Received December 17, 2019, accepted December 31, 2019, date of publication January 6, 2020, date of current version January 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964042

Minimum Throughput Maximization for Multi-UAV Enabled WPCN: A Deep Reinforcement Learning Method

JIE TANG[1], (Senior Member, IEEE), JINGRU SONG[1], JUNHUI OU[1], JINGCI LUO[1], XIUYIN ZHANG[1], (Senior Member, IEEE), AND KAI-KIT WONG[2], (Fellow, IEEE)

[1]School of Electronic of Information Engineering, South China University of Technology, Guangzhou 510641, China
[2]Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K.

Corresponding author: Junhui Ou (oujunhui@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61971194 and Grant 61601186, in part by the Natural Science Foundation of Guangdong Province under Grant 2017A030313383, and in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, under Grant 2019D06.

ABSTRACT This paper investigates joint unmanned aerial vehicle (UAV) trajectory planning and time resource allocation for minimum throughput maximization in a multiple UAV-enabled wireless powered communication network (WPCN). In particular, the UAVs perform as base stations (BS) to broadcast energy signals in the downlink to charge IoT devices, while the IoT devices send their independent information in the uplink by utilizing the collected energy. The formulated throughput optimization problem which involves joint optimization of 3D path design and channel resource assignment with the constraint of flight speed of UAVs and uplink transmit power of IoT devices, is not convex and thus is extremely difficult to solve directly. We take advantage of the multi-agent deep Q learning (DQL) strategy and propose a novel algorithm to tackle this problem. Simulation results indicate that the proposed DQL-based algorithm significantly improve performance gain in terms of minimum throughput maximization compared with the conventional WPCN scheme.

INDEX TERMS Unmanned aerial vehicle (UAV), wireless powered communication network (WPCN), Internet of Things (IoT), trajectory design, deep reinforcement learning (DRL).

I. INTRODUCTION

The Internet of Things (IoT) ensures the data collection and exchange by interconnecting heterogeneous smart devices such as sensors, smart phones, smart transportation system, which makes machine-to-machine (M2M) communication and seamless communication possible [1], [2]. The massive application scenarios in IoT may generate rigorous communication requirements such as low latency, high reliability and safety. The Long-Term Evolution (LTE) can't support Machine-Type communication (MTC) effectively due to the fact that they focus on broadband communication [3]. The fifth generation (5G) mobile network brings higher throughput, lower end-to-end latency and enhanced security mechanism, which is capable of meeting the massive IoT communication demands. Thus, IoT has received significant research attention in 5G era.

In a conventional scene, the IoT devices are battery-constrained and can't handle enormous energy consumption. Radio frequency (RF) based energy harvesting (EH) can be regarded as a prospective scheme to extend the lifetime of energy-constraint IoT devices [4]. Moreover, massive ground IoT devices have large and frequent communication requirements. Wireless powered communication network (WPCN) [5] which integrates wireless power transfer (WPT) and wireless information transfer (WIT), provides a feasible solution for energy-constraint IoT devices. Authors in [6] proposes a classic protocol named "harvest-then-transmit" (HTT). In this protocol, the ground users get charged by the downlink energy flow first, and then transmit their uplink information signals by utilizing the collected energy. Moreover, time division multiple access (TDMA) is adopted as a typical design for WPCN in [6] and sum-throughput is

The associate editor coordinating the review of this manuscript and approving it for publication was Guan Gui.

maximized by optimizing time resource allocation. Furthermore, a multi-antenna energy beamforming and space division multiple access (SDMA) protocol is employed in [7] for higher spectrum efficiency. In [8], the authors combine multi-user multi-in multi-out (MIMO) technology with cognitive radio and WPCN for maximizing the sum throughput. The authors in [9] and [10] introduce backscatter communication mode into HTT-based WPCN for the sake of maximizing the throughput. However, there is still a challenge named ''doubly-near-far'' in conventional WPCN [11], which means that comparing with devices close to base station (BS), devices far away from BS harvest less wireless energy in the downlink but have to consume more to transmit information in the uplink.

Owing to its high maneuverability and flexibility, unmanned aerial vehicle (UAV) can provide greater probability of line-of-sight (LoS) channel and better connectivity comparing to the conventional fixed BS. Therefore, UAV has been applied in many research fields of wireless communication. In [12] and [13], UAVs perform as flying relays in order to achieve the end-to-end throughput maximization by jointly optimizing UAV's trajectory planning and transmit power control. In [14], the authors introduce UAV into a conventional WPCN, and propose a channel-weighted path planning method to maximum the sum throughput, where UAV performs as an assistant of located BS. In [15], a UAV-aided WPCN is considered, in which the UAV performs as the aerial BS in order to provide service to a cluster of ground users. A joint successive hover-and-fly trajectory design and wireless resource allocation protocol is proposed in [15] for throughput maximization. Authors in [16] consider a wireless network in which multiple UAVs provide wireless communication service, and the co-channel interference and transmit power control are discussed. In [17], the authors maximize the number of users in coverage subject to the minimum transmit power by optimizing 3D placement of UAV.

Deep learning (DL) has been proved to be a powerful tool for solving non-convex problems and high complexity issues, which has been widely applied in the optimization of wireless communication system [18]–[22]. As a kind of deep reinforcement learning (DRL), deep Q learning (DQL) makes action strategy by utilizing Deep Neural Network (DNN) and performs well while dealing with dynamic time-variant environments [23]. Therefore, DQL provides a promising technique for UAV's dynamic control. Authors in [24] adopt reinforcement learning (RL) for the purpose of acquiring the optimal hover position of UAVs. In [25], UAVs make decisions based on deep Q network (DQN) for energy-efficient data collection while they are deployed in smart cities. A DRL-based UAV control strategy is proposed by [26] for maximizing both the energy efficiency and communication coverage.

