Cell-Free Networking for Integrated Data and Energy Transfer: Digital Twin based Double Parameterized DQN For Energy Sustainability

Tingyu Shui, Jie Hu, *Senior Member, IEEE*, Kun Yang, *Fellow, IEEE*, Honghui Kang, Hua Rui, Bo Wang

***Abstract*—Cell-free networking enables full cooperation among distributed access points (APs). This paper focuses on reducing the long-term energy consumption of a cell-free network in the downlink integrated data and energy transfer (IDET) for achieving energy sustainability. The resultant design includes both the AP classification on a large time-scale and the beamforming of the APs on a small time-scale in order to simultaneously satisfy the IDET requirements of data users and energy users. For dealing with binary integer actions (AP classification) and continuous actions (beamforming) together, we innovatively propose a stable double parameterized deep-Q-network (DP-DQN), which can be enhanced by a digital twin (DT) running in the intelligent core processor (ICP) so as to achieve faster and more stable convergence. Therefore, the cell-free network may avoid suffering from performance fluctuation during the training process. The simulation results demonstrate that our DP-DQN exceeds in convergence compared to other benchmarks while guaranteeing an optimal solution.**

***Index Terms*—Cell-free networking, integrated data and energy transfer (IDET), mixed time-scale, beamforming, AP classification, deep reinforcement learning (DRL), double parameterized deep-Q-network (DP-DQN), digital twin (DT).**

I. Introduction

A. Backgrounds and Motivations

As future communication systems go to higher frequency bands, wireless networks will become denser than ever before in order to maintain effective coverage. However, powering these densely deployed access points (APs) may consume huge energy. Moreover, miniature devices will be massively deployed for realizing the vision of Internet of Everything (IoE). However, limited battery capacities cannot support their long-life operations [1]–[3]. In order to address these two issues, we have to achieve energy sustainability 1) by providing integrated data and energy transfer (IDET) services to miniature IoE devices and 2) by reducing network energy consumption [4].

Apart from conventional wireless data transfer (WDT), IDET can also provide on-demand wireless energy transfer (WET) towards low-power IoE devices [5]–[7]. Therefore, they may maintain long operational cycles. Traditional heterogeneous networks (HetNet) consist of base stations with different capabilities, which yields small-cells, pico-cells, macro-cells and etc. However, all these base stations cannot be coordinated together, which results in severe mutual interference [8]–[10]. By contrast, cell-free networking is capable of enabling full-cooperation among the densely deployed APs [11]. Its inherent spatial diversity may significantly increase the spectrum efficiency of the WDT [12], while compensating the channel attenuation of the WET [13].

Unfortunately, transceiving design for IDET in cell-free networks results in a large-scale optimization problem with numerous variables. Conventional algorithms with high complexity are not suitable in such fast-changing environment, they cannot achieve long-term performance optimization [14]. Hence, many deep learning (DL) based solutions are relied upon. Once well-trained, DL algorithms output optimal solutions in a polynomial complexity.

Though a well-trained DL algorithm operates efficiently, the time-consuming training process is frustrating. Thankfully as a new infrastructure in 6G [15]–[17], digital twin (DT) can be relied upon for speeding up the training process with the virtual but accurate environmental data generated by itself.

B. Related Works

Cell-free networking has attracted great attention [18]–[22]. Attarifar *et al.* [18] proposed a modified conjugate beamforming design, which effectively canceled the self-interference among users that shared an identical pilot sequence. Therefore, the signal-to-interference-ratio (SIR) was substantially improved. In [19], a more fundamental power allocation problem was solved by an iterative algorithm for maximizing the total EE. In their scheme, a sum-power constraint indicating the coordination cost of the distributed APs should not be violated while quality-of-service (QoS) constraints of all the users were satisfied. Apart from WDT, cell-free networking aided WET and IDET also attracted much attention. Specifically, Zhang *et al.* [20] firstly analyzed the performance of both WET and WDT towards IoE devices in a cell-free network. Moreover, in

The authors would like to thank the financial support of National Natural Science Foundation of China (No. 61971102, No. 62132004), that of Sichuan Science and Technology Program (No. 2022YFH0022), that of MOST Major Research and Development Project (Grant No.: 2021YFB2900204), that of Sichuan Major R&D Project (Grant No.: 22QYCX0168), that of the Municipal Government of Quzhou (Grant No.: 2022D031), and that of Key Research and Development Program of Zhejiang Province (No. 2022C01093).

Tingyu Shui and Jie Hu are with the Yangtze Delta Region Institute (Huzhou) and the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Huzhou, 313001, China, email: grady_uestc@126.com, hujie@uestc.edu.cn.

Kun Yang is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, email:kunyang@uestc.edu.cn.

Honghui Kang, Hua Rui, Bo Wang are with State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation, Shenzhen, 518055, China, e-mail: kang.honghui@zte.com.cn, rui.hua@zte.com.cn, wang.bo40@zte.com.cn.

[21], APs collaboratively offered IDET services to IoE devices. Tractable mathematical expressions were derived for downlink WET and uplink WDT performance. Moreover, optimal power control was found to maximize weighted sum of all the signal-to-interference-plus-noise ratios (SINRs) in the uplink. Furthermore, in [22], APs were classified as transmit-APs for providing IDET services in the downlink and as receive-APs for receiving information from uplink users.

Unfortunately the papers cited above all considered conventional algorithms. Their high complexity cannot cope with the fast-changing environment and they cannot achieve long-term performance optimization. Therefore, deep reinforcement learning (DRL) was relied upon for cell-free networks [23]–[25]. Specifically, Luo *et al.* [23] proposed a model-free method to design a power allocation scheme for maximizing the minimum rate among all the users, where deep deterministic policy gradient (DDPG) based algorithm was adopted to avoid overfitting. Moreover, network clustering and hybrid beamsteering were jointly designed by a DRL algorithm [24]. Furthermore, a scalable uplink beamforming scheme was investigated by exploiting a distributed distributional deterministic policy gradients (D4PG) algorithm in cell-free networks [25]. Nevertheless, in order to obtain good performance, the training of DRL based algorithms may take a very long time. During the training, the attainable performance of cell-free networks may not be satisfactory until the DRL based algorithms converges.

By invoking digital twin (DT) network, we may solve the training dilemma. DT networks may exactly emulate the operation of their physical twins, such as wireless channels and network topologies, and these characteristic can be completely reproduced in the DT network via state of the art techniques, such as generative adversarial network (GAN) [26]–[28]. Comprised of a generator network and a discriminator network, GAN is widely used in image processing [26]. However, the idea that generator produces authentic data while discriminator distinguishes is also suitable for channel prediction. In [27], Ye *et al.* proposed a conditional GAN to estimate instantaneous channel transfer function, and Xiao *et al.* [28] designed a ChannelGAN generating fake channel based on real counterpart. Many works deployed DRL algorithm in the DT to learn from its generating data [29]–[31], but some flaws are listed: 1) A perfect DT without estimation error is considered, which is impractical; 2) The promising data-generating capacity of DT is not expressed, which should have been considered in data-driven DRL algorithm.

Some drawbacks in the existing works are summarized as below: 1) Most of the works about cell-free networking only focused on instantaneous performance optimization, such as [19] and [20]. None of them considered long-term design objectives. 2) All the APs were assumed to have exactly the same functions, either WDT or IDET. However, this setting imposed heavy tele-traffic on the limited fronthaul, since data services had to be delivered to all the APs for enabling full-cooperation. Dynamic AP classification according to their WDT and WET functions was not considered. 3) Although some pioneers introduced DT [27], [28], none of them considered the DT in cell-free networking. 4) The effect of the DT on the system performance was always ignored, and the application of DT was not illustrated clearly. Neither its advantages nor disadvantages were demonstrated.

C. Novel Contributions

Against this background, our novel contributions are summarized as follows:

- A cell-free network supporting both energy users (EUs) and data users (DUs) is studied, which includes distributed APs and a central intelligent core processor (ICP). In order to achieve the energy sustainability in the cell-free network, the long-term network energy consumption of the whole cell-free network for downlink IDET is minimized by optimally classifying all the APs either in a WDT set or in a WET set and by optimally designing the transmit beamformers of all the APs.
- The long-term network energy consumption is comprised of both the energy consumed for updating the cell-free network and that for transmitting downlink signals, where the mixed time-scale design is considered along with long-term users' requirements. A novel double parameterized deep-Q-network (DP-DQN) is proposed to solve the above optimization problem with mixed integer and continuous variables. Moreover, a double-DNN structure is relied upon for improving the convergence.
- An imperfect digital twin (DT) of the physical cell-free network is invoked to assist the DP-DQN. The well-constructed DT generates virtual CSI with prediction errors, which is considered in the signal model and the problem formulation. The virtual CSI is relied upon for training the DP-DQN. Although we suffer from some performance degradation with the imperfect DT, it helps the DP-DQN algorithm to achieve a faster and more stable convergence.

The remainder of this paper is organized as follows. Section II depicts the system model, while the optimization problem is formulated in Section III. After introducing our DT based DP-DQN solution in Section IV, we present the simulation results in Section V. Finally, our paper is concluded in Section VI.

Notation: $(\cdot)^*$ stands for the conjugate operation of a matrix, $\mathrm{Tr}(\cdot)$ stands for the trace operation of a matrix and $|\cdot|$ represents the cardinality of a set.

II. System Model

A. Network Model

In a cell-free network of Fig. 1, L APs constitute a set of $\mathcal{L} = \{AP_l \mid l = 1, \cdots, L\}$. Every AP is equipped with N antennas. All the APs are connected to a centralized intellectual core processor (ICP) via the fronthaul with limited capacity. The locations of APs are fixed. The ICP jointly coordinates all the APs to support a data user (DU) set $\mathcal{K} = \{DU_k \mid k = 1, \cdots, K\}$ and an energy user (EU) set $\mathcal{M} = \{EU_m \mid m = 1, \cdots, M\}$. All the DUs and EUs are equipped wtih a single antenna. We ignore the mobility of DUs and EUs and assume their static location as they are generally some miniature sensors in our WDT and WET scenario [20].

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

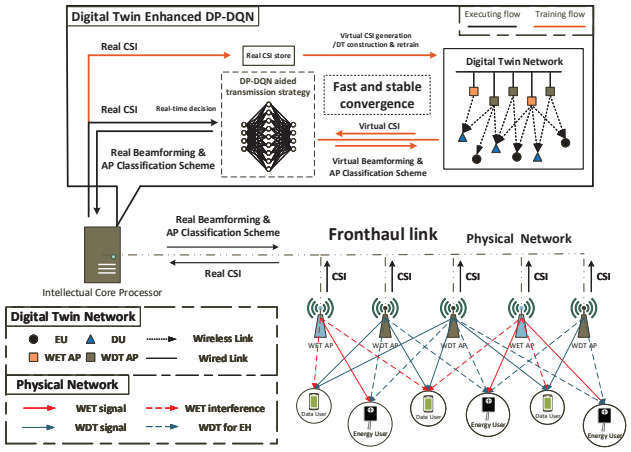


Fig. 1. A cell-free network empowered by Digital Twin based DP-DQN.

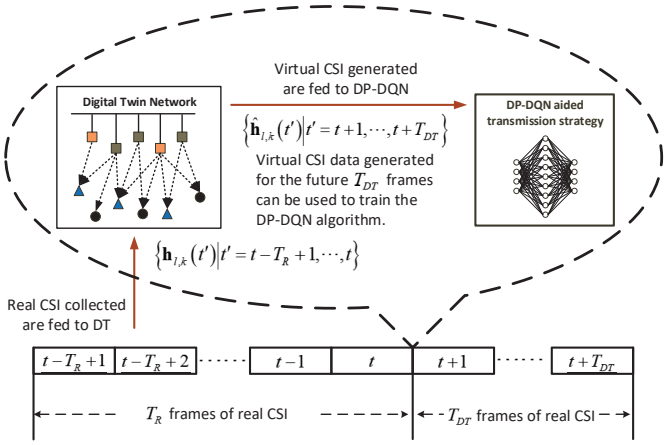


Fig. 2. Operation of DT in enhancing DRL base algorithm.

