Learning-Based Energy-Efficient Resource Management by Heterogeneous RF/VLC for Ultra-Reliable Low-Latency Industrial IoT Networks

Helin Yang, *Student Member, IEEE*, Arokiaswami Alphones, *Senior Member, IEEE*, Wen-De Zhong, *Senior Member, IEEE*, Chen Chen, *Member, IEEE*, and Xianzhong Xie, *Member, IEEE*

***Abstract*— Smart factory under Industry 4.0 and industrial Internet of Things (IoT) has attracted much attention from both academia and industry. In wireless industrial networks, industrial IoT and IoT devices have different quality-of-service (QoS) requirements, ranging from ultra-reliable low-latency communications (URLLC) to high transmission data rates. These industrial networks will be highly complex and heterogeneous, as well as the spectrum and energy resources are severely limited. Hence, this article presents a heterogeneous radio frequency (RF)/visible light communication (VLC) industrial network architecture to guarantee the different QoS requirements, where RF is capable of offering wide-area coverage and VLC has the ability to provide high transmission data rate. A joint uplink and downlink energy-efficient resource management decision-making problem (network selection, subchannel assignment, and power management) is formulated as a Markov decision process. In addition, a new deep post-decision state (PDS)-based experience replay and transfer (PDS-ERT) reinforcement learning algorithm is proposed to learn the optimal policy. Simulation results corroborate the superiority in performance of the presented heterogeneous network, and verify that the proposed PDS-ERT learning algorithm outperforms other existing algorithms in terms of meeting the energy efficiency and the QoS requirements.**

***Index Terms*—Deep reinforcement learning, energy efficiency, heterogeneous radio frequency (RF)/visible light communication (VLC) industrial networks, industrial internet of things, resource management, ultrareliable low-latency communications (URLLC).**

Manuscript received February 27, 2019; revised May 25, 2019 and July 2, 2019; accepted August 5, 2019. Date of publication August 8, 2019; date of current version April 13, 2020. This work was conducted within the Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc. and National Research Foundation (NRF) Singapore under the Corp Lab@University. This work was supported in part by the National Nature Science Foundation of China under Grant 61901065 and Grant 61502067, in part by the Key Research Project of Chongqing Education Commission under Grant KJZD-K201800603, and in part by Chongqing Nature Science Foundation under Grant CSTC2018jcyjAX0432 and Grant CSTC2016jcyjA0455. Paper no. TII-19-0629. *(Corresponding author: Chen Chen.)*

H. Yang, A. Alphones, and W. Zhong are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore (e-mail: hyang013@e.ntu.edu.sg; ealphones@ntu.edu.sg; ewdzhong@ntu.edu.sg).

C. Chen is with School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: c.chen@cqu.edu.cn).

X. Xie is with the Chongqing Key Lab of Computer Network and Communication Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: xiexzh@cqupt.edu.cn).

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

Digital Object Identifier 10.1109/TII.2019.2933867

I. Introduction

With the rapid development of industrial automation, the fourth industrial revolution (Industry 4.0) takes the Internet of Things (IoT) into industrial systems, where smart devices (sensor, actuators, machines, and robots) intelligently send data to realize the real-time industrial control with the minimal human interaction [1], [2]. The future factories and industries expect to replace conventional wired communication networks by wireless networks, in order to improve the flexibility in moving machinery and reducing the infrastructure expenditure [3]–[5]. Massive machine-type communication (mMTC) can effectively support the massive communication connectivity of a large number of IoT devices in industrial networks, by transmitting short packets with low data rates in a short period of time [6]. In practical industrial wireless networks (IWNs), industrial IoT (IIoT) devices generally have the following requirements or challenges: strict latency and reliability requirements [7], high transmission data rate demands, limited energy batteries, and the scarce wireless radio frequency (RF) spectrum resource, all these issues impose challenging requirements to efficient network structures and wireless communication technologies [1]–[4], [8], [9].

Recently, considering the fact that ultra-reliable and low latency communications (URLLC) in fifth generation (5G) is closely related to industrial networks, some advanced resource managements approaches have been proposed to ensure the latency and reliability requirements of IIoT communications [5], [8]–[15]. For instance, industrial automation may require end-to-end latencies in the range of 1–5 ms with the transmission reliability of 99.999% or higher [8], [10], [11]. Ye *et al.* [5] proposed a novel two-phase transmission protocol to guarantee the stringent low delay and high reliability in

device-to-device-enabled industrial networks. Considering the large number of IIoT devices in industrial networks, the current research developments, such as clustered IWNs [12], adaptive routing protocols [13], software defined network, and edge computing [14], have been proposed to provide reliable and low latency wireless links for IIoT communications. Besides reliable and low latency requirements, energy efficiency plays an important role in IWNs, since most of IIoT devices (sensors, actuators, and controllers) are power-constrained in green IWNs [15]. The energy-efficient resource allocation and transmission protocol design were presented for the IIoT communication systems in 5G-enabled IWNs, in order to maximize the network energy efficiency (EE) while satisfying quality of service (QoS) requirements of devices [9], [16]. In [17], a dynamic routing approach was proposed to improve the energy consumption and communication latency performance in large-scale IIoT systems. The authors in [18] investigated the problem of how to optimize the tradeoff between the QoS satisfactions and the EE in IIoT systems.

The practical IWNs may exist in both IoT and IIoT devices, where IoT (consumer usage) mainly focuses on throughput and packet loss rate while IIoT (industrial purpose) emphasizes the latency and reliability, leading to the different QoS requirements [1], [14]. In this case, the different QoS requirements range from low latency and high reliability to high data rates, resulting in heterogeneous industrial networks [14]. Hence, hierarchical structures or designs are widely adopted in industrial networks [8], [14], [16], [19]–[21]. The hierarchical transmission architectures were presented to efficiently complete a large amount of application services based on the priority levels in smart industries [19], [20]. In order to reduce the network complexity, Kalor *et al.* [8] studied how to simplify the manageability of heterogeneous networks by slicing deterministic and packet-switched protocols, and a hierarchical transmission-estimation approach based on 5G enabled codesign was proposed to improve the transmission reliability [16]. Moreover, an ant colony algorithm was employed in industrial heterogeneous networks to improve the network reliability [21].

To achieve the intelligent decision making, the reinforcement learning (RL) tool is applied to learn the optimal policy of resource allocation, energy management, and transmission scheduling for IIoT or IoT [11], [22]–[28]. A Q-learning-based practical duty cycle control was developed to improve the network delay and transmission reliability [22]. He *et al.* proposed a distributed deep RL (DRL) combined with the Ethereum blockchain to create a reliable and safe IIoT communication environment [23], and the authors in [11] and [24] applied DRL to search the optimal solution to minimize the IoT communication delay. Analysis of QoS satisfactions in IoT frameworks using RL was treated in [25]–[27], which also investigated different RL algorithms for resource allocation, access control, and energy saving. Moreover, an efficient transfer RL approach was proposed to guarantee the URLLC requirements of Internet of Vehicles (IoVs) [28]. However, almost all of the above papers [11], [22]–[28] did not investigate how to satisfy the different QoS requirements of devices in dynamic and complex industrial networks.

The above reported works have ability to improve the industrial communication performance, but conventional RF networks may fail to support a large number of communication services (including high data rate) due to the saturation of RF spectrum in industrial networks, and hard to meet the energy-efficient communication due to a large number IIoT or IoT devices [1], [2], [8], [15]. Heterogeneous RF/visible light communication (VLC) network architecture was considered as a promising technique for indoor communication environments with the high energy-efficient utilization and reliable characteristics [29]–[31], where RF is capable of offering long-distance transmission with the wide-area coverage and VLC has the ability to provide high transmission data rate by generating multiple small optical cells. Moreover, the literatures [32] and [33] applied the heterogeneous RF/VLC structure in IoT communication networks to efficiently schedule transmission under high data rate requirements. However, conventional heterogeneous RF/VLC networks reported by the works [29]–[33] did not investigate the URLLC requirements in IWNs.

Motivated by the above observations, in this article, we present an energy-efficient resource management based on the heterogeneous RF/VLC architecture for industrial networks to guarantee the diverse requirements (high reliability, low latency, and high data rate) of IIoT and IoT devices. In order to enable IWNs with high intelligence, a new deep post-decision state (PDS)-based experience replay and transfer (PDS-ERT) RL algorithm is proposed to realize intelligent resource management, with the purpose of maximizing the network EE while ensuring the minimum data rate constraints and the strict URLLC requirements. The major contributions of this work are summarized as follows.

1) A new heterogeneous RF/VLC industrial network architecture is developed to support uplink and downlink communication services, which considers the EE, high reliability, low latency, and high transmission data rates requirements in practical industrial networks.
2) We formulate a joint uplink and downlink resource management (network selection, subchannel assignment, and power management) problem with considering QoS requirements, and the energy-efficient resource management problem is modeled as an RL framework, thus the network is capable of intelligently making decisions based on the instantaneous observations.
3) In order to satisfy different QoS requirements in dynamic industrial networks, a deep PDS-ERT learning algorithm is proposed to learn the optimal policy for the intelligent resource management under the continuous-valued state and action variables, which effectively improves the learning speed, efficiency, and stability.
4) The effectiveness of presented heterogeneous industrial architecture and the proposed deep PDS-ERT learning algorithm-based intelligent resource management have been evaluated by the comprehensive simulations.

