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.2022.3198665

When UAVs Meet Cognitive Radio: Offloading Traffic Under Uncertain Spectrum Environment via Deep Reinforcement Learning

Xuanheng Li, *Member, IEEE*, Sike Cheng, *Student Member, IEEE*, Haichuan Ding, *Member, IEEE*, Miao Pan, *Senior Member, IEEE*, Nan Zhao, *Senior Member, IEEE*

***Abstract*—The emerging Internet of Things (IoT) paradigm makes our telecommunications networks increasingly congested. Unmanned aerial vehicles (UAVs) have been regarded as a promising solution to offload the overwhelming traffic. Considering the limited spectrums, cognitive radio can be embedded into UAVs to build backhaul links through harvesting idle spectrums. For the cognitive UAV (CUAV) assisted network, how much traffic can be actually offloaded depends on not only the traffic demand but also the spectrum environment. It is necessary to jointly consider both issues and co-design the trajectory and communications for the CUAV to make data collection and data transmission balanced to achieve high offloading efficiency, which, however, is non-trivial because of the heterogeneous and uncertain network environment. In this paper, aiming at maximizing the energy efficiency of the CUAV-assisted traffic offloading, we jointly design the Trajectory, Time allocation for data collection and data transmission, Band selection, and Transmission power control (T^{3}B) considering the heterogeneous environment on traffic demand, energy replenishment, and spectrum availability. Considering the uncertain environmental information, we develop a model-free deep reinforcement learning (DRL) based solution to make the CUAV achieve the best decision autonomously. Simulation results have shown the effectiveness of the proposed DRL-T^{3}B strategy.**

***Index Terms*—UAV-assisted network, cognitive radio, traffic offloading, deep reinforcement learning, energy efficiency.**

I. Introduction

Recently, with the development of the emerging Internet of Things (IoT) paradigm, wireless data traffic has witnessed a significant explosion, making our telecommunications network more and more congested. Due to the high flexibility and agility, unmanned aerial vehicle (UAV) assisted traffic offloading has been regarded as a promising solution to mitigate the network congestion. By swiftly deploying UAV-mounted relays at hot spots, these flying taxis can collect data from IoT devices with a short distance, and forward to the base station (BS) through line-of-sight (LoS) connections with reliable transmission performance. In this way, the network congestion can be alleviated effectively [2] [3]. Several recent works have been devoted to the UAV-assisted traffic offloading from different perspectives [4]–[6]. However, most of them ignored an important problem, that is, how to build the wireless backhaul link between the UAV and BS to support the relay of massive data. It is generally simply assumed that there exists abundant spectrum resource to support it. Unfortunately, considering the already overwhelming data traffic in the network, there might not exist sufficient spectrums left for the UAV to relay the large amount of offloaded data traffic [7] [8].

Cognitive radio (CR) technique can enable devices to dynamically access idle spectrums when incumbent users are inactive [9] [10]. Hence, integrating CR capability into UAVs would be a promising solution. Actually, some works have investigated the integration of CR technology with UAVs from a practical implementation aspect [11] [12]. In [11], to mitigate the Doppler frequency shifting due to the high speed of the UAV, H. Reyes *et al.* designed an intelligent modular system to enhance the reliability for the CR enabled UAV. In [12], considering the capability limitation of UAVs, C. W. Bostian *et al.* developed a low-cost cognitive radio to make the implementation feasible. With the rapid development on electronics and UAVs, integrating CR capability into UAVs would become possible, which has been considered as a new paradigm for UAV-assisted networks and attracted many research attentions [13]–[15]. As a result, in this work, we embrace the CR technology to enable the UAV to capture the idle spectrums based on spectrum sensing to construct the wireless backhaul link for data transmission. In such a cognitive UAV (CUAV) assisted network, how to make the CUAV offload as much traffic as possible is the key problem, which, however, is non-trivial, because the traffic that can be actually offloaded depends on not only the uncertain traffic demands, but also the uncertain spectrum environment. Intuitively, for certain hot spots with plenty of traffic to offload, if there does not exist enough bandwidth for harvesting (spectrums are occupied by incumbent users), although the data of grounded IoT devices might be collected by the CUAV through any short-range communication

Manuscript received September 7, 2021; revised January 13, 2022 and June 5, 2022; accepted August 4, 2022. The work of X. Li was supported in part by the Project funded by China Postdoctoral Science Foundation 2020M680038, in part by the Dalian Talents Innovation Support Program under Grant 2021RQ053, and in part by the Doctoral Research Initiation Fund of Liaoning Province under Grant 2019-BS-049. The work of H. Ding was supported in part by the Young Elite Scientists Sponsorship Program by CIC under Grant 2021QNRC001. The work of M. Pan was supported in part by the US National Science Foundation under grant CNS-2107057. The preliminary work of this paper was presented at the IEEE GLOBECOM'21 [1] [DOI: 10.1109/GLOBECOM46510.2021.9685406]. The associate editor coordinating the review of this paper and approving it for publication was L. Duan. *(Corresponding author: Xuanheng Li.)*

X. Li, S. Cheng and N. Zhao are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China. Email: xhli@dlut.edu.cn, 22009074@mail.dlut.edu.cn, zhaonan@dlut.edu.cn.

H. Ding is with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China. Email: hcding@bit.edu.cn.

M. Pan is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204. Email: mpan2@uh.edu.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

technologies, it can be hardly transmitted to the BS if there does not exist enough bandwidth for UAV-BS transmission. As a result, most existing studies on UAV-assisted traffic offloading where the collected data is assumed to be always successfully delivered to the BS cannot be directly applied to CUAV-assisted networks [14] [15]. Considering the spectrum harvesting for data transmission, a sophisticated design on both trajectory and communications is critical to make the data collection (related to traffic demand) and data transmission (related to spectrum availability) balanced to achieve high offloading efficiency.

1) Trajectory Design. In general, due to the limited on-board energy supply, energy consumption has been widely considered as a critical issue for UAV-assisted systems [16]–[18]. Hence, as for the strategy design in this paper, we take the energy-efficiency as the objective to maximize the offload data volume per unit energy consumption (bits/J) of the CUAV. In particular, with the rapid development of solar-powered UAVs, we consider a promising scenario that the CUAV can harvest energy from the environment, e.g., charging its battery from the solar energy [19]–[22]. To achieve an energy-efficient traffic offloading, the trajectory design should comprehensively take the heterogeneous environment on traffic demand, spectrum availability, and energy replenishment into account. First, in general, both traffic demand and spectrum availability are subject to spatial and temporal variations. The former one determines how much data the CUAV can collect, while the latter one determines how much data it can transmit to the BS. Hence, the actual offloaded data volume are limited by the two factors. To offload more data traffic, it is necessary to jointly consider the heterogenous environment on both issues, and well design the trajectory to make the CUAV dynamically serve the areas with not only huge traffic demand but also sufficient available spectrums. Second, from the energy-efficiency perspective, it might not be a good decision to make the CUAV serve the areas that are far away from the current location considering the huge energy consumption for flight, unless massive data traffic could be offloaded there. Furthermore, the heterogeneous environment on energy replenishment should be also taken into account. Serving the areas with more energy to harvest can pro-long the work-time of the CUAV and improve the energy-efficiency. Therefore, a sophisticated design on trajectory with a comprehensive consideration on all the traffic demand, spectrum availability, and energy replenishment is crucial for the CUAV-assisted network. Nevertheless, such a trajectory design is not an easy task, especially considering the fact that the heterogenous environment information is usually uncertain and hardly known precisely in practice.

2) Communications Design. In general, when serving certain area, the CUAV will first collect data from grounded IoT devices, and then forward to the BS through harvested spectrums. As aforementioned, for offloading efficiency, it is essential to balance the two processes of data collection and data transmission. Such a balance relies on not only an effective trajectory considering the heterogenous environment as discussed above, but also a judicious design on communications, such as how to allocate time for two processes, which bands are selected for transmission, how much power is used on each band, etc. *a) Time Allocation.* When the CUAV flies to certain area, it is necessary to effectively determine the time spent on data collection and data transmission. Intuitively, spending too much time for collection will lead to an overwhelming aggregated data volume, making it challenging for the wireless backhaul link to support, and vice versa. Thus, the decision on time allocation for data collection and transmission is very important for the offloading, which, however, is non-trivial due to the uncertainty of the environment on both traffic demand and spectrum availability. *b) Band Selection.* During the data transmission process, the CUAV should determine which bands to use. In general, different bands have different availability depending on the incumbent user's activities [23]–[25]. Hence, they will lead to different data rate for the wireless backhaul link, which will directly affect the offloading efficiency. As a result, it is necessary to well predict the uncertain spectrum environment and select the bands with more idle time in the future to support more data transmission. *c) Power Control.* The transmission power on the selected bands will determine the data rate, as well as the energy efficiency. Which power level should be adopted depends on how much data needs to be transmitted and how much time that can be used for transmission, which are closely related to the decisions on time allocation and band selection. From the perspective of energy-efficiency, it is expected to utilize the minimal power to accomplish the transmission for all the collected data, which, however, is difficult to achieve. An effective joint design is required on all the three issues under the uncertain demand and spectrum environment.

Note that the trajectory design and communications design are closely correlated. Different trajectories will lead to environments with different traffic demand, spectrum availability, and energy replenishment, which will influence the decision on time allocation, band selection, and power control. Therefore, it is imperative to jointly study both trajectory and communications for the CUAV to make it offload as much traffic as possible on the harvested spectrums with high energy-efficiency. Recently, some works have devoted to the CUAV-assisted traffic offloading schematic development [26]–[30]. Some of them employed the underlay mode for the CUAV to access the incumbent users' spectrums [26] [27], where they mainly focused on the power control strategy design to avoid interference to the incumbent users without considering the uncertain and heterogenous spectrum environment. [28]–[30] took the uncertain spectrum environment into account, where the CUAVs capture the spectrum holes to construct the wireless backhaul link via periodic spectrum sensing. In these works, considering the uncertain spectrum availability, the CUAV needs to predict the activities of the incumbent users and decide the transmission duration accordingly, or design the sensing duration to improve the sensing accuracy. Hence, due to the reliable prediction and sensing of spectrum availability, the wireless backhaul link built on the reliable spectrum supply can support more traffic transmission. However, these works endeavored to maximize the throughput of the wireless backhaul link without considering whether there is sufficient traffic to be transmitted. Furthermore, these works employed

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

the model-based optimization approach to achieve the solution. Unfortunately, the perfect knowledge of the uncertain environment on traffic demand, spectrum availability, and energy replenishment is usually hardly obtainable, which even might not follow certain closed-form models, making the traditional optimization method inapplicable.

