Energy-Efficient Multidimensional Trajectory of UAV-Aided IoT Networks With Reinforcement Learning

Silvirianti and Soo Young Shin, Senior Member, IEEE

Abstract—This article proposes a multidimensional search space (or directional space) with more degrees-of-freedom (DOFs) to increase the energy efficiency of limited-battery-powered unmanned aerial vehicle (UAV) in the Internet of Things (IoT) data collection scenario. In this article, the UAV navigates from the initial to the goal point while collecting data from IoT sensors on the ground. Owing to the limited battery power of UAVs, an optimized trajectory is a crucial practical problem. Based on the available directional space, the direction of the UAV related to the navigation trajectory is optimized using reinforcement learning (RL). The objective of RL is to maximize the energy efficiency of the UAV as a long-term reward by selecting the optimal direction. Moreover, a practical energy consumption model and environment are presented in this article. Simulation results verified that the proposed multidimensional trajectory for UAV achieves higher energy efficiency compared with benchmark models.

Index Terms—Directional space, energy efficiency, reinforcement learning (RL), trajectory optimization, unmanned aerial vehicle (UAV).

I. INTRODUCTION

RECENTLY, unmanned aerial vehicle (UAV)-aided communication networks have attracted significant interest because of their potential to provide reliable high-rate connectivity, which is promising for future wireless networks [1]–[10]. On the one hand, UAVs can 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 potential applications in next-generation wireless networks with high mobility, low cost, and flexible deployment [13].

The Internet of Things (IoT) directs a massive connectivity between physical things, e.g., vehicles, and wearable devices such as, various sensors, which are key components of future networks [14]–[16]. However, the transmission reliability of IoT devices remains challenging owing to the long-distance and power constraints [17]. UAVs are expected to overcome these problems because of their high mobility and flexibility. UAVs can navigate over large areas, collecting data from IoT sensors, and transmitting them to other IoT devices or directly to a data center [18]. The high feasibility of LoS connections can improve the transmission efficiency and power of IoT devices [11], [13]. Thus, the service life can be extended.

Nonetheless, the limited battery power of UAVs [19] is a crucial challenge despite the benefits of UAV application in IoT data collection. Moreover, a survey discussing UAVs in public safety communications from an energy efficiency perspective and addressing the limited battery power problem is presented in [20]. Some previous studies proposed trajectory optimization to reduce the energy consumption of UAVs, thus expanding the flight time [21]–[23]. However, most of these studies only considered a fixed-altitude scenario in a two-dimensional (2-D) environment with a limited action space (e.g., up, down, right, and left). The above-mentioned scenario is impractical in an actual environment in which a UAV moves in a three-dimensional (3-D) directional space. Moreover, a limited directional space affects the higher number of shifting actions that the UAV performs and its traveling cost, which results in higher energy consumption. Therefore, a 3-D directional space and 3-D environment for UAVs should be designed to address this problem. Several studies have proposed the optimization of 3-D trajectories to reduce the travel cost and improve the energy efficiency of UAVs [25]. However, the directional space of UAVs in these studies was limited (e.g., ascend, descend, right, left, forward, and backward).

In addition, owing to the high mobility of UAVs which results in change in the environment over time, the trajectory optimization problem of UAVs is difficult to solve using classical optimization techniques (e.g., mathematical derivation) [27]–[30]. Motivated by the advancement in machine learning, this article uses a type of machine learning called reinforcement learning (RL) to optimize the trajectory and maximize the energy efficiency of UAVs. RL is considered dynamic learning, which involves determining optimal actions for an agent that interacts with the environment [31]. RL employs trial-and-error processes to make informed decisions based on imperfect information about the environment. Based on these characteristics, RL is suitable for addressing challenges in dynamic environments and sequential times, such as UAV communication problems. Therefore, a multidimensional trajectory for UAV is proposed and integrated with RL to...
maximize number of collected IoT sensors data per energy consumption of UAV.