Given a transmission frame t, all the APs are dynamically classified into two sets, namely a set $\mathcal{L}_{WDT}(t)$ for downlink WDT towards the DUs, and a set $\mathcal{L}_{WET}(t)$ for downlink WET towards the EUs. Note that AP classification should not be instantaneously changed in every transmission frame so as to avoid excessive computational burdens on the ICP. After the AP classification, all the APs in the WDT set and those in the WET set may carefully form downlink beams for providing IDET towards both the DUs and the EUs. Note that the beamforming design should be dynamically adjusted in line with the instantaneous channel state information in every transmission frame. The instantaneous CSI in the physical network is collected at every transmission frame via the fronthaul. As a result, the ICP responses with a real-time transmission strategy. The APs in $\mathcal{L}_{WET}(t)$ only need to receive control signalling from the ICP for downlink WET, but those in $\mathcal{L}_{WDT}(t)$ further request for the data flows.

B. Time-Varying Channel

Let $\mathbf{h}_{l,k}(t) \in \mathcal{C}^{1\times N}$ denotes the channel between AP_l and DU_k for instance, we have

$$\mathbf{h}_{l,k}(t) = \sqrt{\frac{1}{\mathrm{PL}_{l,k}}}\mathbf{g}_{l,k}(t) = \lambda \mathbf{h}_{l,k}(t-1) + \sqrt{\frac{1-\lambda^2}{\mathrm{PL}_{l,k}}}\boldsymbol{\delta}, \quad (1)$$

which includes the path-loss $\mathrm{PL}_{l,k}$ [1], the small-scale fading $\mathbf{g}_{l,k}(t)$ [11], the temporal correlation factor λ [2] and the random difference $\boldsymbol{\delta} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ with an $N \times N$ identity matrix $\mathbf{I}_N$. The same channel model $\mathbf{h}_{l,m}(t)$ is applied for AP_l and EU_m. Moreover, we assume block fading for all the channels in the cell-free network.

C. Digital Twin

A DT of the physical cell-free network is implemented at the ICP, in which all the APs, DUs and EUs have their own digital versions as depicted in Fig. 1. Up-to-date real CSI is collected and stored by the ICP at each frame. The ICP utilizes the collected historical real CSI data to train the DT. Therefore, the DT is capable of generating virtual CSI data with high accuracy [32] to train the transmission strategy (DP-DQN algorithm), which may quickly converge. If the virtual CSI data generated by the DT becomes inaccurate, the latest real CSI data stored at the ICP can be used to re-train the DT. Moreover, when the ICP collects the instantaneous real CSI data from the distributed APs, they are fed to the well-trained DP-DQN algorithm, which returns us a timely transmission strategy of the physical cell-free network.

Due to its imperfection, the virtual CSI generated by the DT can be expressed as

$$\hat{\mathbf{h}}_{l,k}(t) = \sqrt{1-\zeta}\mathbf{h}_{l,k}(t) + \sqrt{\zeta}\boldsymbol{\delta}_{l,k}^{DT}, \quad (2)$$

where $\sqrt{\zeta}\boldsymbol{\delta}_{l,k}^{DT}$ is the random error between the virtual CSI $\hat{\mathbf{h}}_{l,k}(t)$ and the real CSI $\mathbf{h}_{l,k}(t)$. Note that ζ is the normalized error coefficient and $\boldsymbol{\delta}_{l,k}^{DT} \sim \mathcal{CN}(\mathbf{0}, \sqrt{\frac{1}{\mathrm{PL}_{l,k}}}\mathbf{I}_N)$ is a complex random vector. A well-constructed DT is capable of providing abundant virtual CSI data to accelerate the training of our DRL based algorithms (DP-DQN). The DT may generate T_{DT} frames of the virtual CSI data, when T_R frames of the real CSI data are uploaded from the physical cell-free network. The operation of DT is specified in Fig. 2.

Note that given a specific value of T_R, a longer T_{DT} may result in a higher ζ [33]. However, by exploiting the state of the art technique of time-series data generation, the mismatch between the real CSI and the virtual CSI can be controlled [28]. For example, given $T_R = 200$ frames of the real CSI data, the DT may generate as high as $T_{DT} = 500$ frames of the virtual CSI data with a super low bias [28]. Having the abundant virtual CSI data, the transmission strategy can be quickly trained to converge, since we do not need to wait for the collection of the real CSI data.

[1]Note that the movement of the users result in the dynamic distance between AP_l and DU_k, which can be captured by the path-loss $\mathrm{PL}_{l,k}$ in our channel model.

[2]The temporally correlated small-scale fading reflects the movement of the users. The temporal correlation factor λ is determined by the Doppler shift.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

III. Problem Formulation

A. WDT and WET Performance

Given a certain transmission frame t, we define $\boldsymbol{\omega}_{l,k}(t) \in \mathcal{C}^{N\times 1}$ as the beamformer from AP_l to DU_k. AP_l may be classified into either $\mathcal{L}_{WET}(t)$ or $\mathcal{L}_{WDT}(t)$. Therefore the beamforming matrix $\boldsymbol{\Omega}_l(t)$ of AP_l to be designed is expressed as

$$\boldsymbol{\Omega}_l(t) = \left(\boldsymbol{\omega}_{l,1}(t) \quad ,\cdots, \quad \boldsymbol{\omega}_{l,K}(t)\right), \tag{3}$$

which includes the beamformers towards all the DUs. The signal received by DU_k is expressed as

$$\begin{aligned} y_k(t) =& \sum_{l\in\mathcal{L}}\sum_{k'=1}^{K} \hat{\mathbf{h}}_{l,k}(t)\boldsymbol{\omega}_{l,k'}(t)x_{k'}(t) + n_k(t) \\ =& \underbrace{\sum_{l\in\mathcal{L}_{WDT}(t)} \sqrt{1-\zeta}\mathbf{h}_{l,k}(t)\boldsymbol{\omega}_{l,k}(t)x_k(t)}_{\text{desired signal}} \\ &+ \underbrace{\sum_{l\in\mathcal{L}_{WDT}(t)}\sum_{k'\neq k}^{K} \sqrt{1-\zeta}\mathbf{h}_{l,k}(t)\boldsymbol{\omega}_{l,k'}(t)x_{k'}(t)}_{\text{interference from WDT APs}} \\ &+ \underbrace{\sum_{l\in\mathcal{L}_{WET}(t)}\sum_{k'=1}^{K} \sqrt{1-\zeta}\mathbf{h}_{l,k}(t)\boldsymbol{\omega}_{l,k'}(t)x(t)}_{\text{interference from WET APs}} \\ &+ \underbrace{\sum_{l\in\mathcal{L}_{WDT}(t)}\sum_{k'=1}^{K} \sqrt{\zeta}\boldsymbol{\delta}_{l,k}^{DT}\boldsymbol{\omega}_{l,k'}(t)x_{k'}(t)}_{\text{DT channel error from WDT APs}} \\ &+ \underbrace{\sum_{l\in\mathcal{L}_{WET}(t)}\sum_{k'=1}^{K} \sqrt{\zeta}\boldsymbol{\delta}_{l,k}^{DT}\boldsymbol{\omega}_{l,k'}(t)x(t)}_{\text{DT channel error from WET APs}} + n_k(t), \end{aligned} \tag{4}$$

where $x_k(t) \sim \mathcal{CN}(0,1)$ is the modulated signal requested by DU_k, while $x(t) \sim \mathcal{CN}(0,1)$ denotes the dedicated downlink WET signal. Moreover, $n_k(t) \sim \mathcal{CN}(0,\sigma_k^2)$ represents the additive noise. Based on Eq. (4), the SINR of DU_k can be derived as:

$$\gamma_k(t) = \frac{(1-\zeta)\sum_{l\in\mathcal{L}_{WDT}}\left[\mathbf{h}_{l,k}(t)\boldsymbol{\omega}_{l,k}(t)\boldsymbol{\omega}_{l,k}^*(t)\mathbf{h}_{l,k}^*(t)\right]}{I_k + U_k + \sigma_k^2}, \tag{5}$$

where the signal interference I_k is expressed as

$$\begin{aligned} I_k =& (1-\zeta)\sum_{l\in\mathcal{L}_{WDT}(t)}\sum_{k'\neq k}^{K}\left[\mathbf{h}_{l,k}(t)\boldsymbol{\omega}_{l,k'}(t)\boldsymbol{\omega}_{l,k'}^*(t)\mathbf{h}_{l,k}^*(t)\right] \\ &+ (1-\zeta)\sum_{l\in\mathcal{L}_{WET}(t)}\sum_{k'=1}^{K}\left[\mathbf{h}_{l,k}(t)\boldsymbol{\omega}_{l,k'}(t)\boldsymbol{\omega}_{l,k'}^*(t)\mathbf{h}_{l,k}^*(t)\right]. \end{aligned} \tag{6}$$

Moreover, the uncertainty U_k between the virtual and the real CSI is expressed as

$$U_k = \sum_{l\in\mathcal{L}}\frac{\zeta}{\mathrm{PL}_{l,k}}\sum_{k'=1}^{K}\left[\boldsymbol{\omega}_{l,k'}(t)\boldsymbol{\omega}_{l,k'}^*(t)\right]. \tag{7}$$

Hence, given a bandwidth B, the throughput of DU_k is obtained as

$$R_k(t) = B\log_2\left(1+\gamma_k\right). \tag{8}$$

With a noise $n_m(t) \sim \mathcal{CN}(0,\sigma_m^2)$, the signal received by EU_m is expressed as

$$\begin{aligned} y_m(t) =& \underbrace{\sum_{l\in\mathcal{L}_{WET}(t)}\sum_{k=1}^{K}\hat{\mathbf{h}}_{l,m}(t)\boldsymbol{\omega}_{l,k}(t)x(t)}_{\text{signal from WET APs}} \\ &+ \underbrace{\sum_{l\in\mathcal{L}_{WDT}(t)}\sum_{k=1}^{K}\hat{\mathbf{h}}_{l,m}(t)\boldsymbol{\omega}_{l,k}(t)x_k(t)}_{\text{signal from WDT APs}} + n_m(t). \end{aligned} \tag{9}$$

Moreover, given a single transmission frame with a length of τ_s and a linear energy conversion efficiency μ_e [34], the amount of energy harvested by EU_m during the t-th transmission frame is given by [3]

$$E_m(t) = \tau_s\mu_e\left\{\sum_{l\in\mathcal{L}}\sum_{k=1}^{K}\left[\hat{\mathbf{h}}_{l,m}(t)\boldsymbol{\omega}_{l,k}(t)\boldsymbol{\omega}_{l,k}^*(t)\hat{\mathbf{h}}_{l,m}^*(t)\right] + \sigma_m^2\right\}. \tag{10}$$

The energy harvested by the EUs can be used for data sensing and data transmission [36], [37].

There are two types of information flows in the fronthaul. The first one is the data flow requested by all the DUs. The other one is the control signalling flow for notifying the APs about the AP classification and the beamforming scheme, which is negligibly small compared to the data flow. Note that data information flows requested by all the DUs may be delivered to all the APs in the set of $\mathcal{L}_{WDT}$. Their total throughput in the fronthaul is expressed as [38]

$$R_f(t) = |\mathcal{L}_{WDT}(t)|\sum_{k=1}^{K}R_k(t). \tag{11}$$

Note that $R_f(t)$ should not exceed the maximum capacity R_f^{max}.

