IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 28, 2021, accepted June 9, 2021, date of publication June 15, 2021, date of current version June 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3089521

Deep Reinforcement Learning-Based Adaptive Handover Mechanism for VLC in a Hybrid 6G Network Architecture

LIQIANG WANG, DAHAI HAN, MIN ZHANG, DANSHI WANG, (Member, IEEE), AND ZHIGUO ZHANG

State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Dahai Han (dahaihan@gmail.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Project 61975020.

ABSTRACT Visible light communication (VLC) is considered an important complementary technology for extremely high sixth-generation (6G) data transmission and has become part of a hybrid 6G indoor network architecture with an ultradense deployment of VLC access points (APs) that presents severe challenges to user mobility. An adaptive handover mechanism, which includes a seamless handover protocol and a selection algorithm optimized with a deep reinforcement learning (DRL) method, is proposed to overcome these challenges. Experimental simulation results reveal that the average downlink data rate with the proposed algorithm is up to 48% better than those with traditional RL algorithms and that this algorithm also outperforms the deep Q-network (DQN), Sarsa and Q-learning algorithms by 8%, 13% and 13%, respectively.

INDEX TERMS 6G, visible light communication (VLC), handover, deep reinforcement learning (DRL).

I. INTRODUCTION

The upcoming sixth-generation (6G) networks will no longer use single-frequency bands as channels, and they may have great potential to achieve high throughput, extremely low latency, strong connectivity and high reliability for meeting the requirements of multi-scenario communication [1]. For example, visible light communication (VLC) plays a prominent role in 6G networks [2] due to its extremely high data transmission capacity, reliable security and low energy consumption. However, there are still many problems to be solved despite the numerous advantages of VLC. As mentioned in [3], indoor attenuation and blockage are important challenges for VLC in 6G networking. Accordingly, 6G hotspots must act as central coordinators to implement handover decisions in order to solve the problem that the connection between a user device (UD) and its original access point (AP) becomes gradually unreliable due to attenuation of the visible light signal from a single AP with the user's movement; in addition, these hotspots should serve as reliable wireless APs to maintain a UD's connection to the network when VLC is blocked by human actions or obstacles. In the VLC+6G hybrid network architecture illustrated in Fig. 1, the indoor base stations of the 6G network should work together with a group of APs for VLC to provide an extremely high data rate for UDs. An ultradense deployment of APs is required to achieve effective communication coverage in a large-scale indoor space. However, a UD will not stay within range of a specific AP for long and will need to frequently switch between APs to ensure connectivity in this hybrid architecture. Generally, the more frequently AP switching occurs, the more vulnerable the system performance and user quality of experience (QoE) are; therefore, this scenario requires an adaptive AP handover mechanism to improve handover efficiency and reduce its impact on the VLC+6G hybrid network. The VLC AP handover mechanism should also satisfy 6G QoE requirements, with low latency, high throughput and continuous connectivity. We consider an indoor hybrid network architecture consisting of a large number of APs for VLC coverage and radio frequency (RF) sites for 6G, as discussed in [4]. The APs can perform high-bit-rate downlink transmission, while the indoor 6G hotspots are generally far away from the UDs and cannot provide the required bandwidth. In addition, the indoor 6G hotspots

The associate editor coordinating the review of this manuscript and approving it for publication was Hayder Al-Hraishawi.

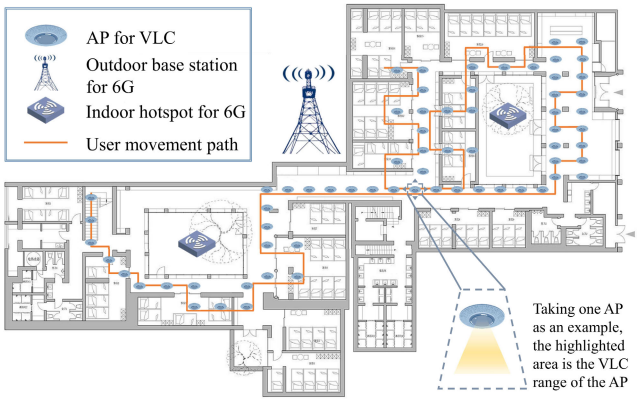

FIGURE 1. Hybrid architecture with VLC and a 6G network in an indoor environment.

play the role of controlling all of the VLC AP selections, as shown in Fig. 1. A general movement path for one UD is marked to illustrate the ultradense deployment of the hybrid architecture.

Due to the asynchronization caused by phasing and delays, researchers should focus on the complicated handover problems encountered in the heterogeneous networking environments of VLC+RF hybrid technologies. In [5], Nguyen *et al*. proposed a hard link-switching scheme based on prescanning and the received signal strength (RSS) that does not require changes to the standard medium access control (MAC) protocol. The ping-pong effect caused by an immediate handover scheme seriously reduces system performance; therefore, Liu *et al*. proposed a dynamic dwell timer design to improve system reliability [6]. Liang *et al*. [7] studied a similar problem using the analytic hierarchy process (AHP) with a cooperative game (CG) model for indoor environments. These traditional methods without artificial intelligence (AI) can be adopted to optimize vertical handover strategies under heterogeneous networking conditions, thereby improving the system reliability and performance in a conventional way. In recent years, AI methods have also come to be considered an important part of 6G networking [8] and have attracted considerable attention in VLC handover mechanisms; in particular, reinforcement learning (RL) algorithms have been applied in AP selection and handover mechanisms. RL can be used to find the optimal scheme in an environment by maximizing the long-term accumulation of rewards. Unlike deep learning (DL), RL does not rely on the sample data in an existing data set. Therefore, it is very suitable for finding the optimal strategy for AP selection and solving the handover problem regarding different channels in heterogeneous VLC networks. In [9], Wang *et al*. proposed a scheme for vertical channel switching between VLC and RF based on a Markov decision process (MDP), which greatly reduces the switching cost while slightly increasing the delay. Bao *et al*. also proposed a vertical switching algorithm based on an MDP in [10] to improve the user QoE and reduce the switching cost. Based on the Q-learning algorithm, in [11], Alenezi *et al*. suggested a scheme that considerably improves system throughput compared with traditional hybrid systems. In [12], Du *et al*. combined the Q-learning algorithm with transfer learning to improve the efficiency of the algorithm itself, thereby further improving the convergence speed and system performance over those achieved with traditional RL. Shao *et al*. [13] proposed a self-optimization algorithm based on Q-learning, where the switching parameters of the APs were optimized by a centralized coordinator. In addition, other AI algorithms have also been used in VLC handover mechanisms; for example, Ji *et al*. [14] used a support vector machine (SVM) approach and Najla *et al*. [15] used a deep neural network (DNN) approach to propose algorithms that effectively improve the network switching performance in VLC- and RF-based heterogeneous networks. However, the performance of the RL algorithms adopted in the above studies is significantly reduced by the restrictions on the Q-table in a large-scale indoor scene with an ultradense deployment of VLC APs. Other AI algorithms mentioned above are difficult to work with because it is challenging to obtain a sufficiently large amount of training data for a particular application scenario. As an alternative, deep reinforcement learning (DRL) has a larger state and action space than RL and does not require the collection of training data sets; therefore, it is suitable for solving highly complex and practical problems in large-scale indoor scenes.

