also a number between 0 and 1 that represents the uncertainty of the learning algorithm on the rewards in the future. A discount factor of 0 will make the transmitter "myopic" (i.e., short-sighted) by only focusing on the current rewards, while a discount factor of 1 will make it take a high long-term reward into account.

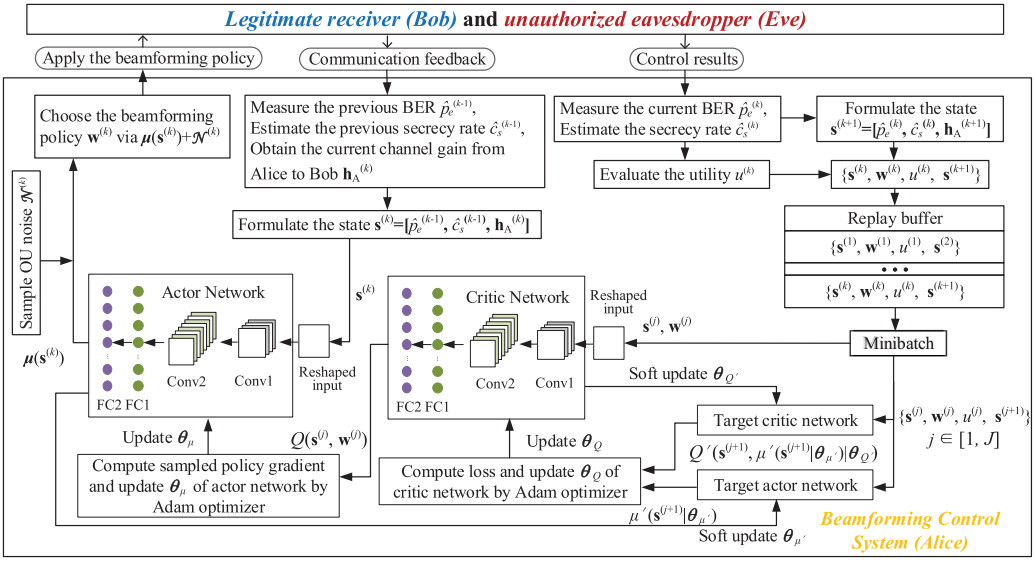


Fig. 3. DRL-based secure MISO VLC beamforming control scheme against eavesdropping.

Algorithm 2 DRL-Based MISO VLC Beamforming Control Scheme Against Eavesdropping (DRL-VB)

Initialize the weights θ_Q and θ_μ for the actor and critic networks
Initialize target networks with weights $\theta_{Q'}$ and $\theta_{\mu'}$
Initialize the OU random process $\mathcal{N}$
for $k = 1, 2, 3, \ldots$ **do**
  Measure the BER $\hat{p}_e^{(k-1)}$ of the previous time slot
  Estimate the previous secrecy rate $\hat{c}_s^{(k-1)}$
  Obtain the current legitimate channel gain $\mathbf{h}_\mathrm{B}^{(k)}$
  Formulate current state $\mathbf{s}^{(k)} = \left[\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, \mathbf{h}_\mathrm{B}^{(k)}\right]$
  Choose beamforming policy $\mathbf{w}^{(k)} = \mu(\mathbf{s}^{(k)}|\theta_\mu) + \mathcal{N}$
  Apply the selected beamforming policy $\mathbf{w}^{(k)}$
  Measure the current BER $\hat{p}_e^{(k)}$
  Estimate the current secrecy rate $\hat{c}_s^{(k)}$
  Calculate the utility $u^{(k)}$
  Formulate next state $\mathbf{s}^{(k+1)} = \left[\hat{p}_e^{(k)}, \hat{c}_s^{(k)}, \mathbf{h}_\mathrm{B}^{(k+1)}\right]$
  Store $\{\mathbf{s}^{(k)}, \mathbf{w}^{(k)}, u^{(k)}, \mathbf{s}^{(k+1)}\}$ in $\mathcal{B}$
  Sample minibatch $\{\mathbf{s}^{(j)}, \mathbf{w}^{(j)}, u^{(j)}, \mathbf{s}^{(j+1)}\}, j \in [1, J]$ from replay memory $\mathcal{B}$
  Update the online networks θ_Q and θ_μ using (8) and (9)
  Soft update the target networks $\theta_{Q'}$ and $\theta_{\mu'}$ using (10)
end

V. DRL-Based MISO VLC Beamforming Control Scheme Against Eavesdropping

Since the action space is high-dimensional and continuous in practice, some quantization error inevitably occurs using the RL-VB algorithm, which might prevent it from approaching the globally optimal solution of the problem given by (4). To further improve the efficiency and convergence rate in practical complicated VLC systems, a DRL-VB algorithm is proposed to avoid the performance loss due to the quantization error, and accelerate the learning speed with a continuous action space enabled by the deep neural networks. Four convolutional neural networks (CNNs) are used to compress the action space, and the ε-greedy policy is replaced by a parameterized actor function denoted by $\mathbf{w} = \mu(\mathbf{s}|\theta_\mu)$.