B. Energy Consumption

The energy consumption of the cell-free network is constituted by signal transmissions and network updates. Note that the energy consumption of all the APs for signal transmissions in a single frame is formulated as

$$E_{AP}(t) = \tau_s\left\{\sum_{l\in\mathcal{L}}\sum_{k=1}^{K}\mathrm{Tr}\left[\boldsymbol{\omega}_{l,k}(t)\boldsymbol{\omega}_{l,k}^*(t)\right]\right\}. \tag{12}$$

When the AP classification and the beamforming are changed, energy is consumed for the corresponding network updates. As exemplified in Fig. 3, we first define a binary

[3]In line with [35], we reasonably assume that the impact of the inaccurate virtual CSI in the DT on the energy harvesting performance can be incorporated into the energy conversion efficiency μ_e.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

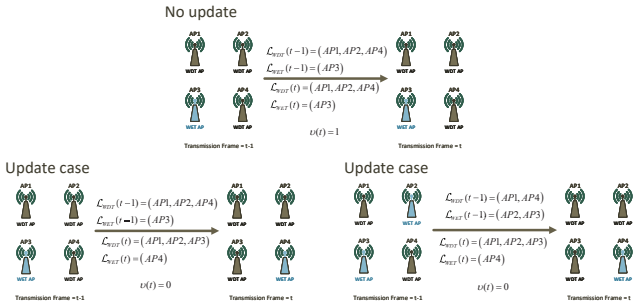


Fig. 3. AP classification and its updating.

indicator $\upsilon(t)$ to represent whether the AP classification is changed or not, which is expressed as

$$\upsilon(t)=\begin{cases}1, & \mathcal{L}_{WDT}(t)=\mathcal{L}_{WDT}(t-1)\\ & \text{and } \mathcal{L}_{WET}(t)=\mathcal{L}_{WET}(t-1),\\ 0, & \text{otherwise.}\end{cases} \tag{13}$$

Generally, we adopt a novel two-dimensional correlation coefficient [39] $r\left[\mathbf{A}(t),\mathbf{A}(t-1)\right]$ to evaluate the correlation between matrices $\mathbf{A}(t)$ and $\mathbf{A}(t-1)$, which is defined as Eq. (14) shown at the bottom of this page, where $\mathbf{A}(t)$ and $\mathbf{A}(t-1)$ have the same size and $(\cdot)^t_{mn}$ represents the element in the m-th row and the n-th column of the corresponding matrix at t-th frame, whereas $\overline{\mathbf{A}}(t)$ and $\overline{\mathbf{A}}(t-1)$ represent the average of all the elements in the corresponding matrices. Based on Eqs. (3) and (14), we define the correlation coefficient $\mu_l(t)$ between the beamforming matrices $\mathbf{\Omega}_l(t-1)$ and $\mathbf{\Omega}_l(t)$ of the l-th AP as

$$\mu_l(t)=\frac{r\left[\mathbf{\Omega}_l(t-1),\mathbf{\Omega}_l(t)\right]+1}{2},\forall l\in\mathcal{L}. \tag{15}$$

As a result, we define the energy consumption for changing the AP classification as ρ_υ in a signal transmission frame. Note that classifying distributed APs into either $\mathcal{L}_{WET}(t)$ or $\mathcal{L}_{WDT}(t)$ sets always results in an NP-hard integer-variable optimization problem. It may totally alter user association, beamforming design and fronthaul transmission for all the APs. Therefore, changing the AP classification consumes more energy in computing the optimization problem, in updating beamforming and delivering data content to APs. By contrast, only updating the beamforming design of an AP without any change on the AP classification thus consume less energy, which is defined as $\rho_{\mu,l}$. This energy consumption is for computing a low-complexity continuous-variable optimization problem, and for updating circuit components, such as converters, mixers and filters [40]. Then, the total energy consumption for the network update in the t-th transmission frame is formulated as

$$E_{update}(t)=\rho_\upsilon\left[1-\upsilon(t)\right]+\sum_{l=1}^{L}\rho_{\mu,l}\left[1-\mu_l(t)\right]. \tag{16}$$

Note that in a cell-free network, all the instantaneous real CSI data are uploaded to the ICP for central coordination and stored for training the DT to generate accurate virtual CSI data. If the virtual CSI data generated by the DT become inaccurate, the historical real CSI data stored at the ICP can be used to re-train the DT. Since the DT is also implemented at the ICP, there is actually no extra cost of transferring the historical CSI data to the DT. Moreover, compared to the data flows requested by the DUs in the fronthaul, the instantaneous real CSI data is negligibly small. As a result, we ignore the cost of uploading real CSI data in the fronthaul [25], [41].

Finally, the network energy consumption is $E_{net}(t)=E_{AP}(t)+E_{update}(t)$.

C. Optimization Problem

We aim for minimizing the long-term network energy consumption by optimizing the AP classification and their beamforming design in every transmission frame, while satisfying the DUs' WDT requirements and the EUs' WET requirements. Moreover, tele-traffic in the fronthaul should not exceed its maximum capacity. The optimization problem is formulated as

$$\text{(P1):}\quad \min_{\left\{\begin{array}{c}\mathcal{L}_{WDT}(t),\mathcal{L}_{WET}(t),\mathbf{W}(t),\\ \forall t=\{1,\cdots,T\}\end{array}\right\}} \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E_{net}(t) \tag{17}$$

$$\text{s.t.}\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R_k(t)\geq R_k^{min},\forall k\in\mathcal{K}, \tag{17a}$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E_m(t)\geq E_m^{min},\forall m\in\mathcal{M}, \tag{17b}$$

$$\begin{array}{c}\left\{\mathcal{L}_{WDT}(t)\cap\mathcal{L}_{WET}(t)=\emptyset\right\}\\ \cap\left\{\mathcal{L}_{WDT}(t)\cup\mathcal{L}_{WET}(t)=\mathcal{L}\right\}\end{array},t=\{1,\cdots,T\}, \tag{17c}$$

$$R_f(t)\leq R_f^{max},t=\{1,\cdots,T\}, \tag{17d}$$

$$\sum_{k=1}^{K}\mathrm{Tr}\left[\boldsymbol{\omega}_{l,k}(t)\boldsymbol{\omega}_{l,k}^{*}(t)\right]\leq P_{max}^{(l)},\forall l\in\mathcal{L},t=\{1,\cdots,T\}, \tag{17e}$$

where we have $\mathbf{W}(t)=[\mathbf{\Omega}_1(t),\cdots,\mathbf{\Omega}_L(t)]$ as an integrated matrix of size $N\times KL$. In (P1), Eqs. (17a) and (17b) represent the long-term average constraints on DUs' WDT requirements and EUs' WET requirements, respectively. Eq.

$$r\left[\mathbf{A}(t),\mathbf{A}(t-1)\right]=\frac{\sum_m\sum_n\left(\mathbf{A}^t_{mn}-\overline{\mathbf{A}}(t)\right)\left(\mathbf{A}^{t-1}_{mn}-\overline{\mathbf{A}}(t-1)\right)}{\sqrt{\left(\sum_m\sum_n\left(\mathbf{A}^t_{mn}-\overline{\mathbf{A}}(t)\right)^2\right)\left(\sum_m\sum_n\left(\mathbf{A}^{t-1}_{mn}-\overline{\mathbf{A}}(t-1)\right)^2\right)}}, \tag{14}$$

(17c) represents that all the APs are classified into either the WDT set $\mathcal{L}_{WDT}(t)$ or the WET set $\mathcal{L}_{WET}(t)$, while these two sets have no intersection. Eq. (17d) represents the throughput constraint on the fronthaul, while Eq. (17e) represents the instantaneous transmit power constraints on all the APs.

D. Problem Reformulation

According to Section III-C, we focused on energy consumption $\{E_{net}(t)\}$ in the long run, while guaranteeing the DUs' and EUs' long-term requirements, namely $\{R_k^{min}\}$ and $\{E_m^{min}\}$. Sequential decisions on the beamforming design $\{\mathbf{W}(t)\}$ and the AP classification $\{\mathcal{L}_{WDT}(t), \mathcal{L}_{WET}(t)\}$ are made according to the temporally correlated channel $\{\mathbf{h}_{l,k}(t)\}$ in Eq. (1). Specifically, when the CSI $\{\mathbf{h}_{l,k}(t-1)\}$ at the $(t-1)$-th transmission frame is given, $\{\mathbf{h}_{l,k}(t)\}$ at the t-th time-slot is only determined by $\{\mathbf{h}_{l,k}(t-1)\}$. This inherent Markov property enables us to formulate our optimization problem as a Markov decision process (MDP) [14] through state, action and reward design [4]. Therefore, a DRL based algorithm can be invoked as a solution. The specific state, action and reward design according to (P1) is given as follow:

1) State: Both the SINRs of K DUs and energy harvested by M EUs in the $(t-1)$-th frame are invoked to model the environmental state in the t-th frame, which is expressed as

$$s_t = \begin{bmatrix} \gamma_1(t-1), \cdots, \gamma_K(t-1), \\ E_1(t-1), \cdots, E_M(t-1) \end{bmatrix} \in \mathcal{S}, \tag{18}$$

where $\mathcal{S}$ represents the state space.

2) Action: In the optimization variables of (P1), namely $[\{\mathcal{L}_{WDT}(t), \mathcal{L}_{WET}(t)\}, \mathbf{W}(t)]$, the AP classification $\{\mathcal{L}_{WDT}(t), \mathcal{L}_{WET}(t)\}$ can be encoded as a binary vector $\mathbf{c}(t) = [c_1(t), \cdots, c_L(t)]$, where $c_l(t) = 0$ represents that AP_l is classified into the WET set $\mathcal{L}_{WET}(t)$, otherwise, it is classified into the WDT set $\mathcal{L}_{WDT}(t)$. Moreover, the indicator variable $v(t)$ equals to 1 when $\mathbf{c}(t-1) = \mathbf{c}(t)$ and 0 otherwise. A hybrid action then is obtained as

$$a_t = [\{\mathcal{L}_{WDT}(t), \mathcal{L}_{WET}(t)\}, \mathbf{W}(t)] = [\mathbf{c}(t), \mathbf{W}(t)] \in \mathcal{A}, \tag{19}$$

where $\mathcal{A}$ represents the hybrid action space. In a practical system, the beamformers $\mathbf{W}(t)$ are selected from a pre-designed discrete codebook. Our DP-DQN can degenerate to a traditional DQN to output discrete beamformers. However, the optimality of the performance is inevitably degraded.

3) Reward: The objective of (P1) is to minimize the long-term network energy consumption. However, predicting the long-term performance is difficult and unreliable. Therefore, a new metric replaces the original objective in the reward design, which is expressed as

$$\Delta E_{net}(t) = E_{net}(t) - \frac{1}{T_0} \sum_{t'=t-T_0}^{t-1} E_{net}(t'), \tag{20}$$

where T_0 is the length of a number of observation frames. Considering the network energy consumption $E_{net}(t)$ at frame t, it is compared to the average $\frac{1}{T_0}\sum_{t'=t-T_0}^{t-1} E_{net}(t')$ over the past T_0 transmission frames, which yields the following reward function:

$$\phi_o(t) = \begin{cases} \psi_o^+ \left|\Delta E_{net}(t)\right|, & \Delta E_{net}(t) \le 0, \\ -\psi_o^- \left|\Delta E_{net}(t)\right|, & \text{otherwise.} \end{cases} \tag{21}$$

As shown in Eq. (21), the cell-free network is encouraged to reduce the energy consumption $E_{net}(t)$ in the current frame t, which may return a positive reward. Therefore, we may heuristically reduce the long-term energy consumption.

Besides, we have to satisfy the series of constraints Eqs. (17a) - (17e) in (P1). We similarly define the reward related to the WDT performance of DU_k as

$$\phi_{WDT,k}(t) = \begin{cases} \psi_{WDT,k}^+ \left|R_k(t) - R_k^{min}\right|, & \text{(17a) holds,} \\ -\psi_{WDT,k}^- \left|R_k(t) - R_k^{min}\right|, & \text{otherwise,} \end{cases} \tag{22}$$

while that related to the WET performance of EU_m is defined as

$$\phi_{WET,m}(t) = \begin{cases} \psi_{WET,m}^+ \left|E_m(t) - E_m^{min}\right|, & \text{(17b) holds,} \\ -\psi_{WET,m}^- \left|E_m(t) - E_m^{min}\right|, & \text{otherwise.} \end{cases} \tag{23}$$

Note that a specific action may win a reward, if it contributes to the long-term objective and constraints, as shown in the first lines of Eqs. (21), (22) and (23), while it may also receive a penalty (negative reward), when temporarily violating these objective and constraints, as shown in the second lines of Eqs. (21), (22) and (23). Note that actions incurring penalties are also acceptable, since temporarily violating the long-term objective and constraints in the current transmission frame can be compensated in the following frames.