In this paper, we propose a seamless AP handover protocol and a DRL-based algorithm to constitute an adaptive VLC handover mechanism for a hybrid 6G network architecture. The proposed protocol can allow a UD to switch to a target AP without interrupting downlink data transmission, thus greatly reducing the handover delay and making the user insensitive to handover behavior. For purposes of comparison, we design a large and complex indoor scene and estimate the performance of the proposed AI-based algorithm through comparative simulations with existing RL algorithms. The simulation results show that the average downlink data rate of the proposed algorithm is better than those of the deep Q-network (DQN), Sarsa and Q-learning algorithms by 8%, 13% and 13%, respectively. Therefore, the proposed DRL-based algorithm with a decreasing experience replay space exhibits the best performance in the training process compared to the other algorithms in the control group.

II. SYSTEM MODEL OF VLC IN A HYBRID 6G NETWORK ARCHITECTURE

An indoor VLC system model for a hybrid 6G network architecture, which combines indoor base stations (BSs) for 6G and VLC APs, is illustrated in Fig. 2. Note that although the VLC handover mechanism proposed in this paper is also applicable to non-line-of-sight (NLOS) links, we focus only on line-of-sight (LOS) links. Since there is no mobility requirement between an indoor BS and its related APs, stable and affordable wire cables can be adopted for controlling the signal transmission and data interaction processes. A VLC channel combines illumination and specific downlink

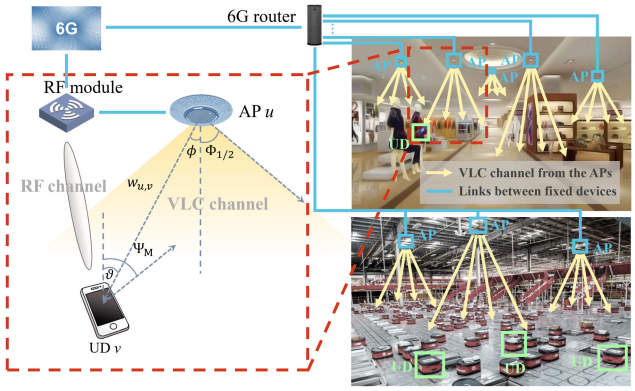

FIGURE 2. Diagram of the system model.

high-speed data transmission requirements. Each indoor BS for 6G is powerful enough to run the handover algorithm to control the signaling between the RF and VLC channels for UDs.

We assume that there are u_{n} UDs that need to access the wireless network in the hybrid 6G architecture and that the channel of each VLC AP is shared by the UDs in a general time-division duplex (TDD) mode [16]. AP_i denotes the ith AP in the environment, the number of UDs accessing AP_i is represented by $u_{i,\mathrm{t}}$ ($u_{\mathrm{m}} \geq u_{i,\mathrm{t}} \geq 1$), and u_{m} is the maximum number of UDs that can access a single AP simultaneously. For VLC systems that allow multiuser access, the downlink data rate R_{m} of each UD accessing the same AP can be expressed as:

$$R_{\mathrm{m}} = \frac{B}{u_{i,\mathrm{t}}} \log_2 (1 + \Gamma), \tag{1}$$

where B is the bandwidth of a single AP and Γ denotes the signal-to-interference-plus-noise ratio (SINR) of the UD, which is given by:

$$\Gamma = \frac{\left(P_{\mathrm{tr}} h_{\mathrm{c}} R_{\mathrm{pd}}\right)^2}{N_{\mathrm{VLC}} B / u_{i,\mathrm{t}} + P_{\mathrm{I}}}, \tag{2}$$

where P_{I} denotes the interference power, P_{tr} represents the transmission power of a single AP, and R_{pd} and N_{VLC} are the detector response and the power spectral density (PSD) of the noise at the UD photodetector (PD), respectively. The channel gain is denoted by h_{c}, which is given by [17]:

$$h_{\mathrm{c}} = \frac{(m+1) A_{\mathrm{da}}}{2\pi w_{u,v}^2} \cos^m \phi g_s (\vartheta) g (\vartheta) \cos \vartheta, \tag{3}$$

where the Lambertian radiant order is expressed as $m = -\ln 2 / \ln \cos \Phi_{1/2}$, with $\Phi_{1/2}$ being the semi-angle of the AP transmitter; A_{da} represents the physical area of the UD PD; $w_{\mathrm{u,v}}$ is the distance between the AP and the UD; and $g_{\mathrm{s}}(\vartheta)$ and $g(\vartheta)$ are the optical signal transmission filter and the optical concentrator gain, respectively. Note that $g(\vartheta)$ is regarded as the gain in an idealized nonimaging concentrator [18] and can be written as follows:

$$g (\vartheta) = \begin{cases} \dfrac{n^2}{\sin^2 \Psi_{\mathrm{M}}}, & 0 \leq \vartheta \leq \Psi_{\mathrm{M}} \\ 0, & \vartheta > \Psi_{\mathrm{M}}, \end{cases} \tag{4}$$

where n is the internal refractive index of the nonimaging concentrator and Ψ_{M} denotes the semi-angle of the field of view (FoV) at the UD PD.

III. USER-ORIENTED SEAMLESS AP HANDOVER PROTOCOL

Considering the need for the intensive deployment of APs due to the limited coverage of a single AP, an adaptive AP selection algorithm is required to keep users online while they move. However, the occurrence of a handover produces additional data overhead, which might diminish the system performance and cause the connection to fail. Therefore, a handover protocol that does not interrupt data streaming is designed in this paper by referring to the seamless handover process for senseless movement in [19], which makes moving users insensitive to the handovers between APs to reduce the impact of handovers on the connection quality of service (QoS). Fig. 3 shows the handover process among a UD, the central coordinator (operating in an indoor 6G hotspot) and the target AP.

As shown in Fig. 3, during the negotiation process, a 6G indoor hotspot acts as the central coordinator for access control and runs the handover algorithm with its computational capabilities. In the handover protocol, due to the introduction of sequence number (SN) status synchronization and handover confirmation processes, the UD may participate in the handover negotiation process while maintaining its original data connection until the handover is successful; this is done to achieve the dual purposes of uninterrupted data transmission in the VLC channel and seamless handover for the user. The main process and advantages of the proposed handover protocol are described below.