A. MAIN CONTRIBUTIONS

The previous research have investigated the UAV and WPCN related system, and provide effective solutions for throughput maximization [15], [16]. However, the work in [15] only considers a single-UAV based WPCN which is not suitable for the scene of massive IoT devices. In [16], a multi-UAV assisted wireless communication network is proposed but the energy supply of ground devices is not considered. Work in [17] investigates the 3D placement of UAV for the purpose of maximizing the coverage, but the flexible trajectory design is not taken into account. In the scenario with a lot of energy-constraint IoT devices located in a large area, multi-UAV and downlink WPT are both worth studying. Furthermore, the 3D trajectory design of UAV is necessary in order to achieve better channel quality. Motivated by the above research, we put forward a minimum throughput maximization problem for multi-UAV enabled WPCN with jointly optimization of 3D trajectory design and time resource assignment. The contributions of this paper are summarized as follows.

1) We come up with a WPCN in which multiple UAVs provide reliable energy supply and communication services to IoT devices. Based on the considered model, our target is to maximize the minimum throughput by jointly scheduling the UAVs' trajectory planning and time resource assignment with the constraint of maximum flight speed, peak uplink power and flight area. Nevertheless, the minimum throughput optimization problem is not convex which is unmanageable. In order to tackle this problem, we introduce the concept of DQL.
2) We put forward a multi-agent DQL based strategy in order to maximize the minimum throughput by jointly optimizing UAVs' path design and time resource assignment. In particular, each UAV owns an independent DQN for making action strategy while the other UAVs are considered as a part of environment. After each epoch, UAVs receive a reward or penalty based on the minimum throughput.
3) The simulation results illustrate that our algorithm accomplishes significant performance improvement in the field of minimum throughput optimization compared with the traditional schemes.

B. ORGANIZATION

The rest of this paper is organized as follows. In Section II, the multi-UAV enabled WPCN model is presented, and we formulate the minimum throughput maximization problem. In Section III, the multi-agent DQL based algorithm is proposed to jointly design UAVs' trajectory and time resource allocation. Our simulation results are provided in Section IV to demonstrate the effectiveness of the proposed algorithm. Finally, conclusions is given in Section V.

II. PRELIMINARIES

In this section, we first introduce the system model of the considered multi-UAV enabled WPCN, and then formulate the corresponding UAVs' path planning and time resource allocation problem.

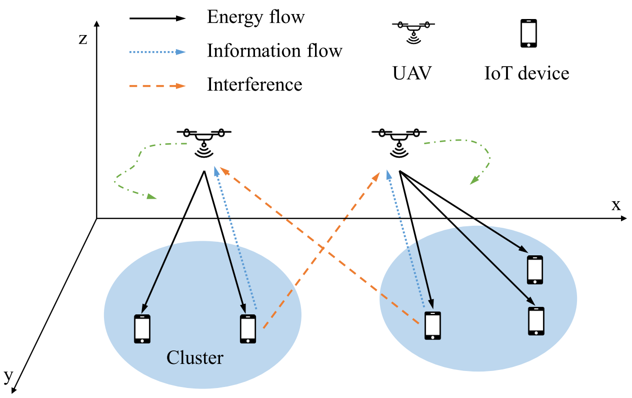


FIGURE 1. A multi-UAV enabled WPCN.

A. SYSTEM MODEL

We consider a WPCN system in which multiple UAVs perform as aerial BSs to support ground IoT devices in a given area as shown in Fig. 1. We repartition the IoT devices into L clusters and each UAV is in charge of a cluster. All the UAVs are equipped with single antenna and share the same frequency band. IoT devices in the particular area are denoted as $\mathcal{K} = \{K_1, \cdots, K_L\}$, where devices in the l-th cluster are denoted as K_l, $l \in \mathcal{L} = \{1, 2, \cdots, L\}$. Then, we have $K_l \cap K_{l'} = \emptyset, l' \neq l, l \in \mathcal{L}$, which means there is no overlap between the clusters. For any cluster l, $l \in \mathcal{L}$, we consider a UAV-enabled TDMA system which adopts HTT protocol, where the UAVs travel through the area periodically to charge the cluster via downlink WPT, and each device utilizes its collected energy to send the information in the uplink.

Let us analyze the system within a specific flight period of the UAVs, represented as $t \in [0, T]$. We describe the locations of IoT devices and UAVs in a 3D Cartesian coordinate system. To be specific, the locations of device $k_l \in K_l$ and UAV l are respectively denoted as $w_{k_l} = (x_{k_l}, y_{k_l}, 0)$ and $q_l(t) = (x_l(t), y_l(t), h_l(t)), h_{min} \leq h_l(t) \leq h_{max}, h_l(t)$ denotes the altitude of UAV l. To facilitate the analysis, the flight period T is discretized into $N+1$ time slots. In order to make sure that the UAVs is approximately stationary in a time slot, the number N is selected to be adequately large. Suppose v_{max} is the maximum speed of UAVs, then the location of UAVs should satisfy

$$\left\| q_l[n] - q_l[n-1] \right\| \leq V_{max} \cdot \delta_N, \tag{1}$$

where $\delta_N = \dfrac{(1-\alpha)T}{N}$ denotes the length of each subslot for uplink, α stands for the proportion of downlink WPT in a period.

The channel condition of UAVs and IoT devices in our system can be regarded as air-to-ground channel, in which the LoS and non-line-of-sight (NLoS) appear randomly. The probability of LoS can be expressed as [27]

$$P_{LoS}(\theta_{l,k_l}) = b_1(\frac{180}{\pi}\theta_{l,k_l} - \zeta)^{b_2}, \tag{2}$$

where $\theta_{l,k_l}[n] = sin^{-1}(\dfrac{h_l[n]}{d_{l,k_l}[n]})$ denotes the elevation angle from IoT device k_l to the UAV l in the n-th time slot, $d_{l,k_l}[n] = \sqrt{(x_l[n] - x_{k_l})^2 + (y_l[n] - y_{k_l})^2 + {h_l}^2[n]}$ is the distance between UAV l and device k_l. Besides, b_1 and b_2 stand for constant values representing the environment influence, ζ is another constant value which is determined by both the antenna and the environment. Note that, the NLoS probability is $P_{NLoS} = 1 - P_{LoS}$.