### A. Related Works

A comparison of related studies is presented in Table I. Some of these studies optimized the trajectory of UAV-assisted communication networks. An optimized 2-D trajectory for a UAV was presented in [21] to minimize energy consumption using the bisection search algorithm and block coordinate descent (BCD) method. Moreover, the energy consumption of UAVs during IoT data collection was minimized by jointly optimizing the communication scheduling, transmission power allocation for IoT devices, and the UAV trajectory [22]. A hovering positions, trajectory, and scheduling of UAVs were optimized for reliable and flexible emergency communication in [34]. In [35] nonorthogonal multiple access (NOMA) precoding and UAV trajectory were jointly optimized to maximize the cumulative collected IoT sensors data.

| Brief Summary of Related Works | Related Works | 3D Direct. Space | RL-based Optim. | Practical Energy Consumption | UAV-aided IoT Net. |
|--------------------------------|---------------|-----------------|----------------|-----------------------------|-------------------|
| [19] A uniform framework of drone energy consumption models and the inter-relationships between key factors and performance measure. | [19] | - | - | ✓ | - |
| [21] A 2D UAV trajectory design to optimize overall energy consumption using bisection search and block coordinate descent (BCD) method. | [21], [22] | - | - | ✓ | - |
| [23] Optimize 2D UAV trajectory to minimize the energy consumption of UAVs during IoT data collection. | [23] | - | ✓ | ✓ | ✓ |
| [24] A 2D UAV trajectory design using deep reinforcement learning to minimize energy consumption during WSNs data collection. | [24], [38] | - | ✓ | - | ✓ |
| [26] A 2D UAV trajectory is designed to improve system capacity using a constrained Deep Q-Network (cDQN) algorithm. | [26], [32], [33], [37] | ✓ | ✓ | - | - |
| [31] A UAV Control policy based on the Deep Deterministic Policy Gradient (UC-DDPG) is proposed to minimize the cumulative collected IoT sensors data and UAV energy efficiency. | - | - | - | - | - |
| [27]-[30] | ✓ | - | - | - | - |

**TABLE I RELATED WORKS**

B. Contributions

The key contributions of this article are summarized as follows.

1) A larger directional space of UAVs is proposed to increase energy efficiency. For more convenient understanding, in practical condition, limited DOFs results a higher travel distance for UAV to reach a specific location. Therefore, energy that consumed by UAV to travel and changing direction is hugely affecting flight time to become less. Furthermore, a less flight time results less collected data by UAV, which means less energy efficiency. Compared with previous studies in [24], [26], and [37], which only considered four and six directions (UC-DDPG) energy efficiency of UAVs and fair coverage service is proposed in [33]. Additionally, in [37], 3-D trajectories with six degree of freedom (DOF) for multi-UAV-assisted communication were optimized using RL (specifically, Q-learning) for maximization of users’ satisfaction. However, none of these studies utilized RL for trajectory optimization problems in UAV-aided IoT networks. Moreover, RL has been used to optimize the trajectory of UAVs in IoT data-collection mission [24]. In [38], the height of UAV is optimized using RL to maximize the collected data throughput of sensor devices and the energy efficiency of UAV with wireless charging scenario. However, none of the aforementioned studies considered the extended directional space of UAVs for 3-D trajectory optimization. Furthermore, none of these studies considered a UAV energy consumption model based on actual practical scenarios (e.g., avionics energy of UAVs).
that affect the higher traveled distance of the UAV, the proposed model provides 26 directions that reduce the travel distance and energy consumption. First, the direction space in the 2-D scenario is increased to eight directions. Subsequently, a 3-D directional space with 26 directions is proposed.

2) A 3-D environment that considers ground contour is designed for UAV flight navigation, which makes the scenario more practical compared with previous studies [24], [26], [37].

3) An energy consumption model is derived based on [19]. The proposed model considers practical scenarios, such as avionics, hovering, traveling, and communication energy consumption. To the best of our knowledge, none of the aforementioned studies considered practical energy consumption.

4) The effects of the directional space and flight environment of UAV toward energy efficiency are investigated. A higher DOF decreases the traveling distance and energy to change direction of UAV that leads to a reduction of cumulative energy consumption. Thus, the energy efficiency of UAV for collecting IoT sensors data is improved. Moreover, a larger dimension of flight environment (3-D space) has less energy efficiency compared to 2-D space, which is make sense in practical, since the energy consumption for UAV to travel is increase along with the travel space.

