Energy Efficient Multi-dimensional Trajectory of UAV-aided IoT Networks with Reinforcement Learning

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

25-01-2022 / 31-01-2022

CITATION

SILVIRIANTI, SILVIRIANTI; Shin, Soo Young (2022): Energy Efficient Multi-dimensional Trajectory of UAV-aided IoT Networks with Reinforcement Learning. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.19028864.v1>

DOI

[10.36227/techrxiv.19028864.v1](https://doi.org/10.36227/techrxiv.19028864.v1)

Energy Efficient Multi-dimensional Trajectory of UAV-aided IoT Networks with Reinforcement Learning

Silvirianti and Soo Young Shin, *Senior Member, IEEE*

Abstract—This paper proposed a multi-dimensional search space (or *directional space*) which has larger degree-of-freedom (DOF) to improve energy efficiency of limited-battery-powered UAV in Internet of Things (IoT) data collection scenario. In this paper, the UAV navigates from initial to goal point while collecting data from IoT sensors on the ground. Due to limited battery-power of UAV, optimized trajectory becomes a crucial practical issue. Based on the available directional space, direction of UAV which related to navigation trajectory is optimized using reinforcement learning. The objective of this reinforcement learning is to maximize energy efficiency of UAV as a long-term reward by selecting optimized direction. Moreover, practical energy consumption model and environment are presented in this paper. The simulation results verify the proposed scheme that has larger directional space achieved higher energy efficiency compared to benchmark models.

Index Terms—Directional space, energy efficiency, reinforcement learning, trajectory optimization, UAV.

I. INTRODUCTION

RECENTLY, Unmanned Aerial Vehicle (UAV)-aided communication networks has hugely attracted attention because of its potential to provide reliable high-rate connectivity which becomes a promising for future wireless networks [1]–[10]. On the one hand, UAVs able to be deployed to support available networks and improve quality-of-service (QoS) for ground users by establishing line-of-sight (LoS) links [11], [12]. On the other hand, UAVs can be beneficially applied to several main potential applications in next generation wireless networks with its high-mobility, low-cost, and flexible deployment [13].

The Internet of Things (IoT) directs a massive connectivity between physical things, e.g. vehicles, wearable devices, various sensors which is a key component of future networks [14]–[16]. However, the transmission reliability of IoT devices remains challenging due to long-distance and power constraints [17]. Fortunately, UAVs are expected to navigate the aforementioned issue with its high-mobility and flexibility power. UAVs capable to navigate around, collect data from IoT sensors, then transmit to other IoT devices or directly to the data center [18]. The high feasibility of LoS connections is able to improve the transmission efficiency and transmit power of IoT devices [19], [20]. Thus, the service life can be improved and longer.

Silvirianti and Soo Young Shin are with the Department of IT Convergence Engineering, Kumoh Institute of Technology, Gumi, Gyeongsangbuk-do, 30332 South Korea e-mail: silvirianti93@kumoh.ac.kr, wdragon@kumoh.ac.kr

Fig. 1. Comparison between proposed multi-dimensional trajectory that has higher DOF and benchmark models which has limited DOF [25], [27].

Nonetheless, limited battery power of UAV [21] becomes a crucial issue despite the benefits of UAV application in IoT data collection. Some prior works have proposed a trajectory optimization to reduce the energy consumption of UAV, thus expand the flight time [22]–[24]. However, most of the aforementioned works only considered a fixed-altitude scenario in 2D environment with limited action space (e.g. *up, down, right, left*), as can be seen in Fig. 1a [25]. The

TABLE I
RELATED WORKS

Related Works	3D Direct. Space	RL-based Optim.	Practical Energy Consumption	UAV-aided IoT Net.
[28]–[31]	✓	-	-	-
[27], [33], [34]	✓	✓	-	-
[21]	-	-	✓	-
[25]	-	✓	-	✓
[24]	-	✓	-	-
[22], [23]	-	-	-	✓
<i>this work</i>	✓	✓	✓	✓

considered scenario which mentioned above is impractical in real environment where UAV moves in three-dimensional (3D) directional space. Moreover, a limited directional space affects to higher number of shifting action that UAV takes and its traveled cost which producing higher energy consumption. Therefore, a 3D directional space and 3D environment for UAV should be designed to address the issue. Several studies have proposed a 3D trajectory optimization to reduce the traveled cost and improve the energy efficiency of UAV [26]. However, the directional space of UAV in aforementioned works is still limited (e.g. *ascend, descend, right, left, forward, backward*), as shown in Fig. 1c [27].

In addition, due to UAV high-mobility that makes the environment changing over time, the trajectory optimization problem of UAV is hard to be solved using classical optimization techniques (e.g. mathematical derivation) [28]–[31]. Motivated by the advancement of machine learning, one of its area called reinforcement learning (RL) is utilized in this paper to optimize trajectory and maximize the energy efficiency of UAV. RL is considered as dynamic learning which concerns finding optimal actions for an agent that interacts with environment [32]. RL employs trial and error processed to make informed decisions based on imperfect information of the environment. Based on these characteristics, RL is suitable for challenges in dynamic environment and sequential time such as UAV communication problem.

A. Related Works

A comparison of related works is presented in Table I. Several studies have considered 3D environment for trajectory optimization in UAV communication networks [27], [33], [34]. However, none of those studies utilized RL for trajectory optimization problem in UAV-aided IoT networks. Moreover, RL is used to optimize the trajectory of UAV in IoT data collection mission [25]. However, none of aforementioned works consider extended directional space of UAV for 3D trajectory optimization. Furthermore, none of aforementioned works are considering UAV energy consumption model based on real practical scenario (e.g. avionics energy of UAV).

B. Contributions

The key contributions of this paper are summarized as follows.

- A larger directional space of UAV is proposed to increase energy efficiency. Compared with prior work in the literature [25], [27] which only have 4 and 6 direction that affects to higher traveled distance of UAV, the proposed model provides 26 direction that reduce the traveled distance and energy consumption. First, the direction space in 2D scenario is increased to 8 directions. Subsequently, the 3D directional space is proposed with 26 directions.
- Three-dimensional environment that consider ground contour is designed for UAV flight navigation which makes the scenario more practical and real compared to prior works [25], [27].
- Energy consumption model is derived based on [21]. The proposed model considered real practical scenario such as avionics, hovering, travelling, and communication energy consumption. To the best of authors knowledge, in the moment of writing, none of the aforementioned works considered practical energy consumption.
- The effects of directional space and flight environment of UAV towards energy efficiency are investigated.

C. Paper Organizations

The remainder of this paper is organized as follows. Section II presents the system model of a UAV with a fixed altitude in a two-dimensional environment. Furthermore, a real environment in which a UAV navigates in a three-dimensional space is presented in Section III. In Section IV, an energy consumption model for a rotary-wing UAV considering a real application is presented. The reinforcement learning model of the energy-efficiency based optimized trajectory for a UAV is presented in Section V. The simulation results are discussed in Section VI, and finally, some concluding remarks are given in Section VII.