For the entire time that a UD has access to the hybrid 6G network, the UD regularly reports the RSS statuses it receives from different APs to the central coordinator, which can then obtain the current location of the AP closest to each UD in accordance with the changes in the reported RSS statuses. This location information can be used as the location parameter data for training the algorithm proposed in this paper. Once the AP selection algorithm executed by the central coordinator determines that a given UD needs to switch to a new AP, it sends a handover request message to the target AP, and the request is confirmed by a handover request acknowledgment (Ack) packet. In this way, the UD is assigned a clear target AP and a precondition for switching. Then, the central coordinator sends an SN status request to the UD to synchronize the UD's current downlink data transmission status and forwards the SN status reported by the UD to the target AP. Because the SN status and Ack represent the SN of each message segment and the SN of the next byte expected to be received by the receiver, the communication

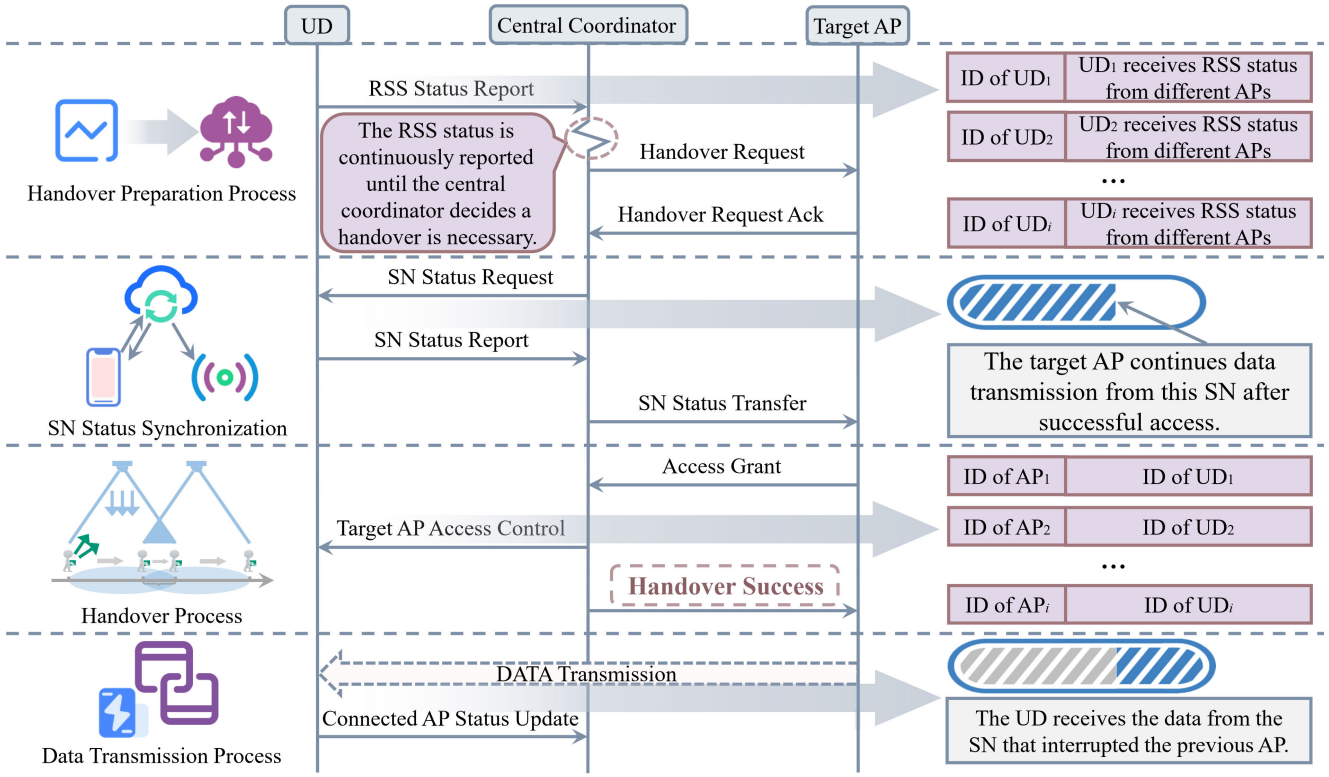

FIGURE 3. Negotiation process of the seamless handover protocol.

process that has been completed before the handover can be accurately recorded. During the handover negotiation process, the UD continues communicating with the original AP until it switches to the target AP. Therefore, the introduction of SN status synchronization largely solves the problem of resuming after interruption through a data synchronization process performed after the selection of the target AP has been negotiated. Finally, the handover process on the UD side is continuously completed until the UD is receiving data from the target AP. The UD reports its connection status to the central coordinator, and the target AP becomes the current AP. In contrast to the traditional VLC handover process, in which a UD disconnects from the current AP before connecting to the target AP, the current AP continues to send downlink data to the UD until handover success is achieved in the proposed handover protocol, thus greatly reducing or even eliminating the delay caused by the UD gaining access to the new target AP.

IV. ADAPTIVE AP SELECTION ALGORITHM WITH DRL

If the UDs' random movements cause too many handovers, this not only increases the amount of unavoidable interruption time but also seriously affects the network performance due to the sharp increase in overhead. Consequently, a fixed AP selection algorithm (i.e., a method based on the principle of proximity or an RSS threshold) is not suitable for large-scale indoor VLC+6G hybrid networks. Therefore, the algorithm proposed in this paper is a dynamic programming method that determines how to act based on the environment for the purpose of maximizing the expected benefits. To this end, DRL is applied for the first time to reduce the delay incurred by unnecessary handovers and increase the downlink data rate.

A. HANDOVER SCHEMES BASED ON TRADITIONAL RL METHODS

To date, several RL methods have been introduced into various handover schemes to determine optimal action strategies so as to obtain the highest possible cumulative rewards and achieve accurate experience estimation.

In [11], Q-learning was used to solve the problem of maximizing throughput when a user selects a new AP. Q-learning is a common model-free learning algorithm that is independent of the transition probability of the environment. Since it is impossible to expand an unknown model with full probability, Q-learning can only observe the transition states and the rewards returned when selected actions are performed in the environment as feedback to guide the selection of the next action. Q-learning, as an off-policy algorithm, provides temporal-difference updates for the sample generated from each state in the form of quadruples $(s_t, a_t, r_{t+1}, s_{t+1})$, where s_t represents the current state, a_t is the action selected in s_t, r_{t+1} denotes the reward returned after performing action a_t, and s_{t+1} is the state after the transition from s_t caused by