In this paper, we propose a joint design of both trajectory and communications for the CUAV-assisted network to achieve an energy-efficient traffic offloading. In our preliminary work [1], we have studied the joint strategy under heterogeneous traffic demand and spectrum availability, but did not consider the energy issue during the traffic offloading. In this work, taking the energy efficiency as the objective, we further consider the environment with heterogeneous energy replenishment, and jointly design the CUAV Trajectory, Time allocation for data collection and data transmission, Band selection, and Transmission power control, named T^{3}B joint strategy. Due to the environmental uncertainty, we develop a model-free solution based on deep reinforcement learning for the T^{3}B joint strategy (a.k.a. DRL-T^{3}B), to make the CUAV adaptive to the uncertain environment and achieve the best strategy autonomously. The main contributions of this paper are summarized as follows.

- Unlike most existing work on UAV-assisted traffic offloading design, where the spectrum for wireless backhaul link construction is assumed to be always sufficient, we consider the practical spectrum limitation problem and employ the CR capability to build the wireless backhaul link through harvesting idle spectrums from the environment. We propose an energy-efficient CUAV-assisted traffic offloading scheme considering the environment with heterogeneous traffic demand, spectrum availability, and energy replenishment to offload as much traffic as possible over the harvested spectrums with high energy-efficiency.
- To achieve an energy-efficient traffic offloading, we jointly design the strategy of the CUAV trajectory, time allocation for data collection and transmission, band selection, and transmission power control. To our best knowledge, this is the first work to comprehensively consider the environment with heterogeneous traffic, spectrum, and energy for the UAV trajectory development, and also the first one to jointly study the trajectory and communications strategy with all of the above issues being considered.
- Since the environment information is usually uncertain and hardly known precisely in practice, the traditional optimization approaches might be inefficient or even infeasible. Hence, we develop a deep reinforcement learning based solution to make the CUAV autonomously learn the best decision under the uncertain environment in a trial-and-error way and adapt to the dynamics. Simulation results have demonstrated the effectiveness of the joint design and the DRL based solution.

The rest of this paper is organized as follows. Related works are reviewed in Section II. Then, we introduce the system model and problem formulation about the T^{3}B joint strategy in Section III. Next, in Section IV, we propose a deep reinforcement learning solution for the T^{3}B joint strategy. Simulation results and analysis are elaborated in Section V. Finally, we conclude our work in Section VI.

II. Related Work

Due to the high mobility and availability of strong LoS communication links, many recent works have investigated the UAV-assisted traffic offloading by taking UAVs as relays in the network [14], [15] [31]–[33]. Several studies on the UAV-assisted network are dedicated to traffic offloading [14], [15]. In [14], Chen *et al.* focused on the trajectory design to maximize the sum rate of UAV-served edge users. In [15], to maximize the minimum throughput of all mobile terminals, Lyu *et al.* jointly optimized the UAV's trajectory, bandwidth allocation, and user partitioning. Considering the limited on-board battery capacity, energy efficiency is also regarded as a key problem for the UAV-assisted traffic offloading [31]–[33]. In [31], Hua *et al.* dispatched a UAV to offload traffic for cell-edge users in hot spots, and proposed a joint strategy by optimizing UAV trajectory, user partitioning, and bandwidth allocation to maximize UAV's energy efficiency. In [32], Ahmed *et al.* jointly investigated the UAV trajectory and transmission power to improve the energy efficiency. In [33], Zeng *et al.* jointly designed user scheduling, UAV trajectory, transmission power, and bandwidth allocation to maximize UAV's energy efficiency while meeting the quality-of-experience of all ground users. These excellent research works have promoted the UAV-relay enabled communication networks. However, most of them considered that there exists sufficient spectrums in the network to support the wireless backhaul link for data transmission from UAVs to base stations, which might be infeasible in practice due to the limited spectrum resource.

To tackle this issue of building the UAV wireless backhaul link, CUAV-enabled traffic offloading has received many research interests [26]–[30]. In [26], regarding CUAVs as secondary users, Huang *et al.* developed a power control strategy to maximize CUAVs' achievable rate while controlling the co-channel interference at the primary receivers. Similarly, in [27], Wu *et al.* jointly optimized the CUAV trajectory and power control to maximize its throughput and meet the the interference constraints when the CUAV accesses the primary users' bands. These works employed the underlay mode for the CUAV and mainly focused on the interference avoidance problem. By harvesting idle spectrums and opportunistically accessing them, [28]–[30] investigated the interweave mode for the CUAV. In [28], in order to enable the CUAV to transmit more data to IoT devices, Almasoud *et al.* proposed an opportunistic spectrum access strategy where the CUAV predicts the activities of the PUs and then determines its transmission duration accordingly. In [29], considering the spectrum sensing accuracy issue, Liang *et al.* designed the spectrum sensing duration to improve the sensing accuracy and maximize the throughput of the CUAVs. In [30], considering the trade-off between the spectrum efficiency and energy efficiency, Hu *et al.* jointly designed the CUAV's trajectory and spectrum sensing duration to maximize its energy efficiency. By accurately predicting or sensing the activities of PUs, CUAVs can capture the uncertain spectrum availability for constructing the wireless backhaul link

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

to transmit more data. However, these works mainly focused on how to improve the throughput of the wireless backhaul link without jointly considering the uncertain traffic demand and spectrum availability. In fact, due to the uncertain traffic and spectrum environment, how to balance the data collection and data transmission for offloading efficiency improvement still calls for an innovative scheme, which depends on a joint design on CUAV trajectory and communications. Furthermore, these works adopted specific models to describe the uncertain spectrum environment, such as the log-normal distribution employed for the OFF period of the spectrum in [28]. However, since the information of the uncertain environment is hardly available in advance and difficult to be expressed with specific closed-form models, the traditional optimization method will face significant challenges in practical applications. Motivated by this observation, we design a T^{3}B joint strategy by jointly optimizing CUAV trajectory, time allocation, band selection, and transmission power control for CUAV energy efficiency maximization, and propose a DRL-based solution to obtain the optimal strategies of the CUAV.

III. System Model And Problem Formulation

A. Network Model

We consider an UAV-assisted IoT network as shown in Fig. 1, where an UAV is deployed in the network acting as a relay to offload the overwhelming data traffic. Assume that there are I candidate visiting points that the UAV can fly to in the network, corresponding to I serving areas. The UAV follows a flying-collection-transmission (F-C-T) procedure. To be specific, at each time t, it will first choose one serving area $i_t \in \mathcal{I}$ and fly to the corresponding location. Then, it will collect data within the area, and relay the aggregated data to the base station (BS). Such an F-C-T procedure will be repeated until the energy of the UAV is exhausted. Then, it will fly back to the charging station for battery re-charging. Due to the short distance, the data collection between the UAV and IoT devices can be supported by many off-the-shelf accessing technologies, such as WiFi, 4G/5G, NB-IoT, *etc*. Thus, in this work, we mainly focus on the wireless backhaul link from the UAV to the BS considering the already congested network with limited spectrum resource. To support the massive data transmission, we equip the UAV with cognitive radio (CR) capability, which is called cognitive-UAV (CUAV), and establish the wireless backhaul link between the UAV and BS based on the white spectrums in the network. In addition, we also assume that the CUAV can harvest energy for battery charging, e.g., solar energy, when serving in the network.

To offload data traffic, the CUAV will implement the F-C-T procedure in a time-slotted way as shown in Fig. 2[1]. For each time slot t, it contains two phases, namely, flying phase and offloading phase, whose length is T_1 and T_2, respectively. During the flying phase, the CUAV will either fly from the current area $\hat{i}_t$ to area $\tilde{i}_t$ or keep hovering at the current area. We denote the location of the CUAV at i_t as $l_{i_t} = (x_{i_t}, y_{i_t})^2$. Then, the flight distance at time slot t can be described as

$$d_t = \sqrt{\left(x_{\tilde{i}_t} - x_{\hat{i}_t}\right)^2 + \left(y_{\tilde{i}_t} - y_{\hat{i}_t}\right)^2}. \tag{1}$$

In general, due to battery power constraint, the maximum speed of the CUAV is v_{max}. Thus, if the CUAV flies to another area, the set of candidate areas served by the CUAV in time slot t is $\tilde{\mathcal{I}} = \left\{\tilde{i}_t \in \mathcal{I}/\hat{i}_t | d_t \le v_{\max} \cdot T_1\right\}$. If it chooses to hover at the current area, there is $\tilde{i}_t = \hat{i}_t$.

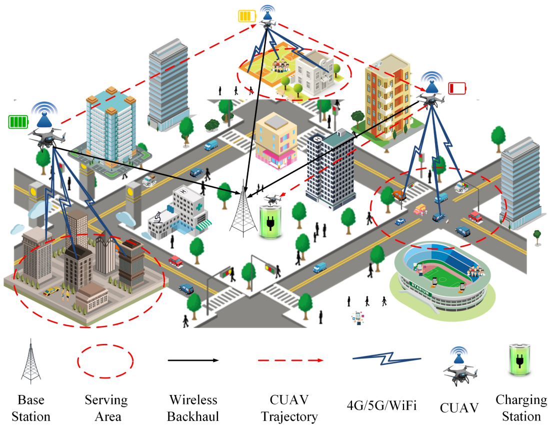


Fig. 1. The cognitive unmanned aerial vehicle assisted network architecture.

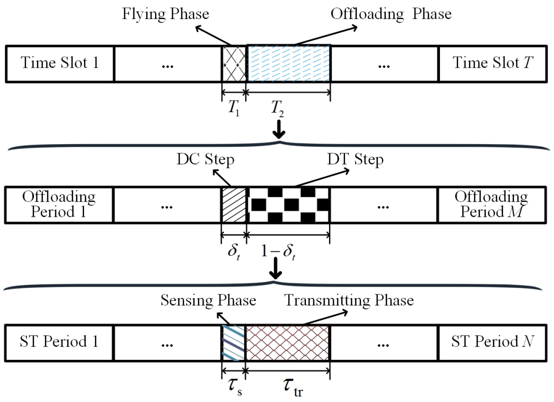


Fig. 2. The F-C-T procedure.

B. CUAV-Assisted Data Offloading Work Flow

During the offloading phase, the CUAV will collect data within the area and transmit to the BS periodically. More explicitly, the offloading phase for each time slot contains M offloading periods, and each period includes a data collection (DC) step and a data transmission (DT) step. In time slot t, the time proportion of DC step is δ_t and that of DT is $1-\delta_t$. In the DC step, the CUAV will collect data from IoT devices within the serving area, and the data in time slot t within area $\tilde{i}_t$ is

[1] Since we mainly focus on the wireless backhaul link in this work, we adopt the time-division mode to study data collection and data transmission separately. The strategy design idea can also apply to other cases with different working modes, such as the frequency-division mode where both two phases are implemented at the same time over different bands.