D. Notations

Notations: In this paper, the Euclidean norm of a vector is denoted by $\|\cdot\|$, $\{\cdot\}$ represents an array. $\mathbb{R}^M$ denotes the space of M -dimensional real vectors.

II. SYSTEM MODEL OF UAV FIXED ALTITUDE IN TWO-DIMENSIONAL SPACE

Consider a limited-battery UAV flying at a fixed altitude H m above $\mathcal{K}$ IoT sensors from the initial point L_I to the goal point L_F as shown in Fig. 2. The sensors, denoted by $\mathcal{K} = \{K_1, K_2, \dots, K_k\}$ are randomly distributed in the region $\mathcal{M} \in \mathbb{R}^3$ between the initial and goal point. As shown in Fig. 2, the UAV communicate with IoT sensors on the ground by providing direct connectivity from the sky. It is assumed that the UAV applies time-division multiple access (TDMA) to collect the uploading data sequentially from the sensors. Specifically, the UAV is only able to receive data-uploaded from a single sensor at each time t . This study focuses on the optimized trajectory of UAV when collecting data for a higher energy efficiency. In addition, it is assumed that the UAV communicates without the assistance of a central controller and partially has global knowledge of the environment. In other words, the sensor locations in the environment are locally known.

Fig. 2. (a). Fixed-altitude UAV enable communication systems to collect IoT sensor data. (b). Optimized-altitude UAV enable communication systems for collecting IoT sensor data considering the terrain environment.

A. Signal Model

The environment of the proposed model $\mathcal{M}$ can be modeled using a three-dimensional (3D) Cartesian coordinate system. The location of sensor K_k can be denoted as P_k at $(a_k, b_k, 0)$. The initial and goal point can be represented as $L_I = (x_I, y_I, H)$ and $L_F = (x_F, y_F, H)$, respectively. Moreover, the location of the UAV at time t is denoted by $L_t = (x_t, y_t, H)$. Subsequently, these locations are projected onto the ground plane which can be written as $p_k = (a_k, b_k)$, $l_I = (x_I, y_I)$, $l_F = (x_F, y_F)$, and $l_t(t) = (x(t), y(t))$, respectively. The Euclidean distance between the UAV and sensor K_k at time t is given by the following:

$$d_k(t) = \sqrt{\|l(t) - p_k\|^2 + H^2}. \quad (1)$$

In contrast to the channel characteristics of terrestrial communications, the UAV-to-ground (A2G) channels are likely dominated by line-of-sight (LoS) links [15]. It is therefore assumed that the communication between the UAV and sensors is monopolized by LoS links. The channel gain between UAV and sensor K_k can be expressed as follows:

$$h_k(t) = \alpha_0 d_k(t)^{-\beta} = \frac{\alpha_0}{\left(\|l(t) - p_k\|^2 + H^2\right)^{\frac{\zeta_{pl}}{2}}}, \quad (2)$$

where α_0 denotes the power loss of the channel at $d_0 = 1$ m where d_0 is the reference distance and $\zeta_{pl} \geq 2$ is the pathloss exponent. The channel gain $h_k(t)$ depends on the location of the UAV at time t .

B. Data Transmissions

Based on the assumption that the UAV collect the sensor data using the TDMA scheme, sensor K_k has a maximum channel gain for each $t \in [0, T]$ when satisfying the following:

$$K_k = \min_{K_k \in \mathcal{K}} \|l(t) - p_k\|. \quad (3)$$

Therefore, the received data rate of the UAV from sensor K_k at time t can be expressed through Eq. (4), where P_k denotes the transmit power of sensor K_k , σ^2 denotes the noise power, and γ_0 represents the signal-to-noise ratio (SNR). Moreover, considering that only a specific sensor is able to communicate with the UAV and upload data at time t , the established connection is indicated by

$$\rho_k(t) = \begin{cases} 1, & C_k(t) \neq 0 \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where $\rho_k(t)$ is an indicator of the communication between the sensor and UAV at time t . In addition, $\rho_k(t) = 1$ if $K_k \in \mathcal{K}$ communicates and uploads its data to the UAV, and $\rho_k(t) = 0$, otherwise. Note that only a specific sensor that can communicate with the UAV at t during flight time T , can be expressed as

$$\sum_{k=1}^K \rho_k(t) \leq 1, \forall t \in [0, T]. \quad (6)$$

Based on Eqs. (3) & (4), the received data rate of UAV at t depends on the minimum location of the UAV and sensor K_k as well as the established communication between them. In other words, the trajectory of the UAV directly affects the achievable data rate. Thus, the total received data rate during flight time T can be expressed as follows:

$$C(t)(\{\rho_k(t), q(t)\}, T) = \int_0^T \rho_k(t) C_k(t) dt, \quad (7)$$

where $C(t)$ denotes the total uploading data rate, and $q(t)$ represents the trajectory of the UAV at t . Furthermore, the reception quality of the data should be ensured when a sensor such as K_k is selected by assuming $C_k(t) \geq r_0$, where r_0 is a predefined minimum target rate. Thus, the total received data rate $C(t)$ at the UAV conforms to the following:

$$\hat{C}(t) = \begin{cases} C(t), & \text{if } D_m \text{ satisfies (3) and } C_k(t) \geq r_0 \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

III. SYSTEM MODEL OF UAV OPTIMIZED ALTITUDE IN THREE-DIMENSIONAL SPACE

A. Signal Model

Under a realistic scenario, the exploration environment is most likely to be in a three-dimensional coordinate space. Therefore, we consider the terrain of the environment in this study. This means that the $\mathcal{K}$ IoT sensors distributed in the environment $\mathcal{M} \in \mathbb{R}^3$ are located at different coordinates and heights. Moreover, the UAV maneuvers along the x, y, z -axes. Similar to a two-dimensional system, the UAV is assumed to fly from the initial point $L_I = (x_I, y_I, H_I)$ to the goal point $L_F = (x_F, y_F, H_F)$ while collecting sensor data that are randomly distributed in the environment. In addition, P_k denotes the location of sensor K_k at (a_k, b_k, c_k) . The location of the UAV at time t can be represented as $L_t = (x_t, y_t, H_t)$. Therefore, the distance between the UAV and sensor K_k at time t can be expressed as follows:

$$d_k(t) = \sqrt{\|L(t) - P_k\|^2} = \sqrt{(x(t) - a_k)^2 + (y(t) - b_k)^2 + (H(t) - c_k)^2}, \quad (9)$$

where $x(t)$ and $y(t)$ denote the coordinates of the UAV at time t . Meanwhile, the height of the UAV at time t is denoted by $H(t)$. Meanwhile, the channel gain between the UAV and sensor K_k is given by the following:

$$h_k(t) = \alpha_0 d_k(t)^{-\beta} = \frac{\alpha_0}{\left(\|L(t) - P_k\|^2\right)^{\frac{\beta}{2}}}, \quad (10)$$

where α_0 denotes the power loss of the channel at $d_0 = 1m$, in which d_0 is the reference distance and $\zeta_{pl} \geq 2$ is the pathloss exponent. The location of the UAV at time t affects the channel gain $h_k(t)$.