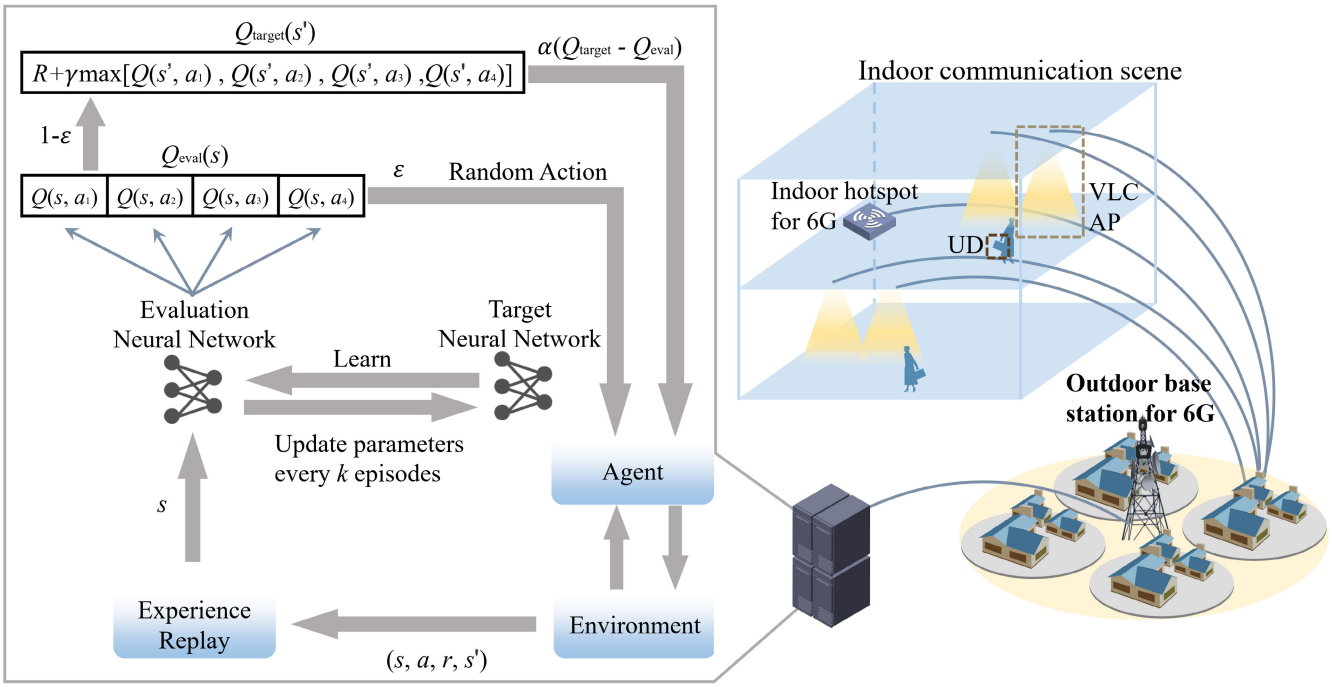

FIGURE 4. DRL process in the hybrid architecture.

the impact of a_t on the environment. The Q-table, composed of the state-action value function $Q(\cdot)$, lies at the core of Q-learning, as it stores the update records of $(s_t, a_t, r_{t+1}, s_{t+1})$. Note that $Q(s_t, a_t)$ is updated based on the Bellman equation, which is as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\,[r_{t\to t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)], \quad (5)$$

where $\alpha \in [0, 1]$ and γ are the learning rate and the discount factor, respectively, and $\max Q(s_{t+1}, a)$ denotes the selection of the largest Q-value among all actions in state s_t. Then, $s_t \leftarrow s_{t+1}$ until s_t is the terminal state.

The Sarsa algorithm was used in [20] to reduce the blocking probability for a handover. Its learning approach is very similar to that of Q-learning; the difference is that in the current state, Sarsa determines the action to be executed in the next state and directly updates the current Q-value with the Q-value of the action to be executed, whereas in Q-learning, the action with the highest Q-value in the next state is not necessarily chosen. Accordingly, Sarsa is called an on-policy algorithm, and its $Q(s_t, a_t)$ is updated as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\,[r_{t\to t+1} + \gamma Q(s_{t+1}, a) - Q(s_t, a_t)], \quad (6)$$

where this equation is repeated for each step of an episode, $s_t \leftarrow s_{t+1}$ and $a_t \leftarrow a_{t+1}$, until s_t is the terminal state.

Due to the limited capacity of the Q-table, it has a very low search efficiency and sometimes cannot even store a full state space. Consequently, traditional RL algorithms (i.e., Q-learning and Sarsa) have difficulty efficiently solving the problem of optimal AP selection for a large-scale indoor VLC hybrid network. Therefore, to address the problem of infinite states and limited actions in the AP selection mechanism, the use of DRL with a neural network is proposed to fit the whole Q-table.

B. ALGORITHM FOR MAXIMIZING THE DATA RATE BASED ON DRL

The DRL process in a large-scale indoor VLC hybrid network environment process is illustrated in Fig. 4. DRL requires the use of the difference between the Q-target and the Q-evaluation to train a neutral network, and the states are updated through backpropagation.

As shown in Fig. 4, the environment includes the physical environment of the VLC network, where APs are densely deployed and the UDs are the agents. The changes in the RSSs from the APs caused by the UDs' random movements produce different types of state feedback, which interact with the agents and produce a set of states $\mathbf{S} = (s_1, s_2, \ldots, s_T)$. An agent uses an ϵ-greedy policy to choose its optimal action with a probability of $1 - \epsilon$ from a set of actions $\mathbf{A} = (a_1, a_2, \ldots, a_N)$, which is obtained from feedback on the state transitions and rewards in the environment. Here, N is the number of APs from which the agent can receive a signal at present. In addition, the agent randomly explores another action with a probability of ϵ to avoid falling into a vicious circle. The APs are evenly distributed in the environment, and the width of the overlapping coverage area of adjacent APs is less than the radius of a single AP coverage area. Therefore,

the agent can receive signals from up to four APs at the same time ($0 < N \leq 4$).

During the DRL process, the reward is a value that can be positive or negative. After a state transition, every state returns a reward with a certain probability distribution, and an additional award or penalty (a positive or negative reward) is given based on the expected estimate of the last action. The relationship between the RSS that a UD receives from an AP and the distance between them follows the Lambert model in equation (3). We refer to the Poisson point process in [21] and assume that the reward R returned by each state and the distance between the UD and the AP are related by a cumulative Poisson distribution, which is given as:

$$R = a_{\mathrm{m}} \sum_{r_i \leq r} \frac{\lambda^{b_{\mathrm{m}} - r_i}}{(b_{\mathrm{m}} - r_i)!} e^{-\lambda}, \tag{7}$$

where a_{m} and b_{m} are the correlation coefficients between the reward and the distance, λ is the average incidence parameter of the Poisson distribution, and r and r_{i} represent the radius of the AP and the distance between the UD and the AP, respectively, the latter of which varies with the UD's movement.