C. Paper Organizations

The remaining of this article is organized as follows. Section II presents the system model of a UAV in a 2-D environment. Furthermore, a system model in 3-D space for UAV is presented in Section III. In Section IV, an energy consumption model for actual rotary-wing UAV is presented. The RL model of optimized trajectory for UAV is presented in Section V. The simulation results are discussed in Section VI. Finally, concluding remarks are given in Section VII.

D. Notations

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

II. SYSTEM MODEL OF FIXED-ALTITUDE UAV IN 2-D SPACE

Consider a limited-battery UAV flying at a fixed altitude $H$ m above K IoT sensors from the initial point $L_I$ to the goal point $L_F$ as shown in Fig. 1. The sensors, denoted by $K = \{K_1, K_2, \ldots, K_k\}$, are randomly distributed in the region $M \in \mathbb{R}^3$ between the initial and goal points. As shown in Fig. 1, the UAV communicates with IoT sensors on the ground by providing direct connectivity from the sky. The UAV is assumed to apply time-division multiple access (TDMA) to collect the uploaded data sequentially from the sensors. Specifically, the UAV can only receive data-uploaded from a single sensor at each time $t$. This article focuses on the optimized trajectory of a UAV when collecting data for higher energy efficiency. In addition, we assume that the UAV communicates without the assistance of a central controller and has partial global knowledge of the environment. In other words, the sensor locations in the environment are locally known.

A. Signal Model

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

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

In contrast to the channel characteristics of terrestrial communications, UAV-to-ground (A2G) channels are likely to be dominated by (LoS) links [15]. Therefore, the communication between the UAV and sensors is considered to be monopolized by the LoS links. The channel gain between the UA V and the sensors is considered to be monopolized by (LoS) links [15]. Therefore, the communication gain $h_k(t)$ can be expressed as follows:

$$
h_k(t) = \alpha_0 d_k(t)^{-\beta} = \frac{\alpha_0}{(||l(t) - p_k||^2 + H_t^2)^{\frac{\beta}{2}}} \tag{2}
$$

where $\alpha_0$ denotes the power loss of the channel at $d_0 = 1$ m where $d_0$ is the reference distance and $\xi_{pl} \geq 2$ is the path-loss exponent. The channel gain $h_k(t)$ depends on the location of the UAV at $t$.

### B. Data Transmissions

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

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

Therefore, the received data rate of the UAV from sensor $K_k$ at $t$ can be expressed using (4), shown at the bottom of the page, where $P_k$ denotes the transmit power of sensor $K_k$, $\sigma^2$ denotes the noise power, and $\gamma_0$ denotes the signal-to-noise ratio (SNR).

Moreover, considering that only a specific sensor can communicate with the UAV and upload data at $t$, the established connection is indicated by

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

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

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

Based on (3) and (4), the received data rate of the 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 UAV trajectory directly affects the achievable data rate. Thus, the total received data rate during $T$ can be expressed as follows:

$$
C(t) = \int_{0}^{T} \rho_k(t)C_k(t) dt \tag{7}
$$

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 $C(t)$ of the UAV satisfies the following:

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

### III. SYSTEM MODEL OF OPTIMIZED UAV ALTITUDE IN 3-D SPACE

#### A. Signal Model

Under realistic scenarios, the exploration environment is most likely to be in a 3-D coordinate space. Therefore, in this article, we consider the terrain of the environment. This means that the $K$ 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 $x$, $y$, and $z$-axes. Similar to a 2-D system, the UAV is assumed to fly from the initial point $L_t = (x_t, y_t, H_t)$ 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 $t$ can be represented as $L_t = (x_t, y_t, H_t)$. Therefore, the distance between the UAV and sensor $K_k$ at $t$ can be expressed as follows:

$$
d_k(t) = \sqrt{||l(t) - p_k||^2} = \sqrt{(x(t) - a_k)^2 + (y(t) - b_k)^2 + (H(t) - c_k)^2} \tag{9}
$$

where $x(t)$ and $y(t)$ denote the UAV coordinates at $t$. The UAV height at $t$ is denoted by $H(t)$. Additionally, the channel gain between the UAV and sensor $K_k$ is given by

$$
h_k(t) = \frac{\alpha_0 d_k(t)^{-\beta}}{(||l(t) - p_k||^2)^{\frac{\beta}{2}}} \tag{10}
$$

where $\alpha_0$ denotes the power loss of the channel at $d_0 = 1$ m, where $d_0$ is the reference distance and $\xi_{pl} \geq 2$ is the path-loss exponent. The location of the UAV at $t$ affects the channel gain $h_k(t)$.

