When Visible Light Communication Meets RIS: A Soft Actor-Critic Approach

Long Zhang, *Member, IEEE,* Xingliang Jia, Choong Seon Hong, *Senior Member, IEEE,* and Zhu Han, *Fellow, IEEE*

Abstract

This letter considers a reconfigurable intelligent surface (RIS) aided indoor visible light communication system, where a mirror array-based RIS is deployed to assist the communication from a light-emitting diode (LED) to multiple user terminals (UTs). We aim to maximize the sum-rate in an entire serving period by jointly optimizing the orientation of the RIS reflecting unit, the time fraction for the UT, and the transmit power at the LED, subject to the communication and illumination intensity requirements. To solve this high-dimensional non-convex problem, we first transform it as a constrained Markov decision process. Then, a soft actor-critic (SAC)-based deep reinforcement learning (DRL) algorithm is proposed with the objective of maximizing both the average reward and the expected policy entropy. Simulation results show that the proposed SAC-based joint optimization design outperforms the existing DRL-based baselines in terms of the sum-rate and long-term average reward.

Index Terms

Visible light communication, reconfigurable intelligent surface, deep reinforcement learning, soft actor-critic.

L. Zhang and X. Jia are with the School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China (e-mail: lzhang0310@gmail.com, xlj896@gmail.com).

C. S. Hong is with the Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Republic of Korea (e-mail: cshong@khu.ac.kr).

Z. Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, Republic of Korea (e-mail: hanzhu22@gmail.com).

I. INTRODUCTION

Motivated by recent advances in both visible light communication (VLC) and reconfigurable intelligent surface (RIS), applying RIS into the design and optimization of VLC systems has been identified as a symbiotic 6G enabler, gaining upsurge of research interests [1], [2]. In the RIS-VLC framework, the ubiquitous light-emitting diodes (LEDs) are used to transmit data through the visible light that carries the message signals to the photodetectors (PDs) of user terminals (UTs), assisted by the RISs for creating favorable propagation conditions. By mitigating skip-zones and configuring the reflections of the incident visible light signals from LEDs to PDs via RISs, the interplay between VLC and RIS has shown as great benefits to improve the performance of VLC, e.g., illumination relaxation, coverage expansion, and signal quality enhancement [1].

Recent research progress has been made to reveal the potentials of a fusion of VLC and RIS. For instance, the authors in [3] designed a joint optimization scheme of power allocation, LED-RIS reflecting unit association, and LED-UT association to maximize the overall spectral efficiency. In [4], the RIS unit assignment for UTs was optimized to maximize the sum-rate. The similar work can be found in [5], where a joint optimization framework of transceiver signal processing and RIS unit alignment with the "LED-PD" pair was presented to minimize the system's mean square error. However, these works in [3]–[5] only considered the RIS configuration via the unit association design, without capturing the *unit orientation*, especially for the mirror array-based RIS. Typically, the RIS unit orientation can be controlled intelligently to better reflect the incident signal towards the UT, which may further exploit its benefits. In this regard, the authors in [6] derived the optimal orientation of the RIS mirror to set up the robust RIS-reflecting path such that the maximal rate was obtained. In [7], the secrecy rate was maximized by optimizing the RIS unit orientation to defend against the eavesdropper. Despite the works in [6], [7] devoted to improving the system performance, the RIS unit orientation optimization design has not been jointly considered with the efficient resource allocation.

To further develop the potentials of the RIS-VLC systems, it is crucial to jointly optimize RIS unit orientation configuration and resource allocation [2]. However, the joint optimization problem involves multiple optimization variables with high dimensionality, which usually suffers from loss of optimality, high computational complexity, and lack of long-term optimization when using traditional optimization methods as in [3]–[7]. Against this problem, the deep deterministic policy gradient (DDPG)-based deep reinforcement learning (DRL) via the actor-critic method was utilized in [8] to jointly optimize the RIS unit orientation and the LED's beamforming weight for maximizing the secrecy rate. The use of DRL in [8] was also shown to be beneficial in real-time deployment and ease of implementation for practical RIS-VLC systems. However, the work in [8] only considered optimizing the RIS unit orientation thus failed to fully explore the resource allocation problem. Meanwhile, the joint optimization problem cannot effectively addressed by traditional DDPG-based DRL algorithm since it may easily trapped in locally optimal solution due to the existence of action space with high dimensions.

Against this background, this letter proposes a DRL-based framework that uses the state-of-the-art soft actor-critic (SAC) algorithm, designed for the joint optimization scheme of RIS unit orientation configuration and resource allocation in the indoor VLC system assisted by a mirror array-based RIS. To the best of our knowledge, this work is the first attempt to investigate the joint optimization problem for an RIS-VLC system with tools from the SAC-based DRL approach. The specific contributions can be listed as below:

- We formulate a joint RIS unit orientation configuration, time fraction assignment, and power allocation optimization problem to maximize the sum-rate of all UTs across an entire serving period, subject to the rotation angle, communication, and illumination constraints.
- To solve this high-dimensional non-convex problem, we reformulate it as a constrained Markov decision process (CMDP) aiming at maximizing the long-term balance between the sum-rate and penalties. An SAC-based DRL algorithm is designed with the goal of maximizing both the average reward and the expected policy entropy.
- We validate the performance of the proposed SAC-based joint optimization scheme by extensive simulations. The results demonstrate that the proposed scheme performs better than existing DRL-based baselines in terms of the sum-rate and long-term average reward.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider the downlink of an indoor VLC system as shown in Fig. 1, where a set of $\mathcal{K}=\{1,2,\cdots,K\}$ UTs, each equipped with a PD, are served by a single LED with the aid of an RIS attached on the wall. The LED transmits data of the UTs in a TDMA manner within a serving period of T. Denote the duration of time slot reserved for UT k by $\tau_k T$, such that $\sum_{k=1}^{K}\tau_k=1$. The RIS is formed with a set of $\mathcal{N}=\{1,2,\cdots,N\}$ reflecting units, configured in the form of an intelligent mirror array. The orientation of each RIS unit can be tuned independently via two rotational degrees of freedom, i.e., yaw and roll angles, denoted by α_k^n and β_k^n of RIS unit n for

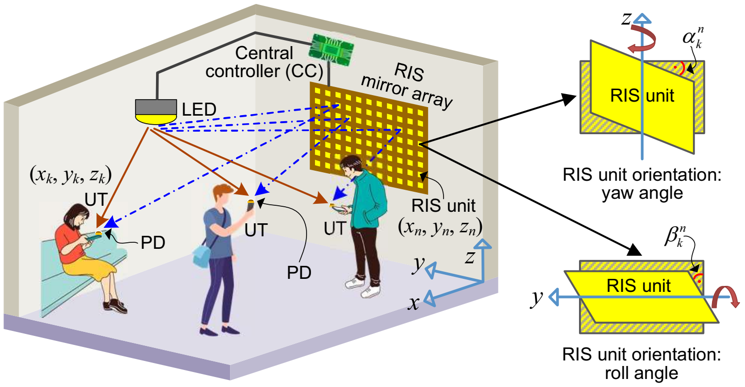

Fig. 1. System model for an RIS-aided indoor VLC system.

UT k, respectively, as shown in Fig. 1. A *central controller (CC)* is mounted at the ceiling to jointly control the RIS and LED. The locations of the UTs and RIS as well as the channel state information (CSI) of all channels are known at this controller, which can be obtained by the VLC positioning and channel estimation techniques [3]. For ease of discussion, the locations of UT k and RIS unit n in the 3D coordinate are specified by (x_k, y_k, z_k) and (x_n, y_n, z_n), respectively.

A. Channel and Signal Model

1) Direct Channel: We employ the Lambertian model [4] to depict the channel gain of direct path between the LED and UT k, which can be determined by

$$h_k = \frac{(m+1)\,A_{\mathrm{p}}}{2\pi d_{l,k}^2} \cos^m\left(\Theta_k^l\right) g_{\mathrm{f}}\left(\psi_k^l\right) \cos\left(\psi_k^l\right) g_{\mathrm{c}}\left(\psi_k^l\right), \tag{1}$$

where $m = -\frac{1}{\log_2 \cos\left(\xi_{1/2}\right)}$ is the Lambertian index with $\xi_{1/2}$ the LED's half-intensity radiation angle, A_{p} is the PD's active aperture area of each UT, $d_{l,k}$ is the distance between the LED and UT k, Θ_k^l and ψ_k^l are the angles of irradiance and incidence for the direct path from the LED to UT k, respectively, and $g_{\mathrm{f}}\left(\psi_k^l\right)$ and $g_{\mathrm{c}}\left(\psi_k^l\right)$ are the gains of the optical filter and concentrator, respectively. Here, $g_{\mathrm{c}}\left(\psi_k^l\right) = \frac{f^2}{\sin^2 \Psi}$, where f is the refractive index and Ψ is the PD's field-of-view of UT.

2) RIS-Reflecting Channel: The visible light reflections via RIS typically include specular reflection and diffuse reflection. However, diffuse reflection can be ignored due to the smooth RIS reflecting surface and the relatively low intensity level compared to the direct path channel gain [1]. We thus focus on specular reflection to depict the RIS-reflecting channel gain, which can be derived as an approximate expression following an additive model under the point source assumption [9]. Specifically, the channel gain of RIS-reflecting path from the LED to UT k reflected by RIS unit n can be calculated as

$$\tilde{h}_{l,n,k}=\frac{\zeta_{\mathrm{u}}\left(m+1\right)A_{\mathrm{p}}A_{\mathrm{u}}}{2\pi\left(d_{l,n}+d_{n,k}\right)^{2}}\cos^{m}\left(\Theta_{n}^{l}\right)\cos\left(\psi_{n}^{l}\right)\times\cos\left(\Theta_{k}^{n}\right)g_{\mathrm{f}}\left(\psi_{k}^{n}\right)\cos\left(\psi_{k}^{n}\right)g_{\mathrm{c}}\left(\psi_{k}^{n}\right),\tag{2}$$

where ζ_{u} is the reflection coefficient of the RIS unit, A_{u} is the physical area of the RIS unit, $d_{l,n}$ is the distance from the LED to RIS unit n, $d_{n,k}$ is the distance from RIS unit n to UT k, Θ_{n}^{l}, ψ_{n}^{l}, Θ_{k}^{n}, and ψ_{k}^{n} are the angles of irradiance and incidence from the LED to RIS unit n and from RIS unit n to UT k, respectively.