As alluded to above, the DQN algorithm, as a landmark algorithm in the field of DRL, uses a deep convolutional neural network in place of the Q-table to estimate the value function. Furthermore, the target network is independently established to handle the time deviation (TD) in the time difference algorithm. Note that breaking the relevance of the data collected in RL through the experience replay technique is the key step for the DQN algorithm to both ensure the stability of the neural network and greatly improve the performance of the algorithm. The experience replay technique uses a memory space of fixed size as a sliding window, with which random sampling is evenly performed. It utilizes additional computations and memory to reduce the cost of interaction between the agent and the environment. However, it has the problem that distant and useless experiences produce negative feedback and reduce the learning efficiency of the algorithm. Considering this, we note that the performance of the algorithm gradually converges with an increasing number of training episodes, while the detrimental actions that cause negative feedback gradually decrease. Therefore, we introduce an experience replay space size parameter, which decreases as the number of training episodes increases based on the DQN algorithm. Thus, as the algorithm gradually converges, more recent experience is considered more valuable. We believe that global exploration is most needed in the early stage of model training. In this stage, we give the model a higher exploration probability and a larger experience replay space than in other stages to avoid the algorithm falling into a local optimum, thus endowing the model with stronger performance in terms of global optimization. By the end of model training, the learning experience tends to show a more obvious trend of approaching the globally optimal value, so it is no longer necessary to explore the early experience space. Thus, during the implementation process, the experience replay space size parameter follows a monotonically decreasing trend to adaptively change the value with an increasing number of episodes and the convergence trend of the results, thereby improving the efficiency of experience data utilization. Specifically, the mapping relationship between the episode number E_{p} and the experience replay space size parameter E_{r} is $E_{\mathrm{r}} \rightarrow a_{\mathrm{e}}{}^{*}E_{\mathrm{p}} + b_{\mathrm{e}}$, where a_{e} and b_{e} are correlation coefficients. This mapping causes the experience replay space size parameter to monotonically decrease with the episode number. On this basis, the experience replay space size parameter and the reward should simultaneously maintain a relationship of the form $E_{\mathrm{r}} \rightarrow a_{\mathrm{r}}{}^{*} \log(b_{\mathrm{r}}{}^{*}R + c_{\mathrm{r}}) + d_{\mathrm{r}}$ to satisfy the dynamically adaptive characteristics of changing with the training situation, where a_{r}, b_{r}, c_{r} and d_{r} are correlation coefficients.

TABLE 1. Key simulation parameters.

Parameter	Value
Bandwidth per AP, B	100 MHz
Transmission power of a single AP, P_{tr}	3 Watt
Detector response, R_{pd}	0.53 A/W
PSD of noise at a UD PD, N_{VLC}	10^{-21} A^2/Hz
Physical area of a UD PD, A_{da}	10^{-4} m^2
Half-intensity radiation angle, $\Phi_{1/2}$	30°
Semi-angle of the FoV at a UD PD, Ψ_{M}	60°
Transmission delay, T	0.7 ms
Radius of the effective coverage region of a single AP, r	2 m

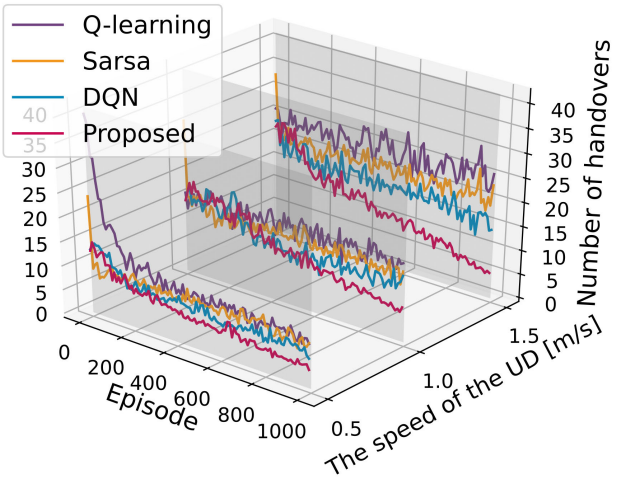

FIGURE 5. Comparison of the numbers of handovers for the four RL-based algorithms.

V. SIMULATION EXPERIMENTS AND RESULTS ANALYSIS

An experimental system was established with experimental VLC data and typical parameters reported for commercially available devices to simulate the indicators for our

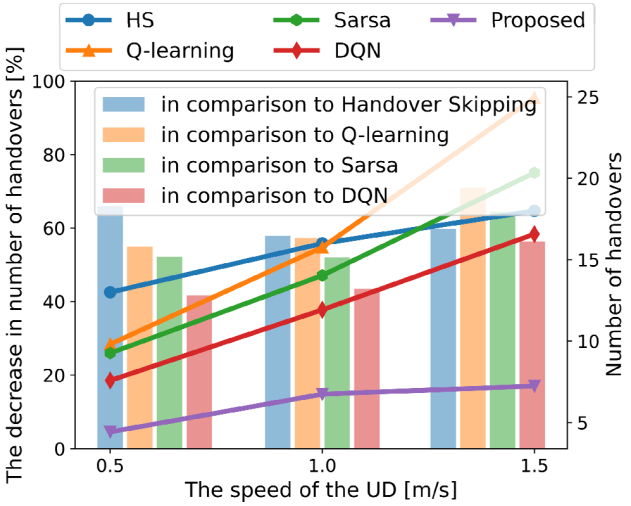

FIGURE 6. Performance improvement of the proposed algorithm compared with the other four methods in terms of the number of handovers.

proposed AP selection algorithm, and the results were compared with the performance of three classic RL algorithms: Q-learning, Sarsa and DQN. The specific VLC parameters are summarized in Table 1, and we set a fixed movement path along which a UD could switch between APs 46 times at most and 4 times at least. In the simulations, we used the number of handovers to measure the performance of the proposed algorithm.

Due to the randomness and diversity of user mobility, using only the RSS or another similar single decision standard would not reduce the number of handovers or mitigate the performance degradation caused by system overhead. Therefore, a lower number of handovers results in better network performance when the UD remains within the effective coverage area of APs and the user QoE is sufficient. We examined the convergence in the above situation with all four RL-based algorithms for various UD movement speeds and one thousand training episodes. As illustrated in Fig. 5, when the UD movement speed is 0.5 m/s, the four algorithms all show an obvious convergence trend. The convergence speed of Q-learning is the slowest, and its limit value is the largest; meanwhile, the proposed algorithm has a fast convergence speed and the lowest limit value (note that the theoretical extreme value is 4 handovers), with the number of handovers continuously oscillating close to the extreme