The path loss model for LoS and NLoS links between UAV l and device k_l is given by [28]

$$L_{l,k_l} = \begin{cases} \mu_1(\dfrac{4\pi f_c d_{l,k_l}}{c})^{\alpha}, & \text{LoS link,} \\ \mu_2(\dfrac{4\pi f_c d_{l,k_l}}{c})^{\alpha}, & \text{NLoS link,} \end{cases} \tag{3}$$

where μ_1 and μ_2 are the attenuation coefficients of the LoS and NLoS links, f_c and c denotes the carrier frequency and the speed of light respectively, α stand for the path loss exponent. Considering (2) and (3), the channel's power gain between UAV l and device k_l can be denoted as

$$g_{l,k_l}[n] = [P_{LoS}\mu_1 + P_{NLoS}\mu_2]^{-1}(K_0 d_{l,k_l}[n])^{-\alpha}, \tag{4}$$

where $K_0 = \dfrac{4\pi f_c}{c}$.

Next, we illustrate the TDMA and HTT transmission protocol of the UAV-enabled WPCN in detail. As mentioned above, there are $N+1$ time slots in each flight period T. Specifically, the 0-th time slot is assigned to the downlink WPT and the n-th time slot, $n \in \mathcal{N} = \{1, 2, \cdots, N\}$ is allocated to the uplink WIT. We use binary variable $a_l[0]$ to denote the downlink WET mode of UAV l, $a_l[0]$ equaling 1 or 0 respectively represent that the energy is transferred or not by the l-th UAV; while $a_{l,k_l}[n]$ is used to represent the uplink WIT allocation between UAV l and IoT device k_l at n-th time slot. Specifically, $a_{l,k_l}[n]$ equaling 1 or 0 means IoT device k_l does communicate or does not with the l-th UAV. Since the TDMA protocol is employed, the following constraints on the time resource allocation should be considered

$$\begin{aligned} a_l[0] &= \{0, 1\}, \forall l \in \mathcal{L}, \\ a_{l,k_l}[n] &= \{0, 1\}, \forall l \in \mathcal{L}, k_l \in \mathcal{K}, n \in \mathcal{N}, \\ \sum_{k_l \in K_l} a_{l,k_l}[n] &\leq 1, \forall K_l \in \mathcal{K}, l \in \mathcal{L}, n \in \mathcal{N}. \end{aligned} \tag{5}$$

At 0-th time slot of each flight period, the UAVs transmit the downlink energy signals with the transmit power P^D. Therefore, the collected energy of each IoT device k_l at period T is expressed as

$$E_{k_l} = \sum_{i=1}^{L} \eta \cdot \alpha \cdot T \cdot a_i[0] \cdot g_{i,k_l}[0] \cdot P^D, \ \ \forall l \in \mathcal{L}, \ k_l \in \mathcal{K}, \tag{6}$$

where $\eta \in (0, 1]$ denotes the RF-to-direct current(DC) energy conversion efficiency of each device.

Then, we consider the WIT mode for IoT device $k_l \in \mathcal{K}$ at time slot n. Let $P^U_{k_l}[n]$ denotes the uplink power of device k_l

at n-th time slot, then the available energy $E_{k_l}[n]$ of device k_l in n-th time slot can be represented as

$$E_{k_l}[n] = E_{k_l} - \sum_{j=1}^{n-1} a_{l,k_l}[j] \cdot \delta_N \cdot P_{k_l}^U[j]. \tag{7}$$

Therefore, the upper bound of uplink power for IoT device k_l should satisfy

$$\begin{aligned} a_{l,k_l}[n]\delta_N P_{k_l}^U[n] &\leq E_{k_l}[n], \\ \sum_{j=1}^{N} a_{l,k_l}[j] \cdot \delta_N \cdot P_{k_l}^U[j] &\leq E_{k_l}. \end{aligned} \tag{8}$$

Accordingly, the received SINR $\gamma_{k_l}[n]$ of UAV l connected to IoT device k_l at time slot n is given by

$$\gamma_{k_l}[n] = \frac{P_{k_l}^U[n] g_{l,k_l}[n]}{I_{k_l}[n] + \sigma^2}, \tag{9}$$

where $\sigma^2 = B_{k_l} N_0$, N_0 represents the power spectral density of the additive white Gaussian noise (AWGN) at the receivers. Moreover, $I_{k_l}[n] = \sum_{j=1, j\neq l}^{L} P_{k_j}^U[n] g_{l,k_j}[n]$ is the inference received by UAV l from cluster $j, j \in \mathcal{L}, j \neq l$.

Then the instantaneous throughput $R_{k_l}[n]$ of IoT device k_l can be represented as

$$R_{k_l}[n] = B_{k_l} log_2(1 + \frac{P_{k_l}^U[n] g_{l,k_l}[n]}{I_{k_l}[n] + \sigma^2}). \tag{10}$$

Therefore, the average throughput R_{k_l} of IoT device k_l of the flight cycle T can be denoted by

$$\begin{aligned} R_{k_l} &= \frac{1}{T} \sum_{n=1}^{N} a_{l,k_l}[n] R_{k_l}[n] \\ &= \frac{1}{T} \sum_{n=1}^{N} a_{l,k_l}[n] B_{k_l} log_2(1 + \frac{P_{k_l}^U[n] g_{k_l}[n]}{I_{k_l}[n] + \sigma^2}). \end{aligned} \tag{11}$$