[2] In this paper, we consider the CUAV serves each area with the same fixed height.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

collected at a rate of $R_{\tilde{i}_t}^t$ in bps. After that, the CUAV will reach the DT step following a periodic sensing-transmission (ST) protocol. We consider the partially observable scenario, where the CUAV can only select $\hat{K}$ bands among all the K available bands to sense and opportunistically access due to the hardware limitation. Specifically, at the beginning of each ST period, the CUAV will execute spectrum sensing on the selected $\hat{K}$ bands. If idle ones exists, it will transmit data accordingly, otherwise, it will keep silence until the next ST period starts. The length for sensing and transmitting in each ST period is denoted as τ_{s} and τ_{tr}, respectively, and there are N ST periods in each offloading period. The transmission power on each band k is assumed to keep constant during the time slot t, which is denoted as $P_{\mathrm{tr}}^{k,t}$. Furthermore, during the offloading phase in each time slot t, in addition to data collection and data transmission, the CUAV will harvest energy as well, and the harvested energy in time slot t is denoted as E_{har}^t.

Following the aforementioned work flow, the CUAV will start at the charging station, implement the F-C-T procedure at each time slot, and fly back to the charging station before the remaining energy falling below a certain level E_{th}. Considering the limited on-board energy, how to make the CUAV offload as much traffic as possible while consuming less energy is critical. Note that the environment of traffic, spectrum, and energy is all temporally-spatially heterogeneous in the network, which determines the data volume to be transmitted, the data volume that can be transmitted, and the energy supply, respectively. Hence, for offloading efficiency, it is necessary to jointly design the trajectory, time allocation between the DC step and the DT step, band selection for data transmission, and transmission power control on each band, which, however, is non-trivial, especially under the dynamic spectrum-energy environment.

C. Energy Consumption Model

The energy consumption of the CUAV is mainly composed of two components associated with its propulsion and communication.

1) Energy Consumption on Propulsion: Recalling the offloading work flow of CUAV as shown in Fig. 2, for each time slot t, the energy consumption on propulsion mainly contains two parts. One is for flying during the flying phase T_1 if the CUAV chooses to move from the current area $\hat{i}_t$ to another one $\tilde{i}_t$. According to the analytical energy model derived in [34], the power consumption for flying can be calculated as (2) shown on the top of the next page. In (2), P_0 represents the blade profile power when the CUAV keeps hovering, and it can be modeled as

$$P_0 = \frac{\phi}{8}\rho\mu\beta\Omega^3\chi^3, \tag{3}$$

and P_1 is the induced power when hovering, denoted as

$$P_1 = (1+\kappa)\frac{G^{3/2}}{\sqrt{2\rho\beta}}. \tag{4}$$

Except for the speed $v_t = \frac{d_t}{T_1}$, all other parameters in (2), (3), and (4) are constants. φ, μ, β, Ω, and χ are parameters related to the blade rotor of CUAV. ω, ϕ, and G are all fuselage size parameters. Moreover, ρ and κ are air density and induced power factor, respectively. Due to the space limitation, these parameters' meanings can refer to [34].

Another consumption is for hovering during the offloading phase for data collection and data transmission, which also exists in the flying phase if the CUAV chooses to stay at the current area in this time slot. By substituting $v_t = 0$ into (2), we can obtain the power consumption for hovering as

$$P_{\mathrm{h}} = P_0 + P_1. \tag{5}$$

As a result, for each time slot t, if the CUAV chooses to fly to another area, the total energy consumption on propulsion can be expressed as

$$E_{\mathrm{pro}}^t = P_{\mathrm{f}}^t \cdot T_1 + P_{\mathrm{h}} \cdot T_2. \tag{6}$$

Otherwise, its propulsion energy can be denoted as

$$E_{\mathrm{pro}}^t = P_{\mathrm{h}} \cdot (T_1 + T_2). \tag{7}$$

2) Energy Consumption on Communications: As for energy consumption on communications, it also contains two components related to spectrum sensing and data transmission, respectively. In general, the power consumed for sensing is much less than that for transmission, and $\tau_{\mathrm{s}} \ll \tau_{\mathrm{tr}}$ [35]. Therefore, we only consider the energy consumption on data transmission here.

As for each ST period n during each offloading period m in time slot t, we adopt a binary indicator $q_{n,m}^{k,t} \in \{0,1\}$ to denote the state of band k. $q_{n,m}^{k,t} = 0$ represents the idle state, otherwise, $q_{n,m}^{k,t} = 1$. For band k, since data can be transmitted on it only when it is in the idle state, the energy consumption for data transmission in time slot t on band k can be calculated by

$$E_{\mathrm{com}}^{k,t} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(1 - q_{n,m}^{k,t}\right)P_{\mathrm{tr}}^{k,t}\tau_{\mathrm{tr}}. \tag{8}$$

3) Total Energy Consumption: In summary, the total energy consumption in time slot t can be calculated as

$$E_{\mathrm{con}}^t = E_{_{\mathrm{pro}}}^t + \sum_{k=1}^{K} b_{k,t}E_{\mathrm{com}}^{k,t}, \tag{9}$$

where $b_{k,t}$ is a binary indicator to indicate whether the CUAV accesses to band k in time slot t.

We denote the remaining energy at the beginning of time slot t as E_t. Considering the energy consumption and energy harvesting in time slot $t-1$, we can obtain that

$$E_t = \min\left\{E_{t-1} - E_{\mathrm{con}}^{t-1} + E_{\mathrm{har}}^{t-1}, E_{\max}\right\}, \tag{10}$$

where $E_{\max}$ represents the battery capacity of the CUAV.

D. Communication Model

First, we consider the data collection in each time slot t. During the offloading phase T_2, since the time proportion for data collection is δ_t, we can derive the collected data volume that to be transmitted in each offloading phase as

$$W_{\mathrm{c}}^t = \sum_{\tilde{i}_t=1}^{I} z_{\tilde{i}_t} \cdot R_{\tilde{i}_t}^t \cdot \delta_t \cdot T_2, \tag{11}$$

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

$$P_{\mathrm{f}}^{t} = P_0\left(1+\frac{3{(v_t)}^2}{\varphi^2}\right) + P_1\left(\sqrt{1+\frac{{(v_t)}^4}{4{v_0}^4}}-\frac{{(v_t)}^2}{2{v_0}^2}\right)^{1/2} + \frac{1}{2}\omega\rho\mu\beta{(v_t)}^3. \tag{2}$$

where $z_{\tilde{i}_t}$ is a binary to denote whether the CUAV serves area $\tilde{i}_t$ in time slot t.

Next, we show the data volume that can be transmitted from the CUAV to BS. As for each time slot t, the data transmission rate on band k can be described as

$$R_{\mathrm{tr}}^{k,t} = b_{k,t}\cdot B_{k,t}\cdot\log_2\left(1+\frac{g_{\mathrm{ub}}^{t}\cdot P_{\mathrm{tr}}^{k,t}}{B_{k,t}\cdot n_0}\right), \tag{12}$$

where $B_{k,t}$ denotes the bandwidth of the band k accessed by the CUAV, n_0 is known as the power spectral density of additive white Gaussian noise, and g_{ub}^{t} denotes the power propagation gain between the CUAV and BS, which can be modeled as

$$g_{\mathrm{ub}}^{t} = g_0\cdot\left(d_{\mathrm{ub}}^{t}\right)^{-\xi}, \tag{13}$$

where g_0 and ξ denote the antenna related constant and the path loss factor, respectively. Since we focus on the wireless backhaul link between UAV and base station in an open field environment with LoS connections, for simplicity, we employ the path loss model for the strategy design. It would not affect the key contribution of this work and the developed joint strategy can also apply to the case if other channel models are employed. Besides, d_{ub}^{t} is the distance between the CUAV and BS located in the center of the network, calculated as

$$d_{\mathrm{ub}}^{t} = \sqrt{\left(x_{\tilde{i}_t}\right)^2+\left(y_{\tilde{i}_t}\right)^2}. \tag{14}$$

Recalling that $q_{n,m}^{k,t}$ denotes the state of band k in time slot t, therefore, the data volume transmitted by the CUAV in the whole offloading phase can be modeled as

$$W_{\mathrm{tr}}^{t} = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(1-q_{n,m}^{k,t}\right)\cdot\tau_{\mathrm{tr}}\cdot R_{\mathrm{tr}}^{k,t}. \tag{15}$$

The actual offloaded data traffic volume by the CUAV is determined by the smaller one between the collected volume as (11) and that can be transmitted as (15), which can be expressed as

$$W_t = \min\left\{W_{\mathrm{c}}^{t}, W_{\mathrm{tr}}^{t}\right\}. \tag{16}$$

to improve the for offloading efficiency, how to balance the data collection and data transmission is of significant important, which depends on the co-design of trajectory, time allocation, transmission power, and band selection.

E. Problem Formulation

To achieve the efficient CUAV-assisted data offloading, we jointly design the trajectory, time allocation, transmission power, and spectrum access for the CUAV to obtain the optimal T^{3}B joint strategy. Specifically, we take energy efficiency as the objective, expressed as $EE_t = \frac{W_t}{E_{\mathrm{con}}^{t}}$. Then, we formulate the energy efficiency optimization problem as

$$\begin{aligned}
\mathbf{P1:}\ &\max_{\{\mathbf{z},\mathbf{b},\mathbf{P},\delta\}}\sum_{t=1}^{T}EE_t\left\{\mathbf{z},\mathbf{b},\mathbf{P},\delta\right\} &&\text{(17a)}\\
&s.t.\ \sum_{\tilde{i}_t\in\tilde{\mathcal{I}}\cup\hat{i}_t} z_{\tilde{i}_t} = 1, \forall t\in\{1,2,...\}, &&\text{(17b)}\\
&\sum_{k\in\mathcal{K}} b_{k,t} = \hat{K}, \forall t\in\{1,2,...\}, &&\text{(17c)}\\
&E_t \geq E_{\mathrm{th}}, \forall t\in\{1,2,...\}, &&\text{(17d)}\\
&P_{\mathrm{tr}}^{k,t} \leq P_{\max}, \forall k\in\hat{\mathcal{K}}, \forall t\in\{1,2,...\}, &&\text{(17e)}\\
&0<\delta_t<1, \forall t\in\{1,2,...\}, &&\text{(17f)}\\
&z_{\tilde{i}_t}\in\{0,1\}, \forall\tilde{i}_t\in\hat{\mathcal{I}}\cup\hat{i}_t, \forall t\in\{1,2,...\}, &&\text{(17g)}\\
&b_{k,t}\in\{0,1\}, \forall k\in\mathcal{K}, \forall t\in\{1,2,...\}. &&\text{(17h)}
\end{aligned}$$

$\mathbf{z}$ denotes which area the CUAV will fly to, $\mathbf{b}$ represents which bands are selected to access, $\mathbf{P}$ indicates the transmission power on the $\hat{K}$ selected bands, and δ is the ratio of DC part and DT part. (17b) means in time slot t the CUAV can only choose one serving area. (17c) indicates that in time slot t the CUAV accesses $\hat{K}$ spectrums from K available spectrums. (17d) constrains the remaining energy of the CUAV to ensure that it can fly back to the charging station before its battery energy is exhausted. Note that the terminal time T that the CUAV returns to the charging station before its energy falls below a threshold E_{th} is not a pre-defined deterministic constant, which is closely related to the energy replenishment at each area. Since the goal is to maximize the accumulative energy efficiency, the environment on energy replenishment would be very important, affecting both trajectory and communication strategies, to prolong the work time T.