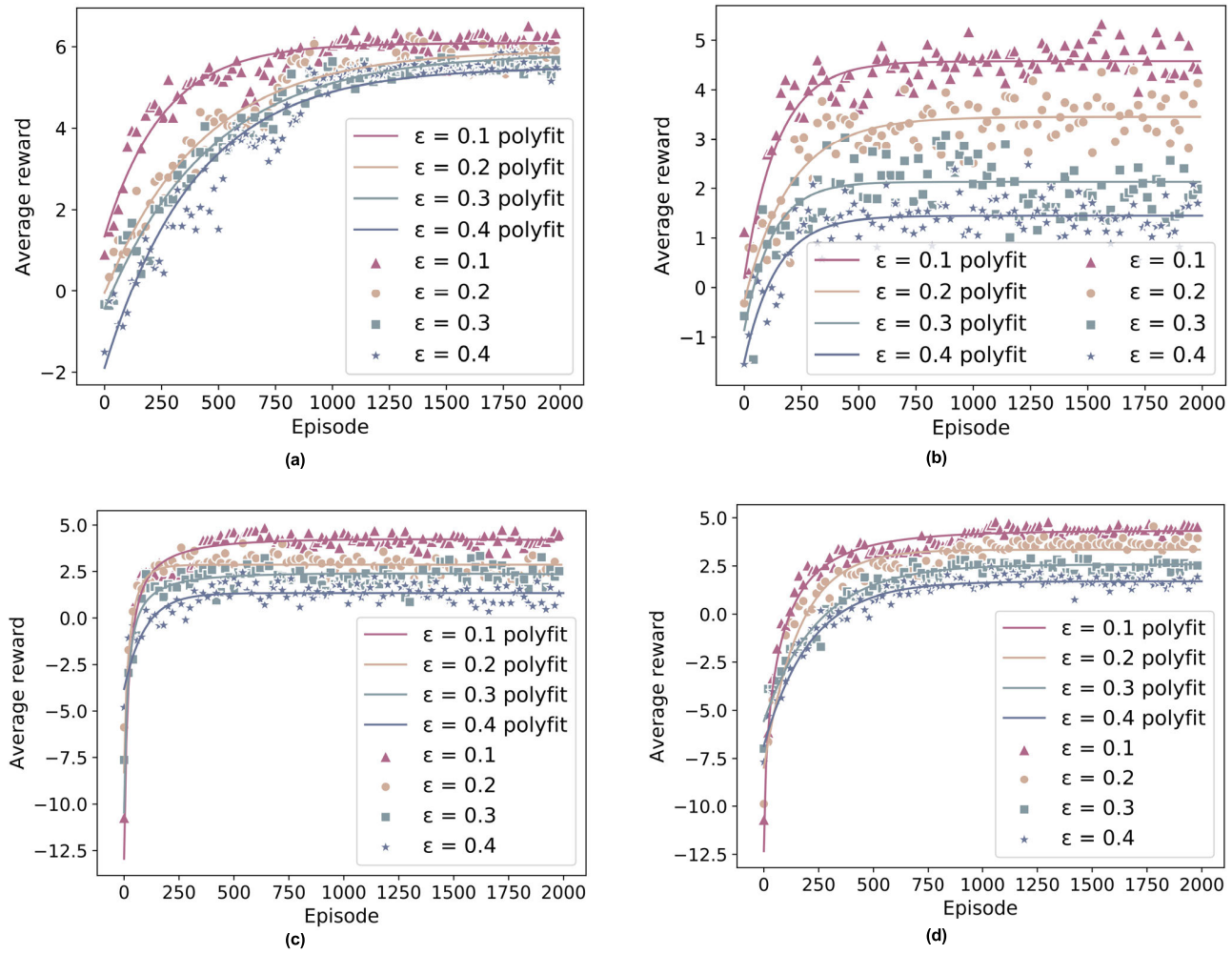

FIGURE 7. Convergence performance comparison of the different RL-based algorithms. (a) Our proposed algorithm. (b) DQN algorithm. (c) Sarsa algorithm. (d) Q-learning algorithm.

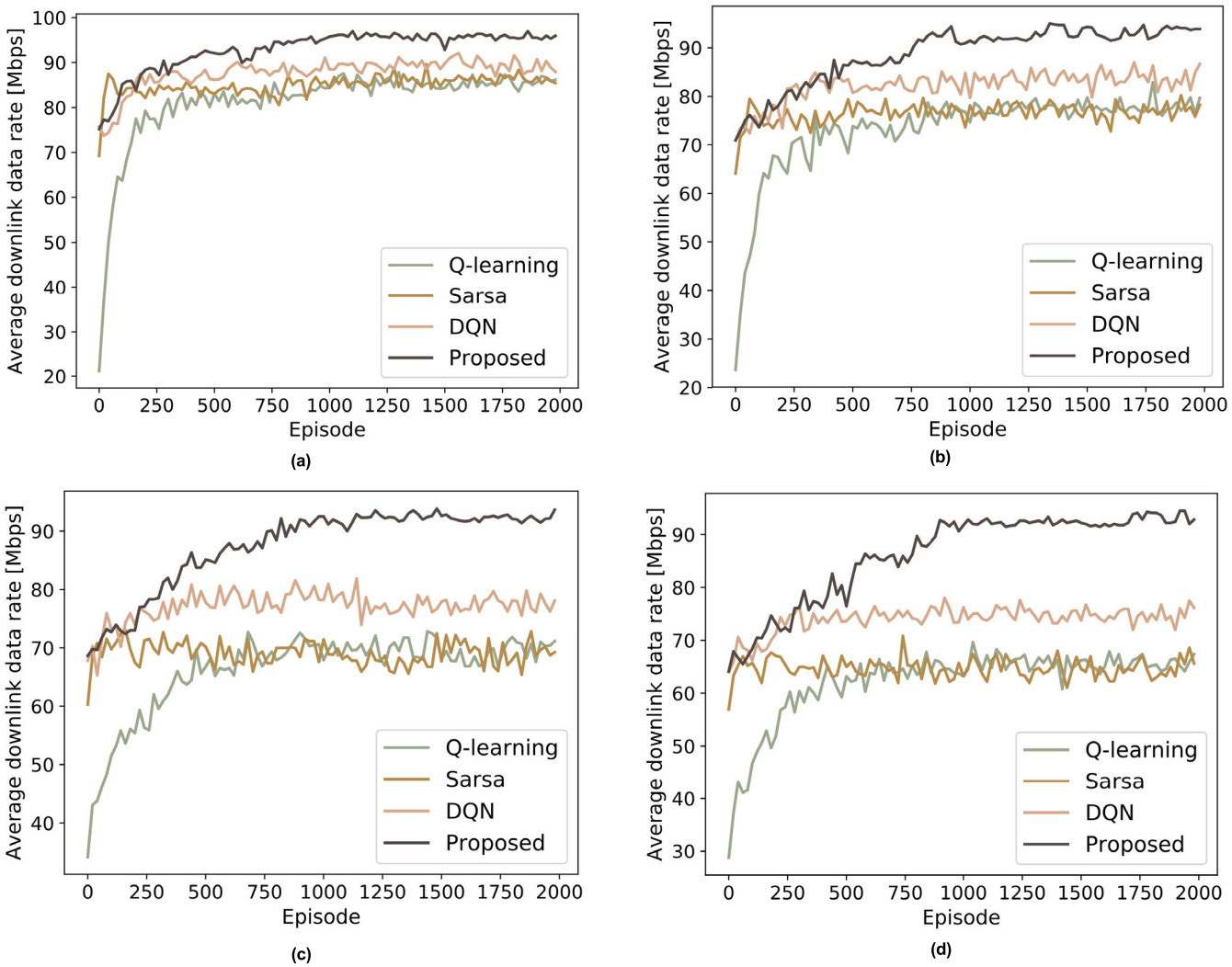

FIGURE 8. Evaluation of the average downlink data rates at four different exploration rates. (a) $\epsilon = 0.1$. (b) $\epsilon = 0.2$. (c) $\epsilon = 0.3$. (d) $\epsilon = 0.4$.