B. Data Transmission

As mentioned before, it is assumed that the UAV collects sensor data using the TDMA scheme. Moreover, sensor K_k has the maximum channel gain when the distance between the UAV and sensor is the minimum for each time $t \in [0, T]$ which can be expressed as follows:

$$K_k = \min_{K_k \in \mathcal{K}} \|L(t) - P_k\|, \quad (11)$$

where P_k and $L(t)$ denote the locations of the sensor and UAV at time t , respectively. Furthermore, the received data rate of the UAV from sensor K_k at time t is given by Eq.(12), where $x(t), y(t), H(t)$ denotes the location of the UAV at time t and a_k, b_k, c_k represents the location of the sensor along the x, y, z axes, respectively. Subsequently, the total received data rate of UAV during flight time T can be calculated using Eqs. (7) and (8), respectively.

IV. ENERGY CONSUMPTION MODEL FOR ROTARY WING UAV

In general, the energy consumption of a UAV is divided into two main components: communication energy and propulsion energy. Propulsion energy consumption is required to support the operation and maneuvering of a UAV. In this study, the propulsion energy consumption is modeled based on real conditions that consider the physical model of the UAV. The propulsion energy is composed of three parts: travel, hovering, and avionics energies [21]. The traveling energy, denoted by E_t represents the amount of power consumed by the UAV when moving from one location to another. The traveling energy is regularly dependent on the distance traveled by the UAV, which can be expressed as in Eq. 13. The n_l denotes the mass of the UAV framework and battery, g denotes the gravitational constant, $q(t)$ denotes the distance traveled by the UAV at time t , r denotes the value of the lift-to-drag ratio, and n denotes the number of rotors. In addition, the energy consumption of the UAV during hovering was also taken into consideration. The hovering energy of the UAV can be expressed as follows:

$$E_h = \frac{((\sum_{l=1}^2 n_l)g)^{3/2}}{\sqrt{2n\rho\zeta}}, \quad (14)$$

where ρ denotes the air density constant, and ζ denotes the spinning area of the blade disc of one rotor. In addition, the energy consumption for avionics components is given by

$$E_a = \int_0^T P_{\text{avio}} \frac{q(t)}{v(t)} dt, \quad (15)$$

where

$$v(t) \triangleq \dot{q}(t). \quad (16)$$

$$C_k(t) = \int_0^T B \log_2 \left(1 + \frac{P_k |h_k(t)|^2}{\sigma^2} \right) dt = \int_0^T B \log_2 \left(1 + \frac{\gamma_0}{[(x(t) - a_k)^2 + (y(t) - b_k)^2 + H^2]^{\zeta_{pl}}} \right) dt. \quad (4)$$

Fig. 3. (a) Illustration of IoT sensor data collection for fixed-altitude UAV with 8 possible flight directions. (b) Illustration of IoT sensor data collection for optimized-altitude UAV with 26 possible flight directions.

P_{avio} denotes the power for avionics components, $q(t)$ denotes the travel distance of the UAV, and $v(t)$ denotes the UAV velocity at time t . In addition to the propulsion energy, the energy consumption of the UAV to receive and transmit data was considered. The communication energy of the UAV can be expressed as follows:

$$E_c = \sum_{k=1}^K P_{\text{UAV}} \int_0^T \rho_k(t) C_k(t) dt, \quad (17)$$

where P_{UAV} denotes the power for UAV, $\rho_k(t)$ denotes indicator of scheduled transmission of sensor k , and $C_k(t)$ denotes the achievable data rate at time t . Therefore, the total energy that consumed by UAV while collecting data from sensors from initial to goal point can be expressed as

$$E_{\text{total}} = E_t + E_h + E_a + E_c, \quad (18)$$

where E_t , E_h , E_a , and E_c denote the traveled, hovering, avionics, and data communication energy of the UAV, respec-

tively. Based on the received data rate $\hat{C}(t)$ from Eq. (8), and the total energy consumption of the UAV in Eq. (13), the energy efficiency of the UAV can be calculated by dividing the received data rate and the total energy consumed by the UAV during flight. The energy efficiency of the UAV can be expressed through Eqs. (20) & (21).

$$\eta_{\text{EE}}(t) = \frac{\hat{C}(t)(\{\rho_k(t), q(t)\}, T)}{E_{\text{total}}}, \quad (19)$$

Based on Eqs. (19)-(21), the energy efficiency of the UAV is affected by the total received data rate and the traveled distance of the UAV.

V. REINFORCEMENT LEARNING MODEL OF ENERGY EFFICIENCY BASED OPTIMIZED TRAJECTORY FOR UAV

The problem formulation in this study is to maximize the energy efficiency $\eta_{\text{EE}}(t)$ in the long term by designing an optimized trajectory of the UAV. The optimization of

$$\begin{aligned} C_k(t) &= \int_0^T B \log_2 \left(1 + \frac{P_k |h_k(t)|^2}{\sigma^2} \right) dt \\ &= \int_0^T B \log_2 \left(1 + \frac{\gamma_0}{[(x(t) - a_k)^2 + (y(t) - b_k)^2 + (H(t) - c_k)^2]^{\zeta_{pl}}} \right) dt. \end{aligned} \quad (12)$$

$$\begin{aligned} E_t &\triangleq \int_0^T \frac{(\sum_{l=1}^2 n_l) g q(t)}{r n} dt, \\ &\triangleq \int_0^T \frac{(\sum_{l=1}^2 n_l) g \sqrt{\sum_{i=t-1}^t (x_i - (x_{i-1}))^2 + (y_i - (y_{i-1}))^2 + (H_i - (H_{i-1}))^2}}{r n} dt. \end{aligned} \quad (13)$$

