Nonlinear Energy-Harvesting for D2D Networks Underlaying UAV with SWIPT Using MADQN

Mohamed Amine Ouamri, Gordana Barb, Daljeet Singh, Abuzar B. M. Adam, M. S. A. Muthanna, Xingwang Li

***Abstract*—Energy Efficiency (EE) has become an essential metric in Device-to-Device (D2D) communication underlaying Unmanned Aerial Vehicles (UAVs) Among the several technologies that provide significant energy, simultaneous wireless information and power transfer (SWIPT) has been proposed as a promising solution to improve EE. However, it is a challenging task to study the EE under nonlinear energy harvesting (EH) due to the limited sensitivity and the composition of the nonlinear circuit. Moreover, when D2D users transmit information using the EH from UAVs, interferences to cellular users occur and deteriorate the throughput. To tackle these problems, we leverage concepts from artificial intelligence (AI) to optimize EE of UAV-assisted D2D communication. Specifically, multi-agent deep reinforcement learning was proposed to jointly maximize throughput and EE, where the reward function is defined in terms of the introduced goal. Simulation results verify the supremacy of proposed approach over traditional algorithms.**

***Index Terms*—Device-to-device, Energy harvesting, Multi-agent DRL, Resource allocation, SWIPT, UAV.**

I. Introduction

IN todays and future wireless communication networks, unmanned aerial vehicles (UAVs) are attracting further attention for enhancing coverage and capacity in diverse scenarios [1]. As such, with the use of flexible placement and reliable line-of-sight (LoS) link in air-to-ground (A2G) connexions, UAVs can assist congested terrestrial networks during sporting events or natural disaster scenarios [2]. Although UAVs offer several advantages such as a high probability of LoS transmission and easy transportation, the constraints related to UAVs communications must be taken into consideration when conceiving device-to-device (D2D) networks namely, interference management, quality of service (QoS), and in particular energy efficiency (EE) [3]. Among the multiple technologies that provide significant energy, simultaneous wireless information and power transfer (SWIPT) is projected to harvest some of the energy carried from the same RF signals [4], [5]. To improve the network EE, the power-splitting (PS) scheme is an effective technique employed in Energy Harvesting (EH) mechanism [6]. Meanwhile, Deep reinforcement learning (DRL) can be an efficient alternative to achieve sustainable EE in complex wireless networks with SWIPT [7], and has been widely used to analyse EH [8]. A handful of studies have incorporated this approach for D2D networks underlying UAVs, e.g. [9], [10]. However, these works do not address nonlinear EH while considering interference. For instance, most existing solutions are only applicable to UAV networks with no underlying D2D. Authors in [11] studied the three-dimension UAVs- enabled wireless communication system with EH to minimize the total energy consumption. Nevertheless, this work assumes that there are no terrestrial networks, and the optimization is not obtained using artificial intelligence (AI). DRL-based multi-UAVs data harvesting was proposed in [12]. This framework maximizes harvested data from internet of things devices and total throughput by optimizing trajectory. However, the authors did not consider the EH and the influence of UAV height on the network. In [13], the authors leveraged a deep neural network architecture to jointly optimize spectral efficiency (SE) and EH in D2D communication. This work adopted an unsupervised method and assumed a discrete-time block fading model. Although this work presents the trade-off between SE and EH, the existence of UAVs and nonlinear EH is not considered. To accommodate EE requirements in UAVs-assisted D2D communication, authors in [14] investigated on-board deep Q-network (DQN) to minimize the overall data packet loss. The objective is to find the device to be loaded and patrolling velocity of the UAVs. According to the above-mentioned literature review, and to the best of our knowledge, most existing contributions in this field focus only on EE and throughput optimization with SWIPT. However, it was noticed that the nonlinear EH schemes were not considered for D2D communication underlaying UAVs. Besides, most previous research rarely considers the influence of PS on energy and throughput together. Motivated by this assessment, we propose a MADRL to address intelligent resource allocation for UAV-assisted D2D networks where a UAV acts as an agent. Firstly, we introduce a nonlinear EH model for SWIPT to include the underlying harvest circuit. We then formulate EE and total throughput in a mmWave scenario while ensuring the minimum QoS requirements for all users according to the environment. Multiple constraints, such as path loss model, UAV height, distance, and minimum transmission rate are employed to describe our mathematical problem.

