IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Manuscript received 16 October 2022; revised 30 March 2023, 30 July 2023, and 22 September 2023; accepted 15 October 2023. Date of publication 30 October 2023; date of current version 18 December 2023.

The associate editor coordinating the review of this article and approving it for publication was Y. Deng.

*Digital Object Identifier 10.1109/TMLCN.2023.3328501*

# Deep Q-Learning-Based Resource Management in IRS-Assisted VLC Systems

**AHMED AL HAMMADI [1] (Member, IEEE), LINA BARIAH [1] (Senior Member, IEEE), SAMI MUHAIDAT [2,3] (Senior Member, IEEE), MAHMOUD AL-QUTAYRI [4] (Senior Member, IEEE), PASCHALIS C. SOFOTASIOS [2,5] (Senior Member, IEEE), AND MEROUANE DEBBAH [4,6] (Fellow, IEEE)**

[1]Technology Innovation Institute, Abu Dhabi, United Arab Emirates
[2]Center for Cyber-Physical Systems, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates
[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada
[4]Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates
[5]Department of Electrical Engineering, Tampere University, 33014 Tampere, Finland
[6]KU 6G Center, Khalifa University, Abu Dhabi, United Arab Emirates

CORRESPONDING AUTHOR: P. C. SOFOTASIOS (p.sofotasios@ieee.org)

**ABSTRACT** Visible Light Communication (VLC) is a promising enabling technology for the next-generation wireless networks, as it complements radio-frequency (RF)-based communications by providing wider bandwidth, higher data rates, and immunity to interference from electromagnetic sources. However, due to its unique characteristics, VLC is highly sensitive to the line-of-sight (LoS) blockage. Recently, intelligent reflecting surface (IRS) has been proposed as an innovative solution that dynamically reconfigures the wireless environment. The present contribution proposes a two-stage resource management framework in an indoor VLC system: In the first stage, a maximum possible fairness (MPF) algorithm is presented in order to maximize the fairness amongst the users. In the second stage, deep Q-learning is exploited in order to maximize the overall spectral efficiency (SE). The corresponding numerical results have shown that the proposed DQL-MPF framework exhibits superior performance in terms of both the overall SE and Jain's Fair Index, achieved at a fast convergence rate. More specifically, when the noise power is high and the number of users is relatively large, the DQL-MPF algorithm achieves a more than tenfold overall SE compared to the Baseline scheme. Moreover, the synergy between the MPF and the DQL algorithms is investigated. To this end, we demonstrate that the MPF algorithm maximizes the fairness amongst the users while the DQL algorithm maximizes the overall SE and improves the robustness against the noise. Our results also highlight the effectiveness of the proposed algorithm in leveraging the increasing number of IRS elements for optimized performance.

**INDEX TERMS** Visible light communications (VLC), intelligent reflecting surfaces (IRS), deep Q-learning (DQL), spectral efficiency.

## I. INTRODUCTION

THE ever-growing demand for high data rate wireless services and the exponential growth of the number of connected devices necessitate the development of new innovative solutions that complement radio frequency (RF) communications. It is well known that RF-based communication has been lately facing several challenges, such as spectrum scarcity and high energy consumption, which results in a significant carbon footprint [1]. Visible light communication (VLC), on the other hand, is a promising technology that has been recognized as an enabler for future networks [2], since in an abundant, open, and unlicensed spectrum, VLC may be employed for both lighting and high-speed data communication.

Yet, VLC is sensitive to line-of-sight (LoS) blockage, i.e. the link between the access point (AP) and the photodetector (PD) is required for reliable communications. However, LoS blockage is particularly common in indoor VLC systems, which might have a negative impact on the corresponding achievable performance. This is due to a fairly common assumption that users' devices point upward towards the ceiling [3], [4], [5], [6]. This assumption is unrealistic because devices are typically subject to random orientation which affects the quality of LoS links, as demonstrated in [7]. Therefore, while designing and analyzing VLC systems, random receiver orientation must be considered.

Intelligent reflecting surface (IRS) is a new technology that holds great potential in improving the performance of wireless communication systems. In VLC systems, where communication performance is heavily reliant on the availability of LoS pathways, a blocked LoS path can be alleviated by modifying the wireless propagation channel. The mirror array-based IRS [8] and the metasurface-based IRS [9] are the most used hardware designs for IRS in VLC systems. The former is based on geometric optics such as Snell's law of reflection, and each unit may spin along two independent and orthogonal axes, much like micro-electro-mechanical systems (MEMS). However, the metasurface-based IRS is composed of arrangements of sub-wavelength metallic or dielectric structures that are used to manipulate the wavelength, the polarization, and the phase of incident light waves. As a result, IRS can be used to mitigate the blockage dilemma in VLC systems. In addition to these two types of hardware designs, a third type has been proposed - the liquid crystal (LC)-based re-configurable intelligent surface (LC-based RIS) VLC, which offers the advantage of tunability and light amplification. According to [10] LC-based RIS can improve the VLC signal detection and transmission range. Recently, Amr et al. studied the temporal characteristics of VLC channels using IRS and radiometric concepts. The study accounts for power delays and shows the impact of system parameters on the temporal characterization of the two IRS-based VLC systems. In another environment, the authors in [11] explores the use of IRS in underwater wireless communication (OWC) systems. More specifically, they derive a closed-form expression for the outage probability under underwater attenuation, pointing error, and turbulence effects. The authors in [12] presented a novel approach of using mirrors to improve the illumination uniformity and throughput of an indoor multi-element VLC system architecture through optimizing the problem of LED-user association using a heuristic technique.

### A. RELATED WORKS

Recently, Q-learning has sparked a rapidly growing interest among researchers and engineers in various fields. Q-learning is a subset of reinforcement learning in which Q tables store optimal sequences of actions that maximizes the future reward. Several studies have used Q-learning to improve the performance of wireless networks [13], [14], [15]. More specifically, the deep reinforcement learning (DRL) algorithm was introduced in several IRS-assisted wireless communication studies to solve difficult non-convex optimization problems and to improve the performance in such networks. Huang et al. [16] proposed a DRL technique for relay selection in IRS-assisted cooperative networks to maximize throughput. More specifically, they proposed a joint relay selection and IRS reflection coefficient optimization for cooperative networks. In [17], a multi-agent deep reinforcement learning-based scheme is investigated, where the DRL is employed to realize a buffer-aided relay selection scheme for an IRS-assisted secure cooperative network. The authors in [18] proposed a dueling double deep Q-network (D3DN) to optimize the performance of the IRS in a multi-robot network, motivated by the benefits of non-orthogonal multiple access and IRS. In this context, they proposed a framework in which the IRS is deployed at an AP and NOMA is used at the AP to serve multiple robots. In a mobile IRS scenario, the authors in [19] proposed a model in which IRSs are mounted on intelligent robots for flexible deployment. A deep deterministic policy gradient (DDPG) framework is used to optimize power allocation and the phase shift. Finally, the works in [20] and [21] investigated the optimization of the total achievable rate of multi-hop multi-user IRS-assisted wireless terahertz (THz) communication systems. To increase the network's capacity, they proposed a DRL algorithm to learn the optimal beamforming.

Several recent studies have investigated the optimization of IRS-assisted VLC systems. The authors in [22] proposed sine-cosine (SC) optimization algorithm to maximize the SE of an indoor VLC system with an IRS mirror array, although only a single-user scenario was considered. In a related study, the authors in [23] explored the use of IRS to improve link reliability in VLC systems using non-orthogonal multiple access (NOMA). They proposed a joint optimization framework based on the genetic algorithm (GA). In addition, a DRL-based framework was proposed in [24] for an IRS-assisted VLC to maximize secrecy capacity. Finally, Sun et al. [25] proposed the frozen variable algorithm and the minorization-maximization algorithm to iteratively maximize the overall SE in an IRS-assisted VLC system.

### B. MOTIVATION AND CONTRIBUTIONS

The integration of IRS technology with VLC systems has gained significant attention in recent years, highlighting the need for an effective optimization framework to maximize their combined potential. This framework should take into account practical scenarios involving multiple users and random device orientation. A key aspect to consider for enhancing system performance is the joint optimization of the LED-user association, power allocation, and IRS mirror orientation. Several notable works have demonstrated the effectiveness of the joint optimization approach

Despite the growing interest in this area, none of the existing studies have specifically focused on optimizing the overall spectral efficiency (SE) of multi-user IRS-assisted

VLC using DQL, while taking into account random device orientation. Compared to DQL, heuristic techniques may not be able to adapt to changing environments or may not consider all possible solutions, leading to suboptimal results. In contrast, a DQL-based optimization framework can dynamically adapt its strategy based on the current environment, which can result in better performance and faster convergence to optimal solutions. Additionally, DQL can handle more complex problems and scenarios where heuristics may not be applicable or may be difficult to design. In this work, our key contributions can be summarized as follows:

In this work, our key contributions can be summarized as follows:

- We present an optimization framework for an IRS-assisted VLC system, which jointly optimizes the IRS mirror orientations, LED-user associations, and LED power allocations. This framework incorporates Jain's fairness index [26] to guarantee equitable resource allocation among users.
- We propose a two-stage approach consisting of the DQL and MPF algorithms, which collaboratively maximize the overall spectral efficiency (SE) of the multi-user IRS-assisted VLC system while maintaining fairness among users, as indicated by Jain's fairness index.
- Our simulation results highlight the superiority of the proposed DQL-MPF algorithm in comparison to several benchmark methods, such as the genetic algorithm (GA) [27]. The synergistic effect between the MPF and DQL algorithms is apparent, as the joint DQL-MPF strategy outshines both the individual DQL and MPF methods in terms of spectral efficiency and fairness. Moreover, our results show that the enhanced performance persists and maintains its superiority even with an increasing number of IRS elements, demonstrating the scalability and effectiveness of the proposed approach.
- By introducing the novel MPF algorithm in tandem with the DQL algorithm, we substantially decrease the computational complexity and establish a more efficient and equitable resource allocation scheme for IRS-assisted VLC systems.

### C. NOTATION

Vectors and matrices are denoted by lower and upper case boldface symbols, respectively. Therefore, $\mathbf{a}_i$ denotes a vector, in which $i$ represents its $i^{th}$ element. Also, with $A_{i,j}$, we denote a matrix with $i$ as an index for the column, and $j$ as an index for the row. $\mathbb{E}(\cdot)$ denotes the statistical expectation, $\mathcal{U}(.)$ denotes the uniform distribution function, $[\mathbf{A}]^T$ is the transpose operation for the matrix $\mathbf{A}$, and $len(\mathbf{A})$ denotes a function that returns the number of elements in the matrix $\mathbf{A}$.