Algorithm 1 Q-learning based optimized UAV-trajectory learning process

Input: Learning parameters: $\alpha \in [0, 1]$, $\gamma \in [0, 1]$, $\epsilon \in [0, 1]$;
 Agent information: $T_s, L_I, L_F, T, \omega, H, \text{ep} = 0$
 Output: Maximized energy efficiency η_{EE} ; Optimized policies $\pi_{Q\text{-learning}}^*$.
Initialization:
 1: For all S_t and A_t , set $Q_{Q\text{-learning}}(S_t, A_t) = 0$. Set $S_t = S_{t+1} = L_I, T_s = 0, \text{ep} = 0, T_{req} = 0$
 2: **while** $\text{ep} \leq \text{maximum episodes}$ **do**
 3: **for** state $S_t \in \mathcal{S}$ **do**
 4: Choose action A_t according to policy π derived from $Q_{Q\text{-learning}}(S_t, A_t)$ of Eq.(30).
 5: UAV takes action A_t .
 6: UAV receives a reward R_{t+1} ,
 UAV moves to the next state S_{t+1} .
 7: Observe elapsed time T_s .
 8: Update Q-values based on Eq. (28) $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_{a \in \mathcal{A}} Q(S_{t+1}, a)]$.
 9: Update state S_t , action A_t , and elapsed time T_s :
 $S_t \leftarrow S_{t+1}, A_t \leftarrow A_{t+1}; T_s \leftarrow T_s + T_{req}$;
 10: **end for** $S_{t+1} == L_F$ or $T_s \geq T$
 11: $\text{ep} \leftarrow \text{ep} + 1$;
 12: **end while**
 13: Find the optimal policy:
 $\pi_{Q\text{-learning}}^*(S_t) = \arg\max_a Q_{Q\text{-learning}}(S_t, A_t)$ for any $S_t \in \mathcal{S}$

Algorithm 2 SARSA-based optimized UAV-trajectory learning process

Input: Learning parameters: $\alpha \in [0, 1]$, $\gamma \in [0, 1]$, $\epsilon \in [0, 1]$;
 Agent information: $T_s, L_I, L_F, T, \omega, H, \text{ep} = 0$
 Output: Maximized energy efficiency η_{EE} ; Optimized policies π_{SARSA}^* .
Initialization:
 1: For all S_t and A_t , set $Q_{\text{SARSA}}(S_t, A_t) = 0$. Set $S_t = S_{t+1} = L_I, T_s = 0, \text{ep} = 0, T_{req} = 0$
 2: **while** $\text{ep} \leq \text{maximum episodes}$ **do**
 3: **for** state $S_t \in \mathcal{S}$ **do**
 4: Choose action A_t according to policy π derived from $Q_{\text{SARSA}}(S_t, A_t)$ of Eq.(30).
 5: UAV takes action A_t .
 6: UAV receives a reward R_{t+1} ,
 UAV moves to the next state S_{t+1} .
 7: Observe elapsed time T_s .
 8: Choose next action A_{t+1} from the state S_{t+1} following the policy derived from $Q_{\text{SARSA}}(S_{t+1}, A_{t+1})$ as in Eq.(30).
 9: Obtain updated Q-values based on Eq. (29) $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$.
 10: Update state S_t , action A_t , and elapsed time T_s :
 $S_t \leftarrow S_{t+1}, A_t \leftarrow A_{t+1}; T_s \leftarrow T_s + T_{req}$;
 11: **end for** $S_{t+1} == L_F$ or $T_s \geq T$
 12: $\text{ep} \leftarrow \text{ep} + 1$;
 13: **end while**

UAV trajectory for maximizing the energy efficiency can be expressed as

$$\max_{\{x, y, H \in q(t)\}} \int_0^T \frac{\hat{C}(t)(\{\rho_k(t), q(t)\}, T)}{E_{\text{total}}} dt, \quad (22)$$

where $v(t)$, E_{total} , and T denote the velocity, traveled distance, total energy consumption, and flight time of the UAV, respectively. Based on the aforementioned formulation, the energy efficiency depends critically on the trajectory of the UAV. Considering a real scenario in which the information related to the locations of the sensors and the amount of data to be transmitted are unavailable to the UAV, the problem is modeled as a Markov decision process (MDP). It is well known that the MDP model can solve the problem of a partially observed environment with uncertainties and challenges. To solve the problem formulation mentioned in Eqs. (20) & (21), the problem is converted into a multi-period decision scheme with finite states and actions.

A. Fixed Altitude UAV

1) *State Space:* With this MDP model, the UAV is considered a learning agent that aims to learn an optimized trajectory through reinforcement learning. The limited flight time of UAV T is discretized into M time slots. Based on this, the step size is defined as $\delta = \frac{T}{M}$. Furthermore, the environment $\mathcal{M}$ can be divided into $M_1 = \frac{L_1}{\delta\omega}$ by $M_2 = \frac{L_2}{\delta\omega}$ tiles, where $\delta\omega$ m is the length of each side. In addition, L_1 and L_2 denote the length and width of the environment (m), respectively, and ω denotes the speed of the UAV (m/s). Fig. 3(a) shows a projection of environment $\mathcal{M}$ on a horizontal ground plane for a fixed altitude UAV. As can be seen in Fig. 3(a), environment $\mathcal{M}$ is divided into $M_1 \times M_2$ tiles and each sensor $K_k \in \mathcal{K}$ is placed at a single tile.

Based on the discretized region of the considered environment, the state set $\mathcal{S}$ of the UAV can be represented as

$$\eta_{EE}(t) = \frac{\int_0^T \rho_k(t) C_k(t) dt}{\int_0^T \frac{(\sum_{l=1}^2 n_l) g q(t)}{rn} dt + \frac{((\sum_{l=1}^2 n_l) g)^{3/2}}{\sqrt{2n\rho\zeta}} + \int_0^T P_{\text{avio}} \frac{q(t)}{v(t)} dt + \sum_{k=1}^K P_{\text{UAV}} \int_0^T \rho_k(t) C_k(t) dt}, \quad (20)$$

$$\eta_{EE}(t) = \int_0^T \frac{\rho_k(t) C_k(t)}{\frac{(\sum_{l=1}^2 n_l) g q(t)}{rn} + \frac{((\sum_{l=1}^2 n_l) g)^{3/2}}{\sqrt{2n\rho\zeta}} + P_{\text{avio}} \frac{q(t)}{v(t)} + \sum_{k=1}^K P_{\text{UAV}} \rho_k(t) C_k(t)} dt. \quad (21)$$

$\mathcal{S} = \{s(1), s(2), \dots, s(M_1 M_2)\}$. States $s(m)$ and $s(m) \in \mathcal{S}$ refer to a small tile. Fig. 3(a) depicts a UAV agent navigating through the tiles from the initial point to the goal point while collecting sensors data randomly distributed within the environment. The finite flight time of a UAV owing to its limited battery energy urges the observation of the optimal trajectory. By considering that the UAV is only able to observe the environment partially, reinforcement learning helps make accurate decisions for an optimized trajectory.

2) *Action Space*: In this scenario, the UAV agent has a maximum of eight actions at each state, that is, $\{\text{West, North West, North, North East, East, South East, South, South West}\}$. For clarity, the action set of the UAV is denoted by $\mathcal{A} = \{W, NW, N, NE, E, SE, S, SW\}$. Because the UAV flies at a fixed altitude (H) m, the UAV moves along the x and y axes. It should be noted that in real practice, the UAV is capable of selecting its direction along any direction θ , that is, $\theta \in [0, 2\pi]$.