B. PROBLEM FORMULATION

Let $A = \{a_l[0], a_{l,k_l}[n], \forall l, k_l, n\}$, $P^U = \{P_{k_l}^U[n], \forall k_l, n\}$, $Q = \{q_l[n], \forall l, n\}$. In this work, our optimization objective is to maximize the minimum average throughput of a multi-UAV enabled WPCN by jointly optimizing the IoT devices' association $\{a_l[0], a_{l,k_l}[n]\}$, the uplink power $\{P_{k_l}^U[n]\}$, and the UAVs' 3D trajectory $\{q_l[n]\}$. Therefore, the throughput optimization problem can be mathematically formulated as follows

$$(P1) \quad \max_{R_{min}, A, P^U, Q} \quad R_{min}$$

$$s.t. \; K_l \cap K_{l'} = \emptyset, l \in \mathcal{L}, \tag{12.1}$$

$$h_{min} \leq h_l[n] \leq h_{max}, \forall l \in \mathcal{L}, \tag{12.2}$$

$$a_l[0], a_{l,k_l}[n] = \{0, 1\}, \forall l \in \mathcal{L}, k_l \in \mathcal{K}, n \in \mathcal{N}, \tag{12.3}$$

$$\sum_{k_l \in K_l} a_{l,k_l}[n] \leq 1, \forall K_l \in \mathcal{K}, l \in \mathcal{L}, n \in \mathcal{N}, \tag{12.4}$$

$$\sum_{j=1}^{N} a_{l,k_l}[j] \cdot \delta_N \cdot P_{k_l}^U[j] \leq E_{k_l}, \tag{12.5}$$

$$R_{k_l} \geq R_{min}, \forall k_l \in \mathcal{K}, \tag{12.6}$$

$$\left\| q_l[n] - q_l[n-1] \right\| \leq V_{max} \cdot \delta_N. \tag{12.7}$$

Constraint (12.1) indicates that each device belongs to a non-overlapping cluster and associates with a specific UAV. Constraint (12.2) indicates the flight range of UAVs. Constraint (12.3) and (12.4) represent the time resource allocation restrictions. Equation (12.5) qualifies the peak uplink power constraint of each IoT device. Constraint (12.6) indicates the minimum rate requirement of each IoT device. Constraint (12.7) represents the maximum speed constraint of UAVs.

It can be observed that there exist two reasons making it difficult to solve problem (P1). First, equations (12.3) and (12.4) have binary constraints on $a_l[0]$ and $a_{l,k_l}[n]$. Besides, constraints (12.5) and (12.6) have complicated energy and rate functions with respect to coupled variables $a_l[0]$, $a_{l,k_l}[n]$, $P_{k_l}^U[n]$, $q_l[n]$. Therefore, problem (P1) is mixed-integer non-convex, and we can't obtain a feasible solution by general methods. As a result, we come up with a DQL based strategy for the purpose of optimizing the minimum throughput.

III. JOINT MULTIPLE UAVS' 3D TRAJECTORY PLANNING AND TIME RESOURCE ASSIGNMENT ALGORITHM

Since the throughput optimization problem is non-convex which is complicated to resolve directly, we bring in the DQL algorithm in this section to solve the minimum throughput maximization problem. In particular, we introduce the background of DQL first and then describe the proposed throughput optimization strategy in detail.

A. DEEP Q LEARNING

A RL problem can be described as a Markov Decision Process (MDP), which is defined by a 4-tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} >$. In particular, $\mathcal{S} = \{s_1, s_2, \cdots, s_m\}$ represents the state space, $\mathcal{A} = \{a_1, a_2, \cdots, a_m\}$ denotes action space. $\mathcal{R}$ denotes the reward function and particularly $R(s, a)$ represents the reward for executing action a at state s. $\mathcal{P}$ is the transition probability matrix. The optimal policy is obtained through the interaction between RL agent and the environment. To be specific, an RL agent observes the environment and then obtains the current state $s_t \in \mathcal{S}$. The next state of agent s_{t+1} can be obtained after choosing and executing an action $a_t \in \mathcal{A}$. At the end of a cycle, the agent receives a reward r_t according to the environment.

RL is designed to find a optimal policy $\pi(s)$ for maximizing the cumulative expectation of rewards. The cumulative reward at t-th step by executing action a at state s on the basis of policy π can be represented by

$$Q^\pi(s, a) = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t, a_t, \pi], \tag{13}$$

where $\gamma \in [0, 1]$ is the discount factor.

As a kind of model free RL, Q learning (QL) evaluates the value of action a executed at state s without building the environment transition model. The Q value determined by the state-action pair is stored in a look-up table and is updated as follow

$$Q(s,a)=(1-\alpha)Q(s,a)+\alpha(R(s,a)+\lambda \max_{a'} Q(s',a')), \quad (14)$$

where $\alpha \in (0, 1]$ is the learning rate, s', a' are respectively the next state and next action. It's proved that Q learning can converge to Q^* in the case where state and action spaces are discrete and finite.

However, since the UAVs fly flexibly in a 3D area, our model has a large and continuous state space. The storage and search of the Q-table becomes impractical and the convergence rate might become slow. Function approximation is adopted in several research to tackle this problem [29]. As a kind of non-linear function approximation, deep neural network (DNN) has been widely applied for large-scale reinforcement learning [23], [30], i.e. $Q(s, a) \approx Q(s, a, w)$, where w represents the weight parameters of neural network. In DQL, the distribution of Q value function is approximated by DNN, and the DNN is trained by means of optimizing the loss function

$$L(w) = E[(y_t - Q(s, a, w))^2], \quad (15)$$

where y_t is the target Q value which is set as a label and can be denoted by

$$y_t = r + \gamma \max_{a'} Q(s', a', w). \quad (16)$$

As a combination of RL and DL, DRL might be unstable because of two reasons. First, the training samples in RL are relevant and can't meet the independent and identical distribution demand of DL. Besides, a slight update of Q parameter may cause a huge oscillation in the strategy, which will bring a variation in the distribution of training samples. Experience replay and target network mechanism are developed in order to solve these issue [31]. In particular, replay buffer is applied to store the state transition samples (s, a, r, s') generated at each episode which can be randomly sampled for learning. Due to the randomness of the samples, the correlation between these data can be eliminated. In addition, target network own the same structure as the online network but different weight parameters. In particular, the parameters in target network remain unchanged and will be duplicated from online network periodically, thus the stability of the target can be ensured.