Due to the specular reflection of concern, we conclude that the angle of incidence is equal to the angle of reflection, i.e., $\psi_{n}^{l}=\Theta_{k}^{n}$, and further represent the cosine of them by [6]

$$\cos\left(\psi_{n}^{l}\right)=\cos\left(\Theta_{k}^{n}\right)=\frac{x_{n}-x_{k}}{d_{n,k}}\sin\alpha_{k}^{n}\cos\beta_{k}^{n}+\frac{y_{n}-y_{k}}{d_{n,k}}\cos\alpha_{k}^{n}\cos\beta_{k}^{n}+\frac{z_{n}-z_{k}}{d_{n,k}}\sin\beta_{k}^{n}.\tag{3}$$

3) Received Signal: Denote by p_k the transmit power of the LED to UT k, and let $s_k\in[-A,A]$ be the transmitted data symbol of UT k with A being a positive value, for $\mathbb{E}\left\{s_k\right\}=0$ and $\mathbb{E}\left\{s_k^2\right\}=1$ [10]. The transmitted signal of the LED to UT k is given as $x_k=\sqrt{p_k}s_k+b$, where b is a constant meaning the direct current (DC) offset. Using the non-negativity of the VLC signal, i.e., $\sqrt{p_k}s_k+b\geq 0$, we obtain $p_k\leq\left(\frac{b}{A}\right)^2$. Moreover, *to satisfy the eye safety and LED illumination requirements*, the transmitted signal has to be bounded by the LED's maximum permissible current I_{c} [10], i.e., $\sqrt{p_k}s_k+b\leq I_{\mathrm{c}}$, which yields $p_k\leq\left(\frac{I_{\mathrm{c}}-b}{A}\right)^2$. Therefore, the transmit power for UT k is upper bounded as $p_k\leq\min\left\{\left(\frac{b}{A}\right)^2,\left(\frac{I_{\mathrm{c}}-b}{A}\right)^2\right\}$.

We combine both the direct and RIS-reflecting channels to determine the received signal at the UT. Integrating (1) and (2), the combined channel gain from the LED to UT k is equal to $\bar{h}_k=h_k+\sum_{n=1}^{N}\tilde{h}_{l,n,k}$. After removing the DC offset at the UT side, the received signal at UT k can be written as

$$y_k=\kappa_{\mathrm{o2e}}\bar{h}_k\sqrt{p_k}s_k+\varpi_k,\tag{4}$$

where κ_{o2e} is the optical-to-electric conversion factor of the PD, and $\varpi_k\sim\mathcal{N}\left(0,\sigma^2\right)$ the additive white Gaussian noise.

B. Sum-Rate Maximization Problem Formulation

In this letter, we study the joint RIS unit orientation configuration and resource allocation problem to maximize the sum-rate of the system. Due to the illumination requirements and the necessity for the non-negativity of transmitted signal, the classic Shannon capacity formula cannot be directly employed to describe the achievable rate of the UT. We thus resort to the tight lower bound of channel capacity for the dimmable VLC systems [11], and particularly employ the closed-form bound to depict the achievable rate of UT k, which is obtained by

$$R_k = \frac{B}{2} \log_2 \left(1 + \frac{e}{2\pi} \cdot \left(\frac{\kappa_{\text{o2e}} \bar{h}_k \sqrt{p_k}}{\sigma} \right)^2 \right), \tag{5}$$

where B is the channel bandwidth of downlink transmission, e is the Euler's number, and $\left(\kappa_{\text{o2e}} \bar{h}_k \sqrt{p_k} / \sigma\right)^2$ is the received signal-to-noise ratio (SNR) of UT k.

Define $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{(\alpha_k^n, \beta_k^n), \forall k, n\}$, $\boldsymbol{\tau} = \{\tau_k, \forall k\}$, and $\mathbf{P} = \{p_k, \forall k\}$. We aim to maximize the sum-rate across the serving period T by jointly optimizing the RIS unit orientation configuration $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, time fraction assignment $\boldsymbol{\tau}$, and power allocation $\mathbf{P}$, subject to the rotation angle, communication, and illumination constraints. The problem of our interest is then formulated as

$$\max_{(\boldsymbol{\alpha}, \boldsymbol{\beta}), \boldsymbol{\tau}, \mathbf{P}} \sum_{k=1}^{K} \tau_k R_k \tag{6a}$$

$$\text{s.t. } R_k \geq R_{\min}, \ \forall k \in \mathcal{K}, \tag{6b}$$

$$\alpha_k^n, \beta_k^n \in \left[-\frac{\pi}{2}, \frac{\pi}{2} \right], \ \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{6c}$$

$$\sum_{k=1}^{K} \tau_k = 1, \ 0 \leq \tau_k \leq 1, \ \forall k \in \mathcal{K}, \tag{6d}$$

$$P_{\min} \leq p_k \leq \min \left\{ \left(\frac{b}{A} \right)^2, \left(\frac{I_{\text{c}} - b}{A} \right)^2 \right\}, \forall k \in \mathcal{K}, \tag{6e}$$

$$\sum_{k=1}^{K} p_k \leq P_l^{\text{total}}, \ \forall k \in \mathcal{K}. \tag{6f}$$

Here, (6b) sets the lower bound $R_{\min}$ of achievable rate for the UT. (6c) implies the bounds of the yaw and roll angles w.r.t. the RIS unit. (6d) details the time fraction allocation constraint. (6e) ensures the lower bound $P_{\min}$ and upper bound of transmit power for the UT to satisfy the *communication and illumination intensity requirements*. Finally, (6f) limits the total transmit power of the LED below an upper bound P_l^{total}.