Moreover, the instantaneous constraints, such as Eqs. (17d) and (17e), have to be strictly obeyed. Therefore, the reward related to the fronthaul capacity constraint Eq. (17d) is expressed as

$$\phi_f(t) = \begin{cases} 0, & \text{(17d) holds,} \\ -\psi_f^- \left|R_f(t) - R_f^{max}\right|, & \text{otherwise.} \end{cases} \tag{24}$$

Moreover, the reward $\phi_{AP,l}(t)$ related to the transmit power constraint Eq. (17e) is expressed as

$$\phi_{AP,l}(t) = \begin{cases} 0, & \text{(17e) holds,} \\ -\psi_{AP,l}^- \left| \sum_{k=1}^K \mathrm{Tr}\left[\boldsymbol{\omega}_{l,k}(t)\boldsymbol{\omega}_{l,k}^*(t)\right] - P_{max}^{(l)} \right|, & \text{otherwise.} \end{cases} \tag{25}$$

To sum up, given a specific action $a_t = [\mathbf{c}(t), \mathbf{W}(t)]$ at the t-th transmission frame, the total reward is formulated as

$$\begin{aligned} r_t = &\sum_{k=1}^K \phi_{WDT,k}(t) + \sum_{m=1}^M \phi_{WET,m}(t) \\ &+ \sum_{l=1}^L \phi_{AP,l}(t) + \phi_o(t) + \phi_f(t). \end{aligned} \tag{26}$$

As a part of the total reward function in Eq. (26), Eq. (21) encourage the cell-free network to pursue a minimum

[4] A MDP can be represented by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected reward of performing action a at state s, and $\mathcal{P}$ is the transition probability function. Note that $\mathcal{P}(s'|s,a) \in [0,1]$ is the probability that state s transits to state s' by selecting action a [25]. γ is the discount factor used for accumulating expected rewards, which aims at a long-term return. Markov property defined as $\mathcal{P}(s_{t+1} = s'|s_t, a_t, s_{t-1}, a_{t-1}, \cdots, s_0, a_0) = \mathcal{P}(s_{t+1} = s'|s_t, a_t)$ is satisfied in our state, action and reward design.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

long-term energy consumption of (P1), while the other parts in Eqs. (22)-(25) enforce the transmission strategy to satisfy all the constraints of (P1). A total reward comprehensively considers both the objective function and the constraints of the original optimization problem (P1). Note that all the reward coefficients $\{\psi_o^+, \psi_{WDT,k}^+, \psi_{WET,m}^+\}$ and $\{\psi_o^-, \psi_{WDT,k}^-, \psi_{WET,m}^-, \psi_f^-, \psi_{AP,l}^-\}$ in Eqs. (21)-(25) are positive values.

IV. DT Based DRL Algorithm

Classic DQN and DDPG can not directly be adopted in our problem since both discrete actions $\{\mathbf{c}(t)\}$ and continuous actions $\{\mathbf{W}(t)\}$ are considered. A novel DRL algorithm named as parameterized-deep-Q-network (P-DQN) is then designed, but it suffers a lot in stability and convergence. Therefore, an enhanced double-parameterized-deep-Q-network (DP-DQN) is further proposed by invoking the double network structure and soft replace method. The application of DT in our DRL algorithm improves the convergence in speed and stability.

A. Basic Principles of DQN and DDPG

Q-leaning is a value based reinforcement learning algorithm, which chooses actions in a greedy manner by estimating the Q-values in advance. Given a state s and an action a at transmission frame t, the corresponding Q-value is defined as the average discounted total reward $r_t^\gamma = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k)$ so far, i.e., $Q(s,a) = \mathbb{E}[r_t^\gamma \mid s_1 = s, a_1 = a]$, where $r(s_k, a_k)$ is the instantaneous reward and $\gamma \in [0,1]$ is the discount factor. Moreover, Bellman equation is widely exploited for iteratively calculating the Q-value, which is expressed as

$$Q(s_t, a_t) = \mathop{\mathbb{E}}_{s_{t+1}}\left[r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a') \mid s = s_t, a = a_t \right]. \tag{27}$$

Optimal action is found for maximizing the Q-value in Eq. (27). When the size of the state space $\mathcal{S}$ and that of the action space $\mathcal{A}$ is very large, a deep-neural-network (DNN) is invoked for approximating the Q-value as $Q(s,a;\boldsymbol{\theta}) \approx Q(s,a)$, where a vector $\boldsymbol{\theta}$ includes all weights in the DNN. This is known as deep-Q-network (DQN) [42]. The vector $\boldsymbol{\theta}_t$ at the t-th transmission frame is updated by minimizing a loss function expressed as

$$L_t(\boldsymbol{\theta}) = \begin{Bmatrix} r(s_t, a_t) + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a'; \boldsymbol{\theta}_{t-1}) \\ -Q(s_t, a_t; \boldsymbol{\theta}) \end{Bmatrix}^2. \tag{28}$$

As shown in Eqs. (27) and (28), in order to minimize $L_t(\boldsymbol{\theta})$ for updating the weight vector $\boldsymbol{\theta}$ in the DNN and to maximize Q-value for finding the optimal action, we need to exhaustively go through the entire discrete action space. When a continuous action space is conceived, Q-function of Eq. (27) is usually non-convex with respect to the action a. Finding its maximum is an NP-hard problem. Therefore, DQN for Q-value maximization cannot deal with a continuous action space.

By contrast, a policy gradient based method is invoked for a continuous action space. A policy function $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps an input state to specific probability density function

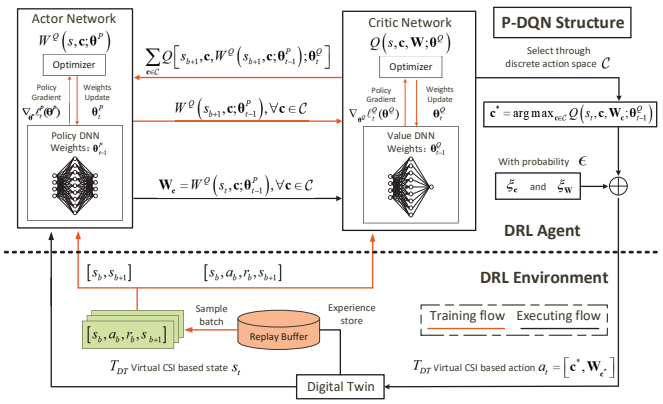


Fig. 4. DT enhanced P-DQN algorithm.

of continuous actions [43]. Let us define $\pi_{\boldsymbol{\theta}}(a \mid s)$ as the probability density of action a conditioned on state s, when a weight vector $\boldsymbol{\theta}$ of the DNN is given. The objective function $J(\pi_{\boldsymbol{\theta}}) = \mathbb{E}[r_1^\gamma \mid \pi_{\boldsymbol{\theta}}]$ can be formulated as [44]

$$\begin{aligned} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \rho^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a \mid s)\, r(s,a) \mathrm{d}a \mathrm{d}s \\ &= \mathbb{E}_{s \sim \rho^{\pi_\theta}, a \sim \pi_{\boldsymbol{\theta}}}[r(s,a)]. \end{aligned} \tag{29}$$

Note that $\mathbb{E}_{s \sim \rho^{\pi_{\boldsymbol{\theta}}}}[\cdot]$ denotes the expected value with respect to discounted state distribution $\rho^{\pi_{\boldsymbol{\theta}}}$ expressed as $\rho^{\pi_{\boldsymbol{\theta}}}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} p_0(s) p(s' \mid s, t, \pi_{\boldsymbol{\theta}}) \mathrm{d}s$, where $p_0(s)$ represents initial state probability density and $p(s' \mid s, t, \pi_{\boldsymbol{\theta}})$ represents the probability density of a state transition from s to s' in t transmission frames with policy $\pi_{\boldsymbol{\theta}}$.

If we invoke the deterministic policy gradient (DPG) theorem, the policy $\mu_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ directly returns the action $a \in \mathcal{A}$ given state $s \in \mathcal{S}$ with a weight vector $\boldsymbol{\theta}$ of the DNN. Similarly, $J(\mu_{\boldsymbol{\theta}})$ of a policy $\mu_{\boldsymbol{\theta}}$ is expressed as

$$J(\mu_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} \rho^{\mu_{\boldsymbol{\theta}}}(s) r(s, \mu_{\boldsymbol{\theta}}(s)) \mathrm{d}s = \mathbb{E}_{s \sim \rho^{\mu_{\boldsymbol{\theta}}}}[r(s, \mu_{\boldsymbol{\theta}}(s))], \tag{30}$$

where $\rho^{\mu_{\boldsymbol{\theta}}} = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} p_0(s) p(s' \mid s, t, \mu_{\boldsymbol{\theta}}) \mathrm{d}s$ is the discounted state distribution. We then adjust $\boldsymbol{\theta}$ in the descent direction of the gradient $\nabla_{\boldsymbol{\theta}} J(\mu_{\boldsymbol{\theta}}) = \mathop{\mathbb{E}}_{s \sim \rho^{\mu_{\boldsymbol{\theta}}}}\left[\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(s) \nabla_a Q^{\mu_{\boldsymbol{\theta}}}(s,a)|_{a=\mu_{\boldsymbol{\theta}}(s)} \right]$, where the Q-value function is defined as $Q^{\mu_{\boldsymbol{\theta}}}(s,a) = \mathbb{E}[r_1^\gamma \mid s_1 = s, a_1 = a; \mu_{\boldsymbol{\theta}}]$. The detailed proof can be found in [44]. In the gradient $\nabla_{\boldsymbol{\theta}} J(\mu_{\boldsymbol{\theta}})$, Q-value function is invoked, which naturally combines policy gradient method and value based method together, leading to the well known DDPG [43].

B. Parameterized-DQN for Hybrid Action

By jointly considering the classic DQN and DDPG algorithms, the framework of parameterized DQN (P-DQN) [45] to solve (P1) with a hybrid action space is illustrated in Fig. 4. Moreover, the application of DT is also depicted.

The Q-value is reformulated as $Q(s,a) = Q(s, \mathbf{c}, \mathbf{W})$, where the discrete action $\mathbf{c} \in \mathcal{C}$ represents the AP classification scheme and the continuous action $\mathbf{W} \in \mathcal{W}$ represents the transmit beamforming matrices of all the APs. The Bellman