As shown in Fig. 3, the four CNNs included are one actor network to produce the actor function $\mu(\mathbf{s}|\theta_\mu)$, one critic network to output the Q-function $Q(\mathbf{s}, \mathbf{w}|\theta_Q)$ and update the weights θ_μ in the actor network, and two corresponding target networks to update the weights θ_Q in the critic network. The function $\mu'(\mathbf{s}|\theta_{\mu'})$ and $Q'(\mathbf{s}, \mathbf{w}|\theta_{Q'})$ are the output of the target critic and target actor networks with $\theta_{\mu'}$ and $\theta_{Q'}$ being the weights of them, and the structure of the target networks is copied from the actor and critic networks.

The proposed DRL-VB algorithm is given in detail in **Algorithm 2**. Similar to the RL-based beamforming control scheme, the transmitter firstly initializes the network parameters and then observes the current state of the communication consisting of the previous secrecy rate, the previous BER of the legitimate receiver and the current channel gain, which formulates the current state $\mathbf{s}^{(k)} = [\hat{p}_e^{(k-1)}, \hat{c}_s^{(k-1)}, \mathbf{h}_\mathrm{B}^{(k)}]$. The transmitter reshapes the state vector and puts it into the actor network which is composed of two convolutional layers (Conv) and two fully connected layers (FC). In the actor network, the channel features are captured by the CNN filters which are influenced by the location of Eve and Bob. Based on the current communication state, the transmitter chooses the beamforming policy $\mathbf{w}^{(k)}$ based on an *exploration policy* by adding a noise sampled from an Ornstein-Uhlenbeck (OU) noise process to the actor policy, i.e., $\mu(\mathbf{s}^{(k)}|\theta_\mu^{(k)}) + \mathcal{N}$, where $\mathcal{N}$ is the OU noise to generate temporally correlated exploration to improve exploration efficiency [40]. Then the beamforming policy $\mathbf{w}^{(k)}$ is chosen according to the noisy actor policy to implement exploration in practice.

After the selected beamforming policy is applied to the MISO VLC system, the transmitter obtains the control results and communication feedback to calculate the utility $u^{(k)}$ and formulate the next state $\mathbf{s}^{(k+1)} = [\hat{p}_e^{(k)}, \hat{c}_s^{(k)}, \mathbf{h}_{\mathrm{B}}^{(k+1)}]$. To memorize the experience, the transmitter assembles the above information into one transition $\psi^{(k)}$, consisting of the current system state $\mathbf{s}^{(k)}$, the current policy $\mathbf{w}^{(k)}$, the utility $u^{(k)}$ and the next state $\mathbf{s}^{(k+1)}$, i.e., $\psi^{(k)} = \{\mathbf{s}^{(k)}, \mathbf{w}^{(k)}, u^{(k)}, \mathbf{s}^{(k+1)}\}$. The transition $\psi^{(k)}$ is stored into a replay buffer $\mathcal{B}$, which is a finite-memory cache to store the previous communication feedbacks and learning experiences, and the earliest samples will be discarded on a rolling basis when the replay buffer is full.

To update the actor and critic networks, the transmitter samples a minibatch uniformly from the replay buffer. More specifically, the transmitter randomly chooses J transitions from the replay buffer $\mathcal{B}$ to formulate the minibatch $\mathcal{J}$ with $\psi^{(j)} = \{\mathbf{s}^{(j)}, \mathbf{w}^{(j)}, u^{(j)}, \mathbf{s}^{(j+1)}\}_{1 \le j \le J}$ being the j-th selected transition including the beamforming policy, the utility, the previous state and next state. The weights of the critic network θ_Q is updated using Adam optimizer to minimize the loss function as

$$\theta_Q = \arg\min_{\theta_Q} \sum_{j=1}^{J} \Big(u^{(j)} + \gamma Q'\left(\mathbf{s}^{(j+1)}, \mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})|\theta_{Q'}\right) - Q\left(\mathbf{s}^{(j)}, \mathbf{w}^{(j)}|\theta_Q\right)\Big)^2, \quad (9)$$

where γ is the discount factor indicating the uncertainty of the transmitter on future rewards, $\mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})$ is the output of the target actor network representing the chosen action with the input of the next state $\mathbf{s}^{(j+1)}$, $Q'\left(\mathbf{s}^{(j+1)}, \mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})|\theta_{Q'}\right)$ is the output of the target critic network with the next state and the chosen action through the target actor network as the input, and $Q\left(\mathbf{s}^{(j)}, \mathbf{w}^{(j)}|\theta_Q\right)$ is the output of the critic network with the input of the current state and the action performed.

Similarly, the weights of the actor network θ_μ is updated based on Adam optimizer to maximize the sampled policy gradient as

$$\theta_\mu = \arg\max_{\theta_\mu} \sum_{j=1}^{J} \nabla_{\mathbf{w}} Q\left(\mathbf{s}, \mathbf{w}|\theta_Q\right)|_{\mathbf{s}=\mathbf{s}^{(j)}, \mathbf{w}=\mu\left(\mathbf{s}^{(j)}\right)} * \nabla_{\theta_\mu}\mu\left(\mathbf{s}|\theta_\mu\right)|_{\mathbf{s}=\mathbf{s}^{(j)}}. \quad (10)$$

where $\nabla_{\mathbf{w}} Q\left(\mathbf{s}, \mathbf{w}|\theta_Q\right)$ is the policy gradient of the Q-function with respect to $\mathbf{w}$, and $\nabla_{\theta_\mu}\mu\left(\mathbf{s}|\theta_\mu\right)$ is the policy gradient of the actor function with respect to θ_μ.