## II. SYSTEM AND CHANNEL MODELS

We consider the downlink of an IRS-assisted time-division-multiple-access (TDMA) VLC system, where $L$ LEDs serve $K$ users and an IRS with $N$ units. It is assumed that each LED serves a single user in a single time slot, resulting in multi-user interferences (MUI) between different LEDs. Without loss of generality, it is assumed that the VLC channel state information (CSI) is known at the system controller, which can be achieved using various channel estimation methods [28].

### A. CHANNEL GAIN OF LoS

Suppose that the $k^{th}$ user is served by the $l^{th}$ LED, the associated LoS direct gain in VLC generally follows the Lambertian model, which is given by [25]

$$h_{k,l}^{(1)} = \begin{cases} \dfrac{(m+1)A_{\mathrm{PD}}}{2\pi d_{k,l}^2}\cos^m(\Phi)T(\xi)G(\xi)\cos(\xi), & 0 \le \xi \le \xi_{\mathrm{FoV}} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $m = -1/\log_2(\cos(\Theta_{1/2}))$ denotes the Lambertian index, and $\Theta_{1/2}$ is the semi-angle at half illuminance of the LED.The physical area of the photodetector (PD) is denoted by $A_{\mathrm{PD}}$, the distance between the $k^{th}$ user and the $l^{th}$ LED is denoted by $d_{k,l}$, whereas $\Phi$ and $\zeta$ are the angles of irradiance and incidence, respectively. The optical filter gain is denoted by $T(\xi)$, and $G(\xi)$ is the optical concentrator gain with respect to field-of-view (FoV), which is given in [29].

The orientation of the user's device is not affected by $\Phi$. On the contrary, $\xi$ is heavily influenced by the device's orientation. The cosine of $\xi$ can be calculated using the device's polar angle, $\alpha$, and azimuth angle, $\beta$, as follows [22]:

$$\cos(\xi) = \left(\frac{x_l - x_k}{d_{k,l}}\right)\sin(\alpha_k)\cos(\beta_k) + \left(\frac{y_l - y_k}{d_{k,l}}\right)\sin(\alpha_k) \times \sin(\beta_k) + \left(\frac{z_l - z_k}{d_{k,l}}\right)\cos(\alpha_k), \tag{2}$$

where $(x_l, y_l, z_l)$ and $(x_k, y_k, z_k)$ are position vectors describing the LEDs and user's locations, respectively. Based on the modeling study in [30], the polar angle can be modeled using the Laplace distribution with a mean and standard deviation of 41° and 9°, respectively. Moreoever, its value is typically restricted to the range [0,($\pi$/2)]. Finally, the yaw angle follows a uniform distribution $\beta \sim \mathcal{U}[-\pi, \pi]$.

In the considered scenario, we assume multiple LEDs and multiple users. Based on this, the channel gain matrix for the LOS components is given by

$$H^{(1)} = \left[h_1^{(1)}, h_2^{(1)}, \ldots, h_K^{(1)}\right], \tag{3}$$

where each column denotes the direct gain vector $k$ between LEDs and the $k^{th}$ user, and is given by

$$h_k^{(1)} = \left[h_{k,1}^{(1)}, h_{k,2}^{(1)}, \ldots, h_{k,L}^{(1)}\right]. \tag{4}$$

### B. CHANNEL GAIN OF NLoS

In general, the NLoS paths in the considered VLC system include reflection, diffraction, scattering, and penetration

paths. The penetration paths are frequently ignored in VLC due to the extremely high visible light penetration loss. The diffraction path is also negligible due to the nanoscale of wavelength. Based on the surface properties of the reflector, light reflection can be divided into two types: diffusely reflected link and specularly reflected link. It has been shown in [8] that the specular reflection is generally considered as the significant NLoS component in IRS-assisted VLC, while the diffuse reflection is ignored. Accordingly, the NLoS channel gain can be rewritten with a focus on the specular reflection component. In our study, we are only considering the first reflections off the mirrors, as this is the most significant contribution to the NLoS channel gain in IRS-assisted VLC systems. Accordingly, the NLoS channel gain can be rewritten as [25]

$$h_{k,n,l}^{(2)} = \begin{cases} \dfrac{\rho_k\,(m+1)}{A_{\mathrm{PD}}} 2\pi\,\left(d_{n,l}+d_{k,n}\right)^2 \times \mathcal{F} \times \cos\left(\xi_k^n\right) \\ \qquad 0 \le \xi_k^n \le \xi_{\mathrm{FoV}} \\ 0, \qquad \text{otherwise}, \end{cases} \tag{5}$$

where $\rho_k$ is the reflection coefficient of the IRS element, $d_{n,l}$ is the distance between the $l^{th}$ LED and the $n^{th}$ reflective surface, $d_{k,n}$ is the distance between the $n^{th}$ reflective surface and the $k^{th}$ user. Based on this, the function $\mathcal{F}$ is given by

$$\mathcal{F} = \cos^m\left(\Phi_n^l\right)\cos\left(\xi_n^l\right)\cos\left(\Phi_k^n\right)\,T\left(\xi\right)G\left(\xi\right), \tag{6}$$

where $\Phi_n^l$ is the angle of irradiance from the $l^{th}$ LED to $n^{th}$ reflective surface, $\xi_n^l$ is the angle of incidence on the $n^{th}$ reflective surface, $\cos\left(\xi_k^n\right)$ accounts for the random device orientation from the $n^{th}$ reflective surface to $k^{th}$ user, and finally, $\Phi_k^n$ is the angle of irradiance from the $n^{th}$ reflective surface to $k^{th}$ user, and is defined by [22]:

$$\cos\left(\Phi_k^n\right) = \frac{(x_k - x_n)}{d_{k,n}}\sin\left(\varphi_n\right)\cos\left(\omega_n\right) + \frac{(y_k - y_n)}{d_{k,n}} \times \cos\left(\varphi_n\right)\cos\left(\omega_n\right) + \frac{(z_k - z_n)}{d_{k,n}}\sin\left(\omega_n\right), \tag{7}$$

where ($x_n$,$y_n$,$z_n$) represent the coordinates of the IRS.

Next, we define a three-dimensional matrix $H^{(2)}$ to denote the NLoS channel gain, which consists of slices as follows:

$$H_k^{(2)} = \mathrm{diag}\left(h_1^{(2)T}, h_2^{(2)T}, \ldots, h_K^{(2)T}\right), \tag{8}$$

in which each column denotes the NLoS gain vector between the $l^{th}$ LED, the $n^{th}$ IRS element, and the $k^{th}$ user, and is given by

$$h_{k,l}^{(2)} = \left[h_{k,1,l}^{(2)}, h_{k,2,l}^{(2)}, \ldots, h_{k,N,l}^{(2)}\right]. \tag{9}$$

### C. INSTANTANEOUS RECEIVED SIGNAL OF IRS-ASSISTED VLC

In this subsection, we calculate the instantaneous received signal in one time slot. To this end, a user association matrix **F**, a power allocation matrix **P** are defined to describe the behavior of transmitters, and finally, the matrices $\Xi$ and $\Omega$ denote the yaw angles of the IRS mirrors, and the roll angles of IRS mirrors, respectively. These four matrices are variables to be jointly optimized at a later stage. Skipping specific enumeration of each since user association and power control are well established.

#### 1) USER ASSOCIATION

The relationship between transceivers is described using the vectors the vectors $y = [y_1, y_2, \ldots, y_K]^T$ transmitting information and $x = [x_1, x_2, \ldots, x_L]^T$ received, as

$$\mathbf{x} = \mathbf{F}\mathbf{y} \tag{10}$$

In $f_k \in \mathbb{R}_+^{L+1}$ if $f_{l,k} = 1$ the $k^{th}$ user accepts the $l^{th}$ LED's service.

#### 2) EMISSION POWER

A diagonal power matrix $\mathbf{P} = \mathrm{diag}\,(P_1, P_2, \ldots, P_L)$ indicates the emission power on LEDs, so that the transmit signal can be expressed as $\tilde{x} = \mathbf{P}\mathbf{x}$, whereas the received signal,$\hat{y}_k^{(1)}$ for the $k^{th}$ user can be expressed as

$$\hat{y}_k^{(1)} = \rho_k \mathbf{h}_k^{(1)T}\tilde{x}\rho_k \mathbf{h}_k^{(1)T}\mathbf{P}\mathbf{f}_k y_k + \rho_k \sum_{i=1, i\neq k}^{K} \mathbf{h}_k^{(1)T}\mathbf{P}\mathbf{f}_i y_i, \tag{11}$$

where the two components denote the useful signal and the MUIs, respectively.

#### 3) IRS CONFIGURATION

Thirdly, an IRS configuration which consists of two matrices, the roll angle matrix, which is defined as $\Omega = [\omega_1, \omega_2, \ldots, \omega_N]$, where $\omega_n$ is the roll angle for the $n^{th}$ unit. The second matrix in the IRS configuation is the yaw matrix, which is given by $\Xi = [\varphi_1, \varphi_2, \ldots, \varphi_N]$ where $\varphi_n$ is the yaw angle for the $n^{th}$ unit.

Considering the channel model in indoor VLC, which is determined mainly by the locations of the transceivers, any movement of the transmitter/receiver/reflector may significantly change the channel gain, resulting in the high spatial resolution of the VLC channel. The specular reflection path can be considered as an extended path emitted from the imaging LED. Therefore, a single IRS unit cannot serve two or more PDs simultaneously. Hence, the NLoS received signal of the $k^{th}$ user can be expressed as

$$\hat{y}_{k,l}^{(2)} = \rho_k \sum_{n=1}^{N} h_{\mathrm{k,n,l}}^{(2)}\left(\omega_n, \varphi_n\right) P_l x_l f_{l,k} \tag{12}$$

$$= \rho_k \mathbf{h}_{\mathrm{k,l}}^{(2)}\left(\Omega,\,\Xi\right) P_l x_l f_{l,k} \tag{13}$$

where $\mathbf{h}_{\mathrm{k,l}}^{(2)}\left(\Omega,\,\Xi\right)$ is the NLoS channel gain for a given IRS configuration. It follows that the overall NLoS signal is obtained using