3) *Reward Formulation*: The rewards of reinforcement learning are designed to drive one to find the best solutions that satisfy the constraints asserted on the learning agent [19]. In this study, the UAV as an agent aims to maximize the energy efficiency during data collection within a finite flight time. During the reinforcement learning process, a score is assigned for an action a that the UAV takes at time t and collects a reward r at future time t' . The score is assigned to estimate the importance of the action in producing the reward. A fly hover-and-communicate design [] is also considered in this study, in which the data stored at sensor K_k will be fetched when communication with UAV is established. Moreover, the UAV hovers when collecting the sensor data. By ensuring the communication quality, the UAV collects the data from the sensor only when the UAV moves to the sensor location at tile $m \in \mathcal{M}$ (see Fig. 3(a)). As previously mentioned, the UAV has a finite flight time, and the maximum required time for a UAV to move from one tile to another can be expressed as follows:

$$T_s = \frac{T_{\max}}{M_1 M_2}, \quad (23)$$

where $T_{\max}$ denotes the maximum flight of the UAV while M_1, M_2 denote the tiles which represent time slot. In addition, the maximum time that UAV required to fetch the sensors data is defined by

$$T_H = \max \left\{ \frac{B_k}{C_k(t)}, T_s \right\}, \quad (24a)$$

$$\text{s.t. } T_H \leq T_s, \quad (24b)$$

where B_k denotes the volume data which stored at sensor K_k while $C_k(t)$ represents the received data rate at the UAV at time t . Therefore, the total time taken for UAV in one tile can be expressed as

$$T_{\text{req}} = \begin{cases} \max \left\{ \frac{B_k}{C_k(t)}, T_s \right\}, & C(t) \neq 0 \\ T_s, & \text{otherwise.} \end{cases} \quad (25)$$

where T_s denotes the time elapsed for the UAV to move from one tile to another. Because the main objective is to maximize

the energy efficiency of the UAV for collecting data within a finite flight time, the reward function of the UAV agent can be expressed as follows:

$$\mathcal{R}(t) = \begin{cases} \eta_{\text{EE}}(t), & \text{if } C_k(t) \geq r_0 \\ 0, & \text{otherwise.} \end{cases} \quad (26)$$

where r_0 is the minimum UAV data rate. In addition, to ensure that the UAV recognizes the goal point L_F , a reward is received by the UAV when it arrives at its destination during the learning process.

B. Optimized Altitude UAV

1) *State Space*: In comparison to the fixed altitude scenario, the optimized altitude UAV learns the optimal three-dimensional trajectory. The environment $\mathcal{M}$ is discretized into $M_1 = \frac{L_1}{\delta\omega} \times M_2 = \frac{L_2}{\delta\omega} \times M_3 = \frac{L_3}{\delta\omega}$ voxels with $\delta\omega$ m the length of each side. Here, L_1, L_2 , and L_3 denote the height, length, and width of the environment (m), respectively, and, ω denotes the speed of UAV (m/s). Fig. 3(b) shows a projection of environment $\mathcal{M}$ on a three-dimensional space for optimized altitude UAV scenario. As can be seen in Fig. 3(b), environment $\mathcal{M}$ is divided into $M_1 \times M_2 \times M_3$ voxels, and each sensor $K_k \in \mathcal{K}$ is placed at a single voxel.

Referring to the discretized region of the considered environment, the state set $\mathcal{S}$ of the UAV can be represented as $\mathcal{S} = \{s(1), s(2), \dots, s(M_1 M_2 M_3)\}$. Each state $s(m)$ and $s(m) \in \mathcal{S}$ refers to a voxel. Fig. 3(b) shows a UAV agent navigating through the voxel from the initial point to the goal point while collecting sensors data randomly distributed within the environment. The finite flight time of a UAV owing to its limited battery energy necessities the observation of an optimal three-dimensional trajectory. Reinforcement learning helps the UAV to make accurate decisions for an optimized 3D trajectory.

2) *Action Space*: Because the UAV is able to navigate in a three-dimensional space, the action space of the UAV is modeled in a Bloch sphere (see Fig. 3(b)). In this case, the UAV agent has a maximum of 26 actions at each state, that is, $\{\text{e.g., Ascend, Descend, West, North West, North, North East, East, South East, South, South West, Ascend West, Descend North West}\}$. For clarity, the action set of the UAV is denoted by $\mathcal{A} = \{A, D, W, NW, N, NE, E, SE, S, SW, AW, DW, ANW, DNW, AN, DN, ANE, DNE, AE, DE, ASE, DSE, AS, DS, ASW, DSW\}$. Because it is known that the UAV flies at an optimized altitude ($H(t)$) m at time t , the UAV moves along the x, y , and z axes. It should be noted that, in real practice, the UAV can select its direction along any direction in θ and ϕ , that is, $\theta \in [0, 2\pi], \phi \in [0, 2\pi]$.

3) *Reward Formulation*: Similar to the fixed-altitude UAV scenario, the maximum required time for a UAV to move from one voxel to another can be expressed as follows:

$$T_s = \frac{T_{\max}}{M_1 M_2 M_3}, \quad (27)$$

where $T_{\max}$ denotes the maximum flight of the UAV, and M_1, M_2, M_3 denote the voxels representing the time slot.

Moreover, the maximum hovering time T_H , total time T_{req} , and reward function $\mathcal{R}(t)$ can be obtained from Eqs. (24)-(26).

C. Reinforcement Learning based Optimized Trajectory Design

Reinforcement learning is adopted in this study to solve sequential decision making of UAV during data-collecting mission. The reinforcement learning addressed this problem by maximizing immediate and long-term rewards. Maximum rewards are achieved from the best action done by agent. The best action is selected under policy¹ and updated using Q-table which recorded experience of agent's action in a state and collected reward. At initial of learning process, the state-action values in the Q-table are initialized to zero. In a state S_t at time t , the UAV as learning agent consecutively implements the following steps: select an action A_t from action set $\mathcal{A}$, move to the next state S_{t+1} , receive a reward R_{t+1} , and update the Q-value in Q-table. In this study, the agent state is represented by the location of UAV (X_t, Y_t, H_t) at time t with action set $\mathcal{A}$ as mentioned in Subsection V-A & V-B. Based on the current state (X_t, Y_t, H_t) and selected action A_t , reward is collected then UAV moves to new state $(X_{t+1}, Y_{t+1}, H_{t+1})$. The reward can be considered as feedback from the environment given action. In terms of maximizing the energy efficiency, the reward is a function of the total achievable throughput divided by total UAV energy consumption over time T , as defined in Eq. 26, Subsection V-A.