To improve the stability of learning and the robustness of the DRL-based beamforming control system, a learning rate denoted by $\tau \ll 1$ is introduced to ensure that the output of the target networks changes slowly when the transmitter updates the weights of the target networks. Therefore, the weights of both the two target networks are soft updated by letting them slowly track the learnt weights of the actor network and critic network as

$$\begin{aligned} \theta_{Q'} &\leftarrow \tau\theta_Q + (1-\tau)\theta_{Q'} \\ \theta_{\mu'} &\leftarrow \tau\theta_\mu + (1-\tau)\theta_{\mu'}. \end{aligned} \quad (11)$$

If the environment such as the location of Eve or Bob changes, the transmitter can use the same neural networks to make decisions. More specifically, because the proposed learning framework is based on a reinforcement mechanism, the transmitter can utilize the system state dynamically observed and updated from the current environment as the input of the CNN, which consists of the feedback of Bob and the estimated secrecy rate. In this way, the CNN can capture the features of the system state to keep track of the location of Eve and Bob and the variant environments, and utilize the stored communication experiences in the replay buffer to update the parameters of the network in real time. Thus, it can be guaranteed that the proposed learning method is capable of adapting to the dynamic communication process and various system environments and configurations.

VI. Performance Analysis

The performance bound of the proposed MISO VLC beamforming control scheme is analyzed using the BER and secrecy rate of the system. For simplicity, the BER of the receiver in the MISO VLC system is considered with 4PAM modulation. Thus the BER is calculated by

$$p_e = \frac{3}{4}\mathrm{erfc}\left(\sqrt{\frac{2P_{\mathrm{T}}\mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{B}}\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}}{5\sigma^2}}\right), \quad (12)$$

where the σ^2 is the noise power and $\mathrm{erfc}(x)$ is the complementary error function given by

$$\mathrm{erfc}(x) = 2/\sqrt{\pi}\int_x^\infty e^{-z^2}dz. \quad (13)$$

The optimal anti-eavesdropping performance bound is formulated by the following theorem.

Theorem 1: The RL-based and DRL-based schemes for MISO VLC beamforming control in **Algorithm 1** *and* **Algorithm 2** *achieve an optimal anti-eavesdropping communication strategy as*

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathbf{h}_B^T\mathbf{w}, \quad (14)$$

with the optimal performance given by

$$u^* = \frac{1}{2}\log\left(\frac{2\alpha^2 I^2 \mathbf{w}^{*T}\mathbf{h}_B\mathbf{h}_B^T\mathbf{w}^*}{\pi e\sigma^2} + 1\right) - \frac{3}{4}\delta erfc\left(\sqrt{\frac{2\eta(I+x)\mathbf{w}^{*T}\mathbf{h}_B\mathbf{h}_B^T\mathbf{w}^*}{5\sigma^2}}\right), \quad (15)$$

if the wiretapper's geometric location is ideally known to the transmitter, and

$$\mathbf{h}_E^T\mathbf{w} = 0, \quad |\mathbf{w}| \preceq \mathbb{1}. \quad (16)$$

Proof: According to (3), (8) and (12), if $\mathbf{h}_{\mathrm{E}}^{\mathrm{T}}\mathbf{w} = 0$, we have

$$u = \frac{1}{2}\log\left(\frac{2\alpha^2 I^2 \mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{B}}\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}}{\pi e\sigma^2} + 1\right) - \frac{3}{4}\delta\mathrm{erfc}\left(\sqrt{\frac{2\eta(I+x)\mathbf{w}^{\mathrm{T}}\mathbf{h}_{\mathrm{B}}\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}}{5\sigma^2}}\right). \quad (17)$$

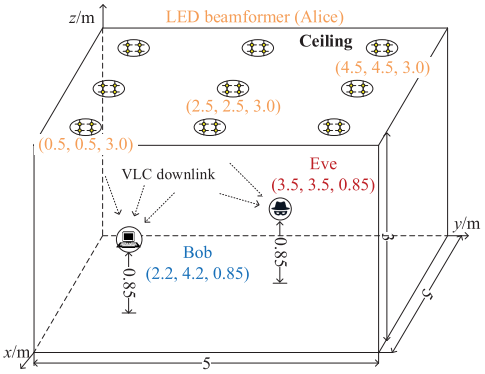


Fig. 4. Simulation setup in a $5 \times 5 \times 3$ m^3 room, consisting of a legitimate receiver (Bob), an eavesdropper (Eve) and the beamforming transmitter (Alice) with 9 light fixtures.