$$\hat{y}_k^{(2)} = \sum_{l=1}^{L} \hat{y}_{k,l}^{(2)} \tag{14}$$

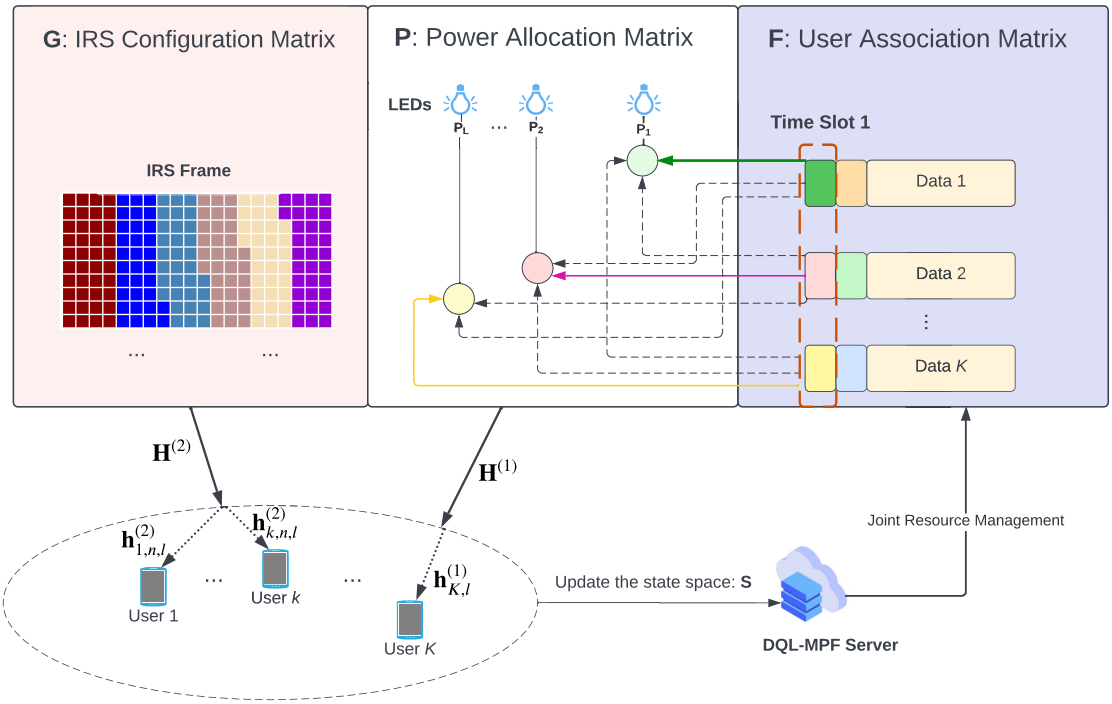


**FIGURE 1.** The IRS-assisted VLC system model with DQL-MPF server.

$$= \rho_k \left[ \mathbf{h}_{\mathrm{k,1}}^{(2)T} (\Omega, \Xi) P_1 x_1, \ldots, \mathbf{h}_{\mathrm{k,L}}^{(2)T} (\Omega, \Xi) P_L x_L \right] \mathbf{f}_k \tag{15}$$

$$= \rho_k \left[ \mathbf{h}_{\mathrm{k,1}}^{(2)T} (\Omega, \Xi), \ldots, \mathbf{h}_{\mathrm{k,L}}^{(2)T} (\Omega, \Xi) \right] \mathbf{P} \mathrm{diag}\,(x)\, \mathbf{f}_k \tag{16}$$

$$= \rho_k \left[ \mathbf{H}_{\mathrm{k}}^{(2)} (\Omega, \Xi) \right]^T \mathbf{P} \mathrm{diag}\,(\mathbf{F}\mathbf{y})\, \mathbf{f}_k. \tag{17}$$

Notably, the equation $\mathrm{diag}\,(\mathbf{F}\mathbf{y})\,\mathbf{f}_k = y_k \mathbf{f}_k$ holds due to the orthogonality among $f_k$. Finally, the above formula can be further rewritten by replacing the last two multiplicators as

$$\hat{y}_k^{(2)} = \rho_k \left[ \mathbf{H}_{\mathrm{k}}^{(2)} (\Omega, \Xi) \right]^T \mathbf{P} \mathbf{f}_k y_k. \tag{18}$$

To sum up, the received signal of the $k^{th}$ user is comprised of the LoS component $\hat{y}_k^{(1)}$, the NLoS component $\hat{y}_k^{(2)}$, and $w_k$ denotes the additive white Gaussian noise at the receiver. Therefore, the resultant received signal $\hat{y}_k$ can be expressed as [25]

$$\hat{y}_k = \hat{y}_k^{(1)} + \hat{y}_k^{(2)} + w_n \tag{19}$$

The proposed system model is shown in Fig. 1. The TDMA-based IRS-assisted system is optimized in real-time through the deep Q-learning with maximum possible fairness (DQL-MPF) algorithm server. More specifically, the users upload the state of the system to the DQL-MPF server, which then applies the joint resource management for the configuration of the IRS mirrors, the power allocation of the LEDs, and the association between the LEDs and the users at a given time slot. The main goal is to maximize the overall SE. In the following sections, the joint optimization problem is formulated, and the proposed optimization framework is proposed.

## III. PROBLEM FORMULATION

Due to several unique constraints, such as the nonnegative and real-valued signal, illumination requirements, and sensitivity to geometric locations, the classic Shannon capacity formula cannot be used to exactly describe the VLC channel capacity [31]. Accordingly, the authors in [32] proposed a tight lower bound for dimmable VLC systems, where the capacity formula is in continuous form as

$$C = \frac{1}{2} W \log_2 \left( 1 + \frac{e \tau^2 P^2}{2\pi \sigma^2} \right) \tag{20}$$

where $W$, $\tau$, $P$, and $\sigma^2$ denote the modulation bandwidth, the responsivity of the PD, the optical power, and the variance of the Gaussian noise, respectively. This formula primarily serves as an introduction to the modified Shannon capacity for VLC systems for a single user. Subsequently, the SE of the $k^{th}$ user is given by for a multi-user scenario considering the additional complexity introduced with this setting is

$$R_k = \frac{1}{2} \log_2 \left( 1 + \frac{e}{2\pi} \delta_k \right) \tag{21}$$

where $\delta_k$ indicates the individual signal-to-interference-plus noise ratio (SINR), which can be expressed as

$$\delta_k = \frac{\rho_k^2 \left[ \left\{ \mathbf{H}_{\mathrm{k}}^{(2)} (\Omega, \Xi) + \mathbf{h}_k^{(1)} \right\}^T \mathbf{P} \mathbf{f}_k \right]^2}{\sigma_k^2 + \tau_k^2 \sum_{i=1, i \neq k}^{K} \left\{ \mathbf{h}_k^{(1)T} \mathbf{P} \mathbf{f}_i \right\}^2 \mathrm{var}\,(y_i)} \tag{22}$$

This term takes into account the idea of interference that is absent in the previous formula. Here in (22), $\text{var}(y_i) = 1$ denotes the variance of the interference signal $y_i$, and $\sigma^2 \in \mathbb{R}^+$ denotes the variance of $w_k$.

After analyzing the individual SINR for each user, the overall SE is then formulated as

$$R = \sum_{k=1}^{K} R_k \left(\Omega, \Xi, \mathbf{P}, \mathbf{F}\right) \tag{23}$$

This optimization framework mainly employs the overall spectral efficiency in its computations to optimize the entire system's performance rather than each user's performance. The goal is to utilize the spectrum as efficiently as possible as the bandwidth is normalized and not a decision variable in this context. Prior to the problem formulation, we introduce Jain's Fairness Index [26], which is a widely used metric for quantifying the fairness of resource allocation among users in communication systems. The index is defined as follows:

$$J = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} R_k^2} \tag{24}$$

where the value of Jain's Fairness Index ranges between 0 and 1, with higher values indicating better fairness among the users. An index value of 1 indicates perfect fairness, where all users have the same data rate, while a value of 0 indicates the worst possible fairness, with one user obtaining all the resources. Incorporating Jain's Fairness Index into the optimization problem (P1) will be beneficial as it enables our proposed algorithm to achieve a balance between maximizing the overall system spectral efficiency and ensuring fair resource allocation among the users. By jointly optimizing the IRS yaw matrix $\Xi$, the IRS roll matrix $\Omega$, the power allocation matrix $\mathbf{P}$, and the user association matrix $\mathbf{F}$, while considering the fairness criterion, the algorithm will not only increase the sum-rate but also allocate resources more equitably among users. This approach will lead to a more efficient and fair system, improving the overall user experience in the visible light communication network, namely

$$\max_{\Omega, \Xi, \mathbf{P}, \mathbf{F}} \; R \cdot J \tag{P1}$$

$$s.t.\; R_k \geq R_{k,min} \quad \forall k \in (1, .., K)\,, \tag{P1.a}$$

$$\sum_{l=1}^{L} P_l \leq P_{max}, \tag{P1.b}$$

$$P_{min} \leq P_l \leq P_{max} \quad \forall l \in (1, .., L)\,, \tag{P1.c}$$

$$f_{l,k} \in \{0, 1\}\,, \quad \forall l \in \{1, \ldots, L\}\,, \quad \forall k \in \{1, \ldots, K\} \tag{P1.d}$$

$$\sum_{k=1}^{K} f_{l,k} \in \{0, 1\}\,, \quad \forall l \in \{1, \ldots, L\} \tag{P1.e}$$

$$\frac{-\pi}{2} \leq \varphi_n \leq \frac{\pi}{2} \quad \forall n \in \{1, \ldots, N\} \tag{P1.f}$$

$$\frac{-\pi}{2} \leq \omega_n \leq \frac{\pi}{2} \quad \forall n \in \{1, \ldots, N\}\,, \tag{P1.g}$$

where the minimum data rate requirement is denoted by the constraint (P1.a). The total power limitation imposed by the VLC system is denoted by the constraint in (P1.b), whereas (P1.c) denotes the individual illumination constraint. Next, constraints (P1.d) and (P1.e) indicate that a single transmitter cannot carry information symbols for multiple users at the same time. The constraints in (P1.f) and (P1.g) result from the IRS yaw matrix and IRS roll matrix definitions, respectively. Finally, it is assumed that the channel state information (CSI), denoted by matrices $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$, are known by using VLC channel estimation techniques. Note that the IRS configuration subproblem is proven to be typically non-deterministic polynomial time (NP)-hard [33]. Therefore, the complexity of pursuing the optimal solution of (P1) suffers from exponential explosion. Accordingly, we solve the above optimization problem by a two-stage DQL-based algorithm. In the following, we present a two-stage algorithm to address the optimization problem (P1). The first stage introduces the Maximum Possible Fairness (MPF) algorithm, while the second stage involves designing and implementing a deep Q-learning framework.

### A. STAGE 1: MAXIMUM POSSIBLE FAIRNESS ALGORITHM

At this stage, the IRS frame is divided into $D$ segments, with each segment optimized and dedicated to a single user. The MPF algorithm can maximize fairness in situations where the number of IRS mirrors is not evenly divisible by the number of users, and also maintain fairness when divisibility occurs. By optimally allocating and dividing mirrors to each user in the IRS-assisted network, the MPF algorithm seeks to maximize fairness among users. The operational steps of the MPF algorithm are outlined below, and a summary is provided in **Algorithm 1**.