The rest of this article is organized as follows. The heterogeneous RF/VLC network architecture is presented in Section II. Section III formulates the energy-efficient resource management problem. The proposed deep PDS-ERT learning algorithm is provided in Section IV. Simulation results are presented in Section V and Section VI concludes this article.

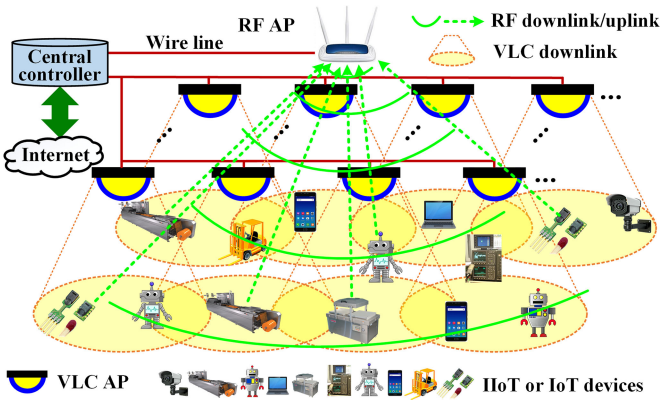

Fig. 1. Indoor heterogeneous RF/VLC industrial network.

II. System Model

A. Heterogeneous RF/VLC Industrial Network Architecture

Smart factories under Industry 4.0 will consist of a large number of IIoT devices (sensors, machines, actuators, robots, etc.) and IoT devices (computers, smartphones, tablets, etc.), resulting in different QoS requirements of communication services, such as ranging from high reliability and low latency to high data rates. Conventional RF networks may fail to support the large number of services due to the limited RF spectrum and energy resources. To address these issues, we present a new heterogeneous RF/VLC network structure to support different QoS requirements in industrial networks.

First of all, we divide the IIoT and IoT devices into two groups based on their different QoS requirements.

Group 1: The URLLC services of the devices (generally IIoT devices) have specific requirements on low latency and high reliability but have much looser constraints on the high date rate. For example, each sensor reports a small amount of collected data to the central controller (uplink) or the central controller sends the low bit rate information to each actuator(downlink) within the strict latency requirements.

Group 2: The normal services of the devices (commonly are IoT devices) have the high data rate requirements but are less interested in the latency and reliability requirements, such as the high quality image, video, and webpage.

We set that the devices in the Group 1 are with a higher priority to access the channel resource to guarantee the high reliability and low latency requirements, while the devices in the Group 2 are with a lower priority to access channel resource.

After that, a heterogeneous RF/VLC industrial network architecture is presented to support the abovementioned different services, as shown in Fig. 1, where a number of VLC access points (APs) (refer to femtocells) are uniformed attached on the room ceiling and one RF AP (refers to microcell) is placed in the center. Each VLC AP contains one light-emitting diode (LED) lamp-based luminaries devices offering both lighting requirements and communications services, and every VLC AP covers a confined area to generate a small optical cell. By contrast, the RF AP provides the coverage for the entire room. Both the VLC and RF APs connect the Internet to perform the communication services, where VLC APs broadcast information to devices through visible light signals and the RF AP provides communication services by the RF signals. Considering the unpractical components and challenges of the wireless VLC uplink [29]–[33], VLC only offers the downlink data streams while RF provides both the uplink and downlink data streams. We would like to mention that due to human activities and device mobility, the VLC line-of-sight (LOS) communication link may be intermittently interrupted or blocked of a number for time slots, called blocked LOS VLC links, and the blocked VLC links may not support general communication services [29]–[33]. Under this heterogeneous network, RF is capable of offering wide-area coverage and VLC has the ability to provide high transmission data rate due to the abundant bandwidth resources across multiple optimal cells. Motivated by the above analysis, the RF AP mainly provides the URLLC services of the devices in Group 1 due to its wide-area coverage, while VLC APs mainly support the normal services of the devices in Group 2 due to its offering high transmission data rate.

In addition, the IoT device's priority depends on its QoS requirements or application services, when the IoT device changes its application services, it will report this information to the central controller in the industrial network by the RF uplink, and hence the IoT device will be assigned to the channel resource based on its current priority. For example, one device in Group 2 with the normal service currently applies the URLLC services with the low latency and high reliability, it will report this information to the central controller and then it will be classified into Group 1 with the higher priority to access the channel resource to guarantee the high reliability and low latency requirements.

In the industrial network, a set of IIoT and IoT devices are randomly distributed on the floor, where the device (mainly IIoT device) requiring the URLLC service is equipped with one RF enabled transceiver, and the device (mainly IoT device) needing the uplink/downlink data rate is equipped with one VLC receiver [called photodetector (PD)], and one RF enabled transceiver. The network selection (RF or VLC) decision-making problem can be formulated as a Markov decision process (MDP) with the goal of maximizing the reward function, and solved with the proposed DRL algorithm, which will be provided in Section III and Section IV.

The number of VLC APs, devices, subchannels per VLC AP, and subchannels per RF AP are denoted by C, K, N^{VLC}, and N^{RF}, respectively. The set of VLC APs and devices are denoted as $\mathcal{C} = \{1, \ldots, C\}$ and $\mathcal{K} = \{1, \ldots, K\}$, respectively. Let $\mathcal{N}^{\mathrm{VLC}} = \{1, \ldots, N^{\mathrm{VLC}}\}$ and $\mathcal{N}^{\mathrm{RF}} = \{1, \ldots, N^{\mathrm{RF}}\}$ represent the subchannel sets of per VLC AP and RF AP, respectively, where the subchannels for VLC are reused across all optical cells. The network employs orthogonal frequency division multiple access (OFDMA) to serve devices.

B. VLC Channel Model

In VLC networks, the VLC LOS links can support the successful communication services while the blocked LOS VLC

links cannot provide the high transmission data rate services [29]–[33]. For the VLC link, the LOS channel gain from one AP to one device is expressed as [28]

$$h^{\mathrm{VLC}} = \frac{(\vartheta+1)A_r}{2\pi (d^{\mathrm{VLC}})^2}\cos^{\vartheta}(\phi)T_s(\psi)g(\psi)cos\psi \quad (1)$$

where A_r is the active area of the PD. d^{VLC} and ψ denote the distance and the angle of incidence between the LED and the device, respectively. ϕ is the angle of irradiance from the LED to the device. ϑ is the order of the Lambertian emission with $\vartheta = -\ln 2/(\ln \cos \phi_{1/2})$ with $\phi_{1/2}$ being the LED's semiangle at half power. $T_s(\psi)$ and $g(\psi)$ are the gain of the optical filter and the optical concentrator gain at the PD, respectively. $g(\psi)$ can be expressed as: $g(\psi) = \eta/\sin^2\psi_c$ when $0 \le \psi \le \psi_c$, and $g(\psi) = 0$ if $\psi_c < \psi$, where ψ_c and η are the semiangle field of view (FOV) of the PD and the refractive index, respectively.

As shown in Fig. 1, due to the multiple VLC APs deployment, the devices locate in the overlapped areas may suffer inter-cell interference (ICI) from adjacent cells. If the kth device ($k \in \mathcal{K}$) is assigned to VLC AP $c \in \mathcal{C}$ on the nth subchannel ($n \in \mathcal{N}^{\mathrm{VLC}}$), the received signal-to-interference-plus-noise-ratio (SINR) of the device is expressed as [29]–[31]

$$\gamma_{k,n}^{\mathrm{VLC}} = \frac{\mu^2 P_{n,c}^{\mathrm{VLC}} (h_{k,n,c}^{\mathrm{VLC}})^2}{\mu^2 \sum_{c' \in \mathcal{C}} P_{n,c'}^{\mathrm{VLC}} (h_{k,n,c'}^{\mathrm{VLC}})^2 + N_0^{\mathrm{VLC}} B_{sub}^{\mathrm{VLC}}} \quad (2)$$

where μ is the PD's responsivity, $P_{n,c}^{\mathrm{VLC}}$ indicates the allocated transmit electrical power on the nth subchannel of the cth VLC AP, $h_{k,n,c}^{\mathrm{VLC}}$ is the VLC channel gain from the cth VLC AP to device k on the nth subchannel, N_0^{VLC} represents the power spectral density (PSD) of noise at the PD, B_{sub}^{VLC} is the subchannel bandwidth $B_{sub}^{\mathrm{VLC}} = B^{\mathrm{VLC}}/N^{\mathrm{VLC}}$ with B^{VLC} being the VLC transmission bandwidth.

Hence, the data rate of kth device associated by VLC AP c can be expressed as

$$R_k^{\mathrm{VLC}} = \sum_{n \in \mathcal{N}^{\mathrm{VLC}}} \rho_{k,n,c}^{\mathrm{VLC}} \frac{B_{sub}^{\mathrm{VLC}}}{2} \log_2(1 + \gamma_{k,n,c}^{\mathrm{VLC}}) \quad (3)$$

where $\rho_{k,n,c}^{\mathrm{VLC}}$ is a binary variable, $\rho_{k,n,c}^{\mathrm{VLC}} \in \{0,1\}$, $\rho_{k,n,c}^{\mathrm{VLC}} = 1$ represents that the kth device assigns the nth subchannel of VLC AP c ; otherwise, $\rho_{k,n,c}^{\mathrm{VLC}} = 0$. The scaling factor 1/2 is due to the Hermitian symmetry [29]–[31].