M. A. Ouamri: Universite´ de Grenoble Alpes, INP Grenoble, CNRS, LIG, Grenoble 38000, France (ouamrimouhamedamine@gmail.com).
Gordana Barb: Faculty of Electronics, Telecommunication and Information Technologies, Politehnica University Timișoara, Romania. (gordana.barb@upt.ro)
Daljeet Singh: Faculty of Medicine, Research Unit of Health Sciences and Technology, University of Oulu, Finland. (daljeetsingh.thapar@gmail.com)
A. B. M. Adam: School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (abuzar@cqupt.edu.cn)
M. S. A. Muthanna: Institute of Computer Technologies and Information Security, Southern Federal University, 347922 Taganrog, Russia. (muthanna@sfedu.ru)
Xi. Li: School of Physics and Electronics Information Engineering, Henan Polytechnic University, Jiaozuo, China (lixingwangbupt@gmail.com).

II. System Model

We consider a D2D communication underlying UAVs network with SWIPT. Our model includes N D2D pairs and a set of UAVs $U \in \{u_1, u_2, u_m\}$ distributed randomly as illustrated in Fig.1. We assume that UAVs are deployed at a particular altitude H. The fading between UAVs and device users is denoted as$h_{u,d}$, and it is assumed to be the small-scale fading. Without loss of generality, D2D user transmitter (D2D-Tx) can harvest energy from the RF energy transmitted by the UAVs [10]. Compared to [9], our framework considers that D2D-Tx applies a power split policy, which means splitting the received power into two parts such as EH power and information power. The PS ratio for D2D-Tx is given as $0 \leq \rho \leq 1$. A power-splitter is used for each D2D-Tx, where the ratio ρ of the power is allocated to the energy receiver and the rest ratio $1-\rho$ is split to the information receiver. Moreover, UAVs and D2D are supposed to transmit in the same mmWave frequency spectrum and therefore interfere with each other. In this work, a typical D2D is associated with UAV or D2D-Tx with an LoS link or a non-line-of-sight (NLoS) caused by blockage effects [15]. The position of each device on the ground and UAV are (x_d^i, y_d^i), and (x_u, y_u, H) where $i \in (D2D-Tx, D2D-Rx)$. Let P_d, P_u denote the power transmission of D2D-Tx and UAVs respectively, the downlink achievable throughput (data rate) of D2D-Rx can be expressed as

$$R_{ToT} = p_u B_u \log_2(1+\gamma_u) + p_d B_d \log_2(1+\gamma_d) \quad (1)$$

Where γ_u and γ_d are the signal-to-interference-plus-noise ratio (SINR), B_u and B_d represent the bandwidth for UAV and D2D respectively, $p_u = \left|(\delta - P_u^{LoS})\right|$ and $p_d = \left|(\delta - P_d^{LoS})\right|$ are the probabilities when the link between UAV/D2D-Tx and D2D-Rx is in LoS or NLoS transmission, where δ is a binary variable that defines the association mode and $|\cdot|$ is the absolute value. In particular, if D2D-Rx is associated with UAV/D2D-Tx according to LoS link, then $\delta = 0$, otherwise $\delta = 1$. We can introduce the following form to describe the LoS probability of A2G link [15]

$$P_u^{LoS}(z) = \frac{1}{1 + b + exp\left(-c\left(\frac{180}{\pi} tan^{-1}\left(\frac{H}{z}\right) - b\right)\right)} \quad (2)$$

where b and c are constants that depend on network environment and $z = \sqrt{(x_u - x_d^{DT_x})^2 + (y_u - y_d^{DT_x})^2 + H^2}$ is the Euclidean distance between the typical UAV and D2D-Tx. Similarly, the LoS probability function when communication is established between D2D-Tx and D2D-Rx is given by

$$P_d^{LoS}(D) = 1 - e^{\beta D} \quad (3)$$

where β is the blockage parameter that defines the average size of obstacles. Here, D corresponds to the distance between D2D pairs.