1) Initialize the IRS-assisted VLC parameters with $N$ total mirrors, $K$ total users, $\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$, and an index $n$ for the IRS segments. Additionally, use the variable *dec* for decrementing the number of segments, *rem* for checking if there is a remainder, and *stop* for terminating the algorithm.
2) Continue running the algorithm until the remainder is 0 and each user is allocated an IRS segment.
3) Initially, check if the number of mirrors $N$ is divisible by the number of users $K$.
4) If not divisible, decrement 1 from $K$ until $N$
5) Divide the IRS frame into $K - dec$ segments.
6) Sort the segments in ascending order based on the CSI.
7) Detach the major segment from the IRS array and assign it the lowest channel gain. This segment will have a higher number of mirrors; hence, assigning it the lowest channel gain will improve the degraded channel.
8) Check if the remaining number of IRS $\mathbf{S}_{rem}$ mirrors is divisible by $K - dec$.
9) If divisible, split the remaining segments into $K - dec$ segments, concatenate them with all the major segments in $\mathbf{S}_J$, and set *stop* to 1.

**Algorithm 1** MPF Algorithm

**Input:** IRS elements array with $N$ mirrors, $K$ number of users, $\mathbf{H}^{(1)}$,$\mathbf{H}^{(2)}$
**Output:** $\mathbf{S}_J$, which is a matrix that includes $D$ IRS segments for $K$ users.

**Initialization:** $n = 0$, $dec = 0$, $rem = 1$, $stop = 0$.
**while** *rem! = 0* **do**
  Set remainder to $N$ **if** *rem ! = 0* **then**
    Set *dec* TO $dec + 1$
  Set $D$ to ($K$ - *dec*) Split IRS array into $D$ segments Sort $D$ segments based on channel gains in ascending order Detach the major segment as $\mathbf{S}_n$
  Assign $\mathbf{S}_n$ with the lowest channel gain
  Set the remaining segments to $\mathbf{S}_{rem}$
  **if** *len*($\mathbf{S}_{rem}$) % ($K$ - *dec*) == *0* **then**
    Split $\mathbf{S}_{rem}$ into ($K$ - *dec*)
    Concatenate ($\mathbf{S}_1$,$\mathbf{S}_2$,...,$\mathbf{S}_n$) to $\mathbf{S}_N$
    Concatenate ($\mathbf{S}_N$, $\mathbf{S}_{rem}$)
    Set ($\mathbf{S}_N$, $\mathbf{S}_{rem}$) to $\mathbf{S}_D$
    Set *stop* to 1
    Return $\mathbf{S}_J$
  **else**
    Update $N$ to [$N$ - len($\mathbf{S}_n$) ]
    Update $n$ to [$n + 1$]
    Update $K$ to $K - 1$
    Continue
  End

10) Otherwise, update the number of mirrors by subtracting the number of mirrors in the major segment $\mathbf{S}_n$, and start over.

We provide a walk-through example for **Algorithm 1** in Appendix A. It is important to note that in $\mathbf{S_D}$, the segments with a higher number of mirrors will be assigned the lowest channel gains from the estimated $\mathbf{H}^{(2)}$, thereby enhancing the overall performance of the IRS-VLC system.

While the MPF algorithm allows for optimizing the distribution of mirrors in a multi-user TDMA-based IRS-VLC system, other parameters still need optimization. Solely relying on the MPF algorithm will not yield optimal performance. In the second stage, we propose an optimization framework that leverages the output of the MPF algorithm and employs a deep Q-learning algorithm to jointly optimize the power allocation of the LEDs, the association between users and LEDs, and the roll angle $\Omega$ and yaw angle $\Xi$ of the IRS frame. Instead of optimizing each mirror's individual orientation, our approach optimizes the segments resulting from the MPF algorithm during stage 1.

### B. STAGE 2: DEEP Q-LEARNING BASED OPTIMIZATION

Before we dive into the specifics of our algorithm that utilizes Deep Reinforcement Learning (DRL) technique, let's first briefly explain the core principles of DRL. Reinforcement Learning (RL) is a specific branch of machine learning that trains an agent to make the best possible decisions in a given environment by maximizing its rewards. Based on its actions, the agent may receive either positive or negative rewards. Typically, RL problems are represented as Markov Decision Processes (MDPs), which illustrate the process of sequential decision-making. On the other hand, DRL uses a Deep Neural Network (DNN) to estimate the value function or the policy in RL.

The key components of RL can be summarized as follows:

1) **State space** ($\mathcal{S}$) - The current state of the environment at time $t$. It is a representation of the environment that captures the relevant information needed to make decisions.
2) **Action space** ($\mathcal{A}$) - The decision made by the agent at time $t$ based on the current state. Actions can change the state of the environment.
3) **Reward** ($u_{s,s',a}$) - A scalar value that reflects the desirability of the current state-action pair. It is a feedback signal that guides the agent towards maximizing long-term cumulative rewards.
4) **Policy** ($\pi$) - A mapping from states to actions that defines the behavior of the agent. The policy can be deterministic or stochastic, depending on the environment.
5) **State-action value** ($Q_\pi(s, a)$) - The future cumulative reward of taking action $a_t$ at a given state $s_t$ and following the policy thereafter.

At time step $t$, the agent starts with an action $a_t$, by which the environment transitions from state $s_t$ to the next state $s'$. As a result, the agent receives an immediate reward $u'$, which is stored in a buffer $\mathcal{B}$.

The primary goal of an agent in RL is to maximize the discounted future cumulative reward, which is given by

$$U_t = \sum_{i=0}^{\infty} \gamma^i u_{t+i}, \tag{25}$$

where $\gamma$ is the discound factor, which is used to account for the future cumulative reward by finding an optimal policy, denoted as $\pi^*$. The optimal policy is expected to map the best actions with all the possible states. In order to achieve this, the Q-value function [34] can be used as a guide, and is obtained using

$$Q_\pi(s_t, a_t) = \mathbb{E}[u_t + \gamma Q_\pi(s', a')|s_t, a_t], \tag{26}$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation, and $a'$ is the next action taken by the agent in state $s'$. The optimal policy $\pi^*$ that maximizes the long-term cumulative discounted reward also maximizes the expected Q-value function, and is obtained using the following equation:

$$Q_{\pi^*}(s_t, a_t) = \mathbb{E}[u_t + \gamma \max_{a'} Q_{\pi_*}(s', a')|s_t, a_t], \tag{27}$$

where the optimal policy, denoted as $\pi_*$, is designed to maximize the expected Q-value function. This policy plays

a crucial role in the training and updating process of the Deep Q-Network (DQN), which aims to approximate the action-value function $Q_{\pi^*}(s, a)$. It is recalled that the experience tuple is defined as $e_t = (s_t, a_t, u_t, s')$. The agent saves its experiences in a buffer $\mathcal{D} = \{e_1\ e_2\ ..\ e_t\}$ that is used to train the DQN using the gradient descent algorithm [35]. While using all data in each iteration is ideal for DQN training, doing so becomes prohibitively expensive when the training set is large. A more effective technique is known as mini-batch, which involves evaluating the gradients for each iteration using a random subset of the replay buffer $\mathcal{D}$. The loss function is defined as follows:

$$\mathcal{L}(\mathbf{W}) = \sum_{e \in D} (\underbrace{u + \gamma \max_{a'} Q_{\pi^*}(s', a', \hat{\mathbf{W}})}_{\text{target}} - Q_{\pi^*}(s, a, \mathbf{W}))^2, \tag{28}$$

where (28) denotes the DQN's loss function for a random mini-batch $\mathcal{D}$ at time slot $t$ and $\hat{\mathbf{W}}$ denotes the quasi-static target parameters that are updated every $t$ time slots. Finally, the optimal weights are obtained using

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}). \tag{29}$$

In order to minimize the loss function defined in (28), the weights of the DQN are updated at every time step $t$ using a stochastic gradient descent (SGD) algorithm on a mini-batch sampled from the replay buffer $\mathcal{D}$. To this effect, the SGD algorithm will update the weights $\mathbf{W}$ in an iterative process with a learning rate of $\mu > 0$ as follows [35]:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu \bigtriangledown \mathcal{L}_t(\mathbf{W}_t). \tag{30}$$

### C. POLICY SELECTION

In general, Q-Learning is regarded as an off-policy algorithm, which means that it estimates the reward for future actions and adds a value to the new state [34] without actually adhering to any greedy policy. We consider a nearly greedy action selection policy in light of this. Two modes of the near-greedy policy are:

1) Exploration: The agent experiments with various actions at each time step $t$ in an effort to find an effective action $at$.
2) Exploitation: The agent selects an action at time step $t$ that maximizes the state-action value function $Q_\pi(s, a; \mathbf{W}_t)$ based on the previous experience.

The agent in the near-greedy policy has an exploitation rate of 1 - $\epsilon$ and an exploration rate of $\epsilon$, where $0 < \epsilon < 1$. The hyper-parameter $\epsilon$ controls the trade-off between the agent's exploitation rate and exploration rate. The agent executes a specific action $a_t$ at a predetermined current state $s_t$ for each time step $t$. As a result, the agent moves into a target state $s' := s_t + 1$ and is rewarded positively or negatively $u_{s,s',a}[t]$.

The period of time during which the agent interacts with the environment is referred to as an episode, with each episode lasting $T$ time steps. The dimension of the input layer is set to the number of states in $\boldsymbol{S}$, and the dimension of the output layer is set to the number of possible actions in $\boldsymbol{A}$. We chose a smaller depth for the hidden layer because it has a significant impact on computational complexity. As a result, we chose a depth that strikes a reasonable balance between performance and computational complexity.

### D. DQL FRAMEWORK SETUP

At each time step $t$, the algorithm calculates the overall SE in the considered IRS-assisted VLC network, which is given in (23). In what follows, we provide some details on the action space, state space, and the reward function.

#### 1) STATE SPACE

All possible states form the state space, denoted as $\boldsymbol{\mathcal{S}}$. In this paper, the state space $\boldsymbol{\mathcal{S}}$ contains the power allocation for the LEDs in the VLC network, the user-LED association matrix, and the yaw angle $\varphi_j$ of the $j^{th}$ segment in the IRS frame, and roll angle $\omega_j$ of the $j^{th}$ segment in the IRS frame. Accordingly, the resultant state space is: $\boldsymbol{\mathcal{S}} = \begin{Bmatrix} p_1 & p_2 & .. & p_L \\ f_1 & f_2 & .. & f_K \\ \omega_1 & \omega_2 & .. & \omega_J \\ \varphi_1 & \varphi_2 & .. & \varphi_J \end{Bmatrix}$.