$$
C_k(t) = \int_{0}^{T} B \log_2 \left( 1 + \frac{P_k h_k(t)^2}{\sigma^2} \right) dt \\
= \int_{0}^{T} B \log_2 \left( 1 + \frac{\gamma_0}{[(x(t) - a_k)^2 + (y(t) - b_k)^2 + H(t)^2]^{\xi_{pl}}} \right) dt \tag{4}
$$
B. Data Transmission

As mentioned previously, the UAV is assumed to collect sensor data using the TDMA scheme. Moreover, sensor \( K_k \) has the maximum channel gain when the distance between the UAV and sensor is minimum for each \( t \in [0, T] \) which can be expressed as follows:

\[
K_k = \min_{K_i \in K} ||L(t) - P_k||
\]

(11)

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

IV. ENERGY CONSUMPTION MODEL FOR ROTARY-WING UAVS

Generally, the energy consumption of a UAV is divided into two main components: 1) the communication and 2) propulsion energy. The consumption of propulsion energy is required to support the operation and maneuvering of UAVs. In this study, the propulsion energy consumption was modeled based on actual conditions that consider the physical model of the UAV. Propulsion energy is composed of three parts: 1) travel; 2) hovering; and 3) avionics energy [19]. 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 (13), shown at the bottom of the page. \( n_l \) denotes the mass of the UAV framework and battery, \( g \) denotes the gravitational constant, \( q(t) \) denotes the distance traveled by the UAV at \( t \), \( r \) denotes the lift-to-drag ratio, and \( n \) denotes the number of rotors. In addition, the energy consumption of the UAV during hovering was considered. The hovering energy of a UAV can be expressed as follows:

\[
E_h = \left( \frac{\sum_{l=1}^{2} n_l g}{2 n \rho \zeta} \right)^{3/2}
\]

(14)

where \( \rho \) denotes the air density constant, and \( \zeta \) denotes the spinning area of the blade disc of one rotor. In addition, the energy consumption of avionic components is given by

\[
E_a = \int_0^T \frac{P_{avio} q(t)}{v(t)} dt
\]

(15)

where

\[
v(t) \triangleq \dot{q}(t).
\]

(16)

\( P_{avio} \) denotes the power for avionic components, \( q(t) \) denotes the travel distance of the UAV, and \( v(t) \) denotes the UAV velocity at \( t \). 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:

\[
E_c = \sum_{k=1}^{K} P_{UAV} \int_0^T \rho_k(t) C_k(t) dt
\]

(17)

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

\[
E_{\text{total}} = E_t + E_h + E_a + E_c
\]

(18)

where \( E_t, E_h, E_a, \) and \( E_c \) denote the travel, hovering, avionics, and data communication energy of the UAV, respectively. Based on the received data rate \( C(t) \) in (8), and the total energy consumption of the UAV in (13), the energy efficiency of the UAV can be calculated by dividing the received data rate by the total energy consumed by the UAV during flight. The energy efficiency of the UAV can be expressed using (20) and (21)

\[
\eta_{EE}(t) = \frac{\dot{C}(t)(\rho_k(t), q(t), T)}{E_{\text{total}}}
\]

(19)

Based on (19)–(21), the energy efficiency of the UAV is affected by the total received data rate and 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 \( \eta_{EE}(t) \) in the long term by designing an optimized trajectory of the UAV. The optimization of the
UAV trajectory for maximizing the energy efficiency can be expressed as