For problem **P1**, to optimize the energy efficiency of the CUAV, the exact information on traffic demand, spectrum availability, and energy replenishment is required. However, such information is usually uncertain and hardly obtainable precisely in advance, which brings challenges to the traditional model-based optimization approaches. Even those issues might be approximately described by certain statistical models in some cases, the formulated optimization problem is an NP-hard problem due to the fractional objective function and the binary decision variables. Furthermore, considering the time-varying environment, if we rely on the traditional optimization approaches, we need to re-solve the optimization problem once the environmental information changes, which would be inefficient, even might be infeasible to track the environment dynamics.

Hence, considering the uncertain and dynamic feature on traffic demand, spectrum availability, and energy replenishment, in the next section, we develop a model-free DRL solution for our T^{3}B joint strategy (DRL-T^{3}B), by which the CUAV can autonomously learn the best decision under the uncertain

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

environment in a trial-and-error way and adapt to the dynamics. Specifically, this model-free solution enables the CUAV to obtain the optimal strategy in a trial-and-error way, where the CUAV pays more attention to exploration at the beginning to learn the environment characteristics and adjusts its strategy based on feedback from interaction with the environment. As the time goes on, the CUAV will exploit the result from exploration to obtain its optimal strategy[3].

IV. A Deep Reinforcement Learning Solution for T^{3}B Joint Strategy

A. Reinforcement Learning Framework of the T^{3}B Strategy

At each time slot t, the CUAV will first observe the current state s_t, and execute action a_t according to a certain policy π. Then it will obtain an immediate reward r_t. The RL method aims to make the CUAV find the optimal policy π^* that maximizes the expected discounted cumulative reward described as

$$Q\left(s_t, a_t\right) = \mathbb{E}\left[\sum_{\lambda=0}^{\infty} \gamma^{\lambda} r_{t+\lambda} \middle| s_t, a_t\right], \tag{18}$$

which is also called Q-function. $\gamma \in [0,1]$ is the discount factor reflecting the influence of future rewards. Next, with regard to problem **P1**, we present the RL models as follows.

1) State: For any time slot t, we define the state observed by the CUAV as

$$s_t = \left\{\hat{i}_t; E_{t-1}; \boldsymbol{\eta}_{t-1}\right\}, \tag{19}$$

where $\hat{i}_t$ indicates the current location of the CUAV in time slot t. We put it as a state element because the traffic demand, the spectrum availability, and the energy supply are all related to it. E_{t-1} denotes the remaining energy of the CUAV at the beginning of time slot t. By perceiving its remaining energy, the CUAV will determine whether to continue to offload traffic or return to the charging station. $\boldsymbol{\eta}_{t-1} = \left\{\eta_1^{t-1}, \eta_2^{t-1}, ..., \eta_{\hat{K}}^{t-1}\right\}$ contains the busy-idle ratio (BIR) information of $\hat{K}$ selected bands in the previous time slot $t-1$. It can reflect the uncertain spectrum environment.

2) Action: At each time t, based on the state s_t, the CUAV will execute an action a_t, which contains all the decisions in the T^{3}B joint strategy, i.e.,

$$a_t = \left\{\mathbf{z}_t; \delta_t; \mathbf{P}_t; \mathbf{b}_t\right\}. \tag{20}$$

$\mathbf{z}_t$, δ_t, $\mathbf{P}_t$, and $\mathbf{b}_t$ denote the decision of trajectory, time allocation, transmission power, and spectrum access, respectively. Due to the δ_t and $\mathbf{P}_t$, the action space is continuous, which is intractable for the discrete control solution, such as deep Q-network (DQN), double deep Q-network (DDQN), etc. To tackle this issue, we divide the δ_t and $\mathbf{P}_t$ into several levels, making the action space discrete. In each time slot t, the CUAV will choose an action from the finite action space.

3) Reward: The reward function can evaluate how good an action a_t is chose in state s_t. An effective reward definition can transform the hard-to-optimized objective into an accumulative reward optimization. With regard to the T^{3}B strategy, the goal is to achieve energy-efficient traffic offloading. Therefore, the reward can be well defined as

$$r_t = \frac{W_t}{\sigma \cdot E_{\text{con}}^t}, \tag{21}$$

where σ is a bias coefficient, with which the CUAV can achieve a trade-off between offloaded traffic and energy consumption. When the remaining energy is higher than the threshold E_{th}, the CUAV can continue to offload traffic, so the reward can be defined as its energy efficiency in time slot t.

B. Proposed DDQN-Based DRL Solution: DRL-T^{3}B

As aforementioned, in RL the agent aims to find the optimal policy π^*, which is a mapping from state to action to maximize the Q-function as (18). According to the Bellman Equation, the optimal Q-function can be modeled as

$$Q^*\left(s_t, a_t\right) = \mathbb{E}_{\pi^*}\left[r_t + \gamma \max_a Q^*\left(s_{t+1}, a\right) \middle| s_t, a_t\right], \tag{22}$$

which can be achieved by iteratively updating the Q-function as in (23) presented on the top of the next page. Since the remaining energy E_{t-1} and BIR factor $\boldsymbol{\eta}_{t-1}$ are continuous variables, the classic RL algorithm built on a look-up table, e.g., Q-Learning, can be hardly adopted here, because the Q-table to evaluate all state-action pairs cannot be constructed, which is also known as the curse of dimensionality for Q-Learning algorithm [38] [39]. Therefore, we develop a double deep Q-network (DDQN) based DRL solution for the T^{3}B joint strategy, where deep neural network is adopted to approximately evaluate Q-values, named DRL-T^{3}B.

In DDQN, there are two deep neural networks (DNNs) with the same structure, namely, main network Q with neuron weight parameters $\boldsymbol{\theta}$ and target network $\hat{Q}$ with neuron weight parameters $\hat{\boldsymbol{\theta}}$. The main network is used to calculate the evaluated Q-value $Q\left(s_t, a_t; \boldsymbol{\theta}\right)$ and select the optimal action as

$$a_{t+1}^* = \arg\max_{\mathbf{a}} Q\left(s_{t+1}, \mathbf{a}; \boldsymbol{\theta}\right). \tag{24}$$

The target network is used to obtain the target Q-value expressed as

$$y_t^{\text{DDQN}} = r_t + \gamma \hat{Q}\left(s_{t+1}, a_{t+1}^*; \hat{\boldsymbol{\theta}}\right), \tag{25}$$

which will be employed to construct the loss function as

$$L\left(\boldsymbol{\theta}\right) = \mathbb{E}\left[\left(y_t^{\text{DDQN}} - Q\left(s_t, a_t; \boldsymbol{\theta}\right)\right)^2\right], \tag{26}$$

using for training the main network. Comparing with the traditional DQN, where the target Q-value is obtained based on the maximum $\hat{Q}$ along with the same policy for action selection, calculated as

$$y_t^{\text{DQN}} = r_t + \gamma \max_a \hat{Q}\left(s_{t+1}, a; \hat{\boldsymbol{\theta}}\right), \tag{27}$$

since DDQN decouples action evaluation from action selection, it can mitigate the overestimation problem and achieve better

[3] Note that the DRL based solution may not reach the optimal strategy if the network environment changes rapidly. Fortunately, as pointed in many research works [36] [37], in general, the traffic demand and spectrum environment in the telecommunication network usually follows certain statistic characteristics on a large time scale and would not change rapidly. Therefore, the developed DRL solution could help the CUAV capture the statistical pattern of the environment to learn the best joint strategy, and adapt to the environmental dynamics.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

$$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha\left\{r_t + \gamma \max Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right\}. \tag{23}$$

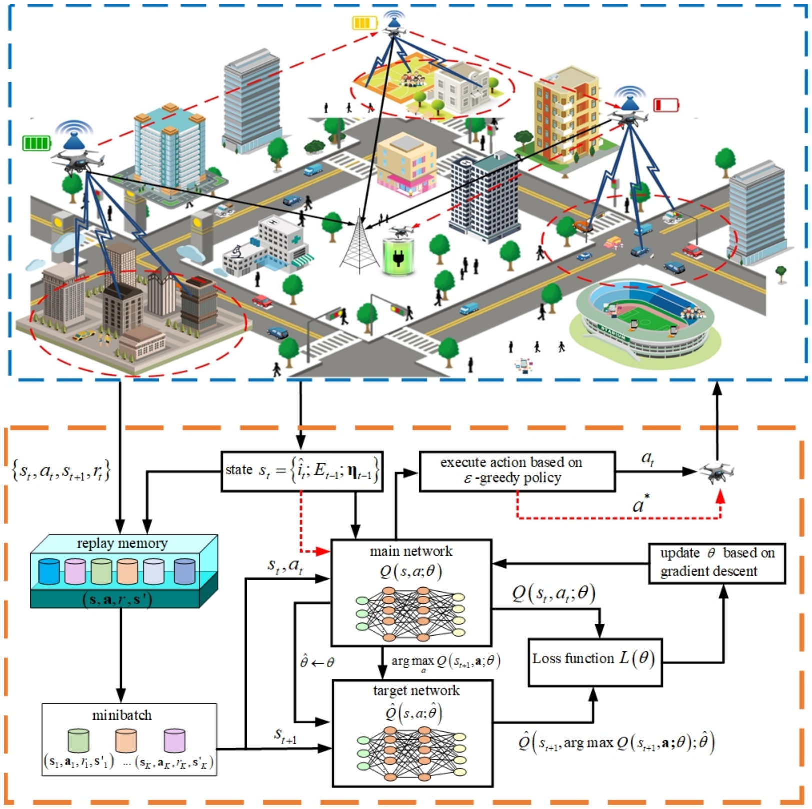

$\{s_t, a_t, s_{t+1}, r_t\}$
state $s_t = \{\hat{i}_t; E_{t-1}; \boldsymbol{\eta}_{t-1}\}$
execute action based on ε-greedy policy
a_t
a^*
replay memory
$(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$
s_t, a_t
main network
$Q(s,a;\theta)$
update θ based on gradient descent
$Q(s_t, a_t;\theta)$
$\hat{\theta} \leftarrow \theta$
$\arg\max_a Q(s_{t+1}, \mathbf{a};\theta)$
Loss function $L(\theta)$
target network
$\hat{Q}(s,a;\hat{\theta})$
minibatch
$(\mathbf{s}_1, \mathbf{a}_1, r_1, \mathbf{s}'_1)$ ... $(\mathbf{s}_K, \mathbf{a}_K, r_K, \mathbf{s}'_K)$
s_{t+1}
$\hat{Q}(s_{t+1}, \arg\max Q(s_{t+1}, \mathbf{a};\theta);\hat{\theta})$

Fig. 3. The framework of proposed DRL-based solution.

performance [40].