A. Channel Model

Regarding propagation, the signal through the wireless channel is characterized by attenuation that depends relatively on distance and antenna components. There is no doubt that an adequate channel model is essential for establishing reliable communication links. As discussed in [1], [7], the A2G

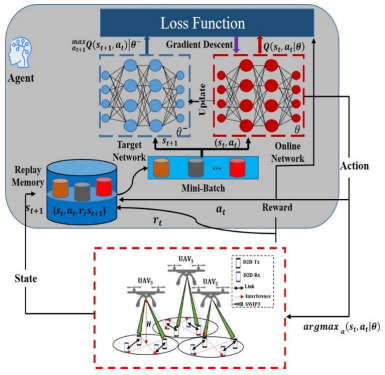


Fig. 1. DQN for D2D communication underlaying UAV.

channel depends strongly on the altitude and environment. However, according to [15], the path loss in LoS and NLoS transmission for A2G can be modeled as

$$PL_u = \begin{cases} \Phi_{LOS}(z)^{\alpha_{LOS}/2}, & \text{for LOS Link} \\ \Phi_{NLOS}(z)^{\alpha_{NLOS}/2} & \text{for NLOS Link} \end{cases} \quad (4)$$

where α_{LOS} and α_{NLOS} are the path loss exponents, $\Phi_{LOS} < \Phi_{NLOS} = (\vartheta_{i\in\{LoS,NLoS\}}(c/4\pi f))^{-1}$ represent the additional path loss for LoS and NLoS links respectively, which depend on environment and frequency. Indeed, an additional path loss is included because, in NLoS links, there is higher signal attenuation compared to LoS transmission. For D2D ground communication, we adopt a standard path loss model with mean additional losses δ_{LoS} and δ_{NLoS} as

$$PL_{SBS} = \begin{cases} \delta_{LOS} D^{-\alpha_{LOS}}, & \text{for LOS Link} \\ \delta_{NLOS} D^{-\alpha_{NLOS}} & \text{for NLOS Link} \end{cases} \quad (5)$$

where D is the distance between D2D-Tx and D2D-Rx. Thus, the channel power gain from the UAV to a particular D2D-Tx can be given as

$$G_{u,d} = P_u^{LoS} \Phi_{LOS}(z)^{\frac{\alpha_{LOS}}{2}} + P_u^{NLoS} \Phi_{NLOS}(z)^{\frac{\alpha_{NLOS}}{2}} \quad (6)$$

In addition, we assume that the channel gain between D2D-Tx and D2D-Rx is given by

$$G_{d,d} = P_d^{LoS} \delta_{LOS} D^{-\alpha_{LOS}} + P_d^{NLoS} \delta_{NLOS} D^{-\alpha_{NLOS}} \quad (7)$$

where $P_d^{NLoS} = 1 - P_d^{LoS}$ is the NLoS probability link.

B. Nonlinear Energy Harvesting and SINR

The energy harvesting process is considered to be a promising solution to power up low-power consuming devices. In particular, EH techniques can be exploited to generate energy from the surrounding environment. This converted energy will then be used in the network devices [16]. Available sources of energy for harvesters can be grouped into linear and nonlinear models. Generally, a practical EH circuit displays a non-linear characteristic due to the non-linear elements implemented in the UAVs and D2D users such as resistances, diodes and capacitance [1]. Here, we consider a nonlinear EH model, where the harvested power is expressed as [17]

$$EH = T\left(\frac{\left(\frac{P_{Max}}{1+e^{-a(P_{r,u}+I_{u'}+I_{d'})}} - \frac{P_{Max}}{1+e^{ab}}\right)}{1 - \frac{1}{1+e^{ab}}}\right) \quad (8)$$