Since multiple agents interact simultaneously with environment and potentially with each other, it's more complex to learn in a multi-agent environment than in the single-agent case. In [32], authors first introduce an independent Q learning (IQL) strategy for multi-agent scenario. Based on this work, the authors in [33] combine DQN and IQL and discuss the phenomena such as cooperation, communication and competition in reinforced multi-agent systems. In [33], each agent learns the action strategy with its independent DQN and executes the action separately, and the other agents are seen as part of environment. The Markov Property becomes invalid in this approach and the environment is not stationary. Despite these disadvantages, IQL achieves great results with low complexity.

B. PROPOSED DQL-BASED SOLUTION

In our proposed multi-agent DQL-based algorithm, the IoT devices are uniformly distributed at an area, which can be partitioned into L clusters by K-means [34]. Each agent stands for a UAV, which owns an independent DQN and performs action respectively. Meanwhile, the agents share the state with others and regard the others as a part of environment. After each epoch, agents get a reward or penalty based on the shared environment.

Let us illustrate the definition of state space, action space and reward function of agents in our algorithm.

- The state space of each agent is made up of three parts:
  1) $q_l[n]$: the location of UAV l;
  2) $\{a_{l,k_l}[n]\}$: the number of times that each device communicates with UAV l;
  3) $\{R_{k_l}[n]\}$: the average throughput of devices in l-th cluster.
- The action space contains 27 elements which are defined by (x, y, z): (x, y, z) varying from $(-1, -1, -1)$ to $(1, 1, 1)$. To be specific, $x = -1$ stands for that the UAV turns to the left; $x = 1$ signifies that the UAV flies towards right; $y = -1$ implies the UAV flies backward; $y = 1$ means the UAV flies forward; $z = -1$ represents the UAV descends; $z = 1$ means the UAV rises; $(x, y, z) = (0, 0, 0)$ indicates the UAV remains still. After flying to the next location, UAV broadcasts energy flow or selects the device that owns the best channel condition in its cluster for uplink communication.
- The reward function is defined as follows:
  1) If UAV flies beyond the border after performing action, then the UAV receives a penalty of -1 and will be located at the boundary.
  2) At each time step, if there is a cross between the trajectory of UAV i and UAV j, then UAV i and UAV j receive a penalty of -1 and stay at the previous location.
  3) After each epoch, if the throughput of device in communication does not increase, which means that the device communications with UAV too many times, its energy is exhausted and thus the UAV only receives the interference, in this situation the UAV receives a penalty of -1; if the device's throughput increases, then the UAV gets a reward of 1.
  4) After each epoch, if the minimum average throughput of devices in a cluster is 0, which means that some devices do not communication with the UAV in this epoch, then the UAV receive a penalty of -2.

5) After each epoch, if the minimum average throughput of all devices does not increase, then all the UAVs receive a penalty of -1; if the minimum averaget throughput increases, all UAVs receive a reward of 1.

Algorithm 1 Proposed 3D Trajectory Design and Time Resource Allocation Solution Based on DQL

```
Initialize target network and online network;
Initialize UAVs' location and IoT devices' location;
for episode = 1, ..., M, do
    for time slot t = 1, ..., T, do
        for UAV i = 1, ..., L, do
            Choose action with $\epsilon$-greedy, while $\epsilon$ increases;
            Get UAV i's next location;
            if UAV i flies beyond the border then
                UAV i stays at the border, and gets a penalty of $-1$;
            end if
        end for
        for UAV i = 1, ..., L, do
            if UAV i and UAV j's trajectory exists cross then
                UAV i and UAV j stay at the previous location, and get a penalty of $-1$;
            end if
            Execute action, and get next state;
            if device t's throughput does not increase then
                UAV i gets a penalty of $-1$;
            end if
        end for
        if time slot t = T then
            if minimum throughput of all devices in a cluster equals zero then
                The UAV get a penalty of $-2$;
            end if
            if minimum throughput of devices does not increase then
                All UAVs get a penalty of $-1$;
            end if
        end if
        Store (s,a,r,s') into replay buffer;
        Randomly select a minibatch of H samples from replay buffer;
        Train the network, and update weight;
    end for
end for
```

The complete algorithm to solve the minimum throughput optimization problem for multi-UAV enabled WPCN system with DQL technique is summarized in Algorithm 1.

IV. SIMULATION RESULTS

In this section, we present numerical results to validate the effectiveness and superiority of our proposed strategy in the field of minimum throughput maximization.

For our simulations, it's assumed that 25 IoT devices are uniformly distributed within a $50m \times 50m$ district. For ease of analysis, the flight period of UAVs is set as $T = 1s$. The transmission power for UAVs' downlink and peak power for IoT devices' uplink transmit are respectively $P^D = 40dBm$ and $P^U_{max} = -20dBm$. The uplink power of IoT device is defined by the available energy and average available time slots. The maximum speed of UAVs is set as $V_{max} = 6m/s$, and the height of UAVs are constraint within $[10, 20]m$. The energy conversion efficiency of devices is set to $\eta = 0.1$ [35]. Other simulation parameters are presented in Table 1.

TABLE 1. Simulation parameters.

Parameter	Description	Value
f_c	Carrier frequency	800MHz
B	Bandwidth	1MHz
σ^2	Noise power spectral	-110dBm
b_1, b_2	Environmental parameters	0.36, 0.21
α	Path loss exponent	2
μ_1	Additional path loss for LoS	3dB
μ_2	Additional path loss for NLoS	23dB
α_L	Learning rate for DQN	0.0001
γ	Discount factor	0.7

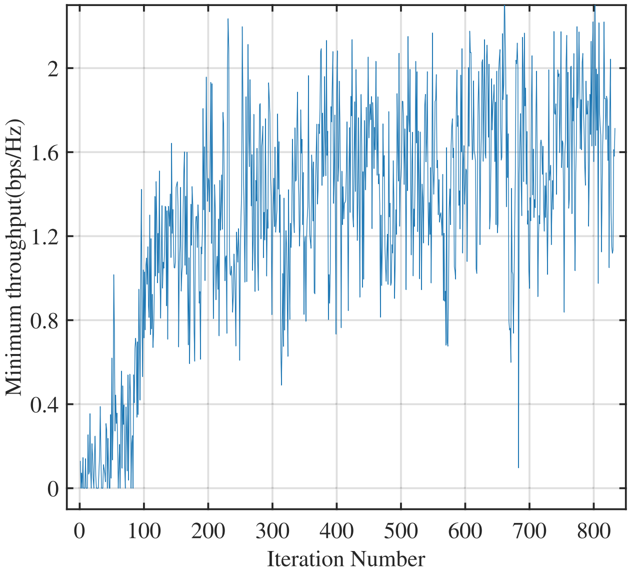


FIGURE 2. Uplink minimum throughput with respect to iteration number.