Algorithm 1 Parameterized Deep Q-Network (P-DQN) Enhanced by DT

Initialize mini-batch size B_s, replay buffer R, learning rate lr_Q and lr_P, learn frame length x and discount factor γ.
Set DT generating frame length T_{DT} and corresponding normalized error coefficient ζ.
Set greedy threshold ϵ, exploration policy $\xi_{\mathbf{c}}$ and $\xi_{\mathbf{W}}$.
Randomly initialize Q-value network $Q\left(s, \mathbf{c}, \mathbf{W}; \boldsymbol{\theta}^Q\right)$ and deterministic policy network $W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}^P\right)$ with weights $\boldsymbol{\theta}_0^Q$ and $\boldsymbol{\theta}_0^P$.
Generate a random action $[\mathbf{c}(0), \mathbf{W}(0)]$ and a random virtual CSI leading to initial state s_1.
for episode = 1 to Max-episode **do**
  Send T_R real CSI to DT and obtain the following T_{DT} virtual CSI.
  for frame t = 1 to T_{DT} **do**
    Choose continuous action $\mathbf{W}_{\mathbf{c}} \leftarrow W^Q\left(s_t, \mathbf{c}; \boldsymbol{\theta}_{t-1}^P\right)$ for each discrete action $\mathbf{c} \in \mathcal{C}$.
    Select discrete action $\mathbf{c}^* \leftarrow \arg\max_{\mathbf{c}\in\mathcal{C}} Q\left(s_t, \mathbf{c}, \mathbf{W}_{\mathbf{c}}; \boldsymbol{\theta}_{t-1}^Q\right)$ with corresponding $\mathbf{W}_{\mathbf{c}^*}$.
    Determine action a_t referring to the ϵ-greedy policy
    $a_t \leftarrow \begin{cases} [\mathbf{c}^* + \xi_{\mathbf{c}}, \mathbf{W}_{\mathbf{c}^*} + \xi_{\mathbf{W}}] & \text{with probability } \epsilon, \\ [\mathbf{c}^*, \mathbf{W}_{\mathbf{c}^*}] & \text{with probability } 1-\epsilon. \end{cases}$
    Take action a_t, observe reward $r_t = r(s_t, a_t)$ and the next state s_{t+1}.
    Store the transition pair $[s_t, a_t, r_t, s_{t+1}]$ into R.
    if $\lfloor \frac{t}{x} \rfloor \in \mathbb{N}$ **and** R is full **then**
      Sample B_s transitions pair $[s_b, a_b, r_b, s_{b+1}]$ randomly from R and separate $[\mathbf{c}(b), \mathbf{W}(b)]$ from a_b.
      Define $y_b \leftarrow r_b + \gamma \max_{\mathbf{c}'\in\mathcal{C}} Q\left[s_{b+1}, \mathbf{c}', W^Q\left(s_{b+1}, \mathbf{c}'; \boldsymbol{\theta}_{t-1}^P\right); \boldsymbol{\theta}_{t-1}^Q\right]$ according to Eq. (35).
      Compute $\nabla_{\boldsymbol{\theta}^Q} \ell_t^Q(\boldsymbol{\theta}^Q)$ and update weights by $\boldsymbol{\theta}_t^Q \leftarrow \boldsymbol{\theta}_{t-1}^Q + lr_Q \nabla_{\boldsymbol{\theta}^Q} \ell_t^Q(\boldsymbol{\theta}^Q)$ according to Eq. (34).
      Compute $\nabla_{\boldsymbol{\theta}^P} \ell_t^P(\boldsymbol{\theta}^P)$ and update weights by $\boldsymbol{\theta}_t^P \leftarrow \boldsymbol{\theta}_{t-1}^P + lr_P \nabla_{\boldsymbol{\theta}^P} \ell_t^P(\boldsymbol{\theta}^P)$ according to Eq. (36).
      Eliminate exploration by adjusting greedy threshold ϵ and exploration policy $\xi_{\mathbf{c}}$ and ξ_W.
    else
      $\boldsymbol{\theta}_t^Q \leftarrow \boldsymbol{\theta}_{t-1}^Q$ and $\boldsymbol{\theta}_t^P \leftarrow \boldsymbol{\theta}_{t-1}^P$.
    end if
  end for
end for

equation of Eq. (27) is reformulated as Eq. (31) shown at the bottom of this page. In Eq. (31), given a certain AP classification scheme $\mathbf{c}_i \in \mathcal{C}$, we try to find the transmit beamforming matrices $\mathbf{W}_i$ that maximizes $Q(s_{t+1}, \mathbf{c}_i, \mathbf{W}')$ as $\sup_{\mathbf{W}'\in\mathcal{W}} Q(s_{t+1}, \mathbf{c}_i, \mathbf{W}') = Q(s_{t+1}, \mathbf{c}_i, \mathbf{W}_i)$. Since the cardinality $|\mathcal{C}|$ of the discrete action space is limited, we may have $Q(s_{t+1}, \mathbf{c}^*, \mathbf{W}^*)$ maximizing $\{Q(s_{t+1}, \mathbf{c}_i, \mathbf{W}_i) \mid \forall i = 1, \cdots, |\mathcal{C}|\}$, with its corresponding AP classification scheme $\mathbf{c}^*$ and the beamforming matrices $\mathbf{W}^*$.

Now we focus on finding the optimal beamforming matrix $\mathbf{W}_i$ for an AP classification scheme $\mathbf{c}_i$, when the environmental state is s. This is expressed as $\mathbf{W}_i = \arg\sup_{\mathbf{W}'\in\mathcal{W}} Q(s, \mathbf{c}_i, \mathbf{W}')$. The resultant optimal beamforming matrix can be regarded as a function $W^Q(s, \mathbf{c}_i) : \mathcal{S} \times \mathcal{C} \to \mathcal{W}$. Therefore, Eq. (31) can be further reformulated as Eq. (32) shown at the bottom of this page. Note that Eq. (32) is equivalent to Eq. (27), if the discrete action space $\mathcal{C}$ is equivalent to $\mathcal{A}$. Moreover, a DNN $Q\left(s, \mathbf{c}, \mathbf{W}; \boldsymbol{\theta}^Q\right)$ with a weights vector $\boldsymbol{\theta}^Q$ is invoked to approximate Eq. (32) as $Q\left(s, \mathbf{c}, \mathbf{W}; \boldsymbol{\theta}^Q\right) \approx Q(s, \mathbf{c}, \mathbf{W})$. Furthermore, a DNN $W^Q\left(s, \mathbf{c}'; \boldsymbol{\theta}^P\right)$ with a weights vector $\boldsymbol{\theta}^P$ is also invoked to approximate the beamforming matrix that leads to the supremum of the Q-function, which is expressed as $W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}^P\right) \approx W^Q(s, \mathbf{c})$. Therefore, Eq. (32) can be reformulated as Eq. (33) shown at the bottom of this page.

Note that $Q\left(s, \mathbf{c}, \mathbf{W}; \boldsymbol{\theta}^Q\right)$ is a Q-value based DNN, while $W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}^P\right)$ is a deterministic policy based DNN. As a result, they are effectively combined in the P-DQN. We then inherit the principles of DQN and DDPG for updating the network weights $\boldsymbol{\theta}_t^Q$ and $\boldsymbol{\theta}_t^P$ at transmission frame t, respectively. Similarly to Eq. (28), we update $\boldsymbol{\theta}_t^Q$ by minimizing the loss function as

$$\boldsymbol{\theta}_t^Q = \arg\min_{\boldsymbol{\theta}^Q} \ell_t^Q(\boldsymbol{\theta}^Q) = \arg\min_{\boldsymbol{\theta}^Q} \left\{ Q\left(s, \mathbf{c}, \mathbf{W}; \boldsymbol{\theta}^Q\right) - y_t \right\}^2, \tag{34}$$

where we have

$$y_t = r(s, a) + \gamma \max_{\mathbf{c}'\in\mathcal{C}} Q\left(s', \mathbf{c}', W^Q\left(s', \mathbf{c}'; \boldsymbol{\theta}_{t-1}^P\right); \boldsymbol{\theta}_{t-1}^Q\right). \tag{35}$$

Note that $\boldsymbol{\theta}_{t-1}^P$ in Eq. (35) is fixed. Then, in order to maximize the output of Q-value DNN $Q\left[s, \mathbf{c}, W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}_{t-1}^P\right); \boldsymbol{\theta}_t^Q\right]$ with fixed weights $\boldsymbol{\theta}_t^Q$, the weights $\boldsymbol{\theta}^P$ of the deterministic policy DNN $W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}^P\right)$ are sequentially updated by minimizing its loss function as

$$\begin{aligned} \boldsymbol{\theta}_t^P &= \arg\min_{\boldsymbol{\theta}^P} \ell_t^P(\boldsymbol{\theta}^P) \\ &= \arg\min_{\boldsymbol{\theta}^P} \left\{ -\sum_{\mathbf{c}\in\mathcal{C}} Q\left[s, \mathbf{c}, W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}^P\right); \boldsymbol{\theta}_t^Q\right] \right\}. \end{aligned} \tag{36}$$

By exploiting Eqs. (35) and (36), we can fix the weights in one DNN and update the other via any gradient descent methods. The training process of the P-DQN algorithm is detailed in Algorithm 1.

C. Enhanced Double P-DQN

As portrayed in Fig. 4, the P-DQN has a deterministic policy DNN $W^Q\left(s, \mathbf{c}; \boldsymbol{\theta}^P\right)$ as an actor network for outputting the

$$Q(s_t, \mathbf{c}, \mathbf{W}) = \mathop{\mathbb{E}}_{s_{t+1}}\left[r(s, a) + \gamma \max_{\mathbf{c}'\in\mathcal{C}} \sup_{\mathbf{W}'\in\mathcal{W}} Q(s_{t+1}, \mathbf{c}', \mathbf{W}') \mid s = s_t, a = (\mathbf{c}, \mathbf{W}) \right] \tag{31}$$

$$Q(s_t, \mathbf{c}, \mathbf{W}) = \mathop{\mathbb{E}}_{s_{t+1}}\left[r(s, a) + \gamma \max_{\mathbf{c}'\in\mathcal{C}} Q\left(s_{t+1}, \mathbf{c}', W^Q(s_{t+1}, \mathbf{c}')\right) \mid s = s_t, a = (\mathbf{c}, \mathbf{W}) \right] \tag{32}$$

$$Q\left(s_t, \mathbf{c}, \mathbf{W}; \boldsymbol{\theta}^Q\right) = \mathop{\mathbb{E}}_{s_{t+1}}\left[r(s, a) + \gamma \max_{\mathbf{c}'\in\mathcal{C}} Q\left(s_{t+1}, \mathbf{c}', W^Q\left(s_{t+1}, \mathbf{c}'; \boldsymbol{\theta}^P\right); \boldsymbol{\theta}^Q\right) \mid s = s_t, a = (\mathbf{c}, \mathbf{W}) \right] \tag{33}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

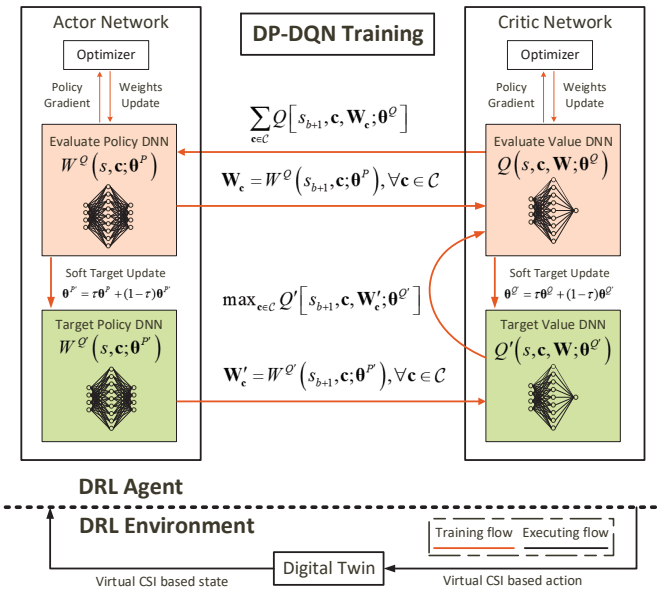


Fig. 5. Digital twin enabled DP-DQN algorithm. For simplicity, the double DNNs structure is highlighted while other part the same as P-DQN is omitted.

action (partially though), while it also has a Q-value DNN $Q\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q}\right)$ outputting the judgment criteria in form of Q-value as a critic network. Therefore, the P-DQN has a natural actor-critic structure.

The weights $\boldsymbol{\theta}^{P}$ and $\boldsymbol{\theta}^{Q}$ in these two coupled DNNs $W^{Q}\left(s,\mathbf{c};\boldsymbol{\theta}^{P}\right)$ and $Q\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q}\right)$ are highly correlated as shown in Algorithm 1, which leads to a poor convergence performance during the iterative weights updating. In order to improve the convergence, we further implement a target-evaluate structure in the P-DQN. The original Q-value DNN $Q\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q}\right)$ operates as an evaluate network cooperating with a target peer $Q'\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q'}\right)$, while the original policy DNN $W^{Q}\left(s,\mathbf{c};\boldsymbol{\theta}^{P}\right)$ also cooperates with a target peer $W^{Q'}\left(s,\mathbf{c};\boldsymbol{\theta}^{P'}\right)$. Specifically, these two evaluate DNNs $Q\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q}\right)$ and $W^{Q}\left(s,\mathbf{c};\boldsymbol{\theta}^{P}\right)$ calculate the Q-values of the input actions and determine the optimal actions by maximizing the Q-values. Moreover, the two target DNNs $Q'\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q'}\right)$ and $W^{Q'}\left(s,\mathbf{c};\boldsymbol{\theta}^{P'}\right)$ estimate the Q-values at the next transmission frame. Note that we do not need to train these two target DNNs, since they can be replaced by their evaluate counterparts. Furthermore, the target DNNs are partially substituted by the evaluate counterparts in an episode of the training process, which is called soft target update. The cooperation between evaluate DNNs and its target peer greatly reduces over-estimations, resulting in more stable and reliable learning [46]. The training process of the double-DNN structure is depicted in Fig. 5 and the resultant DP-DQN is detailed in Algorithm 2.