#### 2) ACTION SPACE

All the actions can be taken by the agent from the action space, denoted as $\boldsymbol{\mathcal{A}}$. The possible actions in the action space $\boldsymbol{\mathcal{A}}$ are:

- *Increase / Decrease* the power of the $l^{th}$ LED by a step size of $\Lambda_l$, where $\Lambda_l$ is a fixed value to be added to (or subtracted from) each $P_l$, such that $\sum_{l=1}^{L} P_l \leq P_{max}$ and $P_{min} \leq P_l \leq P_{max} \forall l \in (1, .., L)$.
- *Change the permutation* of the user-LED association matrix $\mathbf{F}$ such that $f_{l,k} \in \{0, 1\}, \forall l \in \{1, \ldots, L\}$, $\forall k \in \{1, \ldots, K\} \forall l \in \{1, \ldots, L\}$
- *Increase / Decrease* the yaw angle $\omega_k$ of the $j^{th}$ segment by step size $\zeta$, where $\zeta$ is a fixed value to be added to (or subtracted from) the value of the yaw angle of the $j^{th}$ IRS segement, such that the yaw angle is $\frac{-\pi}{2} \leq \varphi_j \leq \frac{\pi}{2} \forall j \in \{1, \ldots, J\}$
- *Increase / Decrease* the roll angle $\omega_j$ of the $k^{th}$ segment by step size $\nu$, where $\nu$ is a fixed value to be added to (or subtracted from) the value of the roll angle of the $j^{th}$ IRS segement, such that the roll angle is $\frac{-\pi}{2} \leq \omega_j \leq \frac{\pi}{2} \forall j \in \{1, \ldots, J\}$.

The total number of actions in the action space $\boldsymbol{\mathcal{A}}$ are calculated using $|A| = (2L) + (2K) + (4J)$.

#### 3) REWARD FUNCTION

The reward function plays an essential role in the RL algorithm. We use the product of the overall spectral efficiency (23) and Jain's Fairness index (24) as a reward. The product of both equations represent the immediate reward $u_t$ returned after choosing action $a_t$ in state $s_t$.

**TABLE 1. IRS assisted VLC system parameters [25].**

| Parameter | Value |
|---|---|
| Number of LEDs, $L$ | 4 |
| Number of IRS mirrors, $N$ | 200 |
| The minimum QoS Requirement, $R_{k,\min}$ | 0.02 bps/Hz |
| $P_{\max}$,$P_{\min}$,$P_{\text{total}}$ | 7 W , 3 W, 10 W |
| PD FOV, $\Psi_{fov}$ | 60° |
| PD responsivity $R_p$ | 0.4 A/W |
| PD detection area, $A$ | 1 $\text{cm}^2$ |
| Reflective index, $q$ | 1.5 |
| Optical filter gain, $G(\xi)$ | 1 |
| The Lambertian index, $m$ | 1 |
| The reflection coefficient, $\rho_{\text{IRS}}$ | 0.9 |

**TABLE 2. Deep Q-Learning hyper-parameters.**

| Parameter | Value |
|---|---|
| Discount factor, $\gamma$ | 0.995 |
| The step size for the yaw angle, $\zeta$ | 5° |
| The step size for the roll angle, $\nu$ | 5° |
| The step size for the power allocation, $\varepsilon$ | 0.1 W |
| Initial exploration rate, $\epsilon$ | 1.000 |
| Number of states | $L + K + 2J$ |
| Deep Q-network width | 24 |
| Exploration decay rate, $\kappa$ | 0.9995 |
| Minimum exploration rate, $\epsilon_{\min}$ | 0.1 |
| Number of actions | $2L + 2K + 4D$ |
| DQN depth | 2 |

Having described the State Space, Action Space, and the Reward Function, in the following, we describe in detail the operational steps of the DQL-MPF algorithm as follows:

1) The IRS-assisted VLC network environment is initialized according to Table 1. The DQL hyper-parameters are initialized as in Table 2. The policy network weights $\mathbf{W}_t$ are randomly initialized.
2) In each episode, the entire state space is reset to the initial values to improve the learning experience. Similarly, the roll angles in $\Omega$ and the yaw angles in $\Xi$ are initialized to the value of 0°.
3) DQL-MPF uses the $\epsilon$-greedy algorithm to select an action from the action space for a given state in our time-sequential decision process.
4) To allow the exploration of the action space, $\tau$ is randomly sampled from a uniform distribution.
   a) If the sampled value is less than or equal to the value of $\epsilon$, then the agent takes a random action.
   b) Otherwise, the agent will select an action based on the learned policy $a_t$=arg max $Q_\pi(s, a; \mathbf{W}_t)$, which aims to maximize the cumulative future reward.
5) In order to maximize the Q-value, which is constructed from the policy network outputs, the agent observes the next state and performs the following set of possible actions:
   a) Increase or decrease the power of the $l^{th}$ LED, by a step size $\varepsilon$.
   b) Change the permutation of the user-LED association matrix.
   c) Increase or decrease the yaw angle $\varphi_j$ of the $j^{th}$ segment, by step size $\zeta$.
   d) Increase or decrease the roll angle $\omega_j$ of the $j^{th}$ segment, by step size $\eta$.
6) Following (23), compute the overall SE based on the new sets of $\Xi$, $\Omega$, $\mathbf{P}$, and $\mathbf{F}$. Next, store the result as a reward $u_t$.
7) If the agent tries to exceed the constraints of the IRS-VLC system, abort the episode.
8) Following that, $s_t$, $s'$, $a_t$, and $u_t$ are stored in the replay memory buffer $\mathcal{D}$, which has a capacity of $\mathcal{M}$.
9) Using the gradient descent algorithm with a learning rate $\mu$, a mini-batch is sampled from the buffer and is used to train the policy network to minimize the loss function, which is given by (28).
10) The resulted loss $\mathcal{L}(\mathbf{W})$ at time step $t$ is recorded and the next state $s'$ is updated as current state $s_t$.

### *E. COMPUTATIONAL COMPLEXITY OF THE PROPOSED ALGORITHM*

It is crucial to analyse the computational complexity of the proposed algorithm. Therefore, we are presenting a theorem that shows how many iterations are needed for **Algorithm 2** to converge, namely

*Theorem 1:* For an indoor IRS assisted VLC system with $K$ users $L$ access points, and $D$segments, the computational complexity of the proposed **Algorithm 2** is given by:

$$\mathcal{O}((8J^2 + 8JK + 8JL + 2K^2 + 4KL + 2L^2) \times \mathcal{H} + K + \mathcal{C}_1). \quad (31)$$

*Proof:* First, the DQL agent observes the state of the system, executes the most valuable action, and calculates the reward based on (23). Assuming that the computational complexity of calculating the reward is $\mathcal{C}_1$, which is directly proportional with $K$, $L$, and $J$. Second, the MPF is responsible for generating IRS segments by iteratively updating the number of elements in each segment. The worst-case scenario occurs when the algorithm updates the number of elements $K$ times, yielding a maximum complexity of $\mathcal{O}(K)$. Finally, it is known that the size of the state space and the size of the action space have a significant role in the complexity of the deep Q-learning algorithm. Following [36], the computational complexity of the Q-learning algorithm with the greedy policy is estimated to be $\mathcal{O}(\mathcal{S} \times \mathcal{A} \times \mathcal{H})$ each iteration, where $\mathcal{S}$ is the number of states, $\mathcal{A}$ is the number of actions, and $\mathcal{H}$

**Algorithm 2** MPF-DQL Algorithm

**Input:** IRS elements array with total number of $N$ mirrors and $K$ number of users,$\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$

**Output:** The optimal roll angle matrix $\Omega$, and the optimal yaw angle matrix $\Xi$ for $D$segments, the optimal power allocation matrix $\mathbf{P}$, and the optimal user association matrix $\mathbf{F}$

Initialize time, actions, states, and replay buffer $\mathcal{D}$
**function** MPF($N$,$K$):
  Execute **Algorithm 1**
  Return $\mathbf{S}_J$
**end function**
$\mathbf{S}_J$ = MPF($N$,$K$)
**while** *No convergence or Not aborted* **do**
  **while** $t < T$ **do**
    t := t+1
    Observe current state $s_t$
    $\epsilon = \max(\epsilon.\text{d}, \epsilon_{min})$
    Sample $\tau \sim$ Uniform(0,1)
    **if** $\tau \leq \epsilon Selectarandomactiona_t$ **then**
    **else**
      Select an action based on
      $a_t$=arg max $Q_\pi(s, a; \mathbf{W}_t)$
    **if** $R_k < R_{k,min} \forall k \in (1, .., K)$ ***or***
    $\frac{-\pi}{2} > \omega_j > \frac{\pi}{2}$ ***or***
      $\frac{-\pi}{2} > \varphi_j > \frac{\pi}{2} \quad \forall \mathbf{S}_j \in \{\mathbf{S}_1, \ldots, \mathbf{S}_J\}$ ***or***
      $\sum_{l=1}^{L} P_l > P_{max}$ ***or***
      $P_{min} > P_l > P_{max} \; \forall P_l \in (P_1, .., P_L)$ **then**
      Abort episode.
    Compute the overall SE based on (23).
    Store experience $e_t = (s_t,a_t,u_{s,s',a}, s')$ in $\mathcal{D}$.
    Minibatch sample from $\mathcal{D}$, $e_i = (s_i,a_i,u_i,s_{i+1})$.
    Set $y_i := u_i + \iota \, max_{a'} \, Q_{\pi^*}(s_{i+1}, a'; \mathbf{W}_t)$.
    Obtain the optimal weights $\mathbf{W}^\star$ by performing SGD on $((y_i - Q_{\pi^*}(s_i,a_i,\mathbf{W}_t))^2$
    Update $\mathbf{W}_t := \mathbf{W}^\star$ in the DQN.
    Record the Loss $\mathcal{L}_t$.
    Update $s_t := s'$.

is the number of steps per episode. It is recalled that the size of the state space is $K + L + 2J$, and the size of the action space is $2K + 2L + 4J$. Therefore, the amount of work per iteration is

$$\mathcal{O}((8J^2 + 8JK + 8JL + 2K^2 + 4KL + 2L^2) \times \mathcal{H}). \quad (32)$$

Based on this and by incorporating $\mathcal{O}(K)$ and $\mathcal{C}_1$ into (32), equation (31) is deduced, which completes the proof. □

## IV. ACHIEVED RESULTS AND DISCUSSION

In this section, the effectiveness of the proposed algorithm in the considered IRS-assisted VLC system is investigated.

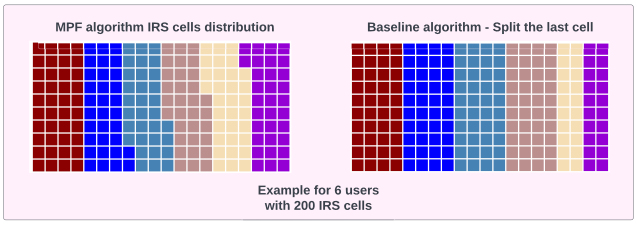


**FIGURE 2. MPF algorithm distribution of mirrors versus Baseline scheme (split the last cell(s)) with $N = 200$ and $K = 6$.**

For the simulation setup, we assume the size of the room is 8m × 8m × 3m, with four LEDs evenly spaced throughout the roof at the following locations: (2 m, 2 m, 3m), (2 m, 6 m, 3 m), (6 m, 2 m, 3 m), and (6 m, 6 m, 3 m). The users are randomly distributed on a plane, which is 1 m above the ground. Additionally, an IRS frame with mirrors is set up against one of the room's walls, and all units are distributed evenly across a rectangle with the coordinates (0 m, 1 m, 1.5 m) and (0 m, 7 m, 2.5 m) as its corners,where each mirror has an area of 30 cm × 10 cm.