C. RF Channel Model

Each indoor industrial factory room deploys one RF AP to be acted as one cell. The device may receive the ICI from adjacent industrial factory rooms with the same technology and the interference from competing technologies operating over the same band [34], when the device locates in the overlapped areas. In the RF network, the channel gain is typically given by [35]

$$g_{k,n}^{\mathrm{RF}} = 10^{-PL_k[dB]/10} \quad (4)$$

where $PL_k[dB]$ is the RF path loss of the kth device in dB, which is expressed as [35]

$$PL_k[dB] = A\log_{10}(d_k^{\mathrm{RF}}) + B + E\log_{10}(f_c/5) + X \quad (5)$$

where d_k^{RF} is the distance from the RF AP to the kth device and f_c denotes the carrier frequency in GHz. A, B, and E are constants depending on the propagation model. For the LOS propagation, $A = 18.7$, $B = 46.8$, and $E = 20$. For NLOS scenario, we have $A = 36.8$, $B = 43.8$, and $E = 20$. X indicates the wall penetration loss in the NLOS scenario, we set $X = 5(N_{\mathrm{wall}} - 1)$ for thin walls or obstacles, where N_{wall} is the number of obstacles between the RF AP and the device.

Let M denote the number of the adjacent industrial factory rooms (or adjacent cells) and let m denote the mth adjacent industrial factory room. For downlink, if the kth device is assigned to the RF AP on the nth subchannel ($n \in \mathcal{N}^{\mathrm{RF}}$), the received SINR of the device is given by

$$\gamma_{k,n}^{\mathrm{RF,D}} = \frac{P_n^{\mathrm{RF,D}} g_{k,n}^{\mathrm{RF}}}{\sum_{m=1}^{M} P_{n,m}^{\mathrm{RF,D}} g_{k,n,m}^{\mathrm{RF}} + N_0^{\mathrm{RF}} B_{sub}^{\mathrm{RF}} + I_{k,n}^{\mathrm{RF,D}}} \quad (6)$$

where $P_n^{\mathrm{RF,D}}$ and $P_{n,m}^{\mathrm{RF,D}}$ are the allocated transmit power on the nth subchannel of the corresponding RF AP and the mth adjacent RF AP, respectively. $g_{k,n,m}^{\mathrm{RF}}$ is the RF interference channel gain from the RF AP in the mth adjacent RF cell to the kth device. N_0^{RF} represents the PSD of noise at the receiver, B_{sub}^{RF} is the subchannel bandwidth $B_{sub}^{\mathrm{RF}} = B^{\mathrm{RF}}/N^{\mathrm{RF}}$ with B^{RF} being the RF AP bandwidth. $I_{k,n}^{\mathrm{RF,D}}$ is the interference from competing technologies operating over the same band. In this article, we assume that there exists one RF transmitter in an adjacent factory room under another competing technology operating over the same band [34].

For uplink, the received signal-to-noise ratio (SNR) at the RF AP from the kth device on the nth subchannel is

$$\gamma_{k,n}^{\mathrm{RF,U}} = \frac{P_{k,n}^{\mathrm{RF,U}} g_{k,n}^{\mathrm{RF}}}{\sum_{m=1}^{M} P_{k',n,m}^{\mathrm{RF,U}} g_{k',n,m}^{\mathrm{RF}} + N_0^{\mathrm{RF}} B_{sub}^{\mathrm{RF}} + I_{k,n}^{\mathrm{RF},U}} \quad (7)$$

where $P_{k,n}^{\mathrm{RF,U}}$ and $P_{k',n,m}^{\mathrm{RF,U}}$ are the transmit power of the kth device on subchannel n in its associated cell and the the k'th device on subchannel n in the mth adjacent RF cell, respectively. $g_{k',n,m}^{\mathrm{RF}}$ is the RF interference channel gain from the k'th device in the mth adjacent RF cell to the current RF AP. $I_{k,n}^{\mathrm{RF},U}$ is the interference from competing technologies operating over the same band [34]. Hence, the achievable data rates of downlink and uplink are defined as

$$R_k^{\mathrm{RF,D}} = \sum_{n \in \mathcal{N}^{\mathrm{RF}}} \rho_{k,n}^{\mathrm{RF,D}} B_{sub}^{\mathrm{RF}} \log_2(1 + \gamma_{k,n}^{\mathrm{RF,D}}) \quad (8)$$

$$R_k^{\mathrm{RF,U}} = \sum_{n \in \mathcal{N}^{\mathrm{RF}}} \rho_{k,n}^{\mathrm{RF,U}} B_{sub}^{\mathrm{RF}} \log_2(1 + \gamma_{k,n}^{\mathrm{RF,U}}) \quad (9)$$

respectively, where $\rho_{k,n}^{\mathrm{RF,D}}$ and $\rho_{k,n}^{\mathrm{RF,U}}$ are the channel assignment indicators, and they are binary values of "1" or "0."

III. Industrial Network Requirements and Problem Formulation

In this section, we formulate the energy-efficient resource management problem (joint network selection, subchannel assignment, and power management) in the heterogeneous RF/VLC industrial network with the objective of maximizing the network EE while guaranteeing the QoS requirements of IIoT or IoT devices. We take these practical requirements into account as constraints in the mathematical way, and the decision making problem is modeled as a MDP [21]–[27].

A. Requirements of IIoT and IoT Devices

1) URLLC Requirements: The real-time industrial control applications (URLLC services) have strict latency and transmission reliability requirements, but they are not interested in the high data rate. This subsection investigates how to model the URLLC requirements in a mathematical way.

For URLLC services, we assume that the kth IIoT device or transmitter follows the independent and identically distributed Poisson distribution with the packet arrival rate λ and the data packet size L^{packet} in bytes [27]. Generally, the total latency (T_{l}) of one packet consists of the waiting time of the packet to be served in the queue (T_{w}), the transmission time (T_{t}), the channel access delay (T_{a}), the backhaul delay (T_{b}), the packet reception delay (T_{r}), and the processing delay (T_{p}), which can be expressed as [36]

$$T_{\text{l}} = T_{\text{w}} + T_{\text{t}} + T_{\text{a}} + T_{\text{b}} + T_{\text{r}} + T_{\text{p}}. \tag{10}$$

In (10), the transmission time of one packet is calculated by $T_{\text{t}} = L^{\text{packet}}/R^{\text{d}}$, where R^{d} is the achievable link data rate.

Due to the latency requirement, each packet in URLLC should be successfully transmitted in a limited time duration. Let $T_{\max}$ denote the maximum tolerable latency threshold of each transmission packet, the latency constraint can be guaranteed by controlling the probability of T_{l} exceeding the threshold value $T_{\max}$, which can be expressed as

$$p^{\text{Lat}} = \Pr\{T_{\text{l}} \geq T_{\max}\} \leq p^{\text{Lat}}_{\max} \tag{11}$$

where $p^{\text{Lat}}_{\max}$ denotes the maximum violation probability.

In this article, the outage probability is used to characterize the reliability requirement and it can be defined as the probability that the received SINR ($\gamma_{k,n}$) at the receiver is lower than the target threshold $\gamma^{\text{tar}}_{k,n}$. Then, the requirement on the reliability is satisfied by controlling the outage probability, $\Pr\{\gamma_{k,n} < \gamma^{\text{tar}}_{k,n}\}$. And the outage probability cannot beyond the violation probability $p^{\text{Rel}}_{\max}$, which can be given by

$$p^{\text{Rel}}_{m} = \Pr\left\{\gamma_{k,n} < \gamma^{\text{tar}}_{k,n}\right\} \leq p^{\text{Rel}}_{\max} \tag{12}$$

2) Minimum Data Rate Requirements: As illustrated above, for the normal services, some IIoT devices and IoT devices may require the high data rates, though the latency is of less significance. Hence, the minimum data rate requirements of these devices should be considered in resource management. Let R_k denote the kth device' current data rate, the minimum data rate requirement can be satisfied by controlling the probability of the unsatisfied normal service (p^{Nor}_k), where R_k fails to achieve its minimum data rate threshold $R^{\min}_k$, which can be given by

$$p^{\text{Nor}}_k = \Pr\{R_k < R^{\min}_k\} \leq p^{\text{Nor}}_{\max} \tag{13}$$

where $p^{\text{Nor}}_{\max}$ denotes the maximum violation probability.

B. Problem Formulation

The total achievable data rate and the total power consumption can be calculated as

$$R = \sum_{k\in\mathcal{K}}\sum_{c\in\mathcal{C}} \alpha_{k,c} R^{\text{VLC}}_{k,c} + \sum_{k\in\mathcal{K}} \beta_k R^{\text{RF,D}}_k + \sum_{k\in\mathcal{K}} \beta_k R^{\text{RF,U}}_k \tag{14}$$

$$P = C P^{\text{VLC}}_{\text{fix}} + P^{\text{RF}}_{\text{fix}} + \sum_{k\in\mathcal{K}}\sum_{c\in\mathcal{C}}\sum_{n\in\mathcal{N}^{\text{VLC}}} \alpha_{k,c}\rho^{\text{VLC}}_{k,n,c} P^{\text{VLC}}_{n,c} \times \sum_{k\in\mathcal{K}}\left(\sum_{n\in\mathcal{N}^{\text{RF}}}\left(\rho^{\text{RF,D}}_{k,n} P^{\text{RF,D}}_{n} + \rho^{\text{RF,U}}_{k,n} P^{\text{RF,U}}_{k,n}\right) + P_{\text{cir}}\right) \tag{15}$$

respectively, where $\alpha_{k,l}$ and β_k denote the association between a device and a VLC AP or a RF AP, respectively, both having binary values of "1" or "0" to indicate that there exists a selection or no selection exists. In addition, $P^{\text{RF}}_{\text{fix}}$ and $P^{\text{VLC}}_{\text{fix}}$ denote the fixed power consumption of the RF AP and each VLC AP, respectively, resulting from the AP hardware power consumption (circuit operation, data processing, backhaul connection, etc.). Note that $P^{\text{VLC}}_{\text{fix}}$ also includes the optical power using for illumination. P_{cir} is the circuit power consumption of one device.