Compared to other classic DRL algorithms, such as the DQN with a discrete action space and the DDPG with a continuous action space, our DP-DQN is capable of solving

Algorithm 2 Double Parameterized Deep Q-Network (DP-DQN) with Experience Replay

Initialize mini-batch size B_s, replay buffer R, learning rate lr_Q and lr_P, learning interval x, discount factor γ and soft target update factor τ.
Set DT generating frame length T_{DT} and corresponding normalized error coefficient ζ.
Set greedy threshold ϵ, exploration policy $\xi_{\mathbf{c}}$ and $\xi_{\mathbf{W}}$.
Randomly initialize Q-value network $Q\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q}\right)$ and deterministic policy network $W^{Q}\left(s,\mathbf{c};\boldsymbol{\theta}^{P}\right)$ with weights $\boldsymbol{\theta}_0^{Q}$ and $\boldsymbol{\theta}_0^{P}$.
Assign target network $Q'\left(s,\mathbf{c},\mathbf{W};\boldsymbol{\theta}^{Q'}\right)$ and $W^{Q'}\left(s,\mathbf{c};\boldsymbol{\theta}^{P'}\right)$ with $\boldsymbol{\theta}^{Q'} \leftarrow \boldsymbol{\theta}_0^{Q}$ and $\boldsymbol{\theta}^{P'} \leftarrow \boldsymbol{\theta}_0^{P}$.
Generate a random action $[\mathbf{c}(0),\mathbf{W}(0)]$ and a random virtual CSI leading to initial state s_1.
for episode = 1 to Max-episode **do**
  Send T_R real CSI to DT and obtain the following T_{DT} virtual CSI.
  for frame t = 1 to T_{DT} **do**
    Choose continuous action $\mathbf{W}_{\mathbf{c}} \leftarrow W^{Q}\left(s_t,\mathbf{c};\boldsymbol{\theta}_{t-1}^{P}\right)$ for each discrete action $\mathbf{c} \in \mathcal{C}$.
    Select discrete action $\mathbf{c}^{*} \leftarrow \arg\max_{\mathbf{c}\in\mathcal{C}} Q\left(s_t,\mathbf{c},\mathbf{W}_{\mathbf{c}};\boldsymbol{\theta}_{t-1}^{Q}\right)$ with corresponding $\mathbf{W}_{\mathbf{c}^{*}}$.
    Determine action a_t referring to the ϵ-greedy policy
    $$a_t \leftarrow \begin{cases} [\mathbf{c}^{*}+\xi_{\mathbf{c}}, \mathbf{W}_{\mathbf{c}^{*}}+\xi_{\mathbf{W}}] & \text{with probability } \epsilon, \\ [\mathbf{c}^{*}, \mathbf{W}_{\mathbf{c}^{*}}] & \text{with probability } 1-\epsilon. \end{cases}$$
    Take action a_t, observe reward $r_t = r(s_t,a_t)$ and the next state s_{t+1}.
    Store the transition pair $[s_t,a_t,r_t,s_{t+1}]$ into R.
    if $\lfloor \frac{t}{x} \rfloor \in \mathbb{N}$ **then**
      Sample B_s transitions pair $[s_b,a_b,r_b,s_{b+1}]$ randomly from R and separate $[\mathbf{c}(b),\mathbf{W}(b)]$ from a_b.
      Define $y_b \leftarrow r_b + \gamma \max_{\mathbf{c}'\in\mathcal{C}} Q'\left[s_{b+1},\mathbf{c}',W^{Q'}\left(s_{b+1},\mathbf{c}';\boldsymbol{\theta}_{t-1}^{P'}\right);\boldsymbol{\theta}_{t-1}^{Q'}\right]$ according to Eq. (35).
      Compute $\nabla_{\boldsymbol{\theta}^{Q}}\ell_t^{Q}(\boldsymbol{\theta}^{Q})$ and update weights by $\boldsymbol{\theta}_t^{Q} \leftarrow \boldsymbol{\theta}_{t-1}^{Q} + lr_Q\nabla_{\boldsymbol{\theta}^{Q}}\ell_t^{Q}(\boldsymbol{\theta}^{Q})$ according to Eq. (34).
      Compute $\nabla_{\boldsymbol{\theta}^{P}}\ell_t^{P}(\boldsymbol{\theta}^{P})$ and update weights by $\boldsymbol{\theta}_t^{P} \leftarrow \boldsymbol{\theta}_{t-1}^{P} + lr_P\nabla_{\boldsymbol{\theta}^{P}}\ell_t^{P}(\boldsymbol{\theta}^{P})$ according to Eq. (36).
      Soft target update by $\boldsymbol{\theta}_t^{Q'} = \tau\boldsymbol{\theta}_{t-1}^{Q} + (1-\tau)\boldsymbol{\theta}_{t-1}^{Q'}$ and $\boldsymbol{\theta}_t^{P'} = \tau\boldsymbol{\theta}_{t-1}^{P} + (1-\tau)\boldsymbol{\theta}_{t-1}^{P'}$.
      Eliminate exploration by adjusting greedy threshold ϵ and exploration policy $\xi_{\mathbf{c}}$ and ξ_{W}.
    else
      $\boldsymbol{\theta}_t^{Q} \leftarrow \boldsymbol{\theta}_{t-1}^{Q}$, $\boldsymbol{\theta}_t^{P} \leftarrow \boldsymbol{\theta}_{t-1}^{P}$ and $\boldsymbol{\theta}_t^{Q'} \leftarrow \boldsymbol{\theta}_{t-1}^{Q'}$, $\boldsymbol{\theta}_t^{P'} \leftarrow \boldsymbol{\theta}_{t-1}^{P'}$.
    end if
  end for
end for

TABLE I
COMPARISON OF DRL ALGORITHMS

	Action space	DNNs structure	DNNs design	Convergence
DQN	Discrete	Single	Simple	Fast but unstable
DDPG	Continuous	Double	Moderate	Slow but stable
P-DQN	Mixed	Single	Complex	Slow and unstable
DP-DQN	Mixed	Double	Complex	Slow but stable

the optimization with mixed action space. Moreover, by considering both the evaluate and the target DNNs, our DP-DQN achieves more stable convergence than the P-DQN without affecting the optimal performance. However, the discount factor γ should be selected carefully, since it may affect the efficiency of the double-DNN structure. We compare DQN, DDPG, P-DQN and DP-DQN in TABLE I. Moreover, the slow convergence of our DP-DQN can be effectively improved by invoking the DT.

D. Complexity of DP-DQN

As a DRL based algorithm, the complexity of our DP-DQN should be analyzed from two aspects, i.e., executing complexity and training complexity. First, it should be noted

that all DNNs in our algorithms are composed of basic fully connected layers, and their complexities are determined by the size of the input and output layers. Therefore, the complexities of operating our policy DNN and value DNN are $O\left(KNL^3\right)$ and $O\left(K+L^2+LN\right)$ respectively. In the execution, the policy DNN is operated L^2 times to find the maximum Q value as shown in Eq. (35), and in Line 10 of Algorithm 2. The executing complexity of our DP-DQN is then obtained as $O\left(KNL^5\right)$. Training complexity is obtained similarly, except that one more gradient-descent algorithm is needed. As a typical gradient-descent algorithm, when the interior-point method is invoked to train our DP-DQN [47], the training complexity is $O\left((KNL^3)^{3.5}\right)$ in the worst case. Note that the DP-DQN can be trained in the digital twin of the cell-free network, which results in a quick convergence. In practice, the well-trained DP-DQN only has a computational complexity of $O\left(KNL^5\right)$ for outputting the optimal solution, which is far lower than $O\left(K^{3.5}N^{3.5}L^{5.5}\right)$ of the conventional mathematical optimization methods in [47] and [48].

V. Simulation Results

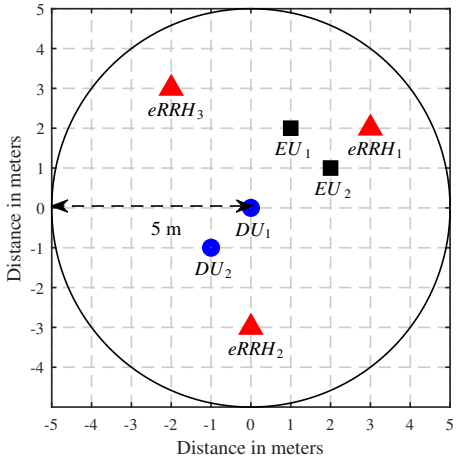


Fig. 6. Topology of the cell-free network in the simulation.

Geographic positions of three APs, two DUs and two EUs in our simulation are portrayed in Fig. 6. All the APs are equipped with three antennas, while DUs and EUs only have a single antenna. The path loss model between AP_l and DU_k is expressed as

$$\mathrm{PL}_{l,k} = 32.45 + 20\lg f + 20\lg d_{l,k}\,(dB)\,,$$

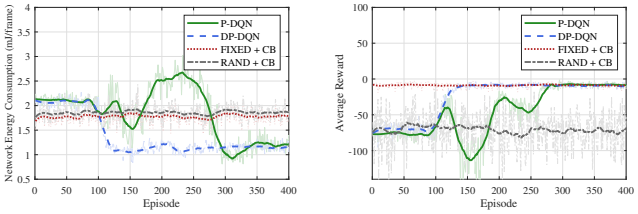
where f is the carrier frequency and $d_{l,k}$ denotes the transmission distance. The path-loss $\mathrm{PL}_{l,m}$ between AP_l and EU_m obeys the same model. All the system parameters are provided in TABLE II [11], [20], [36], [49].

Our DP-DQN is deployed in TensorFlow 1.0. All the results are averaged in an episode, which contains T_{DT} transmission frames and is used for performance evaluation. We set the observation frame length in Eq. (20) as $T_0 = T_{DT}$. Deterministic policy DNNs have dense layers of 256 $\times$ 128 and Q-value DNNs have dense layers of 256 $\times$ 128 $\times$ 64. More parameters and all the reward coefficients can be found in TABLE III. If not stated, all parameters are set according to TABLE II and TABLE III. In all figures, the grey curves stand for the exact performance in every episode, while the colorful ones are obtained by invoking Savitzky-Golay (SG) filter [50].

TABLE II
System Parameters

Parameter	Value
Temporal correlation factor , λ	0.98
Carrier frequency, f	5GHz
System bandwidth, B	30MHz
Transmission frame length, τ_s	1ms
Fronthaul capacity, R_f^{max}	25 Mbps
Energy consumption for updating AP classification, ρ_υ	2 mJ
Energy consumption for beamforming update of AP_l, $\rho_{\mu,l}$	0.2 mJ, $\forall l$
Max Tx power of APs, $P_{\max}^{(l)}$	400 mW, $\forall l$
WDT requirement of DUs, R_k^{min}	5 Mbps, $\forall k$
WET requirement of EUs, E_m^{min}	-27 dbm, $\forall m$
Noise power of DUs / UEs, σ_k^2 / σ_m^2	-90 dbm, $\forall k, m$
Energy conversion efficiency, μ_e	0.8
Collecting frame length, T_R	200 frames
Generating frame length, T_{DT}	400 frames
Normalized error coefficient, ζ	0.1

A. Algorithm Comparison



(a) Network energy consumption (b) Average reward

Fig. 7. Algorithm performance on network energy consumption and average reward.