Note that (6) is a non-convex optimization problem, which is intractable mainly for the non-convex objective function and the tightly coupling of multiple variables with high dimensions. In general, alternative optimization can be used to solve (6) by decoupling the optimization variables of the RIS unit orientation, time fraction, and transmit power, which yields three subproblems that can be alternatively optimized via sine-cosine algorithm (SCA) [6], Lagrangian dual decomposition, and minorization-maximization (MM) algorithm [3], respectively. However, the use of traditional optimization methods cannot obtain the globally optimal solution of (6). Another challenge lies in the computational cost, which grows exponentially with the scale of RIS units. Besides, problem (6) usually benefits from the long-term goal that can be achieved by the solution of sequential decision-making problem suitable for DRL. However, traditional DRL algorithms like DDPG, cannot manage the high-dimensional continuous variables efficiently and may easily get stuck in a local optimum. Therefore, we will resort to the SAC method, as an off-policy maximum entropy actor-critic DRL algorithm [12], to solve (6) with low complexity due to its better exploratory ability and stability compared with other DRL algorithms.

III. SOFT ACTOR-CRITIC BASED SOLUTION

In this section, we first reformulate the original problem into a CMDP, and then develop the SAC-based joint optimization algorithm to achieve the maximum long-term average reward.

A. CMDP Formulation

The problem (6) can be modeled as a CMDP, described by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathbb{P}, r, c \rangle$ with a state space $\mathcal{S}$, an action space $\mathcal{A}$, and a state transition probability $\mathbb{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. We consider the CC serving as an *agent* that learns a stochastic policy $\pi\left(\mathbf{a}_t | \mathbf{s}_t\right) : \mathcal{S} \times \mathcal{A} \to [0, 1]$, by interacting with the RIS-VLC environment. At time t, the agent observes a state $\mathbf{s}_t \in \mathcal{S}$ and then chooses an action $\mathbf{a}_t \sim \pi\left(\mathbf{a}_t | \mathbf{s}_t\right) \in \mathcal{A}$ according to a policy π. After taking action $\mathbf{a}_t$, the environment transits to next state $\mathbf{s}_{t+1} \sim \mathbb{P}\left(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t\right) \in \mathcal{S}$ and returns a reward $r\left(\mathbf{s}_t, \mathbf{a}_t\right)$ via the reward function $r: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the cost function $c: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The agent stores the state transition tuple $\langle \mathbf{s}_t, \mathbf{a}_t, r\left(\mathbf{s}_t, \mathbf{a}_t\right), \mathbf{s}_{t+1} \rangle$ into an experience replay buffer $\mathcal{D}$. Denote by ρ_π the state-action trajectory induced by policy π. The basic elements of the CMDP are designed as follows.

1) Action: The action taken by the agent at time t can be defined by $\mathbf{a}_t = \left(\left\{\left(\alpha_{k,n}^t, \beta_{k,n}^t\right)\right\}_{\forall k,n}, \{\tau_k^t\}_{\forall k}, \{p_k^t\}_{\forall k}\right)$, which consists of the rotation angles of the RIS unit, time fraction for the UT, and transmit power of the LED to the UT.

2) State: Denote by $\eta_k = \sum_{n=1}^N \tilde{h}_{l,n,k}$ the channel gain component via the RIS reflecting for UT k. We organize the state of the agent at time t by $\mathbf{s}_t = (\mathbf{a}_{t-1}, \{h_k^t\}_{\forall k}, \{\eta_k^t\}_{\forall k})$, where $\mathbf{a}_{t-1}$ is the previous action of the agent at time $t-1$, and h_k^t and η_k^t are the direct and reflecting channel gains for UT k at time t, respectively.

3) Reward: We derive the reward obtained by the agent via well capturing both the objective and constraints of (6). Since the objective in (6) is to maximize the sum-rate, the reward function should be correlated positively with (6a), and it thus can be denoted by $r(t) = \sum_{k=1}^K \tau_k^t R_k^t$. Due to the fact that constraints (6c)-(6e) can be easily satisfied through some regulations on action space $\mathcal{A}$, we then examine constraints (6b) and (6f) to design the cost function. For (6b), we define cost $c_1(t) = \sum_{k=1}^K \mathbb{1}\left[R_k^t < R_{\min}\right]$ as the first penalty term implying the total numbers of UTs with the achievable rate being below bound $R_{\min}$. Here, $\mathbb{1}[\cdot]$ refers to an indicator function. For (6f), we model cost $c_2(t) = \mathbb{1}\left[\sum_{k=1}^K p_k^t > P_l^{\text{total}}\right]$ as the second penalty term showing the total transmit power of the LED being out of bound P_l^{total}. Therefore, the reward of the agent taking action $\mathbf{a}_t$ under state $\mathbf{s}_t$ can be achieved by

$$r(\mathbf{s}_t, \mathbf{a}_t) = r(t) - \mu_1 \cdot c_1(t) - \mu_2 \cdot c_2(t), \tag{7}$$

where $\mu_1, \mu_2 > 0$ are the adjustable penalty parameters.