The training process is summarized in Algorithm 1. Note that the data for training the DNN is obtained as the CUAV interacts with environment. As shown in Fig. 3, after the initialization, the CUAV will work following the F-C-T procedure as designed in section III. At each time slot t, the CUAV will observe the state s_t, i.e., its current location $\hat{i}_t$, remaining energy E_{t-1} and the BIR information $\boldsymbol{\eta}_{t-1}$, and then take an action a_t based on ε-greedy policy, where it will randomly choose an action a_t with the probability of ε, while choosing the optimal one with the probability of $1-\varepsilon$ to determine the next served area, the time allocation for DC and DT, the accessing bands, and the transmission power allocated on them. By selecting actions via the ε-greedy policy, the CUAV can achieve a trade-off between exploration and exploitation. At the beginning, ε will be set to a large value to enable the CUAV to explore the environment, which will be gradually decreased as the algorithm converges.

After taking the action a_t, the CUAV can obtain a reward r_t, and get the next state s_{t+1}. This experience will be stored in a repaly memory unit in the form of a tuple as $\{s_t, a_t, s_{t+1}, r_t\}$, which will act as a piece of training data. As the CUAV repeats the F-C-T procedure, the number of tuples in the memory unit will keep growing. Once it exceeds the capacity of the replay memory unit J_{me}, new tuples will replace the previous ones. The DNN will be trained by sampling a mini-batch J_{mi} tuples from the memory unit. Specifically, for each piece of tuple, s_t and a_t will be fed into the main network to calculate the evaluated Q-value. Then r_t and s_{t+1} are used to calculate the target Q-value as (25). Finally, the evaluated Q-value and the target Q-value will be used to construct the loss function as (26), and the neuron weight parameters of the main network $\boldsymbol{\theta}$ will be updated by implementing a gradient step on the loss function. In addition, neuron weight parameters of the target network $\hat{\boldsymbol{\theta}}$ are copied from the main network every F time slots.

Remark 1: Note that although the fairness issue is not considered during the strategy design, the CUAV would fly around the network and serve different areas before hovering on the best place. Specifically, at the beginning, it will fly around the network to explore the environment. As the time goes on, it will learn the environment and hover on the area with the maximal energy-efficiency. When the environment changes, e.g., the traffic demand of the serving area decreases, it will re-explore the environment again by flying to different areas to find the new hot spot and provide services accordingly. Thus, for the proposed DRL-T^{3}B strategy, although it attempts to maximize the energy-efficiency, the CUAV would need to fly among different areas to find the optimal one, especially under the dynamic environment.

Remark 2: At each time slot t, the DRL algorithm is

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

Algorithm 1 A DDQN-based DRL solution: DRL-T^{3}B

```
Initialize: $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}$, $J_{\mathrm{me}}$, $J_{\mathrm{mi}}$, $\gamma$, $\alpha$, $\varepsilon$, Train=true.
for $episode$ = 1,2,... do
  Initialize state $s_1 = \left\{\hat{i}_1; E_0; \boldsymbol{\eta}_0\right\}$.
  for $t$=1,2,... do
    Observe current state $s_t = \left\{\hat{i}_t; E_{t-1}; \boldsymbol{\eta}_{t-1}\right\}$.
    Choose action $a_t = \{\mathbf{z}_t; \delta_t; \mathbf{P}_t; \mathbf{b}_t\}$ based on the state $s_t$ via the $\varepsilon$-greedy policy.
    Execute the action $a_t$, obtain reward $r_t$, update state to $s_{t+1}$.
    If $episode \bmod 100 = 0$ do
      $\varepsilon \leftarrow \max(0.99\varepsilon, 10^{-2})$.
    end if
    If Train do
      Store $\{s_t; a_t; s_{t+1}; r_t\}$ in the memory unit. Count the number of tuples in the memory unit as $j$.
      If $j \geq J_{\mathrm{me}}$ do
        Use new tuple to update the oldest one.
      end if
    end if
    if Train and $j \geq J_{\mathrm{me}}$ do
      Sample a mini-batch tuples to train the main network.
      Use the main network to calculate the evaluated Q-value and obtain the optimal action as (24). Use the target network to calculate the target Q-value as (25).
      Use the evaluated Q-value and the target Q-value to construct the loss function $L(\boldsymbol{\theta})$ as (26).
      Perform a gradient descent method on the loss function to update the neuron weight parameters $\boldsymbol{\theta}$.
      if $t \bmod F = 0$ do
        Copy the main network neuron weight parameters to the target network as $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$.
      end if
    end if
  end for
end for
Algorithm end
```

implemented in two phases, including an operating phase and a training phase. Specifically, during the operating phase, the cognitive UAV will employ the DNN to obtain an action a_t based on the current state s_t. Then, it will execute this action to interact with the environment and get an immediate reward r_t from the environment feedback. Meanwhile, the state will turn to s_{t+1}. Such interaction experience will be stored in the replay memory in the form of a tuple as $\{s_t, a_t, s_{t+1}, r_t\}$. Then, during the training phase, the DNN will be trained by sampling a mini-batch tuples from the memory, and the parameters will be updated accordingly. Note that it might not be necessary to have the training phase in each time slot and the parameters of the DNN could be updated every few time slots.

C. Computational Complexity of the Proposed DDQN-Based DRL Solution

The computational complexity of the proposed DRL algorithm mainly comes from two parts, namely, operating phase and training phase [41] [42]. The operating phase is to using the DNN to generate decisions for the T^{3}B joint strategy. Its computational complexity is determined by the architecture of DNNs. We denote $N_{\max}$ as the number of neurons for the widest layer in a full collected DNN with U layers. Then, according to [42], the complexity can be calculated as

$$O(T_{\mathrm{oper}}) = Q\left(U(N_{\max})^2\right). \tag{28}$$

For the training phase, the computational complexity depends on both forward and backward propagation in the deep neural networks (DNNs). As for the propagation algorithm, the computational complexity is determined by the architecture of DNNs. Considering a full connected DNN with U layers, where the number of neurons in each layer u is N_{neu}^{u}. Then, according to [41], the computational complexity of forward propagation can be calculated as (29) shown on the top of next page. Where $N_{\mathrm{neu}}^{0}N_{\mathrm{neu}}^{1} + \sum\limits_{u=2}^{U} N_{\mathrm{neu}}^{u}N_{\mathrm{neu}}^{u-1}N_{\mathrm{neu}}^{u-2}$ is the number of the multiplications performed in a full connected DNN, $\sum\limits_{u=1}^{U} N_{\mathrm{neu}}^{u}$ is the number of activation function employed in DNN. The computational complexity of the backward propagation can be calculated as (30) shown on the top of next page. Where $O\left(\sum\limits_{u=2}^{U} N_{\mathrm{neu}}^{u}N_{\mathrm{neu}}^{u-1}N_{\mathrm{neu}}^{u-2} + U(U-1)\right)$ is the computational complexity for the gradient operation within the backward propagation. Then, the computational complexity of the training phase can be expressed as

$$O(T_{\mathrm{train}}) = O\left((T_{\mathrm{fwd}} + T_{\mathrm{bwd}})\right). \tag{31}$$

Suppose that there are ϖ_1 iterations in training phases and ϖ_2 operating phases before convergence, the computational complexity of the proposed DRL algorithm can be expressed as

$$O_{\mathrm{DDQN}} = \varpi_1 O(T_{\mathrm{train}}) + \varpi_2 O(T_{\mathrm{oper}}). \tag{32}$$

TABLE I
HYPER-PARAMETERS OF DNNS IN THE PROPOSED DRL-T^{3}B

Parameters	Values
Number of hidden layers	3
Number of neurons in hidden layers	[128,128,64]
Activation function	ReLu
Memory unit size J_{me}	10000
Mini-batch size J_{mi}	300
Discount rate γ	0.9
Learning rate α	0.01
Update frequency of the target network F	200

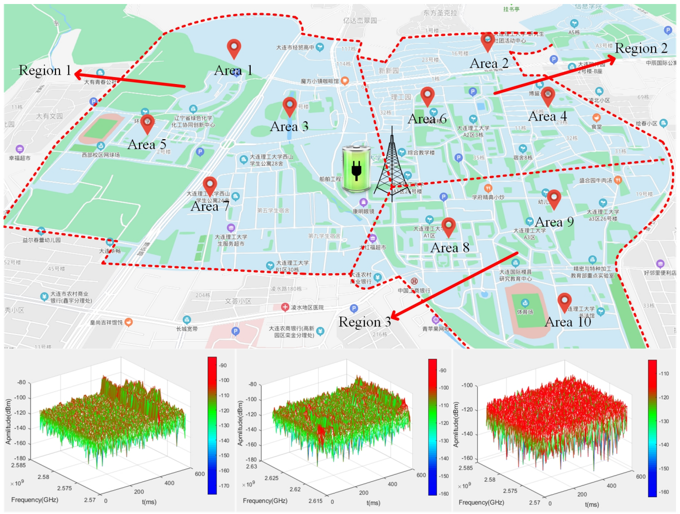
V. SIMULATION RESULTS AND DISCUSSIONS

We take the campus of Dalian University of Technology as the simulation scenario, where $I = 10$ candidate serving points are set as in Fig. 4. The CUAV takes off from the charging station located in the center of the network with the battery capacity as $E_{\max} = 3250$ mAh (corresponding to 152100J

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

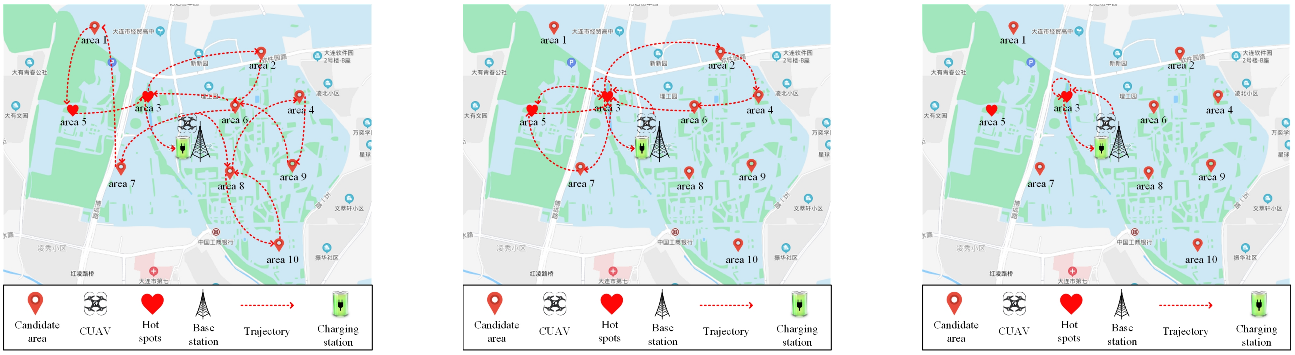
$$O\left(T_{\mathrm{fwd}}\right)=O\left(N_{\mathrm{neu}}^{0}N_{\mathrm{neu}}^{1}+\sum_{u=2}^{U}N_{\mathrm{neu}}^{u}N_{\mathrm{neu}}^{u-1}N_{\mathrm{neu}}^{u-2}+\sum_{u=1}^{U}N_{\mathrm{neu}}^{u}\right), \tag{29}$$

$$O\left(T_{\mathrm{bwd}}\right)=O\left(2\sum_{u=2}^{U}N_{\mathrm{neu}}^{u}N_{\mathrm{neu}}^{u-1}N_{\mathrm{neu}}^{u-2}+U(U-1)+N_{\mathrm{neu}}^{0}N_{\mathrm{neu}}^{1}\right), \tag{30}$$



Spectrum environment in region 1 Spectrum environment in region 2 Spectrum environment in region 3

Fig. 4. Simulation scenario and spectrum measurements.