Our goal is to maximize the network EE (the radio of the overall data rate and the total power consumption: $\eta_{EE} = R/P$) while ensuring the mentioned QoS requirements of devices in Section III-A. In this article, we present a utility function (also called reward function) in the heterogeneous industrial network, which can be expressed as

$$r = \eta_{EE} - \mu_1 \sum_{k\in\mathcal{K}} p^{\text{Lat}}_k - \mu_2 \sum_{k\in\mathcal{K}} p^{\text{Rel}}_k - \mu_3 \left(\sum_{k\in\mathcal{K}} p^{\text{Nor}}_k\right) \tag{16}$$

where the part 1 is the network benefit (the overall EE in Kbit/J), the part 2, part 3, and part 4 are the cost functions in terms of the unsatisfied latency, unsatisfied reliability, and unsatisfied minimum data rate requirements, respectively. The coefficient μ_i, $i \in \{1, 2, 3\}$ are the weights of the last three parts, which are used to balance the benefit and the cost.

Similar to the works [22]–[28], we adopt MDP to model the intelligent resource management decision making in probabilistic or deterministic environments based on the requirements of systems [37]. Generally, MDP can be defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \xi)$, where the main elements of the MDP can be defined as the following:

Agents: The RF AP, VLC APs, and active devices.

State space $\mathcal{S}$: In the heterogeneous industrial network, the network state can be defined as the subchannel occupy status (idle or busy), the channel quality (SINR value), the service

application types (normal services (low priority) and URLLC services (high priority), and service satisfaction (reliability, latency, and minimum data rate).

Action space $\mathcal{A}$: In each time slot, the agent will take the action $a \in \mathcal{A}$ according to the current state s, where the action includes the VLC or RF AP selection, the subchannel assignment and the transmit power management.

Transition probability $\mathbb{P}$: The transition probability $\mathbb{P}(s'|s,a)$ captures the probability that the agent takes the action $a \in \mathcal{A}$ from the state $s \in \mathcal{S}$ to a new state $s' \in \mathcal{S}$.

Reward r: After taking one action, the agent will obtain an immediate reward r from the environment where the learning process is driven by the reward. We have built the reward function shown in (16), which decides that the policy that the agent finds. $\xi \in [0,1)$ is a discount factor.

Policy: The policy is a function that can be deterministic or stochastic, which decides the the action selection with the given state. Let $\pi(s)$ denotes a policy: $\pi(s): \mathcal{S} \rightarrow \mathcal{A}$, which is a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$.

In heterogeneous industrial network, each agent tries to search the policy $\pi(s)$ to improve its immediate reward r. Let $V^{\pi}(s)$ denotes the value function, which is also a cumulative discounted reward, and it can be calculated by

$$V(s) = E_{\pi}\left\{\sum_{t=1}^{\infty}\gamma^{t} r(s_t,a_t)|s_0 = s\right\} = E_{\pi}\left\{r(s,a_t) + \xi\int_{s'\in S}(s'|s,a)V(s')ds'\right\}. \quad (17)$$

The optimal strategy $\pi^*(s)$ can be achieved by satisfying the Bellman equation. $V^*(s) = \max_{a\in A} V(s)$ [36]. Once the optimal strategy $\pi^*(s)$ is achieved by maximizing the cumulative reward from the beginning, it implements the intelligent resource management in heterogeneous industrial networks.

Q-learning is a well-known RL algorithm for policy learning in wireless networks. Let $Q(s,a)$ denote the Q-function of the state–action pair (s, a), which is also the expected utility. The value function $V(s)$ is the maximum Q-function over the feasible actions at the sate s. The Q-function can be updated at the end of each time stage, which is

$$Q_{t+1}(s_t,a_t) = (1-\alpha_t)Q_t(s_t,a_t) + \alpha_t\left[r(s_t,a_t) + \xi V_t(s_{t+1})\right] \quad (18)$$

where $\alpha_t \in (0,1]$ is a time-varying learning rate. When the learning rate α_t admits $\sum_{t=1}^{\infty}\alpha_t = \infty$ and $\sum_{t=1}^{\infty}\alpha_t^2 < \infty$, then the Q-function $Q_t(s,a)$ will converge to the optimal value $Q^*(s,a)$ by $V_t(s_t) = \max_{a\in A} Q_t(s_t,a_t)$ [38].

IV. Proposed Deep PDS-ERT-Based Intelligent Resource Management

As illustrated in the above section, the policy can be numerically learned by adopting the Q-learning, policy gradient schemes, and deep Q-network (DQN) algorithms [38]. However, Q-learning cannot deal with continuous state–action spaces and the policy gradient may converge to the local optimal position. Although DQN has the ability to handle the continuous control problem under high-dimensional sensory inputs, its nonlinear function approximator is known to be unstable or even to diverge. Moreover, it also needs a large number of training samples to guarantee the training efficiency, all the abovementioned factors may limit the application in IIoT networks.

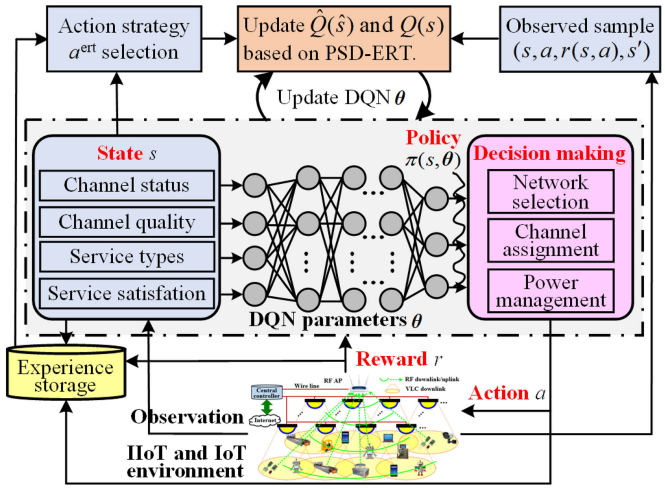

Fig. 2. Deep PDS-ERT learning-based intelligent resource management.

To overcome the above problems, we propose a deep PDS-ERT learning algorithm, as shown in Fig. 2, to accelerate the learning rate, enhance the learning efficiency, and ensure the learning stability toward the optimal policy for the resource management in the heterogeneous industrial network. In details, the agent can utilize the learned strategies from the historical experience and the other agents, and integrate the PDS-learning principle into the conventional DRL to learn the unknown dynamics. The main procedures of the proposed PDS-ERT learning-based intelligent resource management are presented in the following subsections.

A. Experience Replay and Transfer

In RL, the policy strategy $\pi(s,a)$ determines the resource management strategy in heterogeneous industrial networks. In order to improve the learning efficiency, a modified experience replay and transfer mechanism is presented for policy learning by utilizing the historical knowledge or using the learned knowledge from other agents.

1) Policy Strategy Selection: One of the important processes of the experience replay and transfer mechanism is that how to find the most useful learned policy strategy (e.g., network selection, subchannel assignment, and power management) from the historical knowledge, or search one agent as the expert agent to utilize the learned policy strategy from the expert. Instead of blindly searching the expert agent or the historical experience [31], the agent calculates the similarity level between the current agent and other active agents (or the historical knowledge) by evaluating the following three metrics: 1) service information, which refers to URLLC services and the normal services; 2) the device information, which includes the device position and

mobility behavior; 3) the channel information, which contains the channel SINR values and assignment indicators, etc.

The mentioned similarity can be calculated by applying the Bregman ball method [38], where Bregman ball acts as the minimum manifold with a central $Z_{\rm cen}$, and a radius $R_{\rm rad}$. Any information point $Z_{\rm poi}$ is inside this ball, and the agent searches the information point which has the most strong similarity with $Z_{\rm cen}$. The distance between any information point and the central $Z_{\rm cen}$ is expressed as [39]

$$B(Z_{\rm cen}, R_{\rm rad}) = \{Z_{\rm poi} \in Z : D(Z_{\rm poi}, Z_{\rm cen}) \leq R_{\rm rad}\} \quad (19)$$

where $D(a, b)$ is the Bregman divergence [39], which is also the manifold distance between two data points. Once the highest similarity value between the learning agent and the expert agent or historical information is achieved, the learning agent can utilize the policy strategy.

2) Overall Action Strategy: As analyzed above, after finding the most suitable historical policy or transferred policy strategy by adopting the experience replay and transfer mechanism [40], the agent utilizes the learned action strategy $a^{\rm ert}$ and its current native action $a^{\rm na}$ to generate an overall action. Accordingly, the overall action can be selected by

$$a^{\rm ov} = \varsigma a^{\rm ert} + (1 - \varsigma) a^{\rm na} \quad (20)$$

where $\varsigma \in [0, 1]$ denotes the transfer rate, which will be reduced after every learning step to gradually remove the effect of the historical policy information on the new policy.