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

where $v(t)$, $E_{\text{total}}$, and $T$ denote the velocity, travel distance, total energy consumption, and flight time of the UAV, respectively. Based on the above formulation, energy efficiency depends critically on the trajectory of the UAV. Considering an actual scenario in which 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). The MDP is a well-known model that can solve the problem of a partially observed environment with uncertainties and challenges. To solve the problem formulation described in (20) and (21), shown at the bottom of the page, we convert the problem into a multiperiod 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 RL. The limited flight time of the UAV $T$ is discretized into $M$ time slots. Thus, the step size is defined as $\delta = (T/M)$. Furthermore, environment $M$ can be divided into $M_1 = (L_1/\delta)\omega$ using $M_2 = (L_2/\delta)\omega$ tiles, where $\delta m$ is the length of each side. In addition, $L_1$ and $L_2$ denote the length and width of the environment $(m)$, respectively, and $\omega$ denotes the speed of the UAV $(m/s)$. Fig. 2(a) shows a projection of environment $M$ on a horizontal ground plane for a fixed-altitude UAV. As Fig. 2(a) shows, environment $M$ is divided into $M_1 \times M_2$ tiles and each sensor $K_k \in K$ is placed at a single tile.

Based on the discretized region of the considered environment, the state set $S$ of the UAV is represented as $S = \{s(1), s(2), \ldots, s(M_1M_2)\}$. States $s(m)$ and $s(m) \in S$ refer to small tiles. Fig. 2(a) depicts a UAV agent navigating through the tiles from the initial to the goal point while collecting sensors data randomly distributed within the environment. Owing to its limited battery energy, the finite flight time of a UAV requires the observation of the optimal trajectory. By considering that the UAV can only partially observe the environment, RL aids in making accurate decisions for an optimized trajectory.

$$\eta_{\text{EE}}(t) = \frac{\int_0^T \rho_k(t)C_k(t) \; dt}{\int_0^T \left( \sum_{l=1}^2 n_l \right) \rho(t) \; dt + \left( \sum_{l=1}^2 n_l \right)^{3/2} \frac{\sqrt{2n\rho\xi}}{\sqrt{2n\rho\xi}} + \int_0^T \rho_{\text{avio}}(t) \; dt + \sum_{k=1}^K P_{\text{UAV}} \int_0^T \rho_k(t)C_k(t) \; dt} \quad (20)$$

$$\eta_{\text{EE}}(t) = \frac{\int_0^T \rho_k(t)C_k(t) \; dt}{\int_0^T \left( \sum_{l=1}^2 n_l \right) \rho(t) \; dt + \left( \sum_{l=1}^2 n_l \right)^{3/2} \frac{\sqrt{2n\rho\xi}}{\sqrt{2n\rho\xi}} + \int_0^T \rho_{\text{avio}}(t) \; dt + \sum_{k=1}^K P_{\text{UAV}} \rho_k(t)C_k(t) \; dt} \quad (21)$$

Fig. 2. Illustration of IoT sensor data collection for (a) fixed-altitude UAV with eight possible flight directions and (b) optimized-altitude UAV with 26 possible flight directions.
when communication with the UAV is established. Moreover, the UAV hovers when collecting sensor data. By ensuring the communication quality, the UAV collects data from the sensor only when it moves to the sensor location at time \( m \in \mathcal{M} \) [Fig. 2(a)]. As previously mentioned, the UAV has a finite flight time, and the maximum time required for the UAV to move from one tile to another can be expressed as follows:

\[
T_s = \frac{T_{\text{max}}}{M_1 M_2}
\]

where \( T_{\text{max}} \) denotes the maximum flight of the UAV, and \( M_1, M_2 \) denote the tiles that represent the time slot. In addition, the maximum time required by the UAV to fetch sensor data is defined by

\[
T_H = \max \left\{ \frac{B_k}{C_k(t)} \right\}, T_s \quad \text{s.t.} \quad T_H \leq T_s
\]