In this paper, a model-free TD method² is used to update Q-value because of UAV requires a multi-period decision-making. First, the Q-value is updated using off-policy method known as Q-learning. In Q-learning method, the policy can be uncorrelated with the policy that is being evaluated and improved. The Q-value in Q-learning can be updated using

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_{a \in \mathcal{A}} Q(S_{t+1}, a)]. \quad (28)$$

where $\alpha \in [0, 1]$ denotes learning rate (or step size), whereas $\gamma \in [0, 1]$ represents the discount factor determining the importance of future rewards. Furthermore, as comparison, the Q-value is also updated using on-policy method that is known i.e., SARSA, learns the optimal policy by learning the Q-value for the current policy and for all states-actions. Specifically, the policy used in the learning is the same that being evaluated and improved. In this method, the Q-value is updated using

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)], \quad (29)$$

The updated Q value in Eq.(29) depends on the information of the variables $S_t, A_t, S_{t+1}, A_{t+1}$. SARSA iteratively estimates the Q-value under policy Q_π and changes policy π in a greedy manner at the same time based on Q_π .

¹ A policy defines the behavior of a learning agent at a fixed instant in time [7], which denotes a set of rules for that agent.

² The temporal difference (TD) method improves the estimation of future reactions by extracting information from observations of sequential stochastic processes that model the environment of the agent. Q-learning and SARSA are TD-based methods [19].

According to the updated Q-values, UAV as an agent exploits what it has already experienced in order to receive a maximum reward. However, the agent also needs to explore unmapped or altered environment to observe whether it can make a better action selections in the future. The exploration motivates the agent to learn from the environment, where it allows the learning process to get trapped out of a local optimum. Specifically, an exploration gives the UAV chance to take a specific action, observe a reward, and update its action choices. Meanwhile, an exploitation assists the UAV by taking advantage of the knowledge that is already available to obtain the best choice of action. An exploitation drives convergence of the learning process. Thus, selection between exploration and exploitation is a concern that allows UAVs to obtain beneficial actions during the learning process. By introducing a parameter $0 < \epsilon < 1$, an ϵ -greedy exploration is utilized for better balance the exploration and exploitation:

$$A_t = \begin{cases} \operatorname{argmax}_{a \in \mathcal{A}} Q(S_t, a) & \text{with probability } 1 - \epsilon, \\ \text{random selection,} & \text{with probability } \epsilon. \end{cases} \quad (30)$$

The agent selects the action as it believes that the action yields the maximum long-term reward with probability $1 - \epsilon$. In contrast, the agent chooses an action uniformly at random with probability ϵ . The action selection based on Q-values might fall into a local optimum when $\epsilon = 0$.

To elaborate the Q-learning-based trajectory optimization with 3D directional space, an example is presented. *Example:* When UAV starts exploring from initial location at a state $S_0 = (5, 1, 1)$, one of the neighboring states $\{(4, 1, 1), (4, 1, 2), (4, 2, 1), (4, 2, 2), (5, 2, 1), (5, 2, 2)\}$ will be explored by selecting an action from $\mathcal{A}$ according to state-action values, i.e., $\max_{a \in \mathcal{A}} Q(S_0, a)$. Since all the Q-values at the initial of Q-learning are initialized to zero, an action will be randomly chosen or according to some preset criteria. Assumed action "AN" which represented Ascend North direction is selected, thus the next state becomes $S_1 = (4, 1, 2)$. The corresponding reward and Q-value $Q(S_0, A_0 = AN)$ will be updated accordingly. Based on the example, the Q-values is increased by the collected reward value using Eq. 25. The Q-value is iteratively updated during learning based on the chosen action that leads to maximum long-term reward.

VI. PERFORMANCE EVALUATION

In this simulation, the exploration region is considered in 3D space with the upper limit set to $100 \times 100 \text{ m}^2$ for the 2D trajectory scenario and $100 \times 100 \times 100 \text{ m}^3$ for the 3D trajectory scenario. The considered region $\mathcal{M}$ is divided into 5×5 tiles for the 3D trajectory and $5 \times 5 \times 5$ voxels for the 3D trajectory. The initial location of the UAV is set as $L_I = (0, 0, 0)$, whereas the goal location is $L_F = (100, 100, 100)$. Based on real applications [23], the maximum flight time of a quadrotor UAV T is set as 30 min with a speed ω of 5.56 m/s . The mass of the quadrotor UAV, which consists of its framework and battery m_l is 1.46 kg. Because the considered UAV is quadrotary-wing, the number of rotors n is four, and the area of the spinning blade disc of one rotor ζ is 0.0507 m^2 . In addition, the lift-to-drag ratio of UAV r is 3. The

TABLE II
SIMULATIONS PARAMETERS

Parameter	Notation	Value
Considered region	$\mathcal{M}$	$100 \times 100 \times 100 \text{ m}^3$
Number of sensors	$\mathcal{K}$	2,4,6
Maximum flight time of UAV	T	30 mins
UAV speed	ω	5.56 m/s
UAV fixed altitude	H	100 m
Mass of UAV (framework and battery)	n_l	1.46 kg
Gravity acceleration	g	9.807 m/s ²
Lift to drag ratio	r	3
Number of UAV rotors	n	4
Air density	ρ	1.225 kg/m ³
Spinning blade disc area of one rotor	ζ	0.0507 m ²
Transmit power of UAV	P_{UAV}	30 dBm
Transmit power of sensors	P_k	10 dBm
Bandwidth	B	2 MHz
Noise power	σ^2	-110 dBm
Channel power gain	α_0	-50 dBm
Path loss exponent	ζ_{pl}	2
Discount factor	γ	0.98
Learning rate	α	0.5

UAV navigates from L_I to L_F with a gravity acceleration of 9.807 m/s^2 and the air density is 1.225 kg/m^3 . As mentioned in Section II, there are $\mathcal{K}$ sensors are randomly distributed in the region $\mathcal{M}$. The number of sensors $\mathcal{K}$ simulated in this study were divided into three scenarios: 2, 4, and 6. The UAV communicates with sensor K_k by transmitting power P_{UAV} at 30 dBm whereas the transmit power of sensor P_k is 10 dBm. The sensors share the total communication bandwidth $B = 2 \text{ MHz}$ with noise power σ^2 of -110 dBm . The channel power gain α_0 is -50 dBm and the path loss exponent ζ_{pl} is 2. As a reinforcement learning agent, UAV learns through the environment $\mathcal{M}$ with a rate $\alpha = 0.5$. The discount factor of the reward was set as $\gamma = 0.98$. The overall simulation parameter values are listed in Table II.

The performance of the proposed model is evaluated using two baseline RL methods, Q-learning and SARSA.

- Q-learning is an off-policy method in which the learning of the optimal policy is based on the best action, which is taken as greedy, unlike the current policy.
- SARSA is known as an on-policy method that learns the optimal policy by learning the Q-value for the current

policy and for all states and actions.

Moreover, the proposed extended directional space is compared with those of limited DOF in terms of energy efficiency.