3) Experience Collection: In order to avoid storing the unreliable experience, after interacting with the environment, the learned experience $e_t = (s_t, a_t, r_t, s_{t+1})$ with the best reward is recorded in the relay memory. If the capacity of the relay is full, the relay memory will make room for the new collected experience by the following two steps:

1) *Experience combination:* We combine some historical experience data into one data point if they have similar functions by using the the Bregman ball concept [39].
2) *Experience deletion:* If the capacity of the memory is full and the new collected experience needs to be stored in the memory, the least used historical experience is deleted from the memory, because the least used experience makes a tiny contribution to the learning process.

B. Deep PDS-ERT Learning-Based Resource Management

In this subsection, deep PDS-ERT is developed by incorporating the experience replay and transfer mechanism into the deep PDS-learning algorithm. In particular, instead of directly using the selected native action strategy $a_t^{\rm na}$ to update the Q-function $Q(s_t)$, the historical or transferred action strategy $a_t^{\rm ert}$ can be utilized to exploit the extra information to improve learning speed and efficiency. Similar to the classical PDS [27], PDS-ERT can be described as the immediate network state that is achieved after the known information occurs, but before the unknown information takes place.

After achieving the corresponding overall action $a_t^{\rm ov}$ by (20), each deep PDS-ERT learning agent obtains an immediate known reward $r_{\rm k}(s_t, a_t^{\rm ov})$ and then the state s_t transits to the post-decision state $\hat{s}_t$ ($\hat{s}_t \in \hat{S}$ with $\hat{S}$ being the set of PDS-ERT) with a known transition probability $\mathbb{P}_{\rm k}(\hat{s}_t | s_t, a_t^{\rm ov})$. Afterward, PDS-ERT transits the current state $\hat{s}_t$ to the next state s_{t+1} with an unknown transition probability $\mathbb{P}_{\rm u}(s_{t+1} | \hat{s}_t, a_t^{\rm ov})$ and an unknown reward $r_{\rm u}(\hat{s}_t, a_t^{\rm ov})$ feedback to the agent. Mathematically, the transition probability from the current state s_t to the next state s_{t+1} with PDS-ERT admits

$$\mathbb{P}(s_{t+1} | s_t, a_t^{\rm ov}) = \int_{\hat{s} \in \hat{\mathcal{S}}} \mathbb{P}_{\rm u}(s_{t+1} | \hat{s}_t, a_t^{\rm ov}) \mathbb{P}_{\rm k}(\hat{s}_t | s_t, a_t^{\rm ov}) d\hat{s} \quad (21)$$

The reward consists of the known and unknown rewards at $\hat{s}_t$ and s_{t+1}

$$r(s_t, a_t^{\rm ov}) = r_{\rm k}(s_t, a_t^{\rm ov}) + \int_{\hat{s} \in \hat{\mathcal{S}}} \mathbb{P}_{\rm k}(\hat{s}_t | s_t, a_t^{\rm ov}) r_{\rm u}(\hat{s}_t, a_t^{\rm ov}) d\hat{s}. \quad (22)$$

Then, the PDS-ERT quality Q-function with the PDS-ERT state–action pair $(\hat{s}_t, a_t^{\rm ov})$ and the general Q-function can be expressed as

$$\hat{Q}_t(\hat{s}_t, a_t^{\rm ov}) = r_{\rm u}(\hat{s}_t, a_t^{\rm ov}) + \int_{s_{t+1} \in \mathcal{S}} \mathbb{P}_{\rm u}(s_{t+1} | \hat{s}_t, a_t^{\rm ov}) V_t(s_{t+1}) ds \quad (23)$$

$$Q_t(s_t, a_t^{\rm ov}) = r_{\rm k}(\hat{s}_t, a_t^{\rm ov}) + \int_{\hat{s} \in \hat{\mathcal{S}}} \mathbb{P}_{\rm k}(\hat{s}_t | s_t, a_t^{\rm ov}) \hat{Q}_t(\hat{s}_t, a_t^{\rm ov}) d\hat{s}. \quad (24)$$

After obtaining the sample $[s_t, a_t, r_{\rm k}(\hat{s}_t, a_t^{\rm ov}), \hat{s}_t, r_{\rm u}(\hat{s}_t, a_t^{\rm ov}), s_{t+1}]$, the PDS-ERT quality value function is updated

$$\hat{Q}_{t+1}(\hat{s}_t, a_t^{\rm ov}) = (1 - \alpha_t) \hat{Q}_t(\hat{s}_t, a_t^{\rm ov}) + \alpha_t [r_{\rm u}(\hat{s}_t, a_t^{\rm ov}) + \xi V_t(s_{t+1})]. \quad (25)$$

After obtaining $\hat{Q}_{t+1}$ in (25), Q_{t+1} can be updated in (24) by replacing $\hat{Q}_t$ by $\hat{Q}_{t+1}$.

According to the above presented PDS-ERT, the deep PDS-ERT learning algorithm is presented to solve the intelligent resource management problem. As shown in Fig. 2, in the proposed deep PDS-ERT learning algorithm, at each time stage, after updating (23) and (24) on the overall action $a_t^{\rm ov}$ and the observed sample $(s_t, a_t, r(s_t, a_t), s_{t+1})$ by PDS-ERT, the classical DQN is applied to approximate the Q-function $Q(s_t, a_t^{\rm ov}, \boldsymbol{\theta}_t)$ of $Q(s_t, a_t^{\rm ov})$ through minimizing the following loss function at each time stage:

$$L_t(\boldsymbol{\theta}_t) = \{r(s_t, a_t^{\rm ov}) + \xi \max_{a \in \mathcal{A}} Q_t(s_{t+1}, a_{t+1}^{\rm ov}, \boldsymbol{\theta}_t) - Q_t(s_t, a_t^{\rm ov}, \boldsymbol{\theta}_t)\}^2. \quad (26)$$

One key feature of using DQN is to sample the loss functions in (26) at each iteration to reduce the computational cost for the large-scale-state learning problems [25], [26]. The procedures to implement DQN can be found in [25], [26].

The DQN parameters $\boldsymbol{\theta}$ can be achieved by applying the gradient descent method, which is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \beta_{\boldsymbol{\theta}_t} \nabla Loss_t(\boldsymbol{\theta}_t) \quad (27)$$

where $\beta_{\boldsymbol{\theta}_t}$ is the learning rate of the DQN parameters $\boldsymbol{\theta}_t$.

After that, each agent (RF AP, VLC AP, and device) will take the action based on the selected policy strategy $\pi(s_t, \boldsymbol{\theta_t})$, which can be achieved by

$$\pi(s_t, \boldsymbol{\theta}_t) = \arg\max_{a \in \mathcal{A}} \{Q_t(s_t, a_t^{\text{ov}}, \boldsymbol{\theta}_t)\}. \tag{28}$$

Theorem 1: The proposed PDS-ERT learning converges to the optimal point of the MDP when the learning rate α_t admits $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$.

Proof: If each action is executed under an infinite number of iterations, in other words, the learning policy is greedy with the infinite explorations, the function $Q(s, a)$ and the policy strategy $\pi(s, a)$ will gradually converge to the final points, respectively, with a probability of 1 [38], [39]. Due to space limitations, please see [24] and [27] for the full proof. ■

We denote the sets of the historical state space and action space in the memory as $\mathcal{S}'$ and $\mathcal{A}'$, respectively, and denote the current state space and action space as $\mathcal{S}$ and $\mathcal{A}$, respectively. At one decision stage, the sample complexity of the action selection and learning update of the classical Q-learning algorithm and DQN are $O(|\mathcal{S}| \times |\mathcal{A}|)$ and $O(|\mathcal{S}|^2 \times |\mathcal{A}|)$ [23], [24], [27], [36], respectively. As expected, the proposed deep PDS-ERT learning algorithm requires the historical learning experience. Here, the sample complexity of the action selection and learning update of the proposed deep PDS-ERT learning algorithm is $O(|\mathcal{S}'|^2 \times |\mathcal{A}'| + |\mathcal{S}|^2 \times |\mathcal{A}|)$ [23], [27], [36], which is relatively higher than that of the classical Q-learning algorithm and the DQN learning algorithm.

In addition to the abovementioned extra computational complexity, our proposed deep PDS-ERT learning algorithm needs a memory of $|\mathcal{S}'| \times |\mathcal{A}'|$ to store the historical learning knowledge, compared with the classical Q-learning algorithm and the DQN learning algorithm [23],[24], [27]. The proposed deep PDS-ERT learning algorithm-based intelligent resource management in heterogeneous RF/VLC industrial networks is shown in Algorithm 1.

C. Applications of the Presented Network Architecture and the Proposed Deep PDS-ERT Learning Algorithm

For the proposed solution, in addition to the use in the energy-efficient resource management for industrial IoT networks, it can be also applied for the indoor energy harvesting, indoor localization, and connection handover.

1) Indoor Energy Harvesting: In indoor industrial environments, there exists some energy-constrained devices, e.g., sensors for monitoring, humidity, and indoor air quality, etc. Hence, it is important to extend the lifetime of the devices due to their limited energy budget. In our presented heterogeneous RF/VLC industrial network, at each IoT or IIoT device, light energy harvesting is achieved by using PD and the harvested energy is used for sending data over the RF uplink [41].