First, we illustrate the converge property of the proposed joint trajectory design and time resource allocation algorithm in a special case with $L = 3$ UAVs. In order to observe the results more intuitively, we make an average of throughput for every 60 periods. As shown in Fig. 2, the minimum throughput converges to a stable value after 400 iterations for the proposed algorithm.

Afterwards, we investigate the minimum throughput maximization performance of the proposed DQL-based algorithm under different number of UAVs. Moreover, we compare our proposed algorithm with the following strategies [36].

- Static: UAVs are fixed right above the centroid of its cluster $c = (x_c, y_c)$, while the height of UAVs are set as $H = 15m$. The IoT devices communication with its UAV in sequence.
- Circular trajectory: UAVs fly at a plane with altitude of $15m$, and follow a circular trajectory scheme. In this scenario, the center $c = (x_c, y_c)$ is set at the centroid of a cluster and the radius $r = min(r^c, r^v)$, in which $r^c = \frac{1}{K_l}\sum_{k=1}^{K_l} \|c - u_k\|$ and $r^v = \frac{v_{max} \cdot T}{2\pi}$ respectively indicate the average distance between centroid and IoT devices and the maximal radius determined by speed constraint. Same as the static scheme, the IoT devices are served by its UAV in sequence.

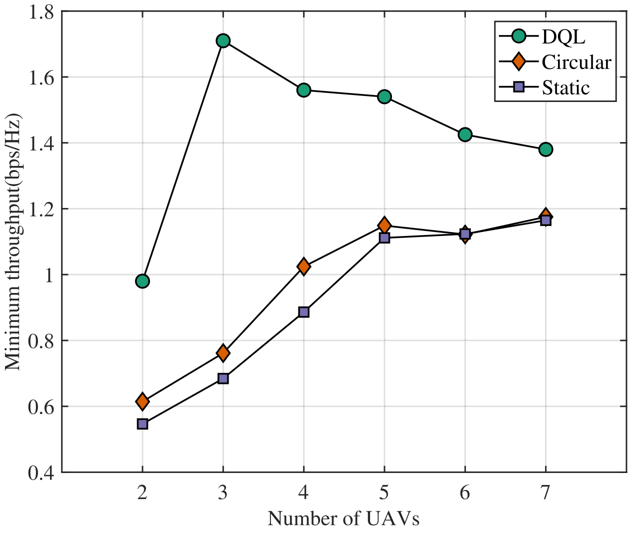


FIGURE 3. Maximum minimum throughput with respect to the number of UAVs.

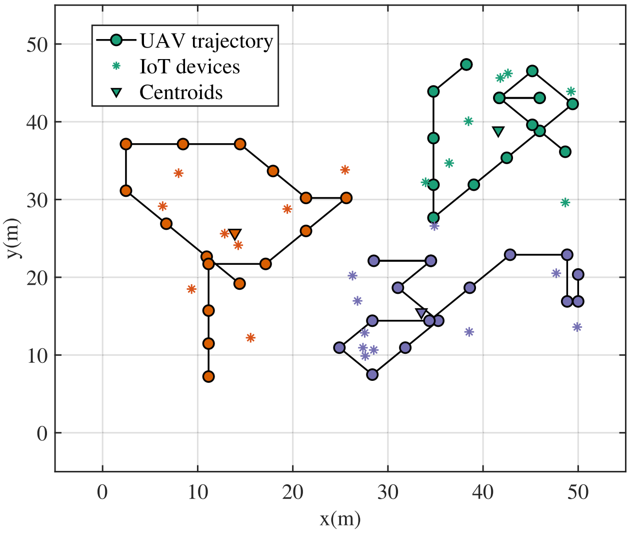


FIGURE 4. Trajectories of UAVs optimized by the proposed algorithm for UAV = 3.

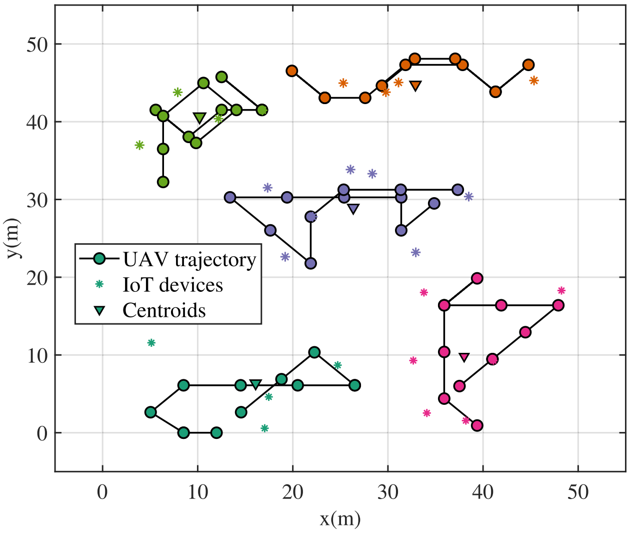


FIGURE 5. Trajectories of UAVs optimized by the proposed algorithm for UAV = 5.