TABLE II
SIMULATION SETUP

Transmitter parameters (typical value)	
Average power of each LED	1000 mW
Modulation index α	10%
Semi-angle of half luminous intensity $\phi_{1/2}$	60°
Receiver parameters (typical values)	
Receiver FoV φ_{F}	60°
Optical concentrator refractive index n_0	1.5
PD responsivity R	0.54 A/W
PD geometrical area A_{P}	1 cm^2
Average noise power σ^2	-98.82 dBm

For simplicity, the beamforming gain of Bob $\mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}$ is denoted by H, then we have $H > 0$ and

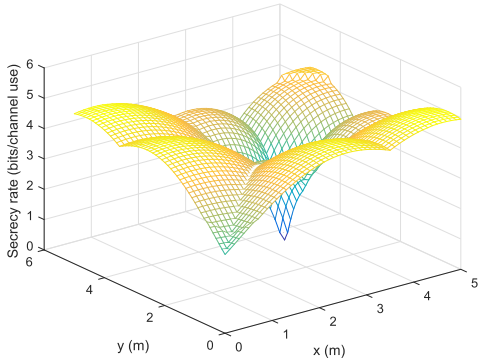
$$\frac{\partial u}{\partial H} = \frac{2\alpha^2 I^2 H}{2\alpha^2 I^2 H^2 + \pi e \sigma^2} + \frac{3\delta}{4\sqrt{\pi}} \sqrt{\frac{2\eta(I+x)}{5\sigma^2}} \exp\left(-\frac{2\eta(I+x)H^2}{5\sigma^2}\right). \quad (18)$$

Since $H > 0$, it is obvious that $\partial u / \partial H \geq 0$, which means the utility performance Therefore, the optimal performance of u^* should be reached when $H = \mathbf{h}_{\mathrm{B}}^{\mathrm{T}}\mathbf{w}$ is maximized, and thus we have (14) and (15). □

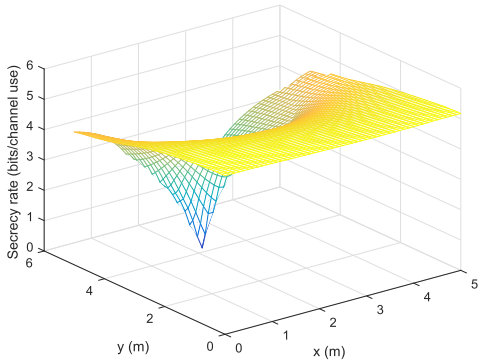
Remark 1: If the CSI of the wiretapper is perfectly available at the transmitter, the transmitter will choose the beamforming control policy as $\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathbf{h}_B^T \mathbf{w}$, *such that the received signal level of the eavesdropper is subject to* $\mathbf{h}_E^T \mathbf{w} = 0$.

VII. SIMULATION RESULTS

Experimental simulations are conducted in a typical indoor MISO VLC case shown in Fig. 4 and simulation setup is given in Table II. The experiment is carried out in a space with the scale of $5 \times 5 \times 3$ m^3, where there are 9 light fixtures facing down on top. Each light fixture contains 4 identical LEDs. Bob and Eve are located at the height of 0.85m above the floor, e.g., on desks. In the experimental simulation of the learning process, we set the learning rate as $\lambda = 0.5$, the discount factor as $\beta = 0.5$, the coefficient as $\delta = 1.2$ and the discount factor for the critic network as $\gamma = 0.5$ based on the logic of each hyper parameter and according to the analytical and empirical configurations in related the previous tasks [18]–[20]. The ε-greedy parameter ε is linearly annealed from 1.0 to 0.1 during the first 400 time slots of the learning process for exploitation and is fixed to 0.1 afterwards for stability. The nine light fixtures are located at the coordinates shown in Fig. 4, and Bob and Eve are located at the coordinates of $(3.5, 3.5, 0.85)$m and $(2.2, 4.2, 0.85)$m.



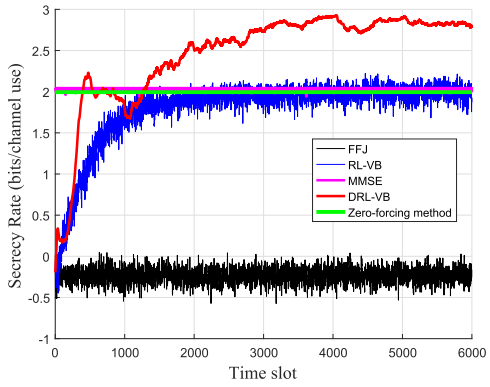
(a) Secrecy rate with Bob's location variant and Eve's location fixed at (3.5, 3.5, 0.85) m



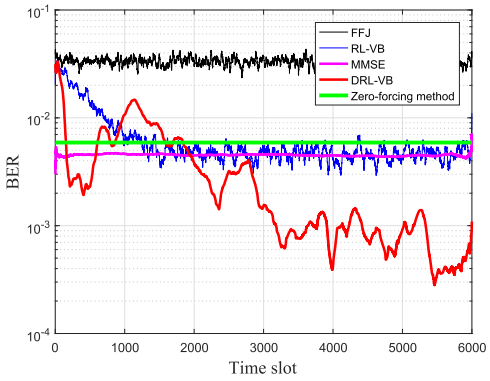
(b) Secrecy rate with Eve's location variant and Bob's location fixed at (2.2, 4.2, 0.85) m

Fig. 5. Secrecy rate of the proposed RL-based MISO VLC beamforming control scheme with variant geometric positions of the Bob and Eve.