value due to the ϵ-greedy strategy. We observe that with an increase in the UD movement speed, the convergence trends of the classic algorithms become less obvious, and the number of handovers continues to oscillate greatly as the number of training episodes increases. The training effect of the proposed algorithm is also obviously affected by the UD movement speed, but it still exhibits the best performance compared with the other algorithms and eventually reaches the theoretical minimum value.

In addition, to reduce the number of handovers caused by the user's movement in the ultradense network to reduce the impact on system performance, some traditional handover methods that are independent of machine learning, instead relying on handover skipping (HS), were proposed in [22] and [23]. These methods are based on the user's trajectory and can directly switch to the next AP when the user only briefly passes through an adjacent cell while entering the coverage area of another cell. We also compared the performance of HS with that of the other four RL-based algorithms in terms of the number of handovers in the simulation environment considered in this paper, as shown in Fig. 6. In this figure, the line chart refers to the y-axis on the right, which shows the number of handovers with each of the five methods under three UD movement speeds from the simulations. We can see that the four RL-based handover strategies all show better performance than the HS methods in terms of the number of handovers when the UD movement speed is 0.5 m/s, which may be too slow for good compliance with HS. As the UD movement speed increases, however, the number of handovers increases more rapidly with the classic RL algorithms. When the UD movement speed reaches 1.5 m/s, the number of handovers with HS is less than those with Q-learning and Sarsa. However, the performance of DQN is still better than that of HS, and the handover algorithm proposed in this paper continues to show great advantages over HS. The histogram in Fig. 6 refers to the y-axis on the left, which shows the performance improvement of the proposed algorithm compared with the other four methods in terms of the number of

handovers under each of the three UD movement speeds from the simulations. The proposed algorithm outperforms the HS, Q-learning, Sarsa, and DQN algorithms by 66%, 55%, 52% and 42%, respectively, in terms of the number of handovers when the UD movement speed is 0.5 m/s. Regarding the number of handovers, these results show that the proposed method significantly outperforms the other four algorithms.

The reward is the feedback signal obtained after performing a series of actions, and it directly reflects the performance of the action (AP) selection strategy. Moreover, the ϵ-greedy strategy can prevent an algorithm from falling into local minima and enable it to obtain the optimal solution. Therefore, we also compared the four RL-based algorithms in terms of the reward convergence with different exploration rates ϵ ($\epsilon = 0.1, 0.2, 0.3, 0.4$), as shown in Fig. 7.

First, all four algorithms are able to converge to their respective optimal AP selection schemes, as seen from the convergence performance comparison in Fig. 7. Note that the convergence values in subgraphs (b), (c) and (d) are similar, thereby proving that similar numbers of handovers are produced for the set movement path of the UD. Compared with the other three subgraphs, it is obvious that the convergence values in (a) are greatest when $\epsilon = 0.1$, 0.2, 0.3, and 0.4, as the continuous decrease of the experience replay space can reduce the amount of distant or useless experience utilized in the training process and yield an optimal solution that is closest to the global optimum. Moreover, Fig. 7 reveals that for each algorithm, a smaller ϵ leads to better convergence performance, indicating that a relatively high exploration rate leads to more punishment for useless actions with higher probability.

The overhead caused by handovers increases the control signaling burden and compresses the space available for data transmission; therefore, the average downlink data rate should also be considered as a metric for handover strategy evaluation. Accordingly, the average downlink rates of the four algorithms mentioned above with various ϵ values are illustrated in Fig. 8.

It is evident that the proposed algorithm achieves a higher average downlink data rate than the other approaches due to its more efficient experience learning and its ability to avoid vicious feedback loops. As shown in Fig. 8, with increasing ϵ, the convergence performance and optimal value of each algorithm decrease. Therefore, although an increase in ϵ might bring more exploration opportunities to avoid converging to a poor local optimal value, it also has a negative impact on the existing positive learning experience. To accurately estimate the performance of the algorithms in terms of their average downlink data rates, Fig. 8 (a) is used as the control group, in which each algorithm is in its best convergence state. The simulation results indicate that the proposed algorithm outperforms the DQN, Sarsa and Q-learning algorithms in terms of the average downlink data rate by 8%, 13% and 13%, respectively. Note that although an increase in ϵ reduces the convergence speed of the proposed algorithm, its convergence value is less affected than those of the other three algorithms; the proposed algorithm outperforms the traditional RL algorithms by up to 48% in terms of the average downlink data rate when $\epsilon = 0.4$, as shown in Fig. 8 (d), which reflects the superior robustness of the proposed algorithm.

VI. CONCLUSION

In this paper, we proposed an adaptive VLC handover mechanism for 6G networks with hybrid architectures for the purpose of reducing the interruption time incurred by handovers. A DRL method was implemented to resolve the problem of network performance degradation caused by unreasonable handovers, and experimental simulations confirmed that the proposed DRL-based algorithm can optimize the AP selection strategy to maximize the downlink data rate. Specifically, the results showed that the average downlink data rate with the proposed algorithm is improved compared with those achieved using the Q-learning, Sarsa and DQN algorithms by 13%, 13% and 8%, respectively, for an indoor scene with ultradense AP deployment; thus, it produces the largest data rate improvements in a 6G network with high throughput while also improving the QoE for users.

REFERENCES

[1] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020, doi: 10.1109/MCOM.001.1900411.

[2] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019, doi: 10.1109/MVT.2019.2921162.

[3] Y. Almadani, D. Plets, S. Bastiaens, W. Joseph, M. Ijaz, Z. Ghassemlooy, and S. Rajbhandari, "Visible light communications for industrial applications—Challenges and potentials," *Electronics*, vol. 9, no. 12, p. 2157, Dec. 2020, doi: 10.3390/electronics9122157.