A. Energy Efficiency of Fixed-altitude UAV under 2D Trajectory

Fig. 4(a) shows the optimal policy π_{SARSA}^* for the optimized trajectory of fixed-altitude UAV with 8-DOF under the SARSA method. Meanwhile, the optimal policy for an optimized-trajectory fixed-altitude UAV with 8-DOF based Q-learning method is presented in Fig. 4(b). As can be seen from the figure, the optimal policy resulting from the SARSA and Q-learning methods under an identical environment are different. A different optimal policy can generate a different sum of the rewards results.

In Fig. 6, the energy efficiency of UAV under 2D trajectory with 2 sensors randomly distributed is presented. The "energy efficiency" refers to average sum rewards of Eq. 27, where a higher rewards indicates better performance. Moreover, "iterations" indicates the iteration step of training which optimizes the action Q-values in Section V-C. In this work, the iteration step is set to 10000. As shown in the figure, the extended directional space with 8-DOF achieves higher energy efficiency than the 4-DOF (Ref. [25]). Moreover, Q-learning converged faster than SARSA in this case because of the limited flight time T of the UAV. Specifically, during the process of learning SARSA requires more actions than Q-learning, as can be seen in **Algorithm 1** for Q-learning and **Algorithm 2** for SARSA. It happened because SARSA learns the optimal policy based on the current policy, which requires more time, whereas Q-learning learns based on the best action taken in a greedy manner. As a result, the UAV fails to reach the optimal result within a finite flight time T .

Fig. 7 shows the comparison of $\mathcal{R}$ values between proposed 8-DOF with related work [25] 4-DOF under Q-learning and SARSA for 4 sensors scenario. The figure shows similar trend results with the 2 sensors in Fig. 6. It clearly shown in the figure that UAV with 8-DOF outperforms the 4-DOF in maximizing energy efficiency. Moreover, the proposed model under Q-learning achieved better performance by converging

Fig. 4. (a). Optimal policy π_{SARSA}^* for the optimized 2D trajectory of the 8-DOF UAV-based SARSA. (b). Optimal policy $\pi_{\text{Q-learning}}^*$ for the optimized 2D trajectory of 8-DOF UAV-based Q-learning.

Fig. 5. (a). Optimal policy π_{SARSA}^* for optimized 3D trajectory of the 26-DOF UAV-based SARSA. (b). Optimal policy $\pi_{\text{Q-learning}}^*$ for the optimized 3D trajectory of 26-DOF UAV-based Q-learning.

Fig. 6. Energy efficiency of fixed-altitude UAV with respect to the iteration steps when the number of sensors $K = 2$.

faster compared to SARSA. Furthermore, the additional 2 other sensors affects higher energy efficiency of UAV since higher amount of data can be collected.

Fig. 8 depicts the $\mathcal{R}$ values of proposed 8-DOF and 4-DOF [25] under Q-learning and SARSA. As can be seen in the figure, 8-DOF with Q-learning achieves higher rewards compared to 8-DOF SARSA and 4-DOFs which has similar trend with previous results (see Fig. 6, 7). Moreover, the result exhibits higher amount of collected rewards directly affected by the number of sensors.

B. Energy Efficiency of Optimized-altitude UAV under 3D Trajectory

Fig. 5(a) presents the optimal policy π_{SARSA}^* for an optimized 3D UAV trajectory with 26-DOF under SARSA method. Meanwhile, the optimal policy for optimized 3D UAV trajectory with 26-DOF based Q-learning method is shown in Fig. 5(b). Similar to the optimized trajectory for fixed-altitude UAV, the optimal policy resulting from the SARSA and Q-learning

Fig. 7. Energy efficiency of fixed-altitude UAV with respect to the iteration step when the number of sensors $K = 4$.

methods under identical environment are different. A different optimal policy is able to generate a different sum of rewards results.

In Fig. 9, the energy efficiency of UAV under 3D trajectory with 2 sensors randomly distributed is presented. The "energy efficiency" refers to average sum rewards of Eq. 27, where a higher reward indicates better performance. Moreover, "iterations" indicates the iteration step of training which optimizes the action Q-values in Section V-C. In this work, the iteration step is set to 10000. As shown in the figure, the extended directional space with 26-DOF achieves higher energy efficiency than the 6-DOF (Ref. [27]). Moreover, Q-learning converged faster than SARSA in this case because of the limited flight time T of the UAV. Specifically, during the process of learning SARSA requires more actions than Q-learning, as can be seen in Algorithm 1 for Q-learning and Algorithm 2 for SARSA. It happened because SARSA learns the optimal policy based on the current policy, which requires more time, whereas Q-learning learns based on the best action taken in a greedy manner. As a result, the UAV fails to reach the optimal result within a finite flight time T .

Fig. 10 shows the comparison of $\mathcal{R}$ values between proposed 26-DOF with related work [27] 6-DOF under Q-learning and SARSA for 4 sensors scenario. The figure shows similar trend results with the 2 sensors in Fig. 9. It clearly shown in the figure that UAV with 26-DOF outperforms the 6-DOF in maximizing energy efficiency. Moreover, the proposed model under Q-learning achieved better performance by converging faster compared to SARSA. Furthermore, the additional 2 other sensors affects higher energy efficiency of UAV since higher amount of data can be collected.

Fig. 11 depicts the $\mathcal{R}$ values of proposed 26-DOF and 6-DOF [27] under Q-learning and SARSA. As can be seen in the figure, 26-DOF with Q-learning achieves higher rewards compared to 26-DOF SARSA and 6-DOFs which has similar trend with previous results (see Fig. 9, 10). Moreover, the result exhibits higher amount of collected rewards directly affected by the number of sensors.

Fig. 8. Energy efficiency of fixed-altitude UAV with respect to the iteration step when number of sensors $K = 6$.

Fig. 9. Energy efficiency of optimized-altitude UAV with respect to the iteration step when number of sensors $K = 2$.

Fig. 10. Energy efficiency of optimized-altitude UAV with respect to the iteration step when number of sensors $K = 4$.

Fig. 11. Energy efficiency of optimized-altitude UAV with respect to the iteration steps when the number of sensors $K = 6$.

VII. CONCLUSION

In this paper, a multi-dimensional search space is proposed which employs higher DOF to optimize the trajectory of UAV. Moreover, the trajectory is also optimized utilizing RL method specifically Q-learning to maximize energy efficiency for UAV-aided IoT networks. In addition, energy consumption model of UAV is designed by considering practical real environments. The performance of proposed model is compared with those of utilizing limited DOF and SARSA method. The proposed extended DOF exhibited higher energy efficiency compared with the limited DOF and SARSA. Moreover, Q-learning achieved faster iteration convergence compared with SARSA. The following points can be considered as future works:

- Inter-sensor interference from other sensors can be considered for a data transmissions scheme.
- A deep neural network RL scheme (called DRL) i.e. [33], [34] can be utilized for to further improve energy efficiency of UAV-aided IoT networks.
- For a better performance, cooperative multi-agent reinforcement learning (MARL) scheme can also be applied to this scenario such as [21], [22].