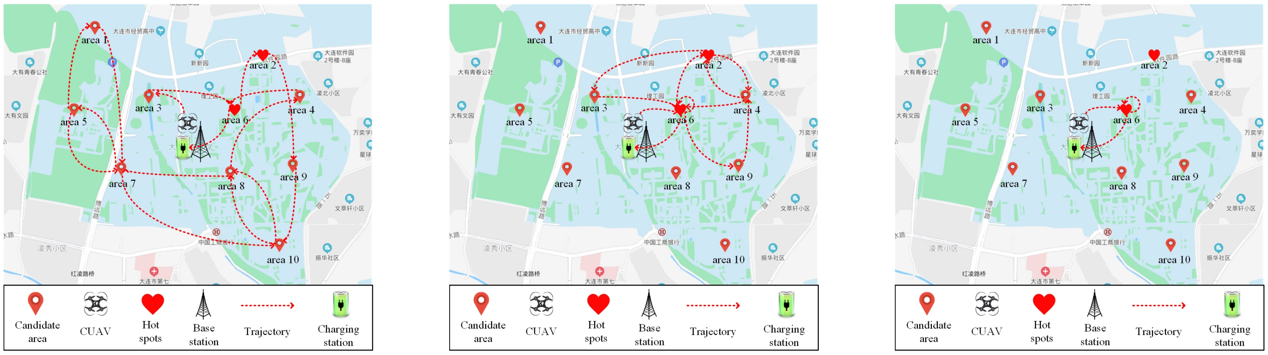
(a) episode=1. (b) episode=1000. (c) episode=3000.

Fig. 5. The trajectory of the CUAV under *CASE 1*.

under 13V) and returns to it before energy falls below the threshold $E_{\mathrm{th}} = 565$ mAh (corresponding to 26442J under 13V), which is the largest energy consumption required for the CUAV to return to the charging station from all the areas. We call the whole serving cycle an episode. In addition, we divide the CUAV-assisted network into 3 regions, and consider the spectrum environment of different areas in each region to be the same. We use SAM-60BX to measure real spectrum data in area $i = 2$, area $i = 5$, and area $i = 8$, respectively. The measurement work is carried out on four bands, ranging from 2576 to 2577 MHz, 2578 to 2579 MHz, 2580 to 2581 MHz, and 2582 to 2583 MHz, respectively. The measurement results can be seen at the bottom of Fig. 4, in which the red part represents occupied state. Assume that the CUAV can choose $\hat{K} = 1$ band from the $K = 4$ candidate bands, and communication parameters are set as $g_0 = 4$, $\xi = 4$. The bias coefficient is set as $\sigma = 1$. In any time slot t, the optional action on transmission power and time ratio of DC step and DT step are discretized into $\mathbf{P}_t = \{1,2,3,4,5,6,7,8,9,10\}$ (in W) and $\delta_t \in \{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9\}$, respectively.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

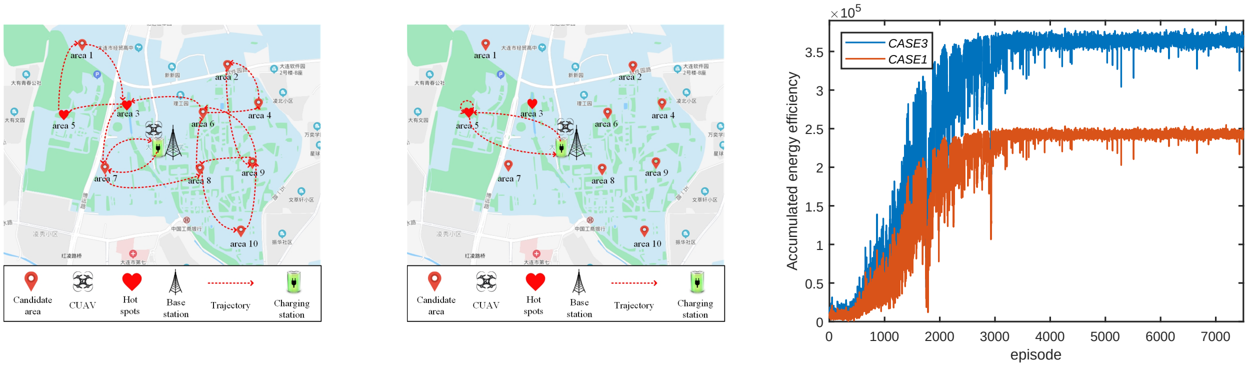


(a) episode=1.

(b) episode=1000.

(c) episode=3000.

Fig. 6. The trajectory of the CUAV under *CASE 2*.



(a) episode=1.

(b) episode=3000.

(c) Accumulated energy efficiency in an episode under *CASE 1* and *CASE 3*.

Fig. 7. The trajectory of the CUAV under *CASE 3* and the accumulated energy efficiency in an episode under two cases.

Referring to [43]–[46], we set the hyper-parameters for DNN as shown in Table I. To be specific, according to [43] and [44], we set $\gamma = 0.9$ and $\alpha = 0.01$. As for the DNN architecture, we construct it with 3 hidden layers and take ReLu as the activation function, where the neurons in each layer follow the settings in [45] and [46].

To demonstrate that the CUAV can learn an energy-efficient offloading strategy, we present the trajectory of the CUAV under three cases with different hot spots. For *CASE 1*, we set the mean of traffic demand distribution in area $i = 3$ and area $i = 5$ as 10Mbps, respectively, representing two hot spots, and that of other areas are set within the range of $(0, 1)$ (in Mbps). For *CASE 2*, we adjust the locations of the hot spots and set the mean of traffic demand distribution in area $i = 2$ and area $i = 6$ as 10Mbps, and that of other areas are also set within the range of $(0, 1)$ (in Mbps). In the two cases, the variance of the traffic demand in each area is set to 1 Kbps. The energy replenishment in each area in the two cases obeys the same normal distribution with the mean as 1000 J and the variance as 100 J. We present the trajectory guided by the DRL-T^{3}B strategy under the two cases in Fig. 5 and Fig. 6, respectively. From the results, we can see that the trajectory of the CUAV can effectively capture the hot spots. At the beginning, the untrained CUAV flies in a random way as shown in Fig. 5(a) and Fig. 6(a). Then, as the time goes on, it will gradually converge to the optimal strategy, i.e., visit the hot spots as shown in Fig. 5(c) and Fig. 6(c). In addition, we can aslo find that some hot spots will not be served by the CUAV, such as area $i = 5$ in *CASE 1* and area $i = 2$ in *CASE 2*. That is because of the remote locations of these areas. If the CUAV flies there to offload traffic, more propulsion energy consumption will be required, making the energy efficiency reduced. In summary, the proposed DRL-T^{3}B strategy can help the CUAV capture the traffic characteristics and tend to serve the hot spots with a high energy efficiency.

Next, we increase the mean of traffic demand distribution of area $i = 5$ in *CASE 1* from 10 Mbps to 15 Mbps, regarded as *CASE 3*, and present the trajectory of the CUAV under *CASE 3*, as well as the accumulated energy efficiency in an episode under these two cases in Fig. 7. Comparing Fig. 7(b) and Fig. 5(c), we can see that the CUAV will change to serve the remote hot spot $i = 5$ in *CASE 3*. This is because the traffic demand in area $i = 5$ is much higher than that in area $i = 3$. Although flying to area $i = 5$ requires more propulsion energy, the CUAV can offload more traffic there. As shown in Fig. 7(c), the offloading efficiency of the CUAV under *CASE 3* is higher than that under *CASE 1*, indicating that the CUAV guided by the DRL-T^{3}B strategy can effectively learn the traffic

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

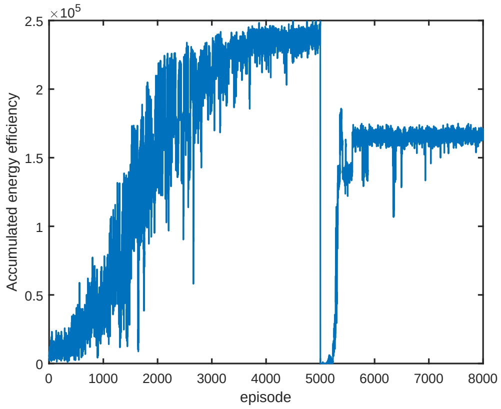


Fig. 8. Accumulated energy efficiency in an episode when the traffic demand changes.

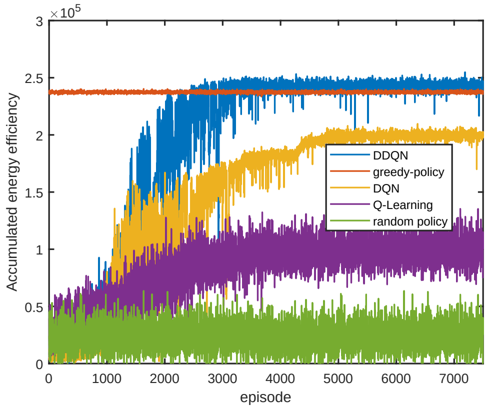


Fig. 9. Accumulated energy efficiency in an episode of different algorithms.

environment and adjust its trajectory accordingly.

Then, to demonstrate the advantage of DRL solution on the adaption to the dynamic environment, we consider that the network environment follows *CASE 1* and changes to *CASE 2* after 5000 episodes (the locations of the hot spots change after episode 5000). From Fig. 8, it can be seen that the accumulative energy efficiency converges gradually during the first 5000 episodes. At the episode 5000, since the network environment changes, the previous optimal strategy becomes bad, making the reward drop significantly. Fortunately, it will increase and re-converge soon, indicating the adaption of the developed algorithm. In other words, the proposed DRL-T^{3}B strategy can help the CUAV capture the environment characteristics with the ability of adaptation to environment changes.