The parameters of constraints are identical as previous in this simulation. The number of UAVs varies from 2 to 7. From Fig. 3, it's seen that comparing to the static WPCN, the proposed method can achieve better minimum throughput performance, which demonstrates that the flexibility of UAV can improve the communication quality of WPCN. Furthermore, for our proposed DQL, the minimum throughput increases when the number of UAV increases from 2 to 3, but decreases afterwards. This is because as the number of agents increase, the cooperation between agents becomes more complicated. On the other hand, as shown in circular and static scheme, the throughput does not increase any more when the number of UAVs is greater than 5. This is because as the number of UAVs increases, the number of devices in a cluster decreases, thus the harvested energy and time of allocated uplink communication increase whereas the distance between UAVs gets closer and thus the co-interference increases. In the end, the gains and interference offset each other. Overall, our proposed DQL-based algorithm provides better performance in maximizing the minimum throughput of UAV-enabled WPCN.

The optimized flight trajectories of multiple UAVs for UAV $= 3$ and 5 are respectively represented in Fig. 4 and Fig. 5. For ease of observation, we choose to observe the trajectory in a 2-dimensional coordinate system. The star represents the IoT devices, and the triangle represents the centroid of a cluster. As it can be seen in Fig. 4, it can be observed that the UAVs attempted to cover all the devices by flying around the centroid of cluster. Moreover, the UAVs hover close with the devices in its cluster to improve the channel quality and stay away from each other as far as possible to reduce the co-interference. In a word, the optimization algorithm tends to make a balance between good channel condition and existing co-interference. As shown in Fig. 5, the above rules are also applicable to the trajectory when 5 UAVs are deployed. These simulation results represent that the proposed algorithm can plan the trajectory excellently no matter how many UAVs there are.

V. CONCLUSION

In this paper, we investigate the throughput maximization problem for multi-UAV enabled WPCN in which UAVs act as wireless charger and information receiver to support ground IoT devices. Our target is to maximize the minimum throughput while satisfying several constraints including maximum flying speed, maximum uplink transmit power, time resource allocation. The formulated jointly UAVs' 3D trajectory design and time resource assignment optimization problem is non-convex, which is difficult to solve straightforward. A multi-agent DQL based algorithm is proposed for a feasible solution. Numerical results illustrate that the proposed strategy surpasses the traditional strategies in the field of maximizing the minimum throughput of multi-UAV enabled WPCN, which confirms the advantage of adopting UAVs' trajectory design and time resource allocation in WPCN system.

REFERENCES

[1] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.

[2] Q. Wu, W. Chen, D. W. K. Ng, and R. Schober, "Spectral and energy-efficient wireless powered IoT networks: NOMA or TDMA?" *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6663–6667, Jul. 2018.

[3] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[4] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 757–789, 2nd Quart., 2015.

[5] F. Yang, W. Xu, Z. Zhang, L. Guo, and J. Lin, "Energy efficiency maximization for relay-assisted WPCN: Joint time duration and power allocation," *IEEE Access*, vol. 6, pp. 78297–78307, 2018.

[6] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Jan. 2014.

[7] L. Liu, R. Zhang, and K.-C. Chua, "Multi-antenna wireless powered communication with energy beamforming," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4349–4361, Dec. 2014.

[8] J. Kim, H. Lee, C. Song, T. Oh, and I. Lee, "Sum throughput maximization for multi-user MIMO cognitive wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 913–923, Feb. 2017.

[9] B. Lyu, Z. Yang, G. Gui, and Y. Feng, "Wireless powered communication networks assisted by backscatter communication," *IEEE Access*, vol. 5, pp. 7254–7262, 2017.

[10] B. Lyu, H. Guo, Z. Yang, and G. Gui, "Throughput maximization for hybrid backscatter assisted cognitive wireless powered radio networks," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2015–2024, Jun. 2018.

[11] S. Bi, Y. Zeng, and R. Zhang, "Wireless powered communication networks: An overview," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 10–18, Apr. 2016.

[12] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4983–4996, Dec. 2016.

[13] G. Zhang, H. Yan, Y. Zeng, M. Cui, and Y. Liu, "Trajectory optimization and power allocation for multi-hop UAV relaying communications," *IEEE Access*, vol. 6, pp. 48566–48576, 2018.

[14] S. Cho, K. Lee, B. Kang, K. Koo, and I. Joe, "Weighted harvest-then-transmit: UAV-enabled wireless powered communication networks," *IEEE Access*, vol. 6, pp. 72212–72224, 2018.

[15] L. Xie, J. Xu, and R. Zhang, "Throughput maximization for UAV-enabled wireless powered communication networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1690–1703, Apr. 2019.

[16] 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, Mar. 2018.

[17] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.

[18] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari, and F. Adachi, "Deep learning for physical-layer 5G wireless techniques: Opportunities, challenges and solutions," *IEEE Wireless Commun. Mag.*, to be published, doi: 10.1109/MWC.2019.1900027.

[19] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Trans. Veh. Technol.*, to be published, doi: 10.1109/TVT.2019.2949122.

[20] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao, "Flight delay prediction based on aviation big data and machine learning," *IEEE Trans. Vehicular Technol.*, to be published.

[21] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[22] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019.

[23] 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, Dec. 2013. [Online]. Available: http://arxiv.org/abs/1312.5602

[24] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7957–7969, Aug. 2019.

[25] B. Zhang, C. H. Liu, J. Tang, Z. Xu, J. Ma, and W. Wang, "Learning-based energy-efficient data collection by unmanned vehicles in smart cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1666–1676, Apr. 2018.

[26] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

[27] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 2898–2904.

[28] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Wireless communication using unmanned aerial vehicles (UAVs): Optimal transport theory for hover time optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8052–8066, Dec. 2017.

[29] X. Xu, L. Zuo, and Z. Huang, "Reinforcement learning algorithms with function approximation: Recent advances and applications," *Inf. Sci.*, vol. 261, pp. 1–31, Mar. 2014.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[31] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Auton Agent Multi-Agent Syst*, vol. 33, no. 6, pp. 750–797, Nov. 2019.