where T is the transmission time, P_{Max} is the maximum harvested energy, a and b are constants that define resistance and capacitance, respectively. $P_{r,u} = \rho P_u h_{u,d} G_{u,d}$, $I_(u') = \sum_{u' \neq u} \rho P_{u'} h_{u',d} G_{u',d}$ and $I_{d'} = \sum_{d' \neq d} \rho P_{d'} h_{d',d} G_{d',d}$ represent the received power, interference from adjacent UAV and D2D-Tx, respectively. In this work, the SINR when the PS SWIPT is applied at the D2D-Tx can be given by

$$\gamma_u = \frac{(1-\rho) P_u h_{u,d} G_{u,d}}{\sum_{u' \neq u} (1-\rho) P_{u'} h_{u',d} G_{u',d} + \sigma^2} \tag{9}$$

where σ^2 refers to additive white Gaussian noise, $(1-\rho)$ designates the part of the received signal intended for information. On the other hand, the instantaneous SINR at the typical D2D-Rx can be computed as

$$\gamma_d = \frac{\rho P_d h_{d,d} G_{d,d}}{I_{d'} I_{u'} + \sigma^2} \tag{10}$$

C. Global Energy Efficiency

One of the objectives that keeps the network alive is energy efficiency. The EE metric is employed to evaluate the total energy consumption for the network, and it is defined as a ratio of the total transferred bits to the total power consumption. Based on the nonlinear harvested energy of D2D-Tx, the global energy efficiency of the UAVs assisted D2D communication can be formulated as

$$GEE = EE_u + EE_d = \frac{SE_u}{P_u + P_c + \sum_{n=1}^{N} EH} + \frac{SE_d}{P_d + P_c} \tag{11}$$

where $0 < P_u \leq P_{u,m}, P_c$ represent the power consumed in the circuit of the transmitter and SE_u is the spectral efficiency that is given by the following equation

$$SE_u = log_2 \left(1 + \frac{(1-\rho) P_u h_{u,d} G_{u,d}}{\sum_{u' \neq u} (1-\rho) P_{u'} h_{u',d} G_{u',d} + \sigma^2} \right) \tag{12}$$

III. Problem Formulation

Due to the discrete nature of data and the high traffic demands in UAV-assisted D2D networks, traditional algorithms for resource allocation do not converge on optimal solutions for improving throughput and achieving large-scale EE. In this section, we present the problem formulation where the performance analysis is calculated in terms of EE and throughput. Our objective is to maximize GEE and throughput. Mathematically the optimization problem is formulated as follows:

$$\max_{SE, P_u, P_d, \rho, \gamma_u, \gamma_d} \quad GEE \tag{13}$$

$$\begin{aligned} s.t. \quad & \mathbf{C1}: SE_u > SE_u^{QoS} \forall u \in U; \mathbf{C2}: SE_d > SE_d^{QoS} \forall d \in N \\ & \mathbf{C3}: 0 < P_u \leq P_u^{max}; \qquad \mathbf{C4}: 0 < P_d \leq P_d^{max} \\ & \mathbf{C5}: 0 < \rho \leq 1; \qquad \mathbf{C6}: \gamma_u > \gamma_u^{th} \forall u \in U \\ & \mathbf{C7}: \gamma_d > \gamma_d^{th} \forall d \in N \end{aligned}$$

Constraints **C1** and **C2** indicate the QoS constraint, where SE_u^{QoS} and SE_d^{QoS} are the spectral efficiency threshold in UAV and D2D communication, respectively. **C3** and **C4** impose that the transmit power P_u and P_d must be in the interval $[0, P_{max}]$. It specifies the upper limit of the power transmission. The constraint in C5 means that the power-splitting ratio is bounded between $[0, 1]$. Finally, constraints in **C6** and **C7** impose that the SINR γ_u, γ_d are more significant than a designated threshold γ_u^{th} and γ_d^{th}. On the other hand, the second objective focuses on total throughput maximization of the network, which is given by

$$\max_{P_u, P_d, \rho, H, \gamma_u, \gamma_d} \quad R_{TOT} \tag{14}$$

$$\begin{aligned} s.t. \quad & \mathbf{C1}: R_{TOT} \geq R_{TOT}^{DL^{min}} \forall d \in N \\ & \mathbf{C2}: H_{min} \leq H \leq H_{max} \forall u \in U \end{aligned}$$