Next, we compare the developed DDQN based algorithm with other existing algorithms, including greedy-policy, DQN, Q-learning, and random policy. The greedy policy here is to obtain the best strategy based on the statistical information of the environment. To be specific, at each time slot t, we discretize the state space and let the CUAV choose the action that could maximize the expectation of the immediate reward based on the state transition probability. Such a greedy policy can be regarded as an ideal case because the accurate statistical information is usually unavailable in practice, which will be used as the benchmark to evaluate the effectiveness of the proposed DDQN algorithm. We show the accumulated energy efficiency within an episode under the T^{3}B strategy solved by these algorithms in Fig. 9. The parameter settings are the same as in *CASE 1*. From Fig. 9, we can see that the developed DDQN algorithm can achieve the performance of the greedy policy, even better after the convergence, because it considers the impact of the current action on the future and does not rely on the transition probability model, which is not accurate considering the continuous state space. Such observation indicates that the proposed solution can help the CUAV achieve the optimal offloading efficiency by learning from the environment. Comparing with DQN algorithm, since two neural networks are employed in DDQN to calculate the target Q-value asynchronously, instead of only relying on one neural network as in DQN, the issue of over estimation in DQN can be avoided by DDQN, making it achieve better convergence performance. As for the Q-learning algorithm, since it can only solve the problem with discrete states, we redefine the state space by discretizing the energy E_{t-1} and the BIR $\boldsymbol{\eta}_{t-1}$ into several levels. From the results, it can be seen that the Q-learning algorithm is not applicable to the case with continuous state space. Even if it might be able to construct a Q-table by discretizing the continuous state space with small discretization levels to approximate the optimal strategy, it would be very difficult to update the table and may converge very slowly because of the huge dimension of it. As a result, the developed DDQN based DRL-T^{3}B strategy can help the CUAV learn the uncertain environment to obtain the best decision effectively.

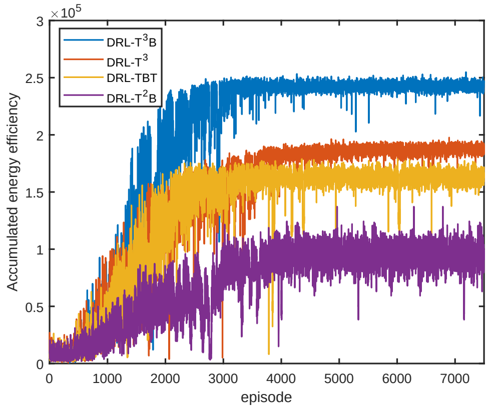
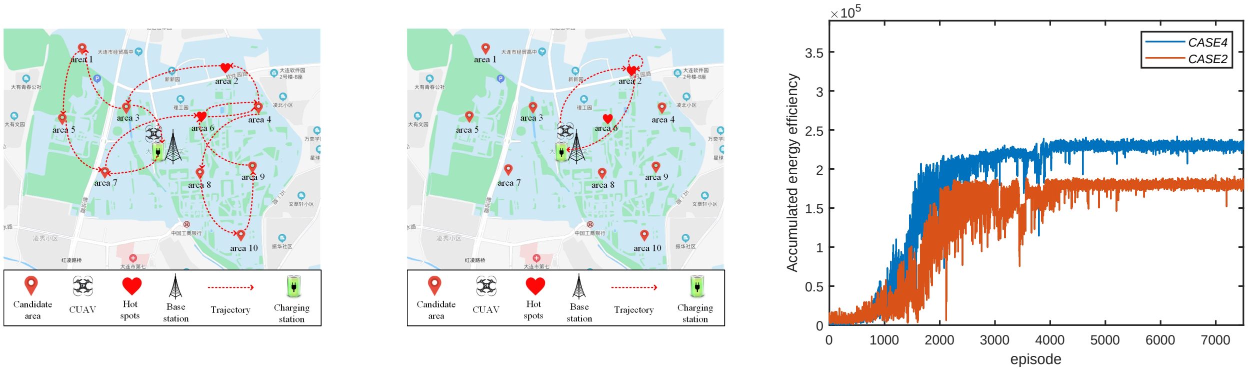


Fig. 10. Accumulated energy efficiency in an episode of different strategies.

To show the effectiveness of the joint design in the proposed T^{3}B strategy, we consider three other strategies where only partial issues (band selection, time allocation, power control) are considered as in [28]–[30]. For the fairness, we compare all the strategies under the same DRL algorithm, and call them DRL-T^3, DRL-TBT, and DRL-T^{2}B here, where band selection, time allocation, and transmission power allocation is not considered, respectively. Fig. 10 shows the accumulated energy efficiency within an episode obtained by the four strategies. As shown in Fig. 10, the proposed T^{3}B joint strategy can assist the CUAV to achieve higher energy efficiency compared to the other three strategies. Since all the decisions on time allocation, power control, and band selection will affect the balance between data collection and data transmission, determining how much

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

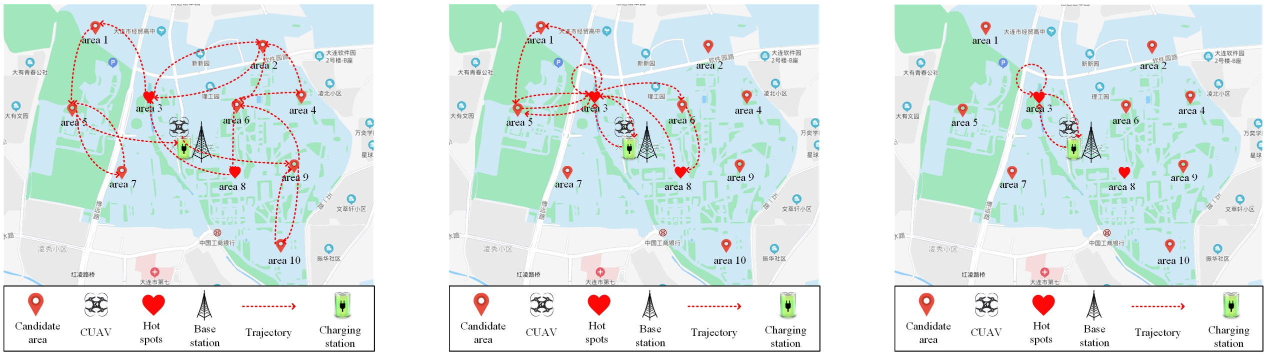


(a) episode=1.

(b) episode=3000.

(c) Accumulated energy efficiency in an episode under *CASE 2* and *CASE 4*.

Fig. 11. The trajectory of the CUAV under *CASE 4* and the accumulated energy efficiency in an episode under two cases.



(a) episode=1.

(b) episode=1000.

(c) episode=3000.

Fig. 12. The trajectory of the CUAV under *CASE 5*.

data traffic that the CUAV can actually offload, the efficiency will be reduced no matter which decision is not considered. In other words, it is necessary to comprehensively address all the issues and co-design the strategy on both trajectory and communications, indicating the effectiveness of the proposed joint strategy design.

Next, we study the effect of energy harvesting on the CUAV trajectory. We increase the mean of energy replenishment distribution of area $i = 2$ in *CASE 2* from 1000 J to 10000 J, and call it as *CASE 4*. The CUAV's trajectory under *CASE 4* and the accumulated energy efficiency in an episode under these two cases can be seen in Fig. 11. Comparing Fig. 11(b) and Fig. 6(c), we can observe that the CUAV will prefer to serve area $i = 2$, instead of area $i = 6$ as shown in Fig. 6(c), because there exists more energy to harvest. The increment on harvested energy will replenish more energy to the CUAV's battery, which can prolong its work-time. Hence, the CUAV can implement the F-C-T procedure in a hot spot more times in one episode, resulting in a higher energy efficiency as shown in Fig. 11(c). As a result, we can conclude that the proposed DRL-T^{3}B strategy can not only help the CUAV learn the traffic environment but also the energy environment, so that it can serve the areas with high traffic demand and energy replenishment to achieve higher energy efficiency.

Finally, we investigate the effect of spectrum availability on the CUAV's trajectory. We set the mean of traffic demand distribution in area $i = 3$ and area $i = 8$ to 9Mbps and 10Mbps, respectively, and that of the other areas are set within the range of (0,1) (in Mbps), and call it *CASE 5*. From Fig. 12(c), we can observe that the CUAV chooses to serve area $i = 3$ despite there has lower traffic demand, that is because the spectrum environment in area $i = 8$ is very poor. From the bottom of Fig. 4, we can find that the spectrums in area $i = 8$ are always occupied, which means that there exists insufficient spectrums for data transmission. Therefore, although the CUAV can collect more traffic in area $i = 8$, the limited capability of the wireless backhaul link built on the insufficient spectrums can hardly transmit these traffic. Whereas, in area $i = 3$, since most spectrums are in the idle state, the actual offloaded traffic is much higher. This indicates that to achieve high offloading efficiency, it is important to balance the data collection (related to traffic demand) and the data transmission (related to spectrum availability). Since the proposed DRL-T^{3}B strategy can help the CUAV capture the uncertain traffic and spectrum environment, it can fly to the hot spot with sufficient spectrums to offload traffic effectively by balancing the data collection and data transmission.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

VI. Conclusions

In this paper, to optimize energy efficiency of the CUAV-assisted network, we propose a T^{3}B joint strategy. By jointly optimizing trajectory design, time allocation, power control and band selection, the CUAV can achieve an optimal energy efficiency. Considering the heterogeneous and uncertain traffic demand, energy replenishiment, and spectrum availability, we develop a DRL solution to make the CUAV learn the best decision autonomously. Simulation results have indicated that the effectiveness of proposed DRL-T^{3}B joint strategy.

References

[1] X. Li, S. Cheng, N. Zhao, and N. Yao, "A joint strategy for CUAV-based traffic offloading via deep reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM'21), Madrid, Spain*, pp. 01–06, 2021.

[2] D. Matolak and R. Sun, "Unmanned aircraft systems: Air-ground channel characterization for future applications," *IEEE Veh. Tech. Mag.*, vol. 10, no. 2, pp. 79–85, 2015.

[3] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, 2014.

[4] M. Dai, Z. Su, Q. Xu, and N. Zhang, "Vehicle assisted computing offloading for unmanned aerial vehicles in smart city," *IEEE Trans. Intell. Transport. Syst.*, vol. 22, no. 3, pp. 1932–1944, 2021.

[5] S. Zhu, L. Gui, N. Cheng, F. Sun, and Q. Zhang, "Joint design of access point selection and path planning for UAV-assisted cellular networks," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 220–233, 2020.

[6] M. A. Ali, Y. Zeng, and A. Jamalipour, "Software-defined coexisting UAV and WiFi: Delay-oriented traffic offloading and UAV placement," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 6, pp. 988–998, 2020.

[7] D. M. Kalathil and R. Jain, "Spectrum sharing through contracts for cognitive radios," *IEEE Trans. Mobile Comput.*, vol. 12, no. 10, pp. 1999–2011, 2013.

[8] A. Q. Bicen, E. B. Pehlivanoglu, S. Galmes, and O. B. Akan, "Dedicated radio utilization for spectrum handoff and efficiency in cognitive radio networks," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 9, pp. 5251–5259, 2015.