It is assumed that each user's NLoS channel is blocked by a homogeneous media with a probability of 50%. Table 1 contains all the simulation parameters. In what follows, we compare the overall SE of the MPF, DQL-MPF and the genetic algorithm [27]. We refer to the "Baseline" scheme, which relies on fixed power allocation and fixed LED-user association, as the benchmark for comparison. The association of LEDs and users in this scheme is determined based on the distance between the LED and the user, with the closest user to a certain LED being associated with it. The power allocation in the Baseline scheme is fixed, meaning that the power is allocated equally among the different LEDs. Additionally, the Baseline scheme uses a fixed IRS mirrors orientation at 0° and follows an IRS mirrors distribution strategy as depicted in Fig. 2. Finally, the DQL-MPF algorithm was realized and trained on a PC equipped with Nvidia GPU 2080Ti and a 20-core 2.6 GHz processor. Note that we have developed our framework using Python and TensorFlow library [37]. The Deep Q-Learning hyper-parameters are shown in Table 2. To begin with, we investigate the effectiveness of the MPF algorithm, in maximizing the fairness among the users in the TDMA-based IRS-VLC system.

Fig. 5 illustrates the comparison of Jain's Fairness Index for the MPF algorithm and the baseline scheme, with $N = 200$. It can be observed that the MPF algorithm consistently achieves higher fairness compared to the baseline scheme when $N$ is not divisible by $K$. For example, when $K = 7$, the MPF algorithm achieves a fairness index of 0.99, while the baseline scheme only achieves 0.93. This demonstrates the effectiveness of the MPF algorithm in maximizing fairness among users by adaptively allocating IRS segments, while also highlighting its superiority over the baseline scheme in scenarios where the divisibility condition is not satisfied.

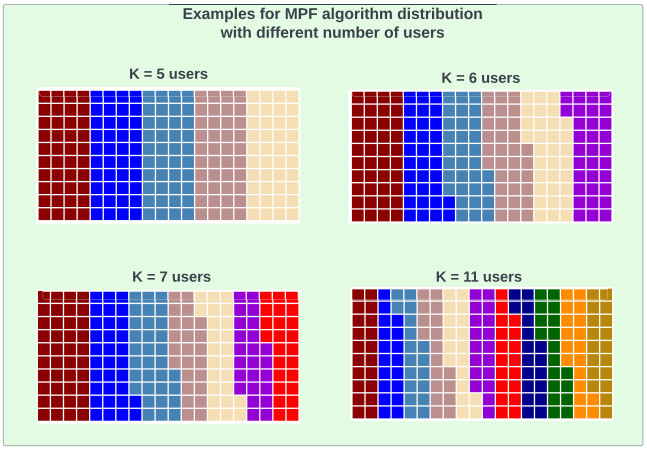


**FIGURE 3.** MPF algorithm distribution of mirrors with $N = 200$ and different values of $K$.

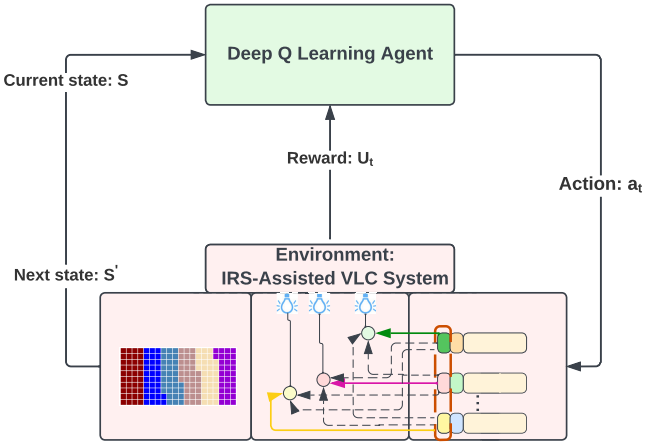


**FIGURE 4.** The interplay between the agent and the IRS-VLC system environment using reinforcement learning.

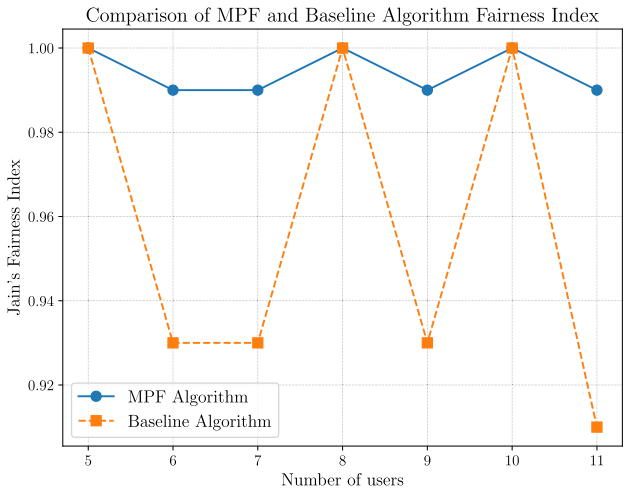


**FIGURE 5.** Jain's Fairness Index using MPF algorithm, and the Baseline scheme, with $N = 200$.

Building on the observations from Fig. 5, Fig. 6 examines the performance of the MPF algorithm by depicting the overall SE versus the noise power for both $K = 7$ and $K = 11$ users. It is demonstrated that the MPF algorithm offers an improved overall SE throughout the entire noise power range for both scenarios. In particular, the MPF algorithm outperforms the Baseline scheme that employs gain-based ordering by more than 2.5 bps/Hz for all the cases where the noise power is greater than $10^{-5}$ W. In contrast, there is a significant performance drop when using the Baseline scheme with no gain-based ordering, and it approaches 0 when the noise power is $10^{-4}$ W, whereas the Baseline with gain-based ordering achieves 2.5 bps/Hz for $K = 6$ and 0.6 bps/Hz for $K = 11$. This further reinforces the effectiveness and superiority of the MPF algorithm in maximizing system performance, both in terms of fairness and spectral efficiency.

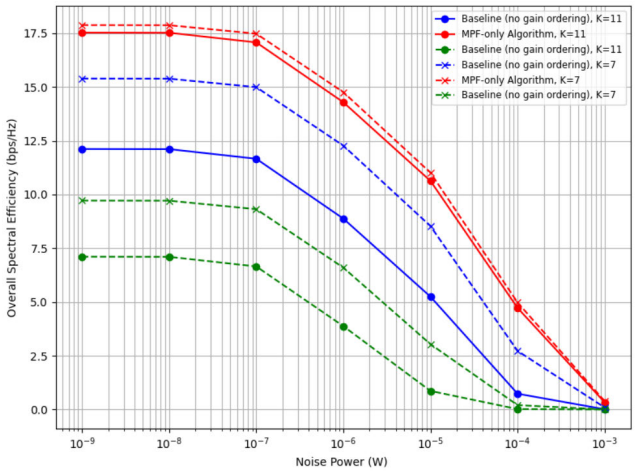


**FIGURE 6.** Overall SE versus noise power using MPF algorithm, and the Baseline scheme, with $N = 200$, $K = 7$ and $K = 11$.

The behavior of the MPF algorithm is almost identical for the two scenarios, but the performance gap between the MPF algorithm and the Baseline schemes has increased significantly due to the increased unfairness among the users in the IRS-VLC system when moving from $K = 7$ to $K = 11$. In particular, when the noise power is $10^{-4}$ W, the MPF algorithm achieves 5 bps/Hz for $K = 11$, while the Baseline scheme with gain-based ordering achieves a lower SE than in the $K = 7$ scenario.

Fig. 7 shows a convergence analysis for the proposed algorithm and the GA with maximum possible fairness (GA-MPF). Please refer to Table 3 for the detailed parameter settings used in the GA benchmark. Additionally, Table 4 shows the time per iteration for each algorithm. It can be observed that the DQL-MPF algorithm converges after 470 iterations, with a maximum overall SE of 20.9 bps/Hz. On the other hand, the GA-MPF algorithm achieves an overall SE of 19.85 bps/Hz, despite taking 2281 iterations. Notably, the MPF algorithm shows a considerable performance enhancement when combined with the DQL algorithm or the GA. For instance, the DQL-MPF algorithm achieves 20.1% higher overall SE, compared to the DQL-only algorithm. Similarly, the GA without the MPF algorithm yields 15.82 bps/Hz, which is 20% lower than the GA-MPF

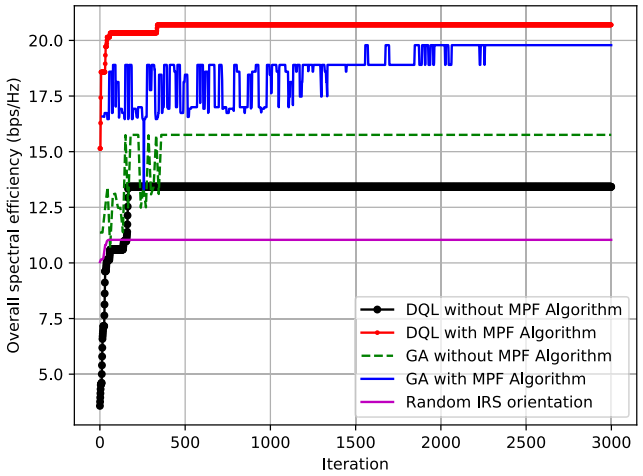


**FIGURE 7.** **Convergence analysis for the proposed algorithm, and the GA algorithm, with $N = 200$ and $K = 7$.**

**TABLE 3.** **Genetic algorithm parameters.**

| Parameter | Value |
|---|---|
| Bits per Variable | 16 |
| Population Size | 50 |
| Crossover Rate | 0.8 |
| Mutation Rate | 0.2 |
| Number of Generations | 100 |
| Selection Method | Tournament Selection |
| Elitism | Enabled |
| Tournament Size | 5 |

counterpart. We can deduce that combining the DQL with the MPF algorithm in the considered system can significantly improve the overall SE. In addition, after the initial convergence, minimal time will be required from the agent to re-optimize the parameters for any changes that may arise from the indoor environment. This feature ensures that the proposed method maintains its effectiveness in real-time applications, even when faced with evolving conditions within the indoor scenario. Furthermore, the Random IRS Orientation curve is introduced as a baseline comparison. As expected, it demonstrates the lowest performance due to its inherent randomness and inability to effectively exploit the benefits of the IRS. The Random IRS Orientation curve achieves a maximum SE of only 11.1 bps/Hz, substantially lower than the other techniques discussed, which further highlights the importance of optimizing the IRS orientation to achieve an improved overall spectral efficiency.