Constraint **C1** indicates that the total throughput should be greater than the minimum required data rate for QoS. Constraint **C2** denotes that the altitude of UAVs ranges between $[H_{min}, H_{max}]$. Obviously, at higher altitudes, the distance between the UAV and D2D-Tx increases, resulting in a considerable path loss. On the other hand, when the UAV is positioned at a certain minimum height, the NLOS conditions are recorded and they can be affected by the throughput, thus the necessity to study this constraint. **C3**, **C4**, **C5**, **C6**, and **C7** are given in the previous paragraph.

IV. Multi Agent Deep Q-Network

In this section, we present MADRL to solve the problem of (13) and (14). The approach of RA is modeled as Markov decision process (MDP). In MADRL algorithms, the problem representation can be divided into three parts according to the nature of the interaction between the agents such as cooperative, competitive and mixed. Moreover, the agent can fully or partially observe the environment. In this letter, we propose an MADQN algorithm based on fully observable settings.

A. Reinforcement Learning

Similar to existing works [9], [10], we consider a tuple (s_t, a_t, r_t, s_{t+1}), where the agent observes the Markov state of the environment $s \in S$, and interacts to take an action $a \in A$. Here, S and A are the state-space and action-space respectively. Based on the transition probability $p(s_{t+1}|s_t, a_t)$ the current network state s_t transits to a new state according to the action a_t selected by the agent at time slot t. r_t denotes the reward function performed by the agent at each time slot t. However, we assume that UAVs act as an agent that continuously interacts with the environment to optimize policy π. For agent j, we denote network state at each time t as s_t^j. After observing the current environment, agent j performs an action according to the policy π_j and s_t^j. Then, at each time policy, the agent receives the reward r_t^j conditioned by an action and moves to the next state s_{t+1}^j. Lastly, j performs the above operations until the maximum episode is completed. We define the components, namely the state space, the action space and the reward function as

State and Observation: In general, the state describes a specific configuration of the network. UAV agents determine state s_t^j from the environment observation. Each UAV j observes the SINR γ_u^j , altitude H_t^j and power transmission

P_u^j. At time slot t, the observation of the agent j can be represented as $O_t^j = \{\gamma_u^j, H_t^j, P_u^j\}$. The environment is fully observable, and the state space is composed by all observations as $s_t^j = \{O_t^1, O_t^2, \cdots, O_t^j\}$.

Action space: The agent interacts with the environment and selects the action a_t^j. At each time slot t, the action of UAV includes the PS ratio ρ and altitude H_t^j. Therefore, the actions can be given as a tuple $a_t^j = \{\rho_t^j, H_t^j\}$.

Reward Function: Reinforcement learning is based on the reward function, stating that the UAV is guided towards an optimal policy. The objective of this work is to jointly maximize the total throughput and GEE. For this purpose, the rewards obtained by UAVs are expressed by vector $r_t^j = \{r_t^1, r_t^2\}$. In our model, we consider a scalarization reward function as

$$r = max\left(\omega_1 GEE + \omega_2 R_{ToT}\right) \tag{15}$$

where ω_1 and ω_2 are weight for each objective and $\sum \omega = 1$. We suppose that $\omega_1 = \omega_2 = 0.5$. A weighted reward criterion is a weighted combination of the average and discounted reward criteria. The agent can give more or less attention to the long-term reward than to the short-term reward by changing the associated weights.