[32] M. Tan, "Multi-agent reinforcement learning: Independent vs. Cooperative agents," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 330–337.

[33] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0172395.

[34] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[35] I. Krikidis, S. Timotheou, S. Nikolaou, G. Zheng, D. W. K. Ng, and R. Schober, "Simultaneous wireless information and power transfer in modern communication systems," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 104–110, Nov. 2014.

[36] J. Park, H. Lee, S. Eom, and I. Lee, "UAV-aided wireless powered communication networks: Trajectory optimization and resource allocation for minimum throughput maximization," *IEEE Access*, vol. 7, pp. 134978–134991, 2019.

JIE TANG (Senior Member, IEEE) received the B.Eng. degree in information engineering from the South China University of Technology, Guangzhou, China, in 2008, the M.Sc. degree (Hons.) in communication systems and signal processing from the University of Bristol, U.K., in 2009, and the Ph.D. degree from Loughborough University, Leicestershire, U.K., in 2012. He held Postdoctoral research positions at the School of Electrical and Electronic Engineering, The University of Manchester, U.K. He is currently an Associate Professor with the School of Electronic and Information Engineering, South China University of Technology, China. His research interests include green communications, NOMA, 5G networks, SWIPT, heterogeneous networks, cognitive radio, and D2D communications.

He was a co-recipient of the Best Paper Awards in IEEE ICNC 2018, CSPS 2018, and IEEE WCSP 2019. He also served as a Track Co-Chair for the IEEE Vehicular Technology Conference (VTC) Spring 2018. He is currently serving as an Editor for IEEE Access, *EURASIP Journal on Wireless Communications and Networking*, *Physical Communications*, and *Ad Hoc & Sensor Wireless Networks*.

JINGRU SONG received the B.Eng. degree from the School of Information Science and Engineering, Shandong University, Jinan, China, in 2018. She is currently pursuing the M.Sc. with the School of Electronic and Information Engineering, South China University of Technology, China, under the supervision of Dr. Jie Tang. Her research interests include machine learning, unmanned aerial vehicle, wireless power transmission, simultaneous wireless information and power transfer, and 5G networks.

JUNHUI OU was born in Guangdong, China. He received the B.E. degree in automation and the M.Sc. and Ph.D. degrees in communication engineering from Sun Yat-sen University, Guangdong, in 2012, 2014, and 2018, respectively.

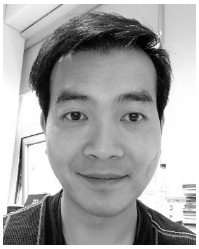
He holds a postdoctoral position with the South China University of Technology, Guangdong. His current research interests include antenna design, RF circuit design, wireless power transmission, and simultaneous wireless information and power transmission.

JINGCI LUO received the B.Eng. degree from the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2017, where she is currently pursuing the M.Sc. degree, under the supervision of Dr. Jie Tang. Her research interests include energy efficiency optimization, machine learning, non-orthogonal multiple access, simultaneous wireless information and power transfer, and 5G networks.

XIUYIN ZHANG (Senior Member, IEEE) received the B.S. degree in communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2001, the M.S. degree in electronic engineering from the South China University of Technology, Guangzhou, China, in 2006, and the Ph.D. degree in electronic engineering from the City University of Hong Kong, Kowloon, Hong Kong, in 2009.

From 2001 to 2003, he was with ZTE Corporation, Shenzhen, China. He was a Research Assistant, from July 2006 to June 2007, and a Research Fellow, from September 2009 to February 2010, with the City University of Hong Kong. He is currently a Full Professor and the Vice Dean with the School of Electronic and Information Engineering, South China University of Technology. He also serves as the Deputy Director of the Guangdong Provincial Engineering Research Center of Antennas and RF Techniques and the Vice Director of the Engineering Research Center for Short-Distance Wireless Communications and Network, Ministry of Education. He has authored or coauthored more than 100 internationally referred journal papers, including 55 IEEE Transaction papers as well as around 60 conference papers. His research interests include microwave circuits and sub-systems, antennas and arrays, and SWIPT.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology. He was a recipient of the National Science Foundation for Distinguished Young Scholars of China, the Young Scholar of the Chang-Jiang Scholars Program of Chinese Ministry of Education, and the Top-notch Young Professionals of National Program of China. He was also a recipient of the Scientific and Technological Award (Hons.) of Guangdong Province. He was supervisor of several conference best paper award winners. He has served as a Technical Program Committee (TPC) Chair/ member and session organizer/Chair for a number of conferences. He is an Associate Editor for IEEE Access.



KAI-KIT WONG (Fellow, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively.

After graduation, he took up academic and research positions at University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, Smart Antennas Research Group of Stanford University, and University of Hull, U.K. He is currently the Chair in wireless communications with the Department of Electronic and Electrical Engineering, University College London, U.K. His current research interests include 5G and beyond mobile communications, including topics such as massive MIMO, full-duplex communications, millimeter-wave communications, edge caching and fog networking, physical layer security, wireless power transfer and mobile computing, V2X communications, and cognitive radios. There are also a few other unconventional research topics that he has set his heart on, including for example, fluid antenna communications systems and team optimization.

Dr. Wong is a Fellow of IET and is also on the editorial board of several international journals. He was a co-recipient of the 2013 IEEE Signal Processing Letters Best Paper Award, the 2000 IEEE VTS Japan Chapter Award at the IEEE Vehicular Technology Conference in Japan, in 2000, and a few other international best paper awards. He served as an Associate Editor for the IEEE Signal Processing Letters, from 2009 to 2012, and an Editor for the IEEE Transactions on Wireless Communications, from 2005 to 2011. He was also a Guest Editor for the IEEE JSAC SI on Virtual MIMO, in 2013, and currently a Guest Editor for the IEEE JSAC SI on physical layer security for 5G. He has been a Senior Editor for the IEEE Communications Letters, since 2012, and for the IEEE Wireless Communications Letters, since 2016. He is also an Area Editor for the IEEE Transactions on Wireless Communications.