In Fig. 8, we depict the overall SE versus the noise power for the DQL-MPF algorithm, MPF-only algorithm, and Baseline with no gain-based ordering, considering both $K = 6$ and $K = 11$ users with $N = 200$. The DQL-MPF algorithm demonstrates noticeable robustness against the increasing noise power, maintaining a consistent overall SE of around 20 bps/Hz for up to $10^{-5}$ W. In contrast, the

**TABLE 4.** **Comparison of time taken per iteration for each algorithm.**

| Algorithm | Time (s per iteration) |
|---|---|
| DQL-MPF | 0.08 |
| GA-MPF | 0.36 |

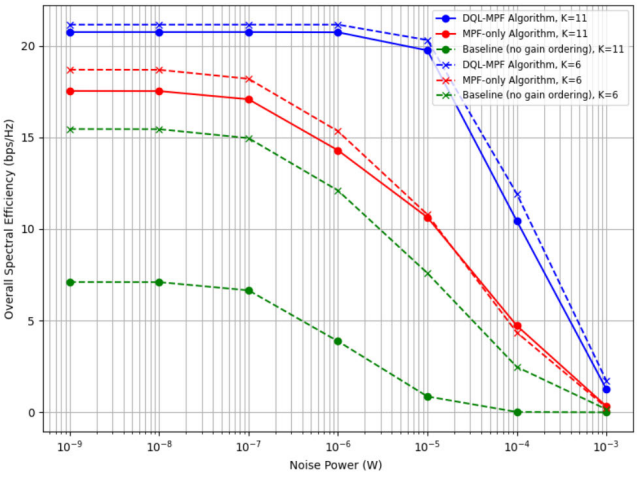


**FIGURE 8.** **Overall SE versus noise power using DQL-MPF algorithm, MPF algorithm, and the Baseline scheme, with $N = 200$, $K = 6$, and $K = 11$.**

MPF-only algorithm starts dropping much earlier, i.e., when the noise power is $10^{-6}$ W. This improved robustness of the DQL-MPF over the MPF algorithm is expected, as the DQL-MPF algorithm benefits from better adaptability and broader control over the considered IRS-VLC system. Moreover, the DQL-MPF algorithm consistently delivers superior performance compared to the other algorithms, even when handling a relatively high number of users, such as $K = 11$. For instance, at a noise power level of $10^{-5}$ W, the DQL-MPF algorithm achieves 20.1 bps/Hz, significantly outperforming the Baseline scheme, which reaches only 1.5 bps/Hz.

Lastly, Fig. 9 presents a comparison of the overall SE against the number of IRS mirrors for the proposed DQL-MPF algorithm and the two baseline algorithms. It is evident that the proposed algorithm surpasses both baseline algorithms, achieving a significantly higher overall spectral efficiency with an increasing number of IRS mirrors. Remarkably, our algorithm with just 200 IRS elements attains the performance level of the "baseline with gain ordering" algorithm with 1000 IRS elements. This implies that the baseline scheme required an additional 800 IRS elements to match the performance of our proposed algorithm. Conversely, the baseline algorithm with gain ordering outperforms the one without gain ordering. This observation underscores the effectiveness of the DQL-MPF algorithm in leveraging the increased number of IRS mirrors to optimize spectral efficiency. Importantly, the performance of all algorithms reaches a saturation point, attributable to the fact

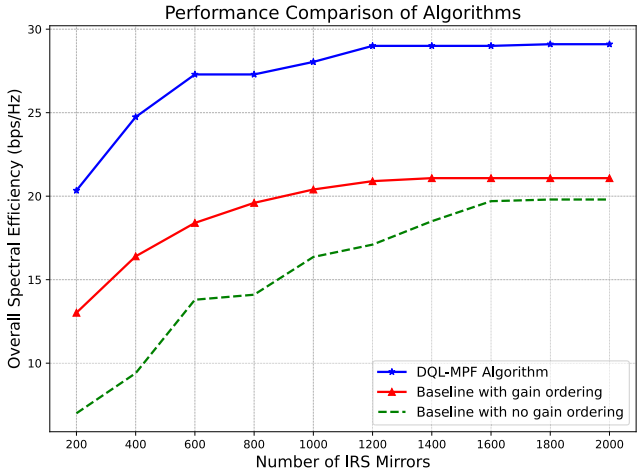


**FIGURE 9. Overall SE versus noise power using DQL-MPF algorithm, MPF algorithm, and the Baseline scheme, with $N = 200$ and $K = 11$.**

that these mirrors merely reflect and are dependent on the active VLC access points, which have power limitations due to illumination constraints.

## V. CONCLUSION

In this work, we proposed a deep Q-learning-based resource management framework to maximize the overall SE of a TDMA-based IRS-assisted VLC network. The proposed framework consists of two main stages: the first stage executes the maximum possible fairness (MPF) algorithm to divide the IRS mirrors into segments optimally, and each segment is optimized and dedicated to serving a single user. Next, we leverage the DQL algorithm to train an agent in order to jointly optimize the orientation of each IRS segment, the power allocation of the LEDs, and the LED-user association matrix. To this effect, the obtained results demonstrated that the DQL-MPF algorithm has a superior performance compared to the other baselines and offers very minimal run-time complexity to converge, at around 94 seconds. It was also shown that using the sole DQL algorithm will not maximize the fairness among the users in the considered system. On the other hand, using the MPF algorithm can maximize the fairness among the users, but the system will be prone to the increased noise power. Therefore, it was shown that it is best to combine both algorithms to maximize the fairness, the overall SE, and the robustness against the noise power.

## APPENDIX A
## A WALK-THROUGH EXAMPLE FOR THE MPF ALGORITHM

Assuming an IRS-assisted VLC system with $K = 7$ users and $N = 200$ mirrors. We present the detailed numerical steps on how the MPF algorithm will reach the final solution.

1) In the first loop:
   - The algorithm will check if $N = 200$ is divisible by $K = 7$. Since it is not divisible, it will decrement one from $K$ and check again if $N = 200$ is divisible by $K = 6$. Since it is not divisible, it will again decrement one from $K$, resulting in $K = 5$, and now 200 is divisible by 5, and we had to decrement 2 times to reach the divisibility.
   - Now, the algorithm will have 5 segments, equally having 40 mirrors. $\mathbf{S}_{rem1} =$

$$\left\{\begin{matrix} len\,(\mathbf{S}_1) = 40, len\,(\mathbf{S}_2) = 40, len\,(\mathbf{S}_3) = 40, \\ len\,(\mathbf{S}_4) = 40, len\,(\mathbf{S}_5) = 40, \end{matrix}\right\}$$

     where len(.) is a function that returns the number of elements in the a matrix.
   - The algorithm will detach $\mathbf{S}_1$, and concatenate the remaining segments in $\mathbf{S}_{rem}$.
   - Next, the algorithm will assign the mirrors of $\mathbf{S}_1$ with the lowest channel gains from $\mathbf{H}^{(2)}$.
   - Check if the number of mirrors in $\mathbf{S}_{rem}$ which is 160 is divisible by $(K - dec) = 5$.
   - Since 160 is divisible by 5, the algorithm will update the total number of mirrors to $N = 160$,, increment the index of the major segment by 1 $n = n + 1$, decrement the number of users $K$ to $K - 1$.

2) In the second loop:
   - The algorithm will check if $N = 160$ is divisible by $K = 6$. Since it is not divisible, it will decrement one from $K$ and check again if $N = 160$ is divisible by $K = 5$.
   - Now it is divisible, the algorithm will have 5 segments, equally having 32 mirrors $\mathbf{S}_{rem2} =$

$$\left\{\begin{matrix} len\,(\mathbf{S}_2) = 32, len\,(\mathbf{S}_3) = 32, len\,(\mathbf{S}_4) = 32, \\ len\,(\mathbf{S}_5) = 32, len\,(\mathbf{S}_6) = 32. \end{matrix}\right\}$$

   - The algorithm will detach $\mathbf{S}_2$, and concatenate the remaining segments in $\mathbf{S}_{rem}$.
   - Next, the algorithm will assign the mirrors of $\mathbf{S}_2$ with the second lowest channel gains from $\mathbf{H}^{(2)}$.
   - Check if the number of mirrors in $\mathbf{S}_{rem}$ which is 128 is not divisible by $(N - dec) = 5$.
   - Obviously 128 is not divisible by 5, therefore, it will update the total number of mirrors to $N = 128$,,increment the index of the major segment by 1 $n = n+1 = 2$, and decrement the number of users $K$ to $K - 1$.

3) In the third loop:
   - The algorithm will check if $N = 128$ is divisible by $K = 5$. Since it is not divisible, it will decrement one from $K$ and check again if $N = 128$ is divisible by $K = 4$. Now it is divisible, and we had to decrement 1 time to reach the divisibility.
   - Now, the algorithm will result into 4 segments, equally having 32 mirrors.

$$\mathbf{S}_{rem3} = \left\{\begin{matrix} len\,(\mathbf{S}_3) = 32, len\,(\mathbf{S}_4) = 32, \\ len\,(\mathbf{S}_5) = 32, len\,(\mathbf{S}_6) = 32. \end{matrix}\right\}$$

- The algorithm will detach $\mathbf{S}_3$, and concatenate the remaining segments in $\mathbf{S}_{rem3}$.
- Next, the algorithm will assign the mirrors of $\mathbf{S}_2$ with the third lowest channel gains from $\mathbf{H}^{(2)}$.
- Check if the number of mirrors in $\mathbf{S}_{rem}$ which is 96 is divisible by $(N - dec) = 4$.
- It follows that 96 is divisible by 4, resulting in the following number of elements for the segments:

$$\mathbf{S}_{rem4} = \begin{Bmatrix} len\,(\mathbf{S}_4) = 24, len\,(\mathbf{S}_5) = 24, \\ len\,(\mathbf{S}_6) = 24, len\,(\mathbf{S}_7) = 24. \end{Bmatrix}$$

- Finally, we concatenate $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$, and $\mathbf{S}_{rem4}$, resulting in the following number of elements for all the IRS array segments: $\mathbf{S}_J =$

$$\begin{Bmatrix} len\,(\mathbf{S}_1) = 40, len\,(\mathbf{S}_2) = 32, len\,(\mathbf{S}_3) = 32, \\ len\,(\mathbf{S}_4) = 24, len\,(\mathbf{S}_5) = 24, len\,(\mathbf{S}_6) = 24, \\ len\,(\mathbf{S}_7) = 24. \end{Bmatrix}.$$

## REFERENCES


[1] F. Hu, B. Chen, and K. Zhu, "Full spectrum sharing in cognitive radio networks toward 5G: A survey," *IEEE Access*, vol. 6, pp. 15754–15776, 2018.