REFERENCES

- [1] M. Asadpour, B. V. den Bergh, D. Giustiniano, K. A. Hummel, S. Pollin, and B. Plattner, "Micro aerial vehicle networks: An experimental analysis of challenges and opportunities," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 141–149, Jul. 2014.
- [2] R. S. Stansbury, M. A. Vyas, and T. A. Wilson, "A survey of UAS technologies for command, control, and communication (C3)," *Unmanned Aircraft Systems. Berlin, Germany: Springer*, 2008, pp. 61–78.
- [3] M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surv. Tut.*, vol. 21, no. 3, pp. 2334–2360, Jul.–Sep. 2019.
- [4] S. ur Rahman, G. Kim, Y. Cho, and A. Khan, "Positioning of UAVs for throughput maximization in software-defined disaster area UAV communication networks," *J. Commun. Netw.*, vol. 20, no. 5, pp. 452–463, Oct. 2019.
- [5] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: Design and optimization for multi-UAV networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1346–1359, Feb. 2019.
- [6] J. Ye, C. Zhang, H. Lei, G. Pan, and Z. Ding, "Secure UAV-to-UAV systems with spatially random UAVs," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 564–567, Apr. 2019.
- [7] V. Sharma, M. Bennis, and R. Kumar, "UAV-Assisted heterogeneous networks for capacity enhancement," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1207–1210, Jun. 2016.
- [8] L. Gupta, R. Jain, and G. Vaszkn, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.
- [9] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.
- [10] J. Wang et al., "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones," *IEEE Veh. Technol. Mag.*, vol. 12, no. 3, pp. 73–82, Sep. 2017.
- [11] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.
- [12] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.
- [13] Z. Wang, L. Duan, and R. Zhang, "Adaptive deployment for UAV-aided communication networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4531–4543, Sep. 2019.
- [14] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

- [15] J. Chen, L. Zhang, Y. Liang, X. Kang and R. Zhang, "Resource Allocation for Wireless-Powered IoT Networks With Short Packet Communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1447-1461, Feb. 2019.
- [16] L. Chettri and R. Bera, "A Comprehensive Survey on Internet of Things (IoT) Toward 5G Wireless Systems," *IEEE IoT Journal*, vol. 7, no. 1, pp. 16-32, Jan. 2020.
- [17] R. Duan, J. Wang, C. Jiang, H. Yao, Y. Ren and Y. Qian, "Resource Allocation for Multi-UAV Aided IoT NOMA Uplink Transmission Systems," *IEEE IoT Journal*, vol. 6, no. 4, pp. 7025-7037, Aug. 2019.
- [18] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Maattanen, M. Bergstrom, S. Euler, and Y.-P.-E. Wang, "The sky is not the limit: LTE for unmanned aerial vehicles," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 204-210, Apr. 2018.
- [19] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569-572, Dec. 2014.
- [20] Z. Wang, L. Duan, and R. Zhang, "Adaptive deployment for UAV-aided communication networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4531-4543, Sep. 2019.
- [21] Juan Zhang, James F. Campbell, Donald C. Sweeney II, and Andrea C. Hupman, "Energy consumption models for delivery drones: A comparison and assessment, Transportation Research Part D: Transport and Environment," *Science Direct*, vol. 90, no. 102668, Jan. 2021.
- [22] H. Yang, R. Ruby, Q. -V. Pham and K. Wu, "Aiding a Disaster Spot via Multi-UAV-based IoT Networks: Energy and Mission Completion Time-Aware Trajectory Optimization," *IEEE IoT Journal*.
- [23] C. Zhan and H. Lai, "Energy Minimization in Internet-of-Things System Based on Rotary-Wing UAV," *IEEE Wireless Commun. Letters*, vol. 8, no. 5, pp. 1341-1344, Oct. 2019.
- [24] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton and J. Henry, "UAV Trajectory Planning in Wireless Sensor Networks for Energy Consumption Minimization by Deep Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9540-9554, Sept. 2021.
- [25] J. Cui, Z. Ding, Y. Deng, A. Nallanathan, and L. Hanzo, "Adaptive UAV-Trajectory Optimization Under Quality of Service Constraints: A Model-Free Solution," *IEEE Access*, vol. 8, pp. 112253-112265, 2020.
- [26] T. Liu, M. Cui, G. Zhang, Q. Wu, X. Chu and J. Zhang, "3D Trajectory and Transmit Power Optimization for UAV-Enabled Multi-Link Relaying Systems," *IEEE Trans. Green Comm. Net.*, vol. 5, no. 1, pp. 392-405, March 2021.
- [27] W. Zhang, Q. Wang, X. Liu, Y. Liu, and Y. Chen, "Three-Dimension Trajectory Design for Multi-UAV Wireless Network With Deep Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 600-612, Jan. 2021.
- [28] M. Hua, L. Yang, Q. Wu and A. L. Swindlehurst, "3D UAV Trajectory and Communication Design for Simultaneous Uplink and Downlink Transmission," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5908-5923, Sept. 2020.
- [29] Y. Sun, D. Xu, D. W. K. Ng, L. Dai and R. Schober, "Optimal 3D-Trajectory Design and Resource Allocation for Solar-Powered UAV Communication Systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4281-4298, June 2019.
- [30] A. Visintini, T. D. P. Perera and D. N. K. Jayakody, "3-D Trajectory Optimization for Fixed-Wing UAV-Enabled Wireless Network," *IEEE Access*, vol. 9, pp. 35045-35056, 2021.
- [31] N. Tang, H. Tang, B. Li and X. Yuan, "Joint Maneuver and Beamwidth Optimization for UAV-Enabled Multicasting," *IEEE Access*, vol. 7, pp. 149503-149514, 2019.
- [32] M. Sun, X. Xu, X. Qin, and P. Zhang, "AoI-Energy-Aware UAV-assisted Data Collection for IoT Networks: A Deep Reinforcement Learning Method," *IEEE Internet of Things Journal*, May 2021.
- [33] R. Ding, F. Gao, and X. S. Shen, "3D UAV Trajectory Design and Frequency Band Allocation for Energy-Efficient and Fair Communication: A Deep Reinforcement Learning Approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796-7809, Dec. 2020.
- [34] H. Qi, Z. Hu, H. Huang, X. Wen, and Z. Lu, "Energy Efficient 3-D UAV Control for Persistent Communication Service and Fairness: A Deep Reinforcement Learning Approach," *IEEE Access*, vol. 8, pp. 53172-53184, March 2020.

PLACE
PHOTO
HERE

Silvirianti Biography text here.

Soo Young Shin Biography text here.