2) Indoor Localization: Recently, VLC-based localization has obtained the attractive attention, because it provides the high positioning accuracy compared with the RF-based indoor positioning systems [42]. Hence, VLC-based localization in our presented heterogeneous RF/VLC industrial network is capable of realizing the indoor localization or navigation with the high positioning accuracy for IoT/IIoT devices in industrial networks [40].

3) Network Handover: In the heterogeneous RF/VLC industrial network, the presented heterogeneous network architecture and the proposed deep reinforcement learning algorithm have the ability to implement the vertical and horizontal handover processes to guarantee both the connectivity and QoS requirements of mobile IoT devices [43].

4) Safety-Critical Systems: Our presented heterogeneous RF/VLC architecture can be applied for the vehicular safety critical networks. Vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-everything (V2X) message exchanges are considered the preferred pattern for safety-critical communication (e.g., in anticollision active systems). LED-based VLC has been also proposed for V2V, V2I, and V2X message delivery [44], which can facilitate the safe driving by adaptive traffic signal control, intersection movement assistance, speed management, and so on.

V. Numerical Results and Analysis

In this section, simulation results are conducted in MATLAB 2017a to evaluate the performance of our presented heterogeneous RF/VLC industrial network and the proposed deep PDS-ERT learning-based intelligent resource management.

We consider an indoor industrial room with the area of $24 \times 24 \times 6$ m, where 6×6 VLC APs (uniform distribution) and a RF AP (locate in the center) are distributed at a height of 5 m. Additionally, the room is entirely covered by the RF AP. A number of devices ($K/2$ IIoT devices and $K/2$ IoT devices) are randomly distributed at four different heights (0.5, 1, 1.5, and 2 m). The RF AP has the carrier frequency of 2.4 GHz, the bandwidth of $B^{\text{RF}} = 10$ MHz, the number subchannels of $N^{\text{RF}} = 32$, the maximum transmit power of 250 mW, the fixed power consumption of $P_{\text{fix}}^{\text{RF}} = 6.7$ W, and the PSD noise of $N_0^{\text{RF}} = 173$ dBm/Hz [29]–[31]. Each VLC AP has the transmission bandwidth of $B^{\text{VLC}} = 20$ MHz (the available bandwidth is 10 MHz due to the Hermitian symmetry [29]–[31]), the number of subchannels $N^{\text{VLC}} = 16$, the maximum transmit electronic power of 250 W, the fixed power consumption of $P_{\text{fix}}^{\text{VLC}} = 4$ W and the PSD noise of $N_0^{\text{VLC}} = 10^{-21}$ A^2/Hz. Each device has the circuit power consumption of $P_{\text{cir}} = 5$ mW and the maximum transmit power of $P_{k,\max}^{\text{RF}} = 30$ mW. The LED lamp semiangle at half power and the Lambertian emission order are 60° and 1, respectively. The active area, the FOV, the concentrator refractive index and the responsivity of the PD are 1 cm^2, 90°, 1.5, and 0.5 A/W, respectively. The gain of the optical filter is 1.

For the URLLC services, we set the maximum latency threshold $T_{\max} = 1$ ms with $T_{\text{a}} + T_{\text{b}} = 0.1$ ms and $T_{\text{r}} + T_{\text{p}} = 0.3$ ms [35], the transmission reliability is 0.999 with each message size being 250 bytes and the SINR threshold is 5 dB. For the normal services, the minimum data rate is set as 3 Mbps in downlink and 0.5 Mbps in uplink. Each time slot is 1 ms. LOS blocking probability for both VLC and RF links is 0.05. We set $\mu_1 = \mu_2 = 10^5$ and $\mu_3 = 2 \times 10^4$ to balance the benefit and the costs in (16) [24], [27]. In RL, the discount parameter

Algorithm 1: Deep PDS-ERT Learning-Based Intelligent Resource Management.

Input: The discount factor ξ, IIoT, and IoT environment simulators and the samples of historical knowledge.

Initialize: The network state s_0, value function $V(s_0)$, policy strategy $\pi(s_0)$, and the DQN with parameters $\boldsymbol{\theta}_0$;
for each time step $t = 0, 1, 2, ...$ **do**
  The agent observes the state s_t;
  if the agent applies new services or has poor performance **then**
    Search the expert agent with the highest similarity;
    Obtain the transferred action strategy a_t^{ert} from the expert;
    Select the overall action by (20) and update the transfer rate ς;
    Perform deep PDS-ER from step 10 to step 17;
  else
  The agent selects the action a_t^{na} with a probability ε or choose a_t^{na} by satisfying $a_t^{\mathrm{na}} = \arg\max_{a\in\mathcal{A}} Q_t(s_t, a, \boldsymbol{\theta})$;
  Search the historical action a_t^{ert} with the highest similarity from the experience replay memory;
  With a_t^{na} and a_t^{ert}, calculate the overall action a_t^{ov} by (20);
  After executing action a_t^{ov}, the agent gets the reward $r(s_t, a_t^{\mathrm{ov}})$ and observes a new state s_{t+1} from the environment;
  The agent stores the experience $e_t = (s_t, a_t^{\mathrm{ov}}, r(s_t, a_t^{\mathrm{ov}}), s_{t+1})$ into its replay memory. If the capacity of the relay memory is full, the least used historical experience is dropped;
  Observe PDS tuple $(s_t, a_t^{\mathrm{ov}}, \hat{s}_t, r(s_t, a_t^{\mathrm{ov}}), s_{t+1})$, the agent updates the Q-function $\hat{Q}_t(\hat{s}_t, a_t^{\mathrm{ov}})$ and $Q_t(s_t, a_t^{\mathrm{ov}})$ by (23) and (24), respectively;
  Update the DQN parameters $\boldsymbol{\theta}_t$ by (27);
  Reset the DQN evaluation network by $\boldsymbol{\theta}_{t+1}$=$\boldsymbol{\theta}_t$;
  end if
end for
Output: RF/VLC network selection, subchannel assignment and power management.

$\xi = 0.98$ and the learning rate $a_t = 0.02$. The deep neural network (DNN) has three hidden layers with each hidden layer being with 50 neurons.

In this section, we present the performance comparisons of the following industrial networks: 1) our presented heterogeneous RF/VLC industrial network (denoted by RF/VLC); 2) the network service is performed using two RF APs (denoted by RF/RF) and the two carrier frequencies are 2.4 and 5 GHz, where the network total bandwidth is 20 MHz to ensure a fair comparison with the RF/VLC network. Moreover, we also compare the performance of our proposed deep PDS-ERT learning algorithm-based intelligent resource management with the following existing algorithms: 1) deep PDS learning [24] (denoted by Deep PDS); 2) Q-learning algorithm with knowledge transfer [31] (denoted by QKT-learning); 3) decomposing the optimization problem into two subproblems: a) network selection and subchannel assignment, b) transmit power management, and solve it iteratively in a centralized way, similar to [29] (denoted as Baseline 1).

Fig. 3 shows the EE per device, the probability of satisfied normal services, the average URLLC latency per packet and the reliability of URLLC services against the device density when the packet arrival rate is $\lambda = 0.12$ packets/slot/per IIoT source. As seen in Fig. 3(a), the higher the number of devices, the lower EE per device achieved, since the ICI becomes more pervasive in the VLC & RF networks which limits the data rate improvement, and the power consumption as well as the subchannel assignment increase in the VLC & RF networks under the high-density scenario of devices, leading to the EE degradation. From Fig. 3(b)–(d), the probability of the satisfied normal services and URLLC reliability decrease and the URLLC latency increases as the number of devices increase. This is because that under the fixed power and bandwidth resource, the large number of services need to be completed and different QoS requirements should to be guaranteed, the network may fail to support all the services' requirements, leading to bring down the performance in the high-density scenario. However, the presented heterogeneous network (RF/VLC) still outperforms the RF–RF network, and the proposed deep PDS-RET learning algorithm achieves the best performance among the existing algorithms.

We study in Fig. 4 how the performances vary with the packet arrival rate (λ) when $K = 160$. We can observe that the EE value increases with λ to a peak due to the increased network throughput when more packets transmit in the network. The power consumption also increases during this process, but the improvement rate of the network throughput is quite bigger than that of the power consumption, leading to EE enhancement. After that, the EE value slightly declines since continuing to increase λ will increase frequent connections and waiting time, which leads to more power consumption. In this case, the throughput enhancement fails to compensate the cost of consuming more total power, which slightly decreases the EE performance. It is worth noting that compared with RF/VLC, the performance of RF/RF is much sensitive to λ due to the limited bandwidth. Even the decreased performances happen with the increase of λ, our proposed deep PDS-RET learning algorithm still achieves the best performance.