[2] L. Bariah et al., "A prospective look: Key enabling technologies, applications and open research topics in 6G networks," *IEEE Access*, vol. 8, pp. 174792–174820, 2020.

[3] T. Tang, T. Shang, and Q. Li, "Impact of multiple shadows on visible light communication channel," *IEEE Commun. Lett.*, vol. 25, no. 2, pp. 513–517, Feb. 2021.

[4] M. S. Demir and M. Uysal, "A cross-layer design for dynamic resource management of VLC networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1858–1867, Mar. 2021.

[5] S. Aboagye, A. Ibrahim, T. M. N. Ngatched, A. R. Ndjiongue, and O. A. Dobre, "Design of energy efficient hybrid VLC/RF/PLC communication system for indoor networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 143–147, Feb. 2020.

[6] S. Aboagye, T. M. N. Ngatched, O. A. Dobre, and A. Ibrahim, "Joint access point assignment and power allocation in multi-tier hybrid RF/VLC HetNets," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6329–6342, Oct. 2021.

[7] Y. S. Eroğlu, Y. Yapıcı, and I. Güvenç, "Impact of random receiver orientation on visible light communications channel," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1313–1325, Feb. 2019.

[8] M. Najafi, B. Schmauss, and R. Schober, "Intelligent reflecting surfaces for free space optical communication systems," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6134–6151, Sep. 2021.

[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, 2021.

[10] A. R. Ndjiongue, T. M. N. Ngatched, O. A. Dobre, and H. Haas, "Reconfigurable intelligent surface-based VLC receivers using tunable liquid-crystals: The concept," *J. Lightw. Technol.*, vol. 39, no. 10, pp. 3193–3200, May 2021.

[11] Y. Ata, H. Abumarshoud, L. Bariah, S. Muhaidat, and M. A. Imran, "Intelligent reflecting surfaces for underwater visible light communications," *IEEE Photon. J.*, vol. 15, no. 1, pp. 1–10, Feb. 2023.

[12] S. Ibne Mushfique, A. Alsharoa, and M. Yuksel, "MirrorVLC: Optimal mirror placement for multielement VLC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 10050–10064, Nov. 2022.

[13] N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[14] X. Kong et al., "Deep reinforcement learning-based energy-efficient edge computing for Internet of Vehicles," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 6308–6316, Sep. 2022.

[15] H. Sharma, I. Budhiraja, N. Kumar, and R. K. Tekchandani, "Secrecy rate maximization for THz-enabled femto edge users using deep reinforcement learning in 6G," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2022, pp. 1–6.

[16] C. Huang, G. Chen, Y. Gong, M. Wen, and J. A. Chambers, "Deep reinforcement learning-based relay selection in intelligent reflecting surface assisted cooperative networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 1036–1040, May 2021.

[17] C. Huang, G. Chen, and K.-K. Wong, "Multi-agent reinforcement learning-based buffer-aided relay selection in IRS-assisted secure cooperative networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4101–4112, 2021.

[18] X. Gao, Y. Liu, and X. Mu, "Trajectory and passive beamforming design for IRS-aided multi-robot NOMA indoor networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[19] R. Zhong, X. Liu, Y. Liu, Y. Chen, and Z. Han, "Mobile reconfigurable intelligent surfaces for NOMA networks: Federated learning approaches," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 10020–10034, Nov. 2022.

[20] C. Huang et al., "Hybrid beamforming for RIS-empowered multi-hop terahertz communications: A DRL-based method," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2020, pp. 1–6.

[21] C. Huang et al., "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.

[22] 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.

[23] H. Abumarshoud, B. Selim, M. Tatipamula, and H. Haas, "Intelligent reflecting surfaces for enhanced NOMA-based visible light communications," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 571–576.

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

[25] 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.

[26] A. K. Jain, D.-M. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Eastern Research Lab., Digit. Equip. Corp., Hudson, MA, USA, Tech. Rep. DEC-TR-301, 1984.

[27] G. Wang, Y. Shao, L.-K. Chen, and J. Zhao, "Subcarrier and power allocation in OFDM-NOMA VLC systems," *IEEE Photon. Technol. Lett.*, vol. 33, no. 4, pp. 189–192, Feb. 2021.

[28] X. Chen and M. Jiang, "Adaptive statistical Bayesian MMSE channel estimation for visible light communication," *IEEE Trans. Signal Process.*, vol. 65, no. 5, pp. 1287–1299, Mar. 2017.

[29] H. Marshoud, P. C. Sofotasios, S. Muhaidat, G. K. Karagiannidis, and B. S. Sharif, "Error performance of NOMA VLC systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[30] M. D. Soltani, A. A. Purwita, Z. Zeng, H. Haas, and M. Safari, "Modeling the random orientation of mobile devices: Measurement, analysis and LiFi use case," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2157–2172, Mar. 2019.

[31] M. Obeed, A. M. Salhab, M.-S. Alouini, and S. A. Zummo, "On optimizing VLC networks for downlink multi-user transmission: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2947–2976, 3rd Quart., 2019.

[32] 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.

[33] J. Hu et al., "Reconfigurable intelligent surface based RF sensing: Design, optimization, and implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2700–2716, Nov. 2020.

[34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[35] L. J. Lin, *Reinforcement Learning for Robots Using Neural Networks*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 1992.

[36] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[37] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Oper. Syst. Design Implement. (OSDI)*. Savannah, GA, USA: USENIX Association, Oct. 2016, pp. 265–283.

**AHMED AL HAMMADI** (Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering and computer science and the Ph.D. degree in electrical and computer engineering from Khalifa University, Abu Dhabi, United Arab Emirates, in 2011, 2015, and 2022, respectively. Currently, he is a Lead Researcher with the Technology Innovation Institute, Abu Dhabi. His research interests include the broad array of critical areas in the field, including but not limited to visible light communications, mmWave massive MIMO, and machine learning-based optimization, reconfigurable intelligent surfaces, holographic MIMO, and 3D wireless cellular networks.

**LINA BARIAH** (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in communications engineering from Khalifa University, Abu Dhabi, United Arab Emirates, in 2015 and 2018, respectively. She was a Visiting Researcher with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, in 2019, and an affiliate Research Fellow with the James Watt School of Engineering, University of Glasgow, U.K. She is currently a Senior Researcher with the Technology Innovation Institute, an Adjunct Faculty with Khalifa University, and an Adjunct Research Professor with Western University, Canada. She is a Senior Member of the IEEE Communications Society, IEEE Vehicular Technology Society, and IEEE Women in Engineering. She is also an Associate Editor of IEEE COMMUNICATION LETTERS, an Associate Editor of IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, and an Area Editor of *Physical Communication* (Elsevier). She is a Guest Editor of *IEEE Communication Magazine*, *IEEE Network Magazine*, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY. She has organized several workshops/special sessions in IEEE flagship conferences, including IEEE VTC and IEEE ICC. She was a member of the technical program committee of a number of IEEE conferences, such as ICC and GLOBECOM. She serves as the session chair and an active reviewer for numerous IEEE conferences and journals. Her research interests include machine learning for wireless communications, large language models, and generative AI for telecom.

**SAMI MUHAIDAT** (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2006. From 2007 to 2008, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, ON, Canada. From 2008 to 2012, he was an Assistant Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a Professor with Khalifa University, Abu Dhabi, United Arab Emirates. He was also a Visiting Reader with the Faculty of Engineering, University of Surrey, Guildford, U.K. He was a recipient of several scholarships during his undergraduate and graduate studies and the Winner of the 2006 Post-Doctoral Fellowship Competition. He was a Senior Editor of IEEE COMMUNICATIONS LETTERS and an Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He is an Area Editor of IEEE TRANSACTIONS ON COMMUNICATIONS.

**MAHMOUD AL-QUTAYRI** (Senior Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Concordia University, Canada, in 1984, the M.Sc. degree in electrical and electronic engineering from The University of Manchester, U.K., in 1987, and the Ph.D. degree in electrical and electronic engineering from the University of Bath, U.K., in 1992. He is currently a Full Professor with the Department of Electrical and Computer Engineering and the Associate Dean for Graduate Studies with the College of Engineering, Khalifa University, United Arab Emirates. Prior to joining Khalifa University, he was with De Montfort University, U.K., and University of Bath. He has authored/coauthored numerous technical papers in peer-reviewed journals and international conferences. He also coauthored a book titled *Digital Phase Lock Loops: Architectures and Applications* and edited a book titled *Smart Home Systems*. This is in addition to a number of book chapters and patents. His current research interests include wireless sensor networks, embedded systems design, in-memory computing, mixed-signal integrated circuits design and test, and hardware security.

**PASCHALIS C. SOFOTASIOS** (Senior Member, IEEE) was born in Volos, Greece, in 1978. He received the M.Eng. degree from Newcastle University, U.K., in 2004, the M.Sc. degree from the University of Surrey, U.K., in 2006, and the Ph.D. degree from the University of Leeds, U.K., in 2011. He was with the University of Leeds; University of California at Los Angeles, CA, USA; Tampere University of Technology, Finland; Aristotle University of Thessaloniki, Greece; and Khalifa University of Science and Technology, United Arab Emirates, where he is currently an Associate Professor with the Department of Electrical Engineering and Computer Science. He received the scholarship from U.K.-EPSRC for the M.Sc. degree and U.K.-EPSRC and Pace plc for the Ph.D. degree. His research interests include digital and optical wireless communications and special functions and statistics. He is a regular reviewer of several international journals and a member of the technical program committee of numerous IEEE conferences. He received the Exemplary Reviewer Award from IEEE COMMUNICATIONS LETTERS in 2012, the Best Paper Award from ICUFN 2013, and IEEE TRANSACTIONS ON COMMUNICATIONS in 2015 and 2016. He is also an Editor of IEEE COMMUNICATIONS LETTERS.

**MEROUANE DEBBAH** (Fellow, IEEE) is currently a Professor with the Khalifa University of Science and Technology, Abu Dhabi, and the founding Director of the 6G Center. He is also a researcher, an educator, and a technology entrepreneur. Over his career, he has founded several public and industrial research centers, start-ups. He is a frequent keynote speaker at international events in the field of telecommunication and AI. His research has been lying at the interface of fundamental mathematics, algorithms, statistics, information, and communication sciences with a special focus on random matrix theory and learning algorithms. In the communication field, he has been at the heart of the development of small cells (4G), massive MIMO (5G), and large intelligent surfaces (6G) technologies. In the AI field, he is known for his work on large language models, distributed AI systems for networks, and semantic communications. He received multiple prestigious distinctions, prizes and best paper awards (more than 35 best paper awards) for his contributions to both fields and according to research.com is ranked as the best scientist in France in the field of electronics and electrical engineering. He is a WWRF Fellow, an EURASIP Fellow, an AAIA Fellow, an Institut Louis Bachelier Fellow, and a Membre émérite SEE.