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

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

where \( T_s \) denotes the time required 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:

\[
R(t) = \begin{cases} \eta_{\text{EE}}(t), & \text{if } C(t) \geq r_0 \\ 0, & \text{otherwise} \end{cases}
\]

where \( r_0 \) denotes 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: Compared with the fixed-altitude scenario, the optimized-altitude UAV learns the optimal 3-D trajectory. The environment \( \mathcal{M} \) is discretized into \( M_1 = (L_1 / \delta m) \times M_2 = (L_2 / \delta m) \times M_3 = (L_3 / \delta m) \) voxels, with a length of \( \delta m \) on each side. Here, \( L_1, L_2, \) and \( L_3 \) denote the height, length, and width of environment \( m \), respectively, and, \( \omega \) denotes the speed of the UAV (m/s). Fig. 2(b) shows a projection of environment \( \mathcal{M} \) on a 3-D space for the optimized-altitude UAV scenario. As shown in Fig. 2(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.

By referring to the discretized region of the considered environment, the state set \( S \) of the UAV can be represented as \( S = \{ s(1), s(2), \ldots, s(M_1 M_2 M_3) \} \). Each state \( s(m) \) and \( s(m) \in S \) refers to a voxel. Fig. 2(b) shows a UAV agent navigating through the voxel from the initial to the goal point while collecting sensors data randomly distributed within the environment. Owing to its limited battery energy, the finite flight
time of the UAV requires the observation of an optimal 3-D trajectory. RL can aid the UAV in making accurate decisions for an optimized 3-D trajectory.

2) Action Space: Because the UAV can navigate in a 3-D space, the action space of the UAV is modeled in a Bloch sphere [Fig. 2(b)]. In this scenario, the UAV agent has a maximum of 26 actions in each state, that is, {e.g., Ascend, Descend, West, North West, North, North East, East, South East, South, South West, Ascend West, and Descend North West}. For clarity, the action set of the UAV is denoted by

\[ A = \{ A, D, W, NW, N, NE, E, SE, S, SW, AW, DW, ANW, DNW, AN, DN, ANE, DNE, AE, DE, ASE, DSE, AS, DS, AW, DN, ANW, DNW, DSW \} \]

Because the UAV flies at an optimized altitude \((H(t))\) at \(t\), it moves along the \(x, y, \) and \(z\) axes. Note that in practice, the UAV can select its path along any direction in \(\theta\) and \(\phi\), that is, \(\theta \in [0, 2\pi]\), \(\phi \in [0, 2\pi]\).

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

\[ T_s = \frac{T_{\text{max}}}{M_1 M_2 M_3} \]

where \(T_{\text{max}}\) 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 \(T_{\text{req}}\), and reward function \(R(t)\) can be obtained using (24)–(26).

C. Reinforcement Learning-Based Optimized Trajectory Design

RL was adopted in this study to solve sequential decision making of a UAV during the data-collecting mission. RL addresses this problem by maximizing the immediate and long-term rewards. The maximum rewards are achieved from the best action performed by the agent. The best action is selected under a policy\(^1\) and updates it using a \(Q\)-table that records the experience of the agent’s action in a state and collects the reward. At the beginning of the learning process, the state–action values in the \(Q\)-table are initialized to zero. In a state \(S_t\) at \(t\), the UAV as the learning agent consecutively implements the following steps: select an action \(A_t\) from action set \(A\), move to the next state \(S_{t+1}\), receive a reward \(R_{t+1}\), and update the \(Q\)-value in the \(Q\)-table. In this article, the agent state is represented by the location of UAV \((X_t, Y_t, H_t)\) at \(t\) with action set \(A\) as mentioned in Sections V-A and V-B. Based on the current state \((X_t, Y_t, H_t)\) and selected action \(A_t\), the reward is collected, and the UAV moves to a 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 the total UAV energy consumption over time \(T_s\), as defined in (26), Section V-A.