[4] M. Z. Chowdhury, M. Shahjalal, M. K. Hasan, and Y. M. Jang, "The role of optical wireless communication technologies in 5G/6G and IoT solutions: Prospects, directions, and challenges," *Appl. Sci.*, vol. 9, no. 20, p. 4367, Oct. 2019, doi: 10.3390/app9204367.

[5] T. Nguyen, Y. M. Jang, and M. Z. Chowdhury, "A pre-scanning-based link switching scheme in visible light communication networks," in *Proc. 5th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Da Nang, Vietnam, Jul. 2013, pp. 366–369, doi: 10.1109/ICUFN.2013.6614843.

[6] R. Liu and C. Zhang, "Dynamic dwell timer for vertical handover in VLC-WLAN heterogeneous networks," in *Proc. 13th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Valencia, Spain, Jun. 2017, pp. 1256–1260, doi: 10.1109/IWCMC.2017.7986465.

[7] S. Liang, Y. Zhang, B. Fan, and H. Tian, "Multi-attribute vertical handover decision-making algorithm in a hybrid VLC-femto system," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1521–1524, Jul. 2017, doi: 10.1109/LCOMM.2017.2654252.

[8] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020, doi: 10.1109/MWC.001.1900488.

[9] F. Wang, Z. Wang, C. Qian, L. Dai, and Z. Yang, "Efficient vertical handover scheme for heterogeneous VLC-RF systems," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 7, no. 12, pp. 1172–1180, Dec. 2015, doi: 10.1364/JOCN.7.001172.

[10] X. Bao, W. Adjardjah, A. A. Okine, W. Zhang, and J. Dai, "A QoE-maximization-based vertical handover scheme for VLC heterogeneous networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 269, Nov. 2018, doi: 10.1186/s13638-018-1284-1.

[11] A. M. Alenezi and K. A. Hamdi, "Reinforcement learning approach for hybrid WiFi-VLC networks," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, Antwerp, Belgium, May 2020, pp. 1–5, doi: 10.1109/VTC2020-Spring48590.2020.9128892.

[12] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33275–33284, 2018, doi: 10.1109/ACCESS.2018.2844882.
[13] S. Shao, G. Liu, A. Khreishah, M. Ayyash, H. Elgala, T. D. C. Little, and M. Rahaim, "Optimizing handover parameters by Q-learning for heterogeneous radio-optical networks," *IEEE Photon. J.*, vol. 12, no. 1, pp. 1–15, Feb. 2020, doi: 10.1109/JPHOT.2019.2953863.
[14] K. Ji, T. Mao, J. Chen, Y. Dong, and Z. Wang, "SVM-based network access type decision in hybrid LiFi and WiFi networks," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Honolulu, HI, USA, Sep. 2019, pp. 1–5, doi: 10.1109/VTCFall.2019.8891341.
[15] M. Najla, P. Mach, and Z. Becvar, "Deep learning for selection between RF and VLC bands in device-to-device communication," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1763–1767, Oct. 2020, doi: 10.1109/LWC.2020.3003786.
[16] Q. Chen, D. Han, M. Zhang, Z. Ghassemlooy, A. C. Boucouvalas, Z. Zhang, T. Li, and X. Jiang, "Design and demonstration of a TDD-based central-coordinated resource-reserved multiple access (CRMA) scheme for bidirectional VLC networking," *IEEE Access*, vol. 9, pp. 7856–7868, 2021, doi: 10.1109/ACCESS.2020.3049004.
[17] J. M. Kahn and J. R. Barry, "Wireless infrared communications," *Proc. IEEE*, vol. 85, no. 2, pp. 265–298, Feb. 1997, doi: 10.1109/5.554222.
[18] X. Ning, R. Winston, and J. O'Gallagher, "Dielectric totally internally reflecting concentrators," *Appl. Opt.*, vol. 26, no. 2, pp. 300–305, Jan. 1987, doi: 10.1364/AO.26.000300.
[19] N. Meng, H. Zhang, and B. Lin, "User-centric mobility management based on virtual cell in ultra-dense networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–6, doi: 10.1109/ICCChina.2016.7636899.
[20] N. Lilith and K. Dogancay, "Dynamic channel allocation for mobile cellular traffic using reduced-state reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Atlanta, GA, USA, vol. 4, Mar. 2004, pp. 2195–2200, doi: 10.1109/WCNC.2004.1311428.
[21] W. Bao and B. Liang, "Stochastic geometric analysis of handoffs in user-centric cooperative wireless networks," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, Apr. 2016, pp. 1–9, doi: 10.1109/INFOCOM.2016.7524592.
[22] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in 5G and beyond: A topology aware skipping approach," *IEEE Access*, vol. 4, pp. 9073–9081, 2016, doi: 10.1109/ACCESS.2016.2642538.
[23] E. Demarchou, C. Psomas, and I. Krikidis, "Mobility management in ultra-dense networks: Handover skipping techniques," *IEEE Access*, vol. 6, pp. 11921–11930, 2018, doi: 10.1109/ACCESS.2018.2810318.

LIQIANG WANG received the B.S. degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2017. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications (BUPT). His research interests include optical wireless communication systems and intelligent optical networks.

DAHAI HAN received the B.S. degree from Jilin University, China, in 2002, and the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2007. He is currently an Associate Professor with the Optical Wireless Communications Group, State Key Laboratory of Information Photonics and Optical Communications, BUPT. His research interests include UV communication and detection and visible light communication.

MIN ZHANG received the Ph.D. degree in optical communications from the Beijing University of Posts and Telecommunications (BUPT), China. He is currently a Professor, the Deputy Director of the State Key Laboratory of Information Photonics and Optical Communications, and the Deputy Dean of the School of Electronic Engineering, BUPT. He holds 45 Chinese patents. He has authored or coauthored more than 300 technical articles in international journals and conferences and 12 books in the area of optical communications. His current research interests include optical communication systems and networks, optical signal processing, and optical wireless communications.

DANSHI WANG (Member, IEEE) received the Ph.D. degree in electromagnetic field and microwave technology from the Beijing University of Posts and Telecommunications (BUPT), in 2016. He is currently an Associate Professor with the State Key Laboratory of Information Photonics and Optical Communications (IPOC), BUPT. He has authored or coauthored over 100 technical articles in international journals and conferences. His research interests include intelligent optical communication and networks, optical signal processing, artificial intelligence, deep learning, digital twins, and data-driven modeling.

ZHIGUO ZHANG received the Ph.D. degree in electronics science and technology from the Beijing University of Posts and Telecommunications (BUPT), China, in 2007. He is currently a Professor with the State Key Laboratory of Information Photonics and Optical Communications (IPOC), BUPT. He holds 60 Chinese patents. He has authored or coauthored more than 130 technical articles in international journals and conferences. His research interests include advanced optical communication systems and networks, convergent broadband multiservice access communication systems, optical signal processing, and fiber sensor technologies.