B. SAC-Based Joint Optimization Algorithm

The objective of SAC is to learn an optimal stochastic policy π^* that maximizes the expected cumulative reward along with the expected entropy of the policy over ρ_π [12], i.e.,

$$\pi^* = \arg\max_{\pi} \mathop{\mathbb{E}}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[\sum_{i=t}^{\infty} \gamma^{i-t} [r(\mathbf{s}_i, \mathbf{a}_i) + \omega \mathcal{H}(\pi(\cdot|\mathbf{s}_i))]\right], \tag{8}$$

where $\gamma \in [0, 1)$ is the discount factor, $\mathcal{H}(\pi(\cdot|\mathbf{s}_t))$ is the policy entropy, given by $\mathcal{H}(\pi(\cdot|\mathbf{s}_t)) = \mathbb{E}_{\mathbf{a}_t \sim \pi}[-\log \pi(\mathbf{a}_t|\mathbf{s}_t)]$, and $\omega > 0$ is the temperature parameter that controls the tradeoff between the policy entropy and the expected return.

The SAC consists of the actor network to generate a policy that decides the actions to be taken, and the critic network to assess the actions taken and guide the actor to learn an optimal policy.

The learning process alternates between optimizing both networks via the policy improvement and evaluation, respectively, for maximizing the expected return and entropy.

1) Critic: For the critic network, the SAC employs two main Q-networks to avoid over-estimation of soft Q-values, and also uses two target Q-networks for enhancing the stability of learning process. Here, one Q-network Q_{θ_j} with parameter θ_j approximates soft Q-function $Q_{\theta_j}\left(\mathbf{s}_t, \mathbf{a}_t\right)$ while maintaining one target Q-network $Q_{\bar{\theta}_j}$ with parameter $\bar{\theta}_j$, for $j \in \{1, 2\}$. The soft Q-function parameters can be trained at time t by minimizing the mean-squared Bellman error (MSBE) between evaluated Q-value $Q_{\theta_j}\left(\mathbf{s}_t, \mathbf{a}_t\right)$ and target Q-value y_t, i.e.,

$$L_Q\left(\theta_j\right) = \mathbb{E}_{\langle \mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1}\rangle \sim \mathcal{D}}\left[\frac{1}{2}\Big(Q_{\theta_j}(\mathbf{s}_t, \mathbf{a}_t) - y_t\Big)^2\right]. \tag{9}$$

Here, $y_t = r_t + \gamma\left(\min_j Q_{\bar{\theta}_j}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \omega \log \pi_\phi\left(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}\right)\right)$, where $r_t \triangleq r\left(\mathbf{s}_t, \mathbf{a}_t\right)$ for notational simplicity, and parameter θ_j of Q_{θ_j} can be updated through the gradient of $L\left(\theta_j\right)$, i.e., $\theta_j \leftarrow \theta_j - \delta_Q \nabla_{\theta_j} L\left(\theta_j\right)$ with δ_Q being the learning rate. Besides, parameter $\bar{\theta}_j$ of $Q_{\bar{\theta}_j}$ can be updated using the soft update method as $\bar{\theta}_j \leftarrow \Gamma\theta_j + (1 - \Gamma)\,\bar{\theta}_j$ with $0 < \Gamma \ll 1$ being the soft update coefficient.

2) Actor: For the actor network, the stochastic policy $\pi_\phi\left(\cdot|\mathbf{s}_t\right)$ is generated as a Gaussian function with parameter ϕ to approximate policy $\pi\left(\cdot|\mathbf{s}_t\right)$. Particularly, parameter ϕ can be learned by minimizing the loss function as follows

$$L_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\Big[\mathbb{E}_{\mathbf{a}_t \sim \pi_\phi}\big[\omega \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) - Q_{\theta_j}(\mathbf{s}_t, \mathbf{a}_t)\big]\Big]\,. \tag{10}$$

Here, policy parameter ϕ of the actor network can be updated via the gradient descent $\phi \leftarrow \phi - \delta_\pi \nabla_\phi L_\pi\left(\phi\right)$ with δ_π being the learning rate.

The loss function of temperature parameter ω is given by

$$L\left(\omega\right) = \mathbb{E}_{\mathbf{a}_t \sim \pi_t}\left[-\omega \log \pi_t(\mathbf{a}_t|\mathbf{s}_t) - \omega\mathcal{H}_0\right], \tag{11}$$

where $\mathcal{H}_0$ is a desired expected entropy constant. By minimizing (11) via the gradient descent, parameter ω can be updated by $\omega \leftarrow \omega - \delta_\omega \nabla_\omega L\left(\omega\right)$ with δ_ω being the learning rate.