In this article, a model-free temporal difference (TD) method\(^2\) is used to update the \(Q\)-value because the UAV requires multiperiod decision-making. First, the \(Q\)-value is updated using an off-policy method known as \(Q\)-learning. In the \(Q\)-learning method, the policy cannot be correlated with the policy being evaluated and improved. The \(Q\)-value in \(Q\)-learning can be updated using

\[ Q(S_t, A_t) = Q(S_t, A_t) + \alpha \left[ R_{t+1} + \gamma \max_{a \in A} Q(S_{t+1}, a) \right] \]

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

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

The updated \(Q\)-value in (29) depends on the information of variables \(S_t, A_t, S_{t+1}, A_{t+1}\). SARSA iteratively estimates the \(Q\)-value under policy \(Q_\pi\) and simultaneously changes policy \(\pi\) in a greedy manner at the same time based on \(Q_\pi\). According to the updated \(Q\)-values, the UAV as an agent exploits what it has already experienced to receive a maximum reward. However, the agent must also explore unmapped or altered environments to determine whether it can execute better action selections in the future. Exploration motivates the agent to learn from the environment, where it enables the learning process to get trapped out of a local optimum. Specifically, exploration provides the UAV the chance to perform a specific action, observe a reward, and update its action choices. Meanwhile, exploitation assists UAVs by utilizing the knowledge that is already available to obtain the best action. Exploitation drives the convergence of the learning process. Thus, the selection between exploration and exploitation is an aspect that enables UAVs to obtain beneficial actions during the learning process. By introducing a parameter \(0 < \epsilon < 1\), an \(\epsilon\)-greedy exploration is utilized to better balance exploration and exploitation.

\[ A_t = \begin{cases} \text{random selection,} & \text{with probability } 1 - \epsilon \\ \arg\max_{a \in A} Q(S_t, a), & \text{with probability } \epsilon \end{cases} \]

The agent selects the action because it considers that the action yields the maximum long-term reward with probability \(1 - \epsilon\). In contrast, the agent selects an action uniformly at random with a probability \(\epsilon\). Action selection based on \(Q\)-values may fall into a local optimum when \(\epsilon = 0\).

To elaborate the \(Q\)-learning-based trajectory optimization with 3-D directional space, we present an example.

Example: When UAV begins exploring from the initial location at a state \(S_0 = (5,1,1)\), one of the neighboring states \((4,1,1), (4,1,2), (4,2,1), (4,2,2), (5,2,1), (5,2,2)\) is explored by selecting an action from \(A\) according to state–action values, i.e., \(\max_{a \in A} Q(S_0, a)\). Because all \(Q\)-values at the initial \(Q\)-learning are initialized to zero, an action will be selected randomly or according to some preset criteria.

\(^1\)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 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 state–action–reward–state–action (SARSA) are TD-based methods [11].
TABLE II
SIMULATIONS PARAMETERS

| Parameter                        | Notation | Value                              |
|----------------------------------|----------|------------------------------------|
| Region considered                | $\mathcal{M}$ | $100 \times 100 \times 100 \text{ m}^3$ |
| Number of sensors                | $K$      | 2, 4, 6                            |
| Maximum flight time of UAV       | $T$      | 30 min                             |
| UAV speed                        | $\omega$ | 5.56 m/s                           |
| UAV fixed altitude               | $H$      | 100 m                              |
| Mass of the UAV (framework and battery) | $m_w$ | 1.46 kg                            |
| Gravity acceleration             | $g$      | 9.807 m/s$^2$                      |
| Lift to drag ratio               | $\tau$  | 3                                  |
| Number of UAV rotors             | $n$      | 4                                  |
| Air density                      | $\rho$  | 1.225 kg/m$^3$                     |
| Spinning blade area of one rotor | $\zeta$ | 0.0507 m$^2$                       |
| Transmit power of UAV            | $P_{\text{UAV}}$ | 30 dBm                             |
| Transmit power of sensors        | $P_k$   | 10 dBm                             |
| Bandwidth                        | $B$      | 2 MHz                              |
| Noise power                      | $\sigma^2$ | -110 dBm                          |
| Channel power gain               | $\alpha_k$ | -50 dBm                           |
| Path loss exponent               | $\zeta_{pl}$ | 2                                 |
| Discount factor                  | $\gamma$ | 0.98                               |
| Learning rate                    | $\alpha$ | 0.5                                |

Assumed action “AN” which represents the Ascend North direction, is selected; thus, the next state becomes $S_1 = (4, 1, 2)$. The corresponding reward and $Q$-value $Q(S_0, A_0 = AN)$ are updated accordingly. Based on the example, the $Q$-values is increased by the collected reward value using (25). The $Q$-value is iteratively updated during learning based on the selected action that results in the maximum long-term reward.