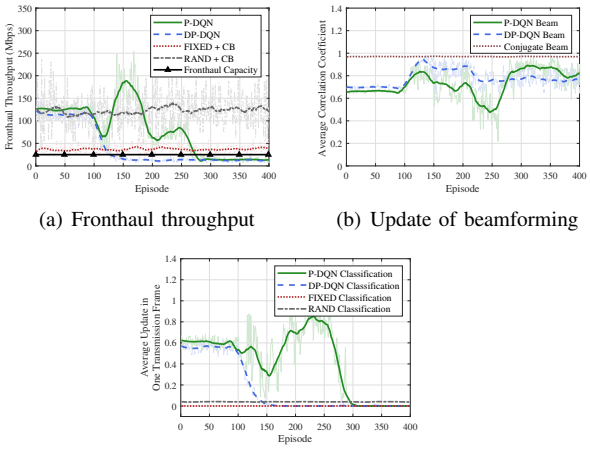
We first compare our DP-DQN algorithm with the following three benchmarks: 1) P-DQN. 2) FIXED + CB: AP classification is fixed, with one WDT AP and two WET AP, while the conjugate beamforming is adopted in the downlink IDET [20]. 3) RAND + CB: AP classification is random and associated with conjugate beamforming as well [20]. Note that DT is not applied in this comparison, so as to focus on the performance of algorithms themselves. All four algorithms interact with the same wireless environment.

The performance of the four algorithms is shown in Fig. 7. Observe from Fig. 7(a) that both P-DQN and the DP-DQN based algorithms experience a learning progress and intelligently adapt to the time-varying channel. Both of them consume 0.5 mJ less energy than the other two static algorithms, namely FIXED + CB and RAND + CB. Moreover, our DP-DQN algorithm converges to the optimal strategy faster than the P-DQN counterpart. By contrast, the FIXED + CB and RAND + CB algorithms do not have training processes. Their energy consumption is much higher than our DP-DQN algorithm. We also evaluate their average reward in Fig. 7(b). Except from the RAND + CB, all the other algorithms reach satisfactory average reward, which indicates that these three algorithms all satisfy the constraints on the fronthaul, the DUs' and EUs' requirements. Fig. 7 demonstrates that, our DP-DQN algorithm outperforms the P-DQN in convergence

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

TABLE III
DRL PARAMETERS

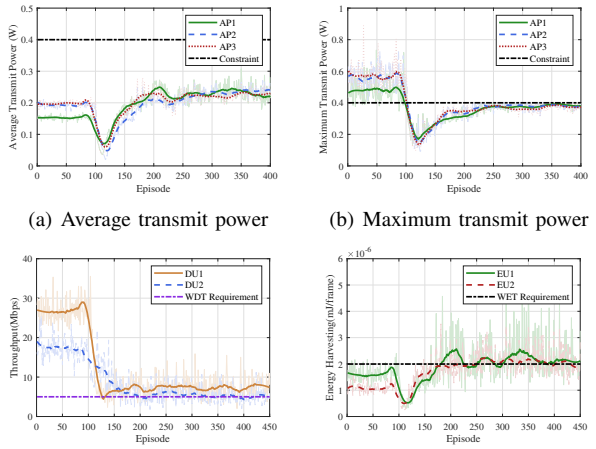
Reward Parameter	Value	DRL Hyper Parameter	Value
Network Energy consumption, $\left(\psi_o^+, \psi_o^-\right)$	$(0.2, 2)$	Replay buffer size, R	$2e^4$
WDT requirement, $\left(\psi_k^+, \psi_k^-\right)$	$(1.5, 15)$, $\forall k$	Mini-batch size, B_s	32
WET requirement, $\left(\psi_m^+, \psi_m^-\right)$	$(0.5, 5)$, $\forall m$	Discount factor, γ	0.5
Tx power of APs, $\psi_{AP,l}^-$	2, $\forall l$	Soft target update factor, τ	$5e^{-2}$
Fronthaul capacity, ψ_f^-	20	Q-value DNN learning rate, lr_Q	$5e^{-4}$
		Policy DNN learning rate, lr_P	$5e^{-4}$
		Learning interval, x	20



(a) Fronthaul throughput (b) Update of beamforming

(c) Update of AP classification

Fig. 8. Algorithm performance on fronthaul throughput, update of beamforming and AP classification.



(a) Average transmit power (b) Maximum transmit power

(c) Throughput of DUs (d) Energy Harvesting of EUs

Fig. 9. Performance of APs, DUs and EUs with DP-DQN.

and stability, since for a double-DNN structure is invoked. Moreover, it consumes much less energy than both the FIXED + CB and the RAND + CB, given its advantage in long-term objective optimization in dynamic environments.

Moreover, we evaluate on the fronthaul throughput and the update of AP classification and beamforming strategies in Fig. 8. As shown in Fig. 8(a), only the DP-DQN algorithms effectively control the fronthaul throughput under its maximum capacity by learning from experience and dynamically adjusting their strategies, while the other three algorithms cannot satisfy the fronthaul constraint. Then we investigate how the AP classification and beamforming are updated with both the P-DQN and DP-DQN in Fig. 8(b) and 8(c). Observe from Fig. 8(b), both of the P-DQN and the DP-DQN algorithms dynamically adjust their beamforming strategies as the environment dynamically change in the cell-free network. Our DP-DQN chooses less correlated beamforming strategies in adjacent transmission frames compared to P-DQN as both of them converge after 300 epochs. Observe from Fig. 8(c) that DP-DQN prefers to change the AP classification less frequently than P-DQN. Therefore, our DP-DQN prefers to keep the AP classification unchanged, but to dynamically update the beamforming strategy, since changing the AP classification consumes more energy. Moreover, observe from Fig. 8(c) that the P-DQN converges more slowly than the DP-DQN, which results in more energy consumption.

B. DP-DQN Performance

Then we investigate the performance of APs, DUs and EUs directed by the strategy of DP-DQN. Observe from Fig. 9(a) that the average transmit power in each episode converges after the training of the DP-DQN in 300 episodes and this metric stays almost unchanged. Observe from Fig. 9(b) that violations on the maximum transmit power constraint of every AP occasionally appear during the training process, since the DP-DQN explores all possible actions to find a better strategy. These constraint violations are avoided after the training of 100 episodes. Observe from Fig. 9(c) and Fig. 9(d) that both the WDT requirements of DUs and WET requirements of EUs are generally satisfied in most of cases, although the constraints are occasionally violated, as shown in the gray curves. The reason is that ideal long-term metrics in constraints Eqs. (17a) and (17b) are considered, but it is hard to evaluate them in practice thus a compromise is made as in the reward design of section IV-B. Moreover, the geographic distributions of DUs and EUs also have impact on their individual performance, when compared to constraints. In a practical deployment, we can appropriately tighten these long-term constraints to provide flexibility and robustness.

We also investigate the impact of different numbers of APs, DUs and EUs on our DP-DQN algorithm. The network consumption from 250-th to 350-th episodes are averaged as in Fig. 10. Observe from Fig. 10 that when we only have 2 APs, the network energy consumption keeps unchanged as we have more DUs and EUs. This is because both of the APs have already operated with full-load. They do not have more

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3258143

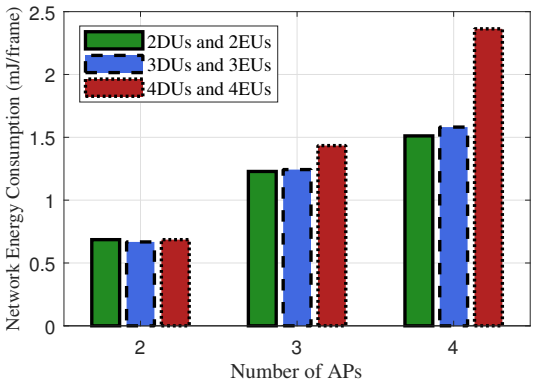
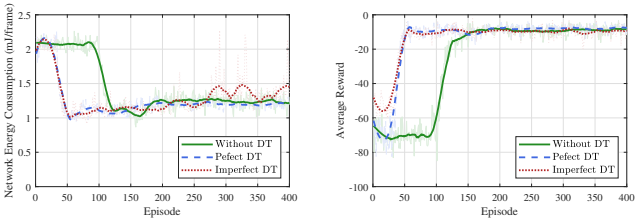


Fig. 10. The impact of different numbers of APs, DUs and EUs on the performance of DP-DQN.



(a) Network energy consumption (b) Average reward

Fig. 11. The impact of DT on the performance of DP-DQN.

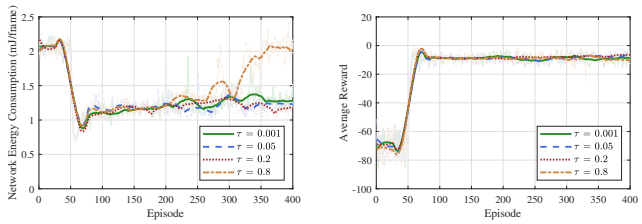
power resources allocated for satisfying the requirements of more DUs and EUs in the network. When we have more than 3 APs, more power resources can be allocated to satisfy the requirements of more DUs and EUs. Therefore, the energy consumption of the cell-free network increases, as we have more DUs and EUs.

C. DT Application

We now investigate how the DT improves our DP-DQN algorithm by evaluating the performance in the following three settings:

- *Perfect DT*: In every actual time-slot, the perfect DT generates virtual CSI data for the future $T_{DT} = 400$ time-slots. There is no error at all between the virtual CSI generated by the DT and the actual CSI, namely $\zeta = 0$. Our DP-DQN is trained with these additional virtual CSI data in every actual time-slot.
- *Imperfect DT*: The imperfect DT generates the same amount of virtual CSI data as the perfect one. However, the error between the virtual CSI and the actual CSI is considered, illustrated by $\zeta = 0.1$. Our DP-DQN is trained with these additional erroneous virtual CSI data in every actual time-slot.
- *Without DT*: The DP-DQN is only trained on actual CSI data fed back from the physical network in a single time-slot. No additional virtual CSI data can be relied upon for training.

Observe from Fig. 11(a) that both of the DP-DQN trained in the perfect DT and imperfect DT converge faster than that without the DT. Specifically, the DP-DQN with the DT converges at the 50-th episode, while the counterpart without the DT only converges at the 125-th episode. However, the the DP-DQN trained in an imperfect DT consumes much more energy than the others. This is because that extra energy is consumed for compensating the influence brought by the error in an imperfect DT. Moreover, Fig. 11(b) also demonstrates the advantage of the DT in the rapid convergence. However the DP-DQN trained in an imperfect DT obtained lower reward, also indicating that the error in an imperfect DT bring degradation in system performance in a comprehensive way.



(a) Network energy consumption (b) Average reward

Fig. 12. Effect of different discount factor γ in DP-DQN.

D. Hyper parameter

Finally, we investigate the impact of the hyper parameters on both the network energy consumption and the average reward. In Fig. 12, all the other hyper parameters are identical except the soft target update factor τ, which represents the percentage of the weights in the evaluate DNN to be substituted by those in the target DNN. Note that when $\tau \to 1$, the DP-DQN gradually degrades into the P-DQN. Hence, a large τ may bring fluctuation and bad convergence, since the prediction relied upon the target DNNs is not stable for a frequent substitutions. Moreover, reducing τ may not always result in a better performance. This is because a very small τ greatly eliminates the training effects and delays the convergence. A carefully selected τ in the DP-DQN may generate the best performance. Observe from Fig. 12(a) and Fig. 12(b) that when we have $\tau = \{0.05, 0.2\}$, the resultant system performance are better than $\tau = \{0.001, 0.8\}$.

VI. Conclusion and Future Work

We aimed for minimizing the long-term energy consumption of a cell-free network for providing integrated data and energy transfer services by jointly designing the AP classification and the transmit beamforming of all the APs. Based on both the classic DQN and DDPG, an enhanced DP-DQN was proposed to achieve more stable convergence. Moreover, an imperfect DT generating virtual CSI with error was relied upon for accelerating the convergence. Thorough simulation results showed that in order to minimize the network energy consumption, the DP-DQN updates the transmit beamforming in a small time-scale, while updating the AP classification in a large time-scale. Moreover, the advantage of the DP-DQN over the P-DQN is demonstrated, in terms of the attainable system performance and average reward. As an important hyper parameter, the soft target update factor τ was carefully selected. Furthermore, a more stable and faster

convergence was achieve by our DT based DP-DQN with subtle performance degradation.