[9] B. Wang and K. Liu, "Advances in cognitive radio networks: A survey," *IEEE J. Sel. Top. Sign. Proces.*, vol. 5, no. 1, pp. 5–23, 2011.

[10] M. Amjad, M. Rehmani, and S. Mao, "Wireless multimedia cognitive radio networks: A comprehensive survey," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 2, pp. 1056–1103, 2018.

[11] H. Reyes and N. Kaabouch, "Wireless multimedia cognitive radio networks: A comprehensive survey," *Communications and Network*, vol. 13, no. 4, pp. 1949–2421, 2013.

[12] C. W. Bostian and A. R. Young, "The application of cognitive radio to coordinated unmanned aerial vehicle (UAV) missions," *Virginia Polytechnic Institute and State University*, 2012.

[13] Z. Ullah, F. Al-Turjman, and L. Mostarda, "Cognition in UAV-aided 5G and beyond communications: A survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 872–891, 2020.

[14] F. Cheng, S. Zhang, Z. Li, Y. Chen, N. Zhao, F. R. Yu, and V. Leung, "UAV trajectory optimization for data offloading at the edge of multiple cells," *IEEE Trans. Veh. Tech.*, vol. 67, no. 7, pp. 6732–6736, 2018.

[15] J. Lyu, Y. Zeng, and R. Zhang, "UAV-aided offloading for cellular hotspot," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 6, pp. 3988–4001, 2018.

[16] Y. Cai, Z. Wei, R. Li, D. Ng, and J. Yuan, "Joint trajectory and resource allocation design for energy-efficient secure UAV communication systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4536–4553, 2020.

[17] H. Yang and X. Xie, "Energy-efficient joint scheduling and resource management for UAV-enabled multicell networks," *IEEE Systems J.*, vol. 14, no. 1, pp. 363–374, 2020.

[18] R. Ding, F. Gao, and X. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 12, pp. 7796–7809, 2020.

[19] Z. Yang, W. Xu, and M. Shikh-Bahaei, "Energy efficient UAV communication with energy harvesting," *IEEE Trans. Veh. Tech.*, vol. 69, no. 2, pp. 1913–1927, 2020.

[20] F. Tang, Z. M. Fadlullah, B. Mao, N. Kato, F. Ono, and R. Miura, "On a novel adaptive UAV-mounted cloudlet-aided recommendation system for LBSNs," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 4, pp. 565–577, 2019.

[21] Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, and A. L. Swindlehurst, "A comprehensive overview on 5G-and-beyond networks with UAVs: From communications to sensing and intelligence," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 2912–2945, 2021.

[22] Z. Yang, W. Xu, and M. Shikh-Bahaei, "Energy efficient UAV communication with energy harvesting," *IEEE Trans. Veh. Tech.*, vol. 69, no. 2, pp. 1913–1927, 2020.

[23] X. Li, R. Xiao, M. Pan, F. Jiang, N. Zhao, and X. Wang, "Green traffic offloading over uncertain shared spectrums with end-to-end QoS guarantee," *IEEE Trans. Veh. Tech.*, vol. 69, no. 9, pp. 9921–9937, 2020.

[24] X. Li, K. Jiao, F. Jiang, J. Wang, and M. Pan, "A service-oriented spectrum-aware RAN-slicing trading scheme under spectrum sharing," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 11303–11317, 2020.

[25] M. Pan, C. Zhang, P. Li, and Y. Fang, "Spectrum harvesting and sharing in multi-hop CRNs under uncertain spectrum supply," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 369–378, 2012.

[26] Y. Huang, W. Mei, J. Xu, L. Qiu, and R. Zhang, "Cognitive UAV communication via joint maneuver and power control," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7872–7888, 2019.

[27] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, 2018.

[28] A. M. Almasoud and A. E. Kamal, "Data dissemination in IoT using a cognitive UAV," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 849–862, 2019.

[29] X. Liang, Q. Deng, J. Lin, and M. Huang, "Joint trajectory optimization and spectrum access for cognitive UAV networks," *IEEE Access*, vol. 8, pp. 144693–144703, 2020.

[30] H. Hu, X. Da, Y. Huang, H. Zhang, L. Ni, and Y. Pan, "SE and EE optimization for cognitive UAV network based on location information," *IEEE Access*, vol. 7, pp. 162115–162126, 2019.

[31] M. Hua, Y. Wang, C. Li, Y. Huang, and L. Yang, "Energy-efficient optimization for UAV-aided cellular offloading," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 769–772, 2019.

[32] S. Ahmed, M. Chowdhury, and Y. Jang, "Energy-efficient UAV relaying communications to serve ground nodes," *IEEE Wireless Commun. Lett.*, vol. 24, no. 4, pp. 849–852, 2020.

[33] F. Zeng, Z. Hu, Z. Xiao, H. Jiang, S. Zhou, W. Liu, and D. Liu, "Resource allocation and trajectory optimization for QoE provisioning in energy-efficient UAV-enabled wireless networks," *IEEE Trans. Veh. Tech.*, vol. 69, no. 7, pp. 7634–7647, 2020.

[34] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 4, pp. 2329–2345, 2019.

[35] F. Shenand, G. Ding, Z. Wang, and Q. Wu, "UAV-based 3D spectrum sensing in spectrum-heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5711–5722, 2019.

[36] L. Yu, M. Li, W. Jin, Y. Guo, Q. Wang, F. Yan, and P. Li, "Step: A spatio-temporal fine-granular user traffic prediction system for cellular networks," *IEEE Trans. Mob. Comput.*, vol. 20, no. 12, pp. 3453–3466, 2021.

[37] X. Ding, L. Feng, Y. Zou, and G. Zhang, "Deep learning aided spectrum prediction for satellite communication systems," *IEEE Trans. Veh. Tech.*, vol. 69, no. 12, pp. 16314–16319, 2020.

[38] X. Wang, Y. Zhang, R. Shen, Y. Xu, and F. C. Zheng, "DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7279–7294, 2020.

[39] Y. Li, X. Hu, Y. Zhuang, Z. Gao, P. Zhang, and N. El-Sheimy, "Deep reinforcement learning (DRL): Another perspective for unsupervised wireless localization,," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6279–6287, 2020.

[40] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence (AAAI'16), Arizona, USA*, 2016.

[41] C. Zhu, Y. H. Chiang, Y. Xiao, and Y. Ji, "Flexsensing: A QoI and latency-aware task allocation scheme for vehicle-based visual crowdsourcing via deep Q-network," *IEEE Internet Things J.*, vol. 8, no. 9, pp. 7625–7637, 2021.

[42] Y. Zhao, I. G. Niemegeers, and S. M. H. D. Groot, "Dynamic power allocation forcell-free massive MIMO: deep reinforcement learning methods," *IEEE Access*, vol. 9, pp. 102953–102965, 2021.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.

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.2022.3198665

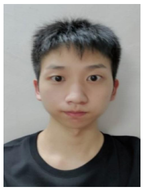
[43] W. Zhang, Q. Wang, J. Li, and C. Xu, "Dynamic fleet management with rewriting deep reinforcement learning," *IEEE Access*, vol. 9, pp. 76921–76937, 2021.

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

[45] J. Liao, T. Liu, X. Tang, X. Mu, B. Huang, and D. Cao, "Decision-making strategy on highway for autonomous vehicles using deep reinforcement learning," *IEEE Access*, vol. 8, pp. 177804–177814, 2020.

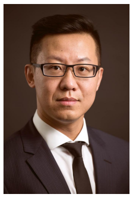
[46] Y. Wang, X. Zhou, H. Zhou, L. Chen, Z. Zheng, Q. Zeng, S. Cai, and Q. Wang, "Transmission network dynamic planning based on a double deep-Q network with deep ResNet," *IEEE Access*, vol. 9, pp. 76921–76937, 2021.

Xuanheng Li (S'13-M'18) received the Ph.D. degree in communication and information system in 2018, from the Dalian University of Technology, Dalian, China. From 2015 to 2017, he was with the Department of Electrical and Computer Engineering at the University of Florida as a visiting scholar. He is currently an Associate Professor at the Dalian University of Technology. He was a recipient of the Best Paper Award at the IEEE Globecom 2015. His current research interests include dynamic spectrum sharing, UAV-assisted networks, mobile edge computing and caching, and cognitive communications.



Sike Cheng (S'21) received the B.S. degree from the Dalian University of Technology, China, where he is currently a graduate student with the School of Information and Communication Engineering. His current research interests include UAV-assisted networks and joint radar and communication (JRC) system.

Haichuan Ding received the B.Eng. and M.S. degrees in electrical engineering from the Beijing Institute of Technology (BIT), Beijing, China, in 2011 and 2014, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2018. From 2012 to 2014, he was with the Department of Electrical and Computer Engineering, the University of Macau, as a Visiting Student. He is currently a professor with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China. His current research interests include covert communications and space-air-ground integrated networks.



Miao Pan (S'07-M'12-SM'18) received his BSc degree in Electrical Engineering from Dalian University of Technology, China, in 2004, MASc degree in electrical and computer engineering from Beijing University of Posts and Telecommunications, China, in 2007 and Ph.D. degree in Electrical and Computer Engineering from the University of Florida in 2012, respectively. He is now an Associate Professor in the Department of Electrical and Computer Engineering at University of Houston. He was a recipient of NSF CAREER Award in 2014. His research interests include Wireless/AI for AI/Wireless, deep learning privacy, cybersecurity, and underwater communications and networking. His work won IEEE TCGCC (Technical Committee on Green Communications and Computing) Best Conference Paper Awards 2019, and Best Paper Awards in ICC 2019, VTC 2018, Globecom 2017 and Globecom 2015, respectively. Dr. Pan is an Editor for IEEE Open Journal of Vehicular Technology, an Associate Editor for ACM Computing Surveys and an Associate Editor for IEEE Internet of Things (IoT) Journal (Area 5: Artificial Intelligence for IoT), and used to be an Associate Editor for IEEE Internet of Things (IoT) Journal (Area 4: Services, Applications, and Other Topics for IoT) from 2015 to 2018. He has also been serving as a Technical Organizing Committee for several conferences such as TPC Co-Chair for Mobiquitous 2019, ACM WUWNet 2019. He is a member of AAAI, a member of ACM, and a senior member of IEEE.

Nan Zhao (S'08-M'11-SM'16) is currently a Professor at Dalian University of Technology, China. He received the Ph.D. degree in information and communication engineering in 2011, from Harbin Institute of Technology, Harbin, China. Dr. Zhao is serving on the editorial boards of IEEE Wireless Communications, IEEE Wireless Communications Letters, IEEE Transactions on Green Communications and Networking. He won the best paper awards in IEEE VTC 2017 Spring, ICNC 2018, WCSP 2018 and WCSP 2019. He also received the IEEE Communications Society Asia Pacific Board Outstanding Young Researcher Award in 2018.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Dalian University of Technology. Downloaded on August 27,2022 at 07:40:21 UTC from IEEE Xplore. Restrictions apply.