3) Algorithm Implementation: The proposed SAC-based joint optimization design is a two-phase procedure enabled by the environment interactions and parameter updates, which can be summarized in Algorithm 1. After initializing the neural network parameters ϕ, θ_1, θ_2, $\bar{\theta}_1$,

Algorithm 1 Proposed SAC-Based Joint Optimization Algorithm.

```
1: Initialize $\phi$, $\theta_1$, $\theta_2$, $\omega$, and experience replay buffer $\mathcal{D}$.
2: Set $\bar{\theta}_j \leftarrow \theta_j$, for $j \in \{1,2\}$.
3: for each episode do
4:     Initialize the environment and obtain initial state $\mathbf{s}_0$.
5:     for each time step do
6:         Observe state $\mathbf{s}_t$ and take action $\mathbf{a}_t \sim \pi\left(\mathbf{a}_t|\mathbf{s}_t\right)$.
7:         Obtain next state $\mathbf{s}_{t+1}$ given action $\mathbf{a}_t$, and calculate reward $r\left(\mathbf{s}_t,\mathbf{a}_t\right)$.
8:         Store transition tuple $\langle \mathbf{s}_t,\mathbf{a}_t,r\left(\mathbf{s}_t,\mathbf{a}_t\right),\mathbf{s}_{t+1}\rangle$ into replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup \{\langle \mathbf{s}_t,\mathbf{a}_t,r\left(\mathbf{s}_t,\mathbf{a}_t\right),\mathbf{s}_{t+1}\rangle\}$.
9:         Randomly sample a mini-batch of transition tuples from replay buffer $\mathcal{D}$ with size of $\Xi$.
10:        Update $\phi$, $\theta_1$, $\theta_2$, $\bar{\theta}_1$, $\bar{\theta}_2$, $\omega$.
11:    end for
12: end for
```

$\bar{\theta}_2$, ω, and experience replay buffer $\mathcal{D}$, each time step consists of two phases: 1) interacting with the environment (Lines 6-7), and 2) updating the network parameters (Lines 8-10). In the first phase, the agent observes system state $\mathbf{s}_t$ and selects action $\mathbf{a}_t$ sampled from current actor network $\pi_\phi\left(\mathbf{a}_t|\mathbf{s}_t\right)$. Afterward, the agent executes the RIS unit orientation configuration and resource allocation policies based on $\mathbf{a}_t$, and then transits to next state $\mathbf{s}_{t+1}$ and gets the reward $r\left(\mathbf{s}_t,\mathbf{a}_t\right)$. In the second phase, the agent stores the transition tuple $\langle \mathbf{s}_t,\mathbf{a}_t,r\left(\mathbf{s}_t,\mathbf{a}_t\right),\mathbf{s}_{t+1}\rangle$ into replay buffer $\mathcal{D}$. Then, we randomly sample a batch of transition tuples from $\mathcal{D}$ with size of Ξ. The parameters of ϕ, θ_1, θ_2, $\bar{\theta}_1$, $\bar{\theta}_2$, ω are updated with the sampled batch accordingly. Through the continuous interactions with the RIS-VLC environment as well as the updates of neural network parameters, the agent can optimizes its policy accordingly, and finally get the optimal policy π^* for achieving the sum-rate maximization.

IV. SIMULATION RESULTS

This section evaluates the performance of the proposed SAC-based joint optimization scheme. For comparison, the following baselines are also performed: 1) SAC with random unit orientation (**Random RIS**), where the rotation angles of RIS uint are randomly generated, following the uniform distribution within $\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$; 2) SAC without RIS (**W/O RIS**) [2], where the wall serves as the reflector with a reflection coefficient of 0.8, and the wall reflective surface is divided into M identical surfaces, each with the same area as the RIS unit, for $M = N$; 3) **DDPG**-based

TABLE I

SUMMARY OF SIMULATION PARAMETERS.

Parameters	Value
Channel model [3]	$m\!=\!1$, $A_{\mathrm{p}}\!=\!1\,\mathrm{cm}^2$, $A_{\mathrm{u}}\!=\!0.01\,\mathrm{m}^2$ $g_{\mathrm{f}}\left(\psi_k^l\right)\!=\!g_{\mathrm{f}}\left(\psi_k^n\right)\!=\!1$, $B\!=\!200\,\mathrm{MHz}$ $f\!=\!1.5$, $\Psi\!=\!80^\circ$, $\zeta_{\mathrm{u}}\!=\!0.95$, $\kappa_{\mathrm{o2e}}\!=\!0.5\,\mathrm{A/W}$
Signal model [10]	$A=2$, $b\!=\!14\,\mathrm{A}$, $I_{\mathrm{c}}=29\,\mathrm{A}$
Problem formulation [3]	$R_{\min}\!=\!1\,\mathrm{Mbps}$, $P_{\min}\!=\!3\,\mathrm{W}$, $P_l^{\mathrm{total}}\!=\!25\,\mathrm{W}$
SAC hyperparameters [12]	$\lvert\mathcal{D}\rvert\!=\!1,000,000$, $\delta_Q\!=\!\delta_\pi\!=\!\delta_\omega\!=\!0.0001$ $\gamma\!=\!0.95$, $\Gamma\!=\!0.005$, $\Xi\!=\!256$, $\omega\!=\!0.036$

algorithm for solving our formulated CMDP; 4) the deterministic policy based SAC (**DP-SAC**) algorithm, where the temperature parameter $\omega\!=\!0$.