There are still some open issues in digital twin aided cell free networking. First of all, we need to establish an accurate digital twin for a physical cell-free network. This digital twin should be able to generate virtual CSI data with a high accuracy by capturing multi-dimensional correlation of actual CSI. Second, a digital twin should be able to regularly interact with its physical counterpart in order to capture actual dynamic of a physical cell-free network. Third, multi-agent DRL may be invoked to form a distributed system for reducing algorithmic complexity. Heavy tele-traffic in the fronthaul can thus be substantially reduced.

REFERENCES

[1] Z. Chu, F. Zhou, Z. Zhu, R. Q. Hu, and P. Xiao, "Wireless powered sensor networks for internet of things: Maximum throughput and optimal power allocation," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 310–321, 2018.

[2] G. Kwon, H. Park, and M. Z. Win, "Joint beamforming and power splitting for wideband millimeter wave swipt systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 5, pp. 1211–1227, 2021.

[3] Z. Hu, T. Zhang, and J. Loo, "Power allocation for coordinated multi-cell systems with imperfect channel and battery-capacity-limited receivers," *IEEE Communications Letters*, vol. 21, no. 12, pp. 2746–2749, 2017.

[4] J. Hu, Q. Wang, and K. Yang, "Energy self-sustainability in full-spectrum 6g," *IEEE Wireless Communications*, vol. 28, no. 1, pp. 104–111, 2021.

[5] T. Bai, C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for intelligent reflecting surface aided wireless powered mobile edge computing in ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 8, pp. 5389–5407, 2021.

[6] J. Hu, K. Yang, G. Wen, and L. Hanzo, "Integrated data and energy communication network: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3169–3219, 2018.

[7] Y. Wang, K. Yang, W. Wan, Y. Zhang, and Q. Liu, "Energy-efficient data and energy integrated management strategy for iot devices based on rf energy harvesting," *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13 640–13 651, 2021.

[8] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, 2014.

[9] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in lte-advanced hetnets," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 113–122, 2013.

[10] I. Guvenc, "Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination," *IEEE Communications Letters*, vol. 15, no. 10, pp. 1084–1087, 2011.

[11] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive mimo versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[12] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Spectral efficiency maximization for uplink cell-free massive mimo-noma networks," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[13] E. Shi, J. Zhang, S. Chen, J. Zheng, Y. Zhang, D. W. Kwan Ng, and B. Ai, "Wireless energy transfer in ris-aided cell-free massive mimo systems: Opportunities and challenges," *IEEE Communications Magazine*, vol. 60, no. 3, pp. 26–32, 2022.

[14] W. Li, W. Ni, H. Tian, and M. Hua, "Deep reinforcement learning for energy-efficient beamforming design in cell-free networks," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2021, pp. 1–6.

[15] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for hybrid 5g services in mobile edge computing systems: Learn from a digital twin," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4692–4707, 2019.

[16] L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-twin-enabled 6g: Vision, architectural trends, and future directions," *IEEE Communications Magazine*, vol. 60, no. 1, pp. 74–80, 2022.

[17] Y. Lu, S. Maharjan, and Y. Zhang, "Adaptive edge association for wireless digital twin networks in 6g," *IEEE Internet of Things Journal*, vol. 8, no. 22, pp. 16 219–16 230, 2021.

[18] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive mimo," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 616–619, 2019.

[19] T. C. Mai, H. Q. Ngo, and L.-N. Tran, "Energy efficiency maximization in large-scale cell-free massive mimo: A projected gradient approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6357–6371, 2022.

[20] Y. Zhang, W. Xia, H. Zhao, W. Xu, K.-K. Wong, and L. Yang, "Cell-free iot networks with swipt: Performance analysis and power control," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13 780–13 793, 2022.

[21] G. Femenias, J. García-Morales, and F. Riera-Palou, "Swipt-enhanced cell-free massive mimo networks," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5593–5607, 2021.

[22] X. Xia, P. Zhu, J. Li, H. Wu, D. Wang, Y. Xin, and X. You, "Joint user selection and transceiver design for cell-free with network-assisted full duplexing," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 7856–7870, 2021.

[23] L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai, and D. W. K. Ng, "Downlink power control for cell-free massive mimo with deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6772–6777, 2022.

[24] Y. Al-Eryani and E. Hossain, "Self-organizing mmwave mimo cell-free networks with hybrid beamforming: A hierarchical drl-based design," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 3169–3185, 2022.

[25] F. Fredj, Y. Al-Eryani, S. Maghsudi, M. Akrout, and E. Hossain, "Distributed beamforming techniques for cell-free wireless networks using deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1186–1201, 2022.

[26] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[27] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional gans as unknown channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3133–3143, 2020.

[28] H. Xiao, W. Tian, W. Liu, and J. Shen, "Channelgan: Deep learning-based channel modeling and generating," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 650–654, 2022.

[29] W. Yang, W. Xiang, Y. Yang, and P. Cheng, "Optimizing federated learning with deep reinforcement learning for digital twin empowered industrial iot," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2022.

[30] G. Shen, L. Lei, Z. Li, S. Cai, L. Zhang, P. Cao, and X. Liu, "Deep reinforcement learning for flocking motion of multi-uav systems: Learn from a digital twin," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 141–11 153, 2022.

[31] Y. Dai, K. Zhang, S. Maharjan, and Y. Zhang, "Deep reinforcement learning for stochastic computation offloading in digital twin networks," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4968–4977, 2021.

[32] H. X. Nguyen, R. Trestian, D. To, and M. Tatipamula, "Digital twin for 5g and beyond," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 10–15, 2021.

[33] D. Jyotishi and S. Dandapat, "An lstm-based model for person identification using ecg signal," *IEEE Sensors Letters*, vol. 4, no. 8, pp. 1–4, 2020.

[34] Y. Zheng, J. Hu, and K. Yang, "Average age of information in wireless powered relay aided communication network," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 311–11 323, 2022.

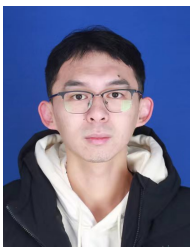
[35] X. Li, Q. Wang, M. Liu, J. Li, H. Peng, M. J. Piran, and L. Li, "Cooperative wireless-powered noma relaying for b5g iot networks with hardware impairments and channel estimation errors," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5453–5467, 2021.

[36] H. Yang, X. Xia, J. Li, P. Zhu, and X. You, "Joint transceiver design for network-assisted full-duplex systems with swipt," *IEEE Systems Journal*, vol. 16, no. 1, pp. 1206–1216, 2022.

[37] T. C. Mai, H. Q. Ngo, M. Egan, and T. Q. Duong, "Pilot power control for cell-free massive mimo," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 264–11 268, 2018.

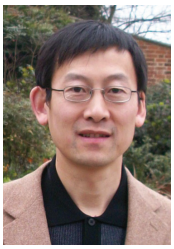
[38] P. Luong, C. Despins, F. Gagnon, and L.-N. Tran, "A fast converging algorithm for limited fronthaul c-rans design: Power and throughput

trade-off," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.
[39] F. Dikbaş, "A novel two-dimensional correlation coefficient for assessing associations in time series data," *International Journal of Climatology*, vol. 37, no. 11, pp. 4065–4076, 2017. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4998
[40] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive mimo," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 25–39, 2018.
[41] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "Decentralized beamforming for cell-free massive mimo with unsupervised learning," *IEEE Communications Letters*, vol. 26, no. 5, pp. 1042–1046, 2022.
[42] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013. [Online]. Available: http://arxiv.org/abs/1312.5602
[43] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1509.02971
[44] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, p. I–387–I–395.
[45] J. Xiong, Q. Wang, Z. Yang, P. Sun, L. Han, Y. Zheng, H. Fu, T. Zhang, J. Liu, and H. Liu, "Parametrized Deep Q-Networks Learning: Reinforcement Learning with Discrete-Continuous Hybrid Action Space," *arXiv e-prints*, p. arXiv:1810.06394, Oct. 2018.
[46] H. v. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 2094–2100.
[47] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing uav communications via joint trajectory and power control," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1376–1389, 2019.
[48] M. Cui, G. Zhang, Q. Wu, and D. W. K. Ng, "Robust trajectory and transmit power design for secure uav communications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 9042–9046, 2018.
[49] S. Chakraborty, E. Björnson, and L. Sanguinetti, "Centralized and distributed power allocation for max-min fairness in cell-free massive mimo," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 576–580.
[50] A. John, J. Sadasivan, and C. S. Seelamantula, "Adaptive savitzky-golay filtering in non-gaussian noise," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5021–5036, 2021.



Tingyu Shui received the B.Eng. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2020, where he is currently working toward the M.Sc. degree. His research interests include simultaneous wireless information and power transfer, cell-free network, resource management, and machine learning.

Jie Hu [S'11, M'16, SM'21] (hujie@uestc.edu.cn) received his B.Eng. and M.Sc. degrees from Beijing University of Posts and Telecommunications, China, in 2008 and 2011, respectively, and received the Ph.D. degree from the School of Electronics and Computer Science, University of Southampton, U.K., in 2015. Since March 2016, he has been working with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). He is now a Research Professor and PhD supervisor. He won UESTC's Academic Young Talent Award in 2019. Now he is supported by the "100 Talents" program of UESTC. He is an editor for *IEEE Wireless Communications Letters*, *IEEE/CIC China Communications* and *IET Smart Cities*. He serves for *IEEE Communications Magazine*, *Frontiers in Communications and Networks* as well as *ZTE communications* as a guest editor. He is a technical committee member of ZTE Technology. He is a program vice-chair for IEEE TrustCom 2020, a technical program committee (TPC) chair for IEEE UCET 2021 and a program vice-chair for UbiSec 2022. He also serves as a TPC member for several prestigious IEEE conferences, such as IEEE Globecom/ICC/WCSP and etc. He has won the best paper award of IEEE SustainCom 2020 and the best paper award of IEEE MMTC 2021. His current research focuses on wireless communications and resource management for B5G/6G, wireless information and power transfer as well as integrated communication, computing and sensing.



Kun Yang [M'00, SM'10, F'23]received his PhD from the Department of Electronic & Electrical Engineering of University College London (UCL), UK. He is a Chair Professor in the School of Computer Science & Electronic Engineering, University of Essex, leading the Network Convergence Laboratory (NCL), UK. He is also an affiliated professor at UESTC, China. Before joining in the University of Essex at 2003, he worked at UCL on several European Union (EU) research projects for several years. His main research interests include wireless networks and communications, IoT networking, data and energy integrated networks and mobile computing. He manages research projects funded by various sources such as UK EPSRC, EU FP7/H2020 and industries. He has published 400+ journal papers and filed 20 patents. He serves on the editorial boards of both IEEE (e.g., IEEE TNSE, WCL, ComMag) and non-IEEE journals. He is an IEEE ComSoC Distinguished Lecturer (2020-2021) and a Member of Academia Europaea (MAE).

Honghui Kang received the B.S. and M.S. degree from Huazhong University of Science and Technology, China in 1997 and 2001. He is the chief architect of ZTE Corporation wireless network intelligence,and focus on 6G network AI architecture and application.

Hua Rui received the B.S. and Ph.D degree from Nanjing University of Aeronautics and Astronautics, China in 1999 and 2005. He then joined ZTE as an Algorithm Engineer. He now is the director of Laboratory of Intelligence and Future Technologies in ZTE. His research interests include digital twin, artificial intelligence, blockchain and wireless communication.

Bo Wang received the B.S. degree from Nanchang University, China, and M.S. and Ph.D degrees from Shanghai Jiao Tong University, China. He is a senior expert in wireless algorithm of Laboratory of Intelligence and Future Technologies in ZTE. His research interests focuses on digital twin network, RAN MAC layer algorithm, network intelligence, artificial intelligence, etc.