B. Learning Algorithm Process

In this part, inspired by [10] we propose a MADQN to optimize the EE and throughput resource allocation framework. In particular, we extend the single-agent DQN algorithm to a multiple-agent approach. In a non-linear deep neural network (DNN), the Q function is defined as $Q(s_t^j, a_t^j)\theta \approx Q^*(s, a)$, where θ represents the weights of the neural networks. As illustrated in Fig. 1, the action a_t is taken according to the ϵ-greedy policy, and the transition tuple $(s_t^j, a_t^j, r_t, s_t^j)$ is stored in a replay memory denoted by D. To remove the correlation between the samples, the DQN agent will randomly sample minibatch from D to adjust θ of the DNN. Another target network model $Q(s_t^j, a_t^j; \theta^-)$ with weight θ^- is used in the DQN procedure to ensure DQN stability. At each episode, the optimal state-action function is formulated as

$$Q(s_t^j, a_t^j|\theta) = E_{s_{t'}}\left[r + \gamma \max_{a_t^{j'}} Q^*\left(s_t^{j'}, a_t^{j'}\right)\middle| s_t^j, a_t c\right] \tag{16}$$

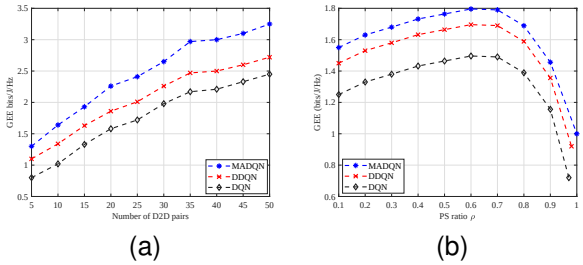
To train the DQN, the Q-network updates θ according to $L(\theta)$ to minimize the following loss function given as

$$L(\theta) = E_{s_t^j, a_t^j, r_t, s_{t+1}^j \in D}\left[r_t\left(s_t^j, a_t^j\right) + \gamma \max_{a^{t+1}} Q\left(s_t^j, a_t^j\right)\middle|\theta^- - Q\left(s_t^j, a_t^j\right)\middle|\theta\right]^2 \tag{17}$$

Our proposed pseudo code is outlined in Algorithm 1.

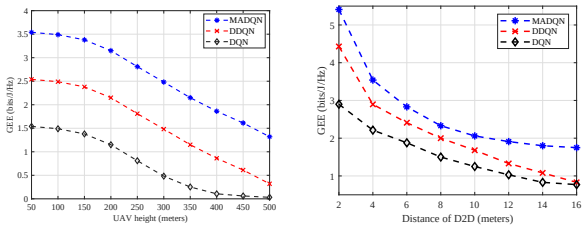
V. Simulation Results

We evaluate our proposed model and verify the effectiveness of the introduced algorithm-based energy harvesting for UAV-assisted D2D communication, comprising 3 UAVs and 25 D2D pairs uniformly distributed over an area of $3 \times 3km^2$. We assume that the maximum power transmission for UAVs is set as $P_u^{max} = 30dBm$. In addition, the power consumed in the circuit of the transmitter$P_c = 40dBm$. The additive white Gaussian noise $\sigma^2 = -114dBm$. In the MADQN algorithm, the DNN of each agent is a four-layer fully connected neural network with two hidden layers: 64 and 32 neurons in each hidden respectively. We compare the convergence of our algorithm to that of the DQN and Double DQN. First, Fig. 2(a) illustrates the effect of the number of D2D pairs on the GEE for different algorithms. A common observation in Fig. 2(a) is that by increasing the number of D2D pairs can lead to an increase in GEE, since the degree of improvement in EH becomes more important. As shown in Fig. 2(b), the power splitting ratio ρ has an impact on the GEE. Indeed, when the PS ratio is low, GEE increases significantly to the value GEE=1.8(bits/J/Hz) and then starts to decrease to a value of $\rho = 0.72$. This can be explained by the fact that at a low PS ratio, the EH by D2D-Tx increases. We can also observe from the two previous figures that the performance is clearly outperformed in MADQN compared to DDQN and DQN. Fig. 3(a) and Fig. 3(b) illustrate the variation of GEE



(a) (b)

Fig. 2. GEE (bits/J/Hz) versus (a) number of D2D pairs with different algorithms, and (b) PS ratio ρ.