Let us now quantify the effect of the blocking probability of RF&VLC links on the network performance, when $K = 160$ and λ =0.12 packets/slot/per IIoT source, as shown in Fig. 5. As seen in Fig. 5(a), when the blocking probability is increased, the EE performance obviously declines in RF&VLC networks while it is slightly reduced for RF/RF networks. This is because the blocked links in the VLC network unsuccessfully provides the high transmission data rate, while the effect of blocked links can be negligible in the RF network. From Fig. 5(b)–(d), the probability of the satisfied normal services and URLLC reliability decrease, and the URLLC latency increases during this

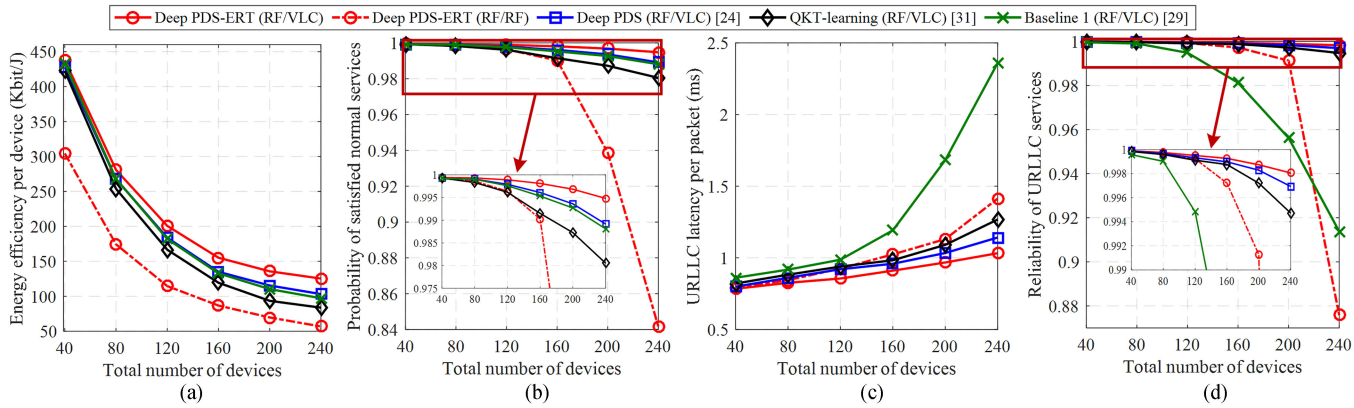

Fig. 3. Performance evaluations and comparisons with varying total numbers of devices.

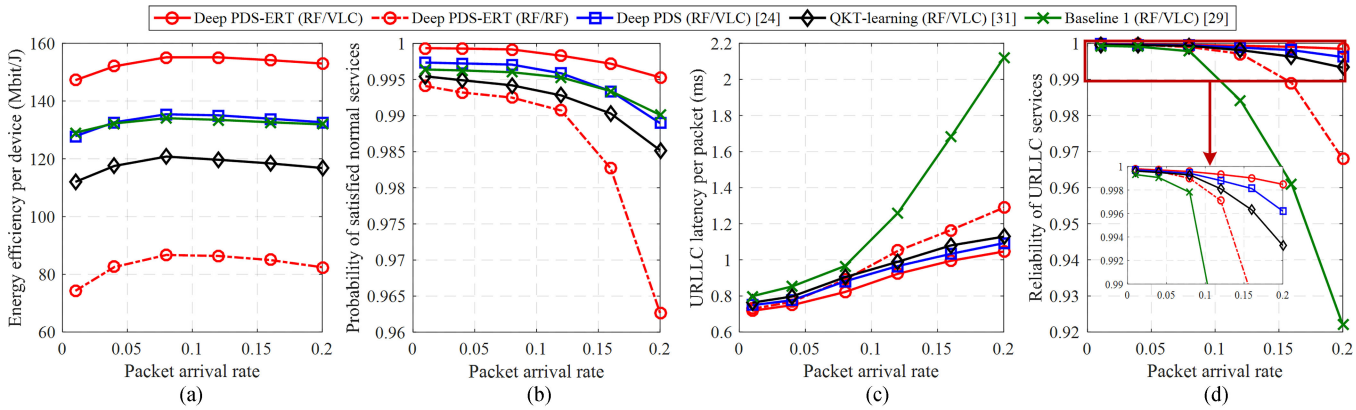

Fig. 4. Performance evaluations and comparisons versus packet arrival rate of URLLC services (packets/slot/per IIoT source).

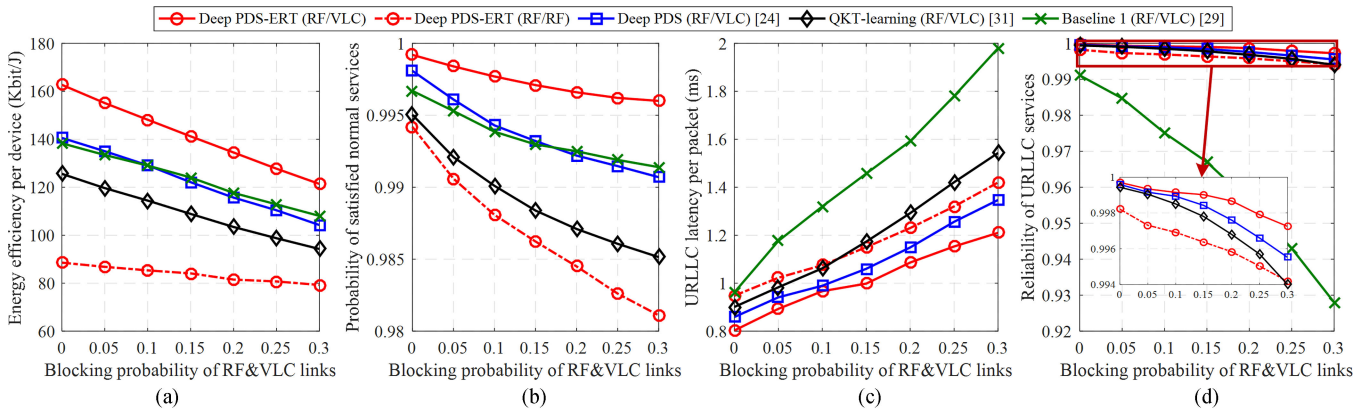

Fig. 5. Performance evaluations and comparisons versus blocking probability of RF&VLC links.

process, because the blockage degrades the received SINR value results in failing to guarantee the different QoS requirements of devices. However, for all blocking probabilities, our proposed solution still outperforms other solutions (network architecture and algorithms).

In Fig. 6, we show the learning process of the RL algorithms in terms of the reward when $K = 120$ and $\lambda = 0.12$ packets/slot/per IIoT source. Clearly, the deep PDS-RET and QKT-learning algorithms achieve the faster convergence than that of the deep PDS learning algorithm, but QKT-learning has the lowest performance in large-scale networks. The deep PDS-ERT learning algorithm achieves the best reward value, the fastest convergence and the most stability (less fluctuations) by utilizing the historical experience strategy to improve the learning efficiency and convergence speed, compared with other RL algorithms.

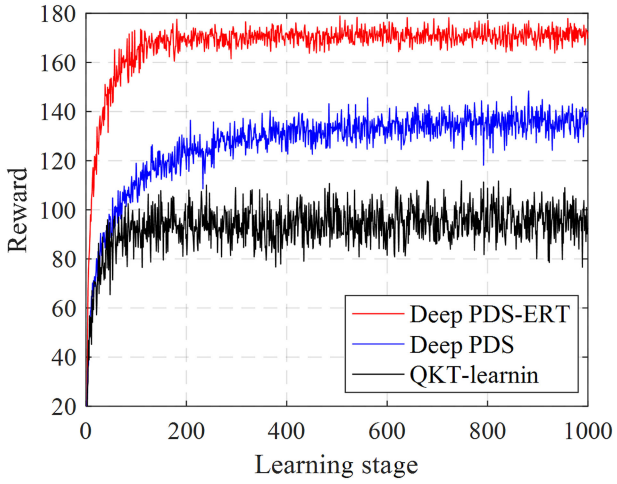

Fig. 6. Learning process comparisons of RL algorithms.

From Figs. 3–6, the proposed deep PDS-RET learning algorithm-based heterogeneous RF/VLC can effectively meet the energy-efficient communications, guarantee the strict URLLC requirements and ensure the high data rate demands at different scenarios in industrial networks.

VI. CONCLUSION

In this article, we presented a heterogeneous RF/VLC network architecture for wireless industrial networks to support different QoS requirements [ranging from high reliability and low latency (URLLC requirements) to high data rates] of IIoT and IoT devices. Based on the heterogeneous industrial network, we formulated an energy-efficient resource management decision-making problem (joint network selection, subchannel assignment, and power management) as a MDP, and a new deep PDS-ERT learning algorithm was proposed to learn the optimal policy for the intelligent resource management in heterogeneous industrial networks, which accelerates the learning rate and improves the learning efficiency. Simulation results verified the effectiveness of the presented heterogeneous RF/VLC industrial network and also showed that the proposed deep PDS-ERT learning algorithm outperforms other existing algorithms.

REFERENCES

[1] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the Internet of Things and Industry 4.0," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 17–27, Mar. 2017.

[2] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial Internet of Things: Challenges, opportunities, and directions," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018.

[3] C. Lu *et al.*, "Real-time wireless sensor-actuator networks for industrial cyber-physical systems," *Proc. IEEE*, vol. 104, no. 5, pp. 1013–1024, May 2016.

[4] A. A. Kumar S., K. Ovsthus, and L. M. Kristensen, "An industrial perspective on wireless sensor networks—A survey of requirements, protocols, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1391–1412, 2014.

[5] L. Liu and W. Yu, "A D2D-based protocol for ultra-reliable wireless communications for industrial automation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5045–5058, Aug. 2018.

[6] S. K. Sharma and X. Wang, "Towards massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning assisted solutions," *IEEE Commun. Surveys Tuts*.

[7] M. Weiner, M. Jorgovanovic, A. Sahai, and B. Nikolić, "Design of a low latency, high-reliability wireless communication system for control applications," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2014, pp. 3829–3835.