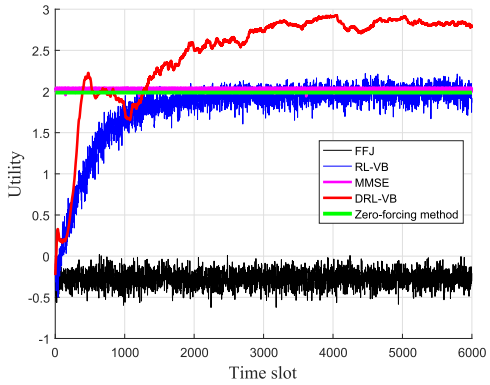
Firstly, the secrecy rate of the proposed RL-based MISO VLC beamforming control scheme is derived for variant geometric locations of Bob or Eve, with the location of Eve or Bob fixed, respectively. As shown in Fig. 5, it can be noted that the secrecy rate of the smart beamforming system is generally satisfactory and sufficiently high in the overall room scale. It is also observed from Fig. 5(a) that if the location of Eve was fixed at $(3.5, 3.5, 0.85)$m, the secrecy rate may fall into a local minimum as the location of Bob changes, but it is still much greater than the global minimum when Bob happens to fall into the location of Eve. On the other hand, as indicated by Fig. 5(b), if the location of Bob is fixed at $(2.2, 4.2, 0.85)$m, the secrecy rate with variant location of Eve is overall quite high for the system in the entire room, and



(a) Secrecy rate of the MISO VLC beamforming system



(b) BER of the legitimate receiver Bob



(c) Utility of the MISO VLC beamforming system

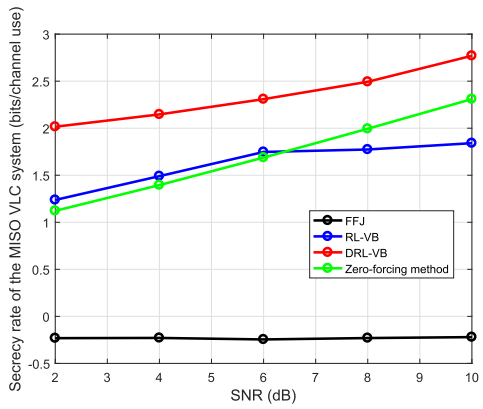
Fig. 6. Performance of the MISO VLC beamforming control scheme.

only has one global minimum, which is when Eve falls into the location the same with Bob. Fortunately, it is much easier to spot the eavesdropper directly in practice if Eve is very close to the legitimate user Bob, so Eve is most likely to choose a location not so close to Bob. In this case, the proposed smart beamforming approach is able to learn and reach a sufficiently high and satisfactory secrecy rate performance as indicated by the results in Fig. 5.

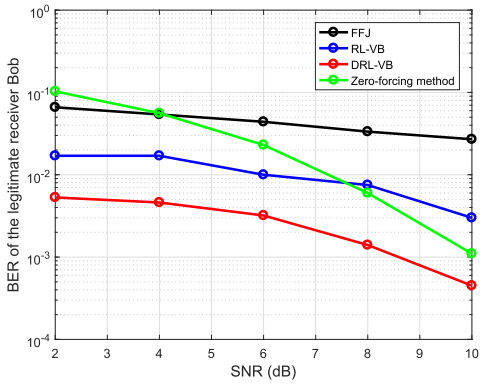
The learning process and the performance of the proposed RL-based and DRL-based smart beamforming schemes, i.e., RL-VB and DRL-VB, are reported in Fig. 6, compared with the state-of-the-art benchmarks. In the communication process, we assume that Alice can exploit the prior geometric information of the surrounding VLC LoS environment, which is easily available in the devised VLC system, to estimate the possible location of Eve, and thus obtain an estimate of the wiretapping channel gain. Then, the prior information of the VLC transmission environment as well as the statistical wiretapping model can be utilized to calculate the secrecy rate and the utility used in the devised smart beamforming VLC system in the line-of-sight (LoS) VLC transmission scenario [10]. The SNR is set as 8 dB. It can be observed from Fig. 6 that the proposed RL-VB scheme converges to the performance derived by zero-forcing beamforming in (6). Moreover, the proposed RL-VB scheme outperforms the benchmark scheme using the state-of-the-art fixed friendly jamming (FFJ) method [9], yielding a lower BER, higher secrecy rate and higher utility. Moreover, it can be noted that in the framework of the actor-critic enabled deep neural networks, the proposed DRL-VB scheme can avoid the quantization error imposed on the RL-VB algorithm, and the overall system performance is further improved compared with the RL-VB algorithm. Besides, it can be observed that the learnt policy using the DRL-VB algorithm outperforms the zero-forcing method and the MMSE method, indicating the superior performance of the proposed learning based smart beamforming framework with respect to the conventional optimization methods in VLC anti-eavesdropping systems.