(a) (b)

Fig. 3. (a) UAVs Height with different algorithms, and (b) GEE (bits/J/Hz) vs. Distance of D2D.

according to UAVs height and D2D distance, respectively. As can be seen from Fig. 3(a), the GEE decreases as the UAVs' height increases. This is due to the increase in the distance between UAV and D2D-Tx, resulting in a considerable path loss. Fig. 3(b) presents the GEE as function of the maximum distance between D2D. We conclude that when the distance between devices increases, the GEE gradually decreases. This is explained by the fact that increasing distance leads to greater path loss. Fig. 4(a) shows the variation of total throughput versus UAVs height for $\rho = 0.2$ and $\rho = 0.4$, with different AI algorithms. As can be easily seen in Fig. 4(a), total throughput

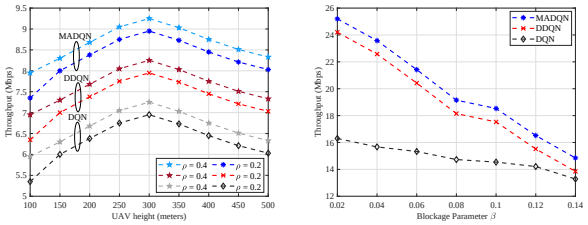


Fig. 4. (a) Throughput vs. UAV height with different ρ, and (b) Throughput as function of blockage β.

increases as the UAVs height increases and then decreases after $H = 300m$. This is because UAVs experience LoS conditions when $120 < H \leq 300m$ and the deterioration of the throughput after the $300m$ is due to increased pathloss. Finally, we plot the throughput as a function of blockage parameter β (blockage between D2D) in Fig. 4(b). It can be observed that when β increases, the total throughput of the network decreases. Thus, with the increases in obstacle density, more UEs are served by NLoS conditions. We can also observe from Fig. 4(b) that the proposed MADQN algorithm converges to highly satisfactory results compared to the other approaches. This is because our algorithm handles interference perfectly.

Algorithm 1 MADQN for UAVs assisted D2D

INITIALIZATION
Initialize parameters: learning rate, ϵ-greedy, discount factor γ^t, memory buffer D
for agent $j = 1$ to N **do**
Replay memory D to capacity N
Initialize action-value function $Q_j(s_j, a_j|\theta_j)$ with random weight θ_j
Target Network $Q_j(s'_j, a'_j|\theta_j^-)$, $\theta_j^- \rightarrow \theta_j$ (θ_j^- equals to θ_j)
end for
LEARNING
for iteration $1, 2, \ldots, L$ **do**
s_1: Initialize the first state from $s_j^t = (z_1^t, z_2^t \ldots z_j^t)$
for episode $1, E$ **do**
Each agent (UAV) j execute random actions a_j^t according to $\epsilon-$greedy
otherwise, choose $a_j^t = \max_{a^{t+1}} Q(s_j^{t+1}, a_j^t|\theta^-)$
Get the immediate reward $r(s_j^t, a_j^t)$ and next state s_j^{t+1}
Store transition $(s_j^t, a_j^t, r_t, s_j^{t+1})$ in D
Update SD-WAN information from the controllers
for $j = 1$ to N **do**
Controllers randomly sample minibatch of $(s_j^t, a_j^t, r_t, s_j^{t+1})$ in D
Set $r_t(s_j^t, a_j^t) + \gamma_{a^{t+1}}^{max} Q(s_j^{t+1}, a_j^t|\theta_j^-)$
Apply gradient descent step on $[r_t(s_j^t, a_j^t) + \gamma_{a^{t+1}}^{max} Q(s_j^{t+1}, a_j^t|\theta^-) - Q(s_j^t, a_j^t|\theta_j)]^2$
end for
end for
UAV j replaces target parameters $\theta_j^- \rightarrow \theta_j$;
Empty D;
end for

VI. Conclusion

This paper investigated the energy efficiency and throughput optimization in D2D communication underlying UAVs with SWIPT. The objective was to maximize both EE and sum-rate under power splitting. A distributed MADQN was applied and compared to traditional approaches such as DDQN and DQN. The simulation results demonstrated that our algorithm outperforms EE and throughput with different parameter variations. Our results also indicated that the EE is affected by the number of D2D pairs to be deployed in the coverage area, as well as the maximum altitude variation. Moreover, it is important to obtain an optimal value of PS ratio for efficient resource allocation. As a future research direction, we will extend our work to combine power splitting and time EH while considering UAV mobility.