[8] A. E. Kalor, R. Guillaume, J. J. Nielsen, A. Mueller, and P. Popovski, "Network slicing in Industry 4.0 applications: Abstraction methods and end-to-end analysis," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5419–5427, Dec. 2018.

[9] S. Li, Q. Ni, Y. Sun, G. Min, and S. Al-Rubaye, "Energy-efficient resource allocation for industrial cyber-physical IoT systems in 5G era," *IEEE Trans. Ind. Informat.*, vol. 14, no. 6, pp. 2618–2628, Jun. 2018.

[10] "5G: Study on Scenarios and Requirements for Next Generation Access Technologies" (Release 14) document 38.913, 2017.

[11] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled IoT using natural actor-critic deep reinforcement learning," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2061–2073, Apr. 2019.

[12] C. Esposito, M. Ficco, A. Castiglione, F. Palmieri, and H. Lu, "Loss-tolerant event communications within industrial Internet of Things by leveraging on game theoretic intelligence," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1679–1689, Jun. 2018.

[13] P. Bartolomeu, M. Alam, J. Ferreira, and J. A. Fonseca, "Supporting deterministic wireless communications in industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 4045–4054, Sep. 2018.

[14] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, "Adaptive transmission optimization in SDN-based industrial IoT with edge computing," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1351–1360, Jun. 2018.

[15] K. Wang, Y. Wang, Y. Sun, S. Guo, and J. Wu, "Green industrial Internet of Things architecture: An energy-efficient perspective," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 48–54, Dec. 2016.

[16] L. Lyu, C. Chen, S. Zhu, and X. Guan, "5G enabled codesign of energy-efficient transmission and estimation for industrial IoT systems," *IEEE Trans. Ind. Informat.*, vol. 14, no. 6, pp. 2690–2704, Jun. 2018.

[17] N. B. Long, H. Tran-Dang, and D. -S. Kim, "Energy-aware real-time routing for large-scale industrial Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2190–2199, Jun. 2018.

[18] L. Song, K. K. Chai, Y. Chen, J. Schormans, J. Loo, and A. Vinel, "QoS-aware energy-efficient cooperative scheme for cluster-based IoT systems," *IEEE Syst. J.*, vol. 11, no. 3, pp. 1447–1455, Sep. 2017.

[19] Y. Luo, Y. Duan, W. Li, P. Pace, and G. Fortino, "A novel mobile and hierarchical data transmission architecture for smart factories," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3534–3546, Aug. 2018.

[20] D. A. Chekired, L. Khoukhi, and H. T. Mouftah, "Industrial IoT data scheduling based on hierarchical fog computing: A key for enabling smart factory," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4590–4602, Oct. 2018.

[21] T. Qiu, B. Li, W. Qu, E. Ahmed, and X. Wang, "TOSG: A topology optimization scheme with global-small-world for industrial heterogeneous Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3174–3184, Jun. 2019.

[22] Y. Li, K. K. Chai, Y. Chen, and J. Loo, "Smart duty cycle control with reinforcement learning for machine to machine communications," in *Proc. IEEE Int. Conf. Commun. Workshop*, 2015, pp. 1458–1463.

[23] C. H. Liu, Q. Lin, and S. Wen, "Blockchain-enabled data collection and sharing for industrial IoT with deep reinforcement learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3516–3526, Jun. 2019.

[24] X. He, R. Jin, and H. Dai, "Deep PDS-learning for privacy-aware offloading in MEC-enabled IoT," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4547–4555, Jun. 2019.

[25] M. Chu, H. Li, X. Liao, and S. Cui, "Reinforcement learning based multi-access control and battery prediction with energy harvesting in IoT systems," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2009–2020, Apr. 2019.

[26] X. He, K. Wang, H. Huang, T. Miyazaki, Y. Wang, and S. Guo, "Green resource allocation based on deep reinforcement learning in content-centric IoT," *IEEE Trans. Emerging Topics Comput.*, 2019.

[27] N. Mastronarde and M. van der Schaar, "Joint physical-layer and system-level power management for delay-sensitive wireless communications," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 694–709, Apr. 2013.

[28] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-Reliable and Low-Latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.

[29] H. Zhang, N. Liu, K. Long, J. Cheng, V. C. M. Leung, and L. Hanzo, "Energy efficient subchannel and power allocation for software-defined heterogeneous VLC and RF networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 658–670, Mar. 2018.
[30] J. Wang *et al.*, "Learning-aided network association for hybrid indoor LiFi-WiFi systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3561–3574, Apr. 2018.
[31] 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.
[32] P. K. Sharma, Y. Jeong, and J. H. Park, "EH-HL: Effective communication model by integrated EH-WSN and hybrid LiFi/WiFi for IoT," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1719–1726, Jun. 2018.
[33] L. I. Albraheem, L. H. Alhudaithy, A. A. Aljaser, M. R. Aldhafian, and G. M. Bahliwah, "Toward designing a Li-Fi-based hierarchical IoT architecture," *IEEE Access*, vol. 6, pp. 40811–40825, 2018.
[34] S. Y. Shin, H. S. Park, and W. H. Kwon, "Mutual interference analysis of IEEE 802.15.4 and IEEE 802.11b," *Comput. Netw.*, Vol. 51, no. 12, pp. 3338–3353, Aug. 2007.
[35] "IST-4-027756 WINNER II D1.1.2 V1.2. WINNER II Channel Models," Feb. 2008, [Online]. Available: https://www.cept.org/files/8339/winner2%20-%20final%20report.pdf
[36] M. Doudou, D. Djenouri, and N. Badache, "Survey on latency issues of asynchronous MAC protocols in delay-sensitive wireless sensor networks," *IEEE Commun. Surv. Tuts.*, vol. 15, no. 2, pp. 528–550, Apr. 2013.
[37] D. Zhao and Y. Zhu, "MECA near-optimal online reinforcement learning algorithm for continuous deterministic systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 346–356, Apr.–Jun. 2015.
[38] M. Wiering and M. van Otterlo, *Reinforcement learning: State-of-the-art*. Berlin Heidelberg, Germany: Springer, 2014.
[39] Y. Wu, F. Hu, S. Kumar, J. D. Matyjas, Q. Sun, and Y. Zhu, "Apprenticeship learning based spectrum decision in multi-channel wireless mesh networks with multi-beam antennas," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 314–325, Feb. 2017.
[40] H. Yang, P. Du, W. Zhong, C. Chen, A. Alphones, and S. Zhang, "Reinforcement learning based intelligent resource allocation for integrated VLCP systems," *IEEE Wireless Commun. Lett*.
[41] G. Pan, H. Lei, Z. Ding, and Q. Ni, "3-D hybrid VLC-RF indoor IoT systems with light energy harvesting," *IEEE Trans. Green Commun. Netw.*
[42] M. F. Keskin, A. D. Sezer, and S. Gezici, "Localization via visible light systems," *Proc. IEEE*, vol. 106, no. 6, pp. 1063–1088, Jun. 2018.
[43] X. Bao, X. Zhu, T. Song, and Y. Ou, "Protocol design and capacity analysis in hybrid network of visible light communication and OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1770–1778, May 2014.
[44] T. Yamazato *et al.*, "Image-sensor-based visible light communication for automotive applications," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 88–97, Jul. 2014.

Helin Yang (S'15) is currently working toward the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore.

His current research interests include wireless communication, visible light communication, Internet of Things, and resource management.

Mr. Yang serves as a Reviewer for IEEE international journals such as IEEE COMMUNICATIONS MAGAZINE, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, etc.

Arokiaswami Alphones (M'92–SM'98) received the B.Tech. degree in electronics from the Madras Institute of Technology, Chennai, India, in 1982, the M.Tech. degree in microwave and optical communication engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 1984, and the Ph.D. degree in optically controlled millimeter wave circuits from the Kyoto Institute of Technology, Kyoto, Japan, in 1992.

Since 2001, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include electromagnetic analysis on planar RF circuits and integrated optics, microwave photonics, metamaterial-based leaky wave antennas, and wireless power transfer technologies.

Wen-De Zhong (SM'03) received the B.E. degree in electronic engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1982, and the M.E. and Ph.D. degrees in electronics and information engineering from the University of Electro Communications, Tokyo, Japan, in 1990 and 1993, respectively.

He was a Senior Research Fellow with the Department of Electrical and Electronic Engineering, University of Melbourne, Australia, from 1995 to 2000. He joined Nanyang Technological University, Singapore, in 2000, as an Associate Professor and became a Full Professor in 2009, and is currently with the School of Electrical and Electronic Engineering. His current research interests include visible light communication/positioning, optical fiber communication systems and networks, optical access networks, and signal processing.

Chen Chen (S'13–M'19) received both the B.S. and M.Eng. degrees in optical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2010, and 2013, respectively, and the Ph.D. degree in optical communication from Nanyang Technological University, Singapore, in 2017.

He is currently a Tenured-Track Assistant Professor of Communication Engineering with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China. His research interests include visible light communications, Li-Fi, visible light positioning, optical access networks, and digital signal processing.

Xianzhng Xie (M'18) received the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2000.

He is currently a Professor of School of Optoelectronic Engineering, and Director of Chongqing Key Lab of Computer Network and Communication Technology, with the Chongqing University of Posts and Telecommunications (CQUPT), China. His research interests include multiple-input multiple-output (MIMO) precoding, cognitive radio networks, and cooperative communications.