To investigate the simulation results more specifically, as shown in Fig. 6(a), the secrecy rate with the proposed RL-VB scheme grows rapidly over time and converges to 2.03 after 5000 time slots, which has increased by 107.3% compared with the beginning of the learning process. It is noted that the secrecy rate of the RL-VB scheme is approximately 116.3% larger compared with the FFJ method at the 5000th time slot. Furthermore, the proposed DRL-VB scheme increases the secrecy rate by 29.7% higher than that of the RL-VB scheme thanks to the high-dimensional and continuous action and state spaces effectively handled by the deep networks. It is shown by Fig. 6 that, the RL-VB and DRL-VB schemes even outperform the zero-forcing method which assumes that the position of the wiretapper is available. This is because the utility in (8) to drive the learning process has taken the BER into consideration, and has successfully learnt to minimize the BER meanwhile maximizing the secrecy rate. The zero-forcing method is aimed at nulling the channel gain of Eve so the degree of freedom of the beamforming vector $\mathbf{w}$ is inevitably reduced. Thus the channel gain of Bob as well as the BER performance is constrained especially for a low SNR, which constraints the utility given by (8). According to (3), the secrecy rate of zero-forcing might be constrained by the limited channel gain of Bob, while the proposed learning algorithms are able to overcome this constraint and learn a better policy for secrecy rate. However, due to the quantization error introduced by the discrete states and actions, there exists a performance gap between the proposed RL-VB method and the theoretical upper bound given by the solution of (4), which can be converged to by the further enhanced deep RL based DRL-VB algorithm as an improvement.

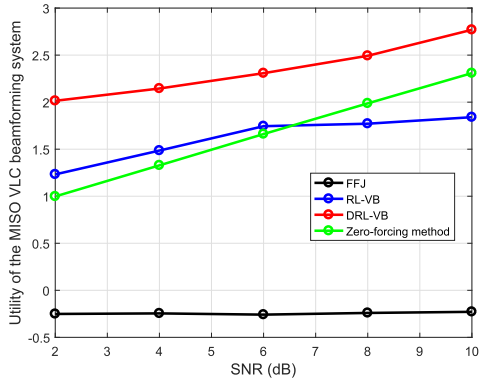
As illustrated by Fig. 6(b), the BER of the legitimate receiver decreases rapidly using the RL-VB scheme, and



(a) Secrecy rate of the MISO VLC beamforming system



(b) BER of the legitimate receiver Bob



(c) Utility of the MISO VLC beamforming system

Fig. 7. Performance of the MISO VLC beamforming control system with respect to SNR.

finally approaches 2.8×10^{-3}, which is much lower than the BER at the beginning. It is shown by Fig. 6(b) that the BER of the RL-VB scheme is about 9×10^{-4}, which is far much lower than that of the FFJ method at the 5000th iteration. Furthermore, the BER of the legitimate receiver of the proposed DRL-VB scheme finally reaches 3.9×10^{-4}, which is even lower than that of the RL-VB scheme.

Considering the utility of the VLC system, as reported by Fig. 6(c), the utility of the proposed RL-VB scheme increases over time and converges to about 2.03 after approximately 5000 time slots, which is approximately 107.3% higher compared with the beginning of the learning process. Besides, the utility of the proposed RL-VB scheme significantly exceeds the benchmark FFJ scheme by 2.39 at the 5000th time slot, which validates the good performance of the proposed RL-VB scheme in anti-eavesdropping. The proposed DRL-VB scheme further improved the utility of the system by 29.8% compared by the RL-VB method.

The performance of the proposed schemes are investigated through simulations with respect to different SNR values, with the results reported in Fig. 7. The average performance over 5000 time slots shows that the BER of the legitimate receiver decreases with the SNR while the utility of the system and secrecy rate of the transmitter increase with the SNR. For instance, if observed at the target SNR of 10 dB, the secrecy rate of the system increases by 37.4% and the utility of the system increases by 37.5%, respectively, compared with those at SNR of 0 dB. From Fig. 7(b), it is noted that at the target BER of 2×10^{-3}, the proposed DRL based DRL-VB scheme has the SNR gain of about 8 dB and more than 10 dB compared with the proposed RL-based RL-VB scheme and the conventional FFJ scheme, respectively. If observed the target SNR of 10 dB, it can be shown from Fig. 7 that the RL-VB scheme has 112.4% higher utility and 112.0% higher secrecy rate compared with FFJ. The DRL-VB scheme further increases the utility by 50.1% and the secrecy rate by 50.5% compared with the RL-VB scheme, which validates the effectiveness of the proposed DRL framework for VLC smart beamforming against eavesdropping. Moreover, the DRL-VB scheme increases the utility by 20.0% and the secrecy rate by 20.0% compared with the zero-forcing method. The performance of the zero-forcing method reported in Fig. 7 shows that, the solution space of the zero-forcing method is seriously limited and it cannot find the global optimal solution of the system. Compared with the zero-forcing method, the proposed RL-based framework, especially the DRL-VB algorithm without quantization error, is able to outperform the zero-forcing one and converge to the optimal solution in the VLC anti-eavesdropping system.

VIII. Conclusion

In this paper, a MISO VLC wiretap scenario has been studied, where an eavesdropper attempts to wiretap the information which is originally sent for the legitimate receiver. A learning-based anti-eavesdropping framework via smart beamforming over the MISO VLC wiretap channel has been proposed to prevent the eavesdropper from wiretapping the secret signals. To derive the optimal beamforming policy, an RL-based MISO VLC beamforming control scheme has been designed for the MDP in a dynamic environment. To cope with the high-dimensional and continuous action and state spaces more effectively and efficiently, a DRL-based MISO VLC beamforming control scheme has been introduced to further increase the convergence speed and the learning performance of the smart beamforming based anti-eavesdropping system. Simulation results verify that the proposed DRL-based scheme can significantly increase the secrecy rate and utility, and decrease the BER of the legitimate receiver compared with the

existing benchmark scheme. Moreover, the proposed schemes are able to approach or even outperform the performance of the zero-forcing beamforming.