References

[1] R. Jiang, K. Xiong, H.-C. Yang, P. Fan, Z. Zhong, and K. B. Letaief, "On the coverage of uav-assisted swipt networks with nonlinear eh model," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4464–4481, 2021.

[2] P. S. Bithas, V. Nikolaidis, A. G. Kanatas, and G. K. Karagiannidis, "Uav-to-ground communications: Channel modeling and uav selection," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5135–5144, 2020.

[3] B. Liu, H. Xu, and X. Zhou, "Resource allocation in unmanned aerial vehicle (uav)-assisted wireless-powered internet of things," *Sensors*, vol. 19, no. 8, p. 1908, 2019.

[4] W. Wang, J. Tang, N. Zhao, X. Liu, X. Y. Zhang, Y. Chen, and Y. Qian, "Joint precoding optimization for secure swipt in uav-aided noma networks," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5028–5040, 2020.

[5] I. Budhiraja, N. Kumar, S. Tyagi, S. Tanwar, and M. Guizani, "Swipt-enabled d2d communication underlaying noma-based cellular networks in imperfect csi," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 692–699, 2021.

[6] R. Ma, H. Wu, J. Ou, S. Yang, and Y. Gao, "Power splitting-based swipt systems with full-duplex jamming," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9822–9836, 2020.

[7] A. Alwarafy, M. Abdallah, B. S. Ciftler, A. Al-Fuqaha, and M. Hamdi, "The frontiers of deep reinforcement learning for resource management in future wireless hetnets: Techniques, challenges, and research directions," *IEEE Open Journal of the Communications Society*, 2022.

[8] S. Muy, D. Ron, and J.-R. Lee, "Energy efficiency optimization for swipt-based d2d-underlaid cellular networks using multiagent deep reinforcement learning," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3130–3138, 2021.

[9] K. K. Nguyen, N. A. Vien, L. D. Nguyen, M.-T. Le, L. Hanzo, and T. Q. Duong, "Real-time energy harvesting aided scheduling in uav-assisted d2d networks relying on deep reinforcement learning," *IEEE Access*, vol. 9, pp. 3638–3648, 2020.

[10] H. Wang, J. Wang, G. Ding, L. Wang, T. A. Tsiftsis, and P. K. Sharma, "Resource allocation for energy harvesting-powered d2d communication underlaying uav-assisted networks," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 14–24, 2017.

[11] Z. Yang, W. Xu, and M. Shikh-Bahaei, "Energy efficient uav communication with energy harvesting," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1913–1927, 2019.

[12] H. Bayerlein, M. Theile, M. Caccamo, and D. Gesbert, "Multi-uav path planning for wireless data harvesting with deep reinforcement learning," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1171–1187, 2021.

[13] M. Sengly, K. Lee, and J.-R. Lee, "Joint optimization of spectral efficiency and energy harvesting in d2d networks using deep neural network," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 8361–8366, 2021.

[14] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep q-network for uav-assisted online power transfer and data collection," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 215–12 226, 2019.

[15] Z. Kuang, G. Liu, G. Li, and X. Deng, "Energy efficient resource allocation algorithm in energy harvesting-based d2d heterogeneous networks," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 557–567, 2018.

[16] D. Alkama, M. A. Ouamri, M. S. Alzaidi, R. N. Shaw, M. Azni, and S. S. Ghoneim, "Downlink performance analysis in mimo uav-cellular communication with los/nlos propagation under 3d beamforming," *IEEE Access*, vol. 10, pp. 6650–6659, 2022.

[17] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for swipt systems," *IEEE Communications Letters*, vol. 19, no. 12, pp. 2082–2085, 2015.