In our simulation, we configure an indoor VLC system that $K = 5$ UTs are randomly distributed at their different heights, confined within $[1\,\mathrm{m}, 1.6\,\mathrm{m}]$, inside a $5\,\mathrm{m}\!\times\!5\,\mathrm{m}\!\times\!3\,\mathrm{m}$ room. An LED is mounted at the center of the ceiling, and a mirror array-based RIS is installed on the wall. We define the area of each RIS unit as $10\!\times\!10\,\mathrm{cm}^2$, and set the interval between the adjacent units to $0.25\,\mathrm{cm}$ [3]. The 3D coordinate of the RIS unit's center point for the first row and the first column of the mirror array is set as $(0\,\mathrm{m}, 4\,\mathrm{m}, 2\,\mathrm{m})$. For Algorithm 1, we adopt a five-layer fully connected neural network structure, which consists of three hidden layer, each with 256 neurons. The ReLU is used as the activation function in all hidden layers. During the training process, both the policy network and the Q-networks are trained by the Adam optimizer. Unless otherwise stated, the remaining system parameters and the hyperparameter settings of Algorithm 1 are given in Table I.

Fig. 2(a) shows the learning curves of the proposed scheme and the baselines for $12,000$ episodes of training. As shown, the proposed scheme needs more episodes to achieve convergence in comparison with the baselines. The reason is that stochastic policy can enable the agent to obtain enhanced exploratory abilities, resulting in a longer learning process. However, the proposed scheme exhibits superior performance than the baselines when all the schemes reach the convergence. Noteworthy, since the baselines of Random RIS and W/O RIS fail to optimize the RIS unit orientation, they need less time to reach their maximum rewards than other schemes.

Fig. 2(b) plots the sum-rate of the system w.r.t. N. As shown, the proposed scheme can obtain an average sum-rate gain of about 40.43% when comparing to the Random RIS scheme. Besides, even for the case of the wall as reflector, it still outperforms the case of Random RIS without optimizing the RIS unit orientation under any values of N. The results show the sum-rate gains benefit a lot from the optimization of the RIS unit orientation. We can also see that the proposed