REFERENCES

[1] X. Ma, J. Gao, F. Yang, W. Ding, H. Yang, and J. Song, "Integrated power line and visible light communication system compatible with multi-service transmission," *IET Commun.*, vol. 11, no. 1, pp. 104–111, 2017.

[2] 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.

[3] J. Song, W. Ding, F. Yang, H. Yang, B. Yu, and H. Zhang, "An indoor broadband broadcasting system based on PLC and VLC," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 299–308, Jun. 2015.

[4] F. Yang and J. Gao, "Dimming control scheme with high power and spectrum efficiency for visible light communications," *IEEE Photon. J.*, vol. 9, no. 1, Feb. 2017, Art. no. 7901612.

[5] H. Haas, L. Yin, Y. Wang, and C. Chen, "What is LiFi?" *J. Lightw. Technol.*, vol. 34, no. 6, pp. 1533–1544, Mar. 15, 2016.

[6] Y. Sun, F. Yang, and J. Gao, "Comparison of hybrid optical modulation schemes for visible light communication," *IEEE Photon. J.*, vol. 9, no. 3, Jun. 2017, Art. no. 7904213.

[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] W. Xu, J. Wang, H. Shen, H. Zhang, and X. You, "Indoor positioning for multiphotodiode device using visible-light communications," *IEEE Photon. J.*, vol. 8, no. 1, pp. 1–11, Feb. 2016.

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

[10] 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.

[11] R. Liu and W. Trappe, *Securing Wireless Communications at the Physical Layer*. New York, NY, USA: Springer, 2010.

[12] Q. Yan, H. Zeng, T. Jiang, M. Li, W. Lou, and Y. T. Hou, "Jamming resilient communication using MIMO interference cancellation," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1486–1499, Jul. 2016.

[13] G. Pan, J. Ye, and Z. Ding, "Secure hybrid VLC-RF systems with light energy harvesting," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4348–4359, Oct. 2017.

[14] G. Pan, C. Tang, X. Zhang, T. Li, Y. Weng, and Y. Chen, "Physical-layer security over non-small-scale fading channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1326–1339, Mar. 2016.

[15] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, 1975.

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

[17] L. Xiao, J. Liu, Q. Li, N. B. Mandayam, and H. V. Poor, "User-centric view of jamming games in cognitive radio networks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2578–2590, Dec. 2015.

[18] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.

[19] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3420–3430, Apr. 2018.

[20] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087–4097, May 2018.

[21] C. Lu, W. Xu, H. Shen, J. Zhu, and K. Wang, "MIMO channel information feedback using deep recurrent network," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 188–191, Jan. 2019.

[22] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*. [Online]. Available: https://arxiv.org/abs/1509.02971

[23] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 387–395.

[24] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

[25] A. Chaaban, J.-M. Morvan, and M.-S. Alouini, "Free-space optical communications: Capacity bounds, approximations, and a new sphere-packing perspective," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 1176–1191, Mar. 2016.

[26] A. Chaaban, Z. Rezki, and M. S. Alouini, "Fundamental limits of parallel optical wireless channels: Capacity results and outage formulation," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 296–311, Jan. 2017.

[27] F. Oggier and B. Hassibi, "The secrecy capacity of the MIMO wiretap channel," *IEEE Trans. Inf. Theory*, vol. 8, no. 57, pp. 4961–4972, Aug. 2011.

[28] M. A. Arfaoui, A. Ghrayeb, and C. M. Assi, "Secrecy performance of multi-user MISO VLC broadcast channels with confidential messages," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7789–7800, Nov. 2018.

[29] A. Mostafa and L. Lampe, "Optimal and robust beamforming for secure transmission in MISO visible-light communication links," *IEEE Trans. Signal Process.*, vol. 64, no. 24, pp. 6501–6516, Dec. 2016.

[30] M. A. Arfaoui, H. Zaid, Z. Rezki, A. Ghrayeb, A. Chaaban, and M. S. Alouini, "Artificial noise-based beamforming for the MISO VLC wiretap channel," *IEEE Trans. Commun.*, vol. 67, no. 4, pp. 2866–2879, Dec. 2018.

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

[32] F. Wang *et al.*, "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.

[33] C. Zhang, W. Xu, and M. Chen, "Robust MMSE beamforming for multiuser MISO systems with limited feedback," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 588–591, Jul. 2009.

[34] H. Ma, L. Lampe, and S. Hranilovic, "Robust MMSE linear precoding for visible light communication broadcasting systems," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Atlanta, GA, USA, Dec. 2013, pp. 1081–1086.

[35] S. O. Somuyiwa, A. György, and D. Gündüz, "A reinforcement-learning approach to proactive caching in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1331–1344, Jun. 2018.

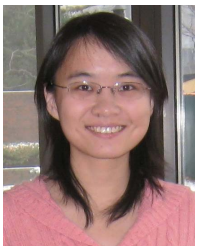
[36] A. Sadeghi, F. Sheikholeslami, and G. B. Giannakis, "Optimal and scalable caching for 5G using reinforcement learning of space-time popularities," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 180–190, Feb. 2018.

[37] 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.

[38] L. Zeng *et al.*, "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.

[39] J. M. Kahn and J. R. Barry, "Wireless infrared communications," *Proc. IEEE*, vol. 85, no. 2, pp. 265–298, Feb. 1997.

[40] P. Cheridito, H. Kawaguchi, and M. Maejima, "Fractional Ornstein-Uhlenbeck processes," *Electron. J. Probab.*, vol. 8, no. 3, pp. 1–14, Jan. 2003.



Liang Xiao (M'09–SM'13) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, China, in 2000, the M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and the Ph.D. degree in electrical engineering from Rutgers University, NJ, USA, in 2009. She was a Visiting Professor with Princeton University, Virginia Tech, and the University of Maryland, College Park. She is currently a Professor with the Department of Communication Engineering and the Department Head of the Department of Cybersecurity, Xiamen University, Fujian, China. Her research interests include wireless security, smart grids, and wireless communications. She is a member of the IEEE Technical Committee on Big Data. She was a recipient of the Best Paper Award for the 2017 IEEE ICC and the 2016 IEEE INFOCOM Bigsecurity WS. She has served in several editorial roles, including an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE TRANSACTIONS ON COMMUNICATION.

Geyi Sheng received the B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2017, where she is currently pursuing the M.S. degree with the Department of Communication Engineering. Her research interests include network security and wireless communications.

Sicong Liu (S'15–M'17) received the B.S.E. and Ph.D. degrees (Hons.) in electronic engineering from Tsinghua University, Beijing, China, in 2012 and 2017, respectively. He was a Visiting Scholar with the City University of Hong Kong in 2010. He served as a Senior Research Engineer with Huawei Technologies. He joined Xiamen University in 2018, where he is currently an Assistant Professor. His current research interests lie in sparse signal processing, machine learning, and wireless communications. He has served as an editor, the track chair, or a TPC member of several IEEE and other academic journals and conferences.

Huaiyu Dai (F'17) received the B.E. and M.S. degrees from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 2002, all in electrical engineering.

He was with Bell Labs, Lucent Technologies, Holmdel, NJ, USA, in 2000, and with AT&T Labs-Research, Middletown, NJ, USA, in 2001. He is currently a Professor of electrical and computer engineering with NC State University, Raleigh, holding the title of University Faculty Scholar. His research interests are in the general areas of communication systems and networks, advanced signal processing for digital communications, and communication theory and information theory. His current research focuses on networked information processing and cross-layer design in wireless networks, cognitive radio networks, network security, and associated information-theoretic and computation-theoretic analysis.

Dr. Dai was a co-recipient of the Best Paper Awards at the 2010 IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS 2010), the 2016 IEEE INFOCOM BIGSECURITY Workshop, and the 2017 IEEE International Conference on Communications (ICC 2017). He co-chaired the Signal Processing for Communications Symposium of the IEEE GLOBECOM 2013, the Communications Theory Symposium of the IEEE ICC 2014, and the Wireless Communications Symposium of the IEEE GLOBECOM 2014. He has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He is currently an Area Editor-in-Charge of wireless communications for IEEE TRANSACTIONS ON COMMUNICATIONS.

Mugen Peng (M'05–SM'11) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005.

He joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering since 2012. In 2014, he was also an Academic Visiting Fellow with Princeton University, USA. He leads a Research Group focusing on wireless transmission and networking technologies in BUPT. He has authored or coauthored over 90 refereed IEEE journal papers and over 300 conference proceeding papers. His main research areas include wireless communication theory, radio signal processing, cooperative communication, self-organization networking, heterogeneous networking, cloud communication, and the Internet of Things. He was a recipient of the 2018 Heinrich Hertz Prize Paper Award, the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in the JCN 2016, the IEEE WCNC 2015, the IEEE GameNets 2014, the IEEE CIT 2014, ICCTA 2011, IC-BNMT 2010, and the IET CCWMC 2009. He is currently or has been on the Editorial/Associate Editorial Board of *IEEE Communications Magazine*, IEEE ACCESS, IEEE INTERNET OF THINGS JOURNAL, *IET Communications*, and *China Communications*.

Jian Song (M'06–SM'10–F'16) received the B.Eng. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1990 and 1995, respectively.

He was with The Chinese University of Hong Kong and the University of Waterloo, Canada, in 1996 and 1997, respectively. He has been with Hughes Network Systems, USA, for seven years. He joined the Faculty Team, Tsinghua University, in 2005, as a Professor, where he is currently the Director of the Tsinghua's DTV Technology R&D Center. He has been working in quite different areas of fiber-optic, satellite, wireless, power-line communication, and visible light communication. He has published more than 260 peer-reviewed journal and conference papers. He holds 2 U.S. and more than 40 Chinese patents. His current research interest is in the area of digital TV broadcasting. He is a fellow of IET.