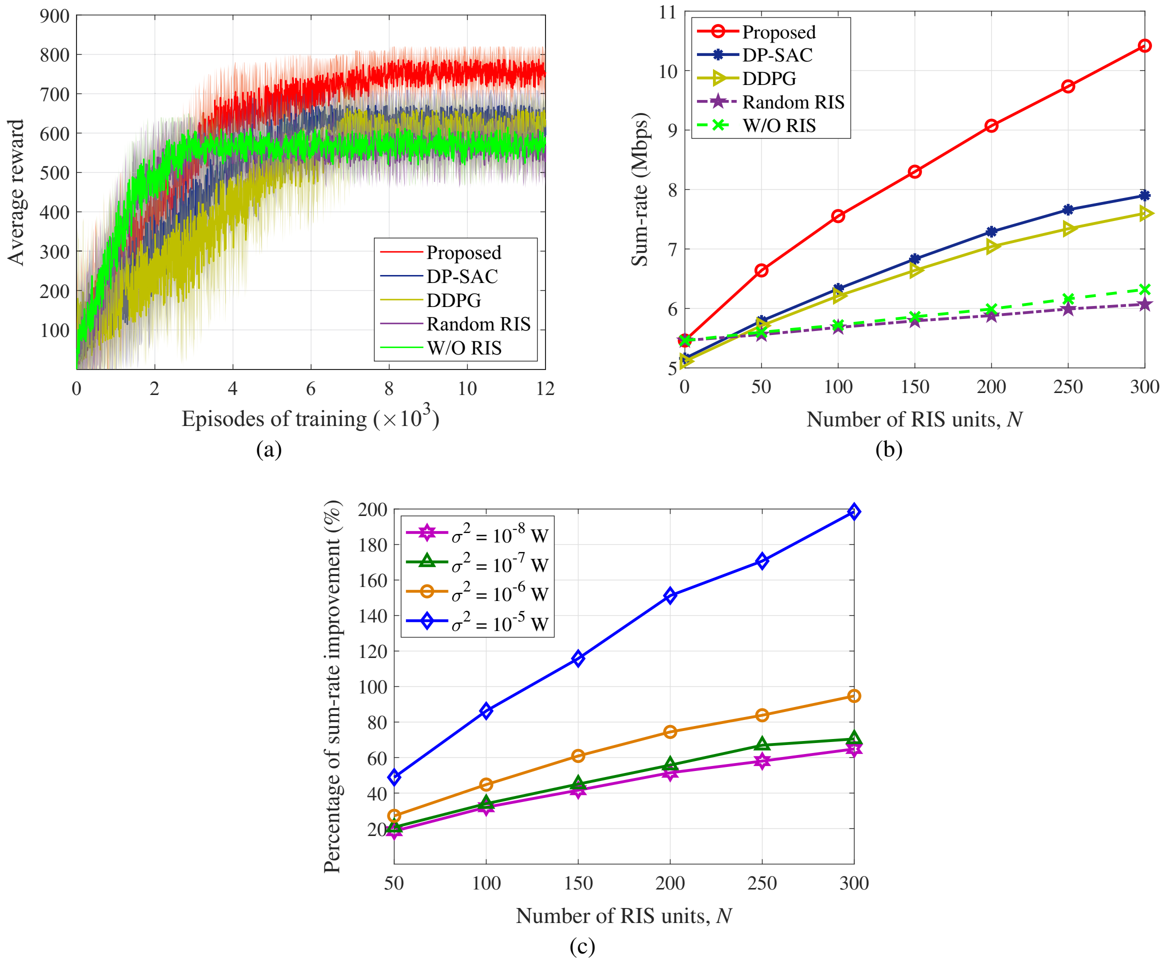

Fig. 2. (a) The average reward versus the episodes of training, with $N = 100$ and $\sigma^2 = 10^{-8}$ W; (b) The sum-rate versus the number of RIS units N, with $\sigma^2 = 10^{-8}$ W; (c) The percentage of sum-rate improvement versus the number of RIS units N under different noise power σ^2.

scheme outperforms the baselines, and the gap between them becomes larger as N grows. The reason is that the stochastic policy allows for the exploration of a wider range of actions and obtain better performance than the deterministic policy as the variable dimension becomes larger.

Fig. 2(c) depicts the percentage of sum-rate improvement w.r.t. N under different noise power σ^2. The percentage of interest is used to reflect the relationship between the proposed scheme and the W/O RIS scheme. It can be seen that the proposed scheme outperforms the W/O RIS scheme for any values of N and σ^2. Besides, more improvements can be significantly obtained as σ^2 becomes larger under the same value of N. Noteworthy, the percentage of the sum-rate improvement can reach up to 198.53% with $N = 300$ when $\sigma^2 = 10^{-5}$ W. This important observation indicates that the proposed scheme can still achieve superior performance than the W/O RIS scheme under the relatively high noise power.

V. CONCLUSION

In this letter, we proposed an SAC-based framework for maximizing the sum-rate of all UTs in the mirror array-based RIS aided indoor VLC system. Specifically, we formulated an optimization problem that jointly considers the RIS unit orientation configuration, time fraction assignment, and power allocation. To solve this problem efficiently, we reformulated it as a CMDP, and further presented the SAC-based joint optimization algorithm that aims to obtain the maximum long-term average reward. The simulation results indicated that our proposed scheme improves the sum-rate significantly and obtains higher average reward compared with other baselines. Besides, our proposed scheme is still able to achieve superior sum-rate performance when the noise power is relatively large owing to the deployment of RIS in VLC systems.

REFERENCES

[1] S. Sun, T. Wang, F. Yang, J. Song, and Z. Han, "Intelligent reflecting surface-aided visible light communications: Potentials and challenges," *IEEE Veh. Technol. Mag.*, vol. 17, no. 1, pp. 47–56, Mar. 2022.

[2] S. Aboagye, A. R. Ndjiongue, T. M. N. Ngatched, O. A. Dobre, and H. V. Poor, "RIS-assisted visible light communication systems: A tutorial," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 1, pp. 251–288, Firstquarter 2023.

[3] S. Sun, F. Yang, J. Song, and Z. Han, "Joint resource management for intelligent reflecting surface–aided visible light communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6508–6522, Aug. 2022.

[4] Z. Liu, F. Yang, S. Sun, J. Song, and Z. Han, "Sum rate maximization for NOMA-based VLC with optical intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 12, no. 5, pp. 848–852, May 2023.

[5] S. Sun, F. Yang, J. Song, and R. Zhang, "Intelligent reflecting surface for MIMO VLC: Joint design of surface configuration and transceiver signal processing," *IEEE Trans. Wireless Commun.*, Jan. 2023, Early Access.

[6] S. Aboagye, T. M. N. Ngatched, O. A. Dobre, and A. R. Ndjiongue, "Intelligent reflecting surface-aided indoor visible light communication systems," *IEEE Commun. Lett.*, vol. 25, no. 12, pp. 3913–3917, Dec. 2021.

[7] L. Qian, X. Chi, L. Zhao, and A. Chaaban, "Secure visible light communications via intelligent reflecting surfaces," in *Proc. IEEE ICC*, Montreal, QC, Canada, Jun. 2021.

[8] 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, Dec. 2022.

[9] A. M. Abdelhady, A. K. S. Salem, O. Amin, B. Shihada, and M.-S. Alouini, "Visible light communications via intelligent reflecting surfaces: Metasurfaces vs mirror arrays," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1–20, Dec. 2020.

[10] S. Ma, F. Zhang, H. Li, F. Zhou, Y. Wang, and S. Li, "Simultaneous lightwave information and power transfer in visible light communication systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 19, pp. 5818–5830, Dec. 2019.

[11] J.-B. Wang, Q.-S. Hu, J. Wang, M. Chen, and J.-Y. Wang, "Tight bounds on channel capacity for dimmable visible light communications," *J. Lightw. Technol.*, vol. 31, no. 23, pp. 3771–3779, Dec. 2013.

[12] T. Haarnoja *et al.*, "Soft actor-critic algorithms and applications," arXiv preprint arXiv:1812.05905v2, Jan. 2019.