VI. PERFORMANCE EVALUATION

In this simulation, the exploration region was considered in a 3-D space with the upper limit set to $100 \times 100$ m$^2$ for the 2-D trajectory scenario and $100 \times 100 \times 100$ m$^3$ for the 3-D trajectory scenario. The considered region $\mathcal{M}$ was divided into $5 \times 5$ tiles for the 2-D trajectory and $5 \times 5 \times 5$ voxels for the 3-D trajectory. The initial location of the UAV was set as $L_I = (0, 0, 0)$, whereas the goal location was $L_F = (100, 100, 100)$. Based on actual applications [22], the maximum flight time of a quadrotor UAV $T$ was set to 30 min with a speed $\omega$ of 5.56 m/s. The mass of the quadrotor UAV, which consisted of its framework and battery $m_w$ was 1.46 kg. Because the UAV considered was a quadrotary-wing UAV, whose number of rotors $n$ was four, and the area of the spinning blade disc of one rotor $\zeta$ was 0.0507 m$^2$. In addition, the lift-to-drag ratio of UAV $r$ was 3. The UAV navigated from $L_I$ to $L_F$ with a gravity acceleration of 9.807 m/s$^2$, and the air density was 1.225 kg/m$^3$. As described in Section II, $K$ sensors were randomly distributed in region $\mathcal{M}$. The number of sensors $K$ simulated in this study was divided into three scenarios: 2, 4, and 6. The UAV communicated with sensor $K_k$ by transmitting power $P_{\text{UAV}}$ at 30 dBm, whereas the transmit power of sensor $P_k$ was 10 dBm. The sensors shared a total communication bandwidth of $B = 2$ MHz with a noise power $\sigma^2$ of $-110$ dBm. The channel power gain $\alpha_k$ was $-50$ dBm and the path loss exponent $\zeta_{pl}$ was 2.

1) $Q$-learning is an off-policy method in which the learning of the optimal policy is based on the best action, which is implemented greedily, unlike the current policy.
2) 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 was compared with those of limited DOF in terms of energy efficiency.

A. Energy Efficiency of Fixed-Altitude UAV Under 2-D Trajectory

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

Fig. 5 shows the energy efficiency of a UAV under a 2-D trajectory with two sensors randomly distributed. “Energy efficiency” refers to average sum rewards of (27), where...
higher rewards indicate a better performance. Moreover, “iterations” indicates the training iteration steps that optimized the action Q-values (Section V-C). In this study, the iteration steps were set to 10,000. As shown in the figure, the extended directional space with eight-DOFs achieved a higher energy efficiency than that with benchmark model four-DOFs ([24]). Moreover, Q-learning converged faster than SARSA in this scenario owing to the limited T of the UAV. Specifically, during the learning process, SARSA required more actions than Q-learning, as shown in Algorithm 1 for Q-learning and Algorithm 2 for SARSA. This occurred because SARSA learns the optimal policy based on the current policy, which requires more time, whereas Q-learning learns based on the best action performed in a greedy manner. Consequently, the UAV fails to reach the optimal result within a finite T.

Fig. 6 shows the comparison of R values between the proposed 8-DOF model with a related study [24] for four-DOFs under Q-learning and SARSA for four sensor scenarios. The figure shows similar trend results for the two sensors shown in Fig. 5. The figure clearly shows that 8-DOF UAV outperformed the 4-DOF UAV in maximizing the energy efficiency. Moreover, the proposed model under Q-learning achieved a better performance by converging faster than SARSA. Furthermore, the additional two other sensors resulted in a higher energy efficiency of the UAV because a higher amount of data could be collected.

Fig. 7 depicts the R values of proposed 8-DOF and 4-DOF [24] models under Q-learning and SARSA. As the figure shows, the 8-DOF model with Q-learning achieved higher rewards than 8-DOF SARSA and 4-DOFs models, which had a similar trend to the previous results (see Figs. 5 and 6). Moreover, the result indicated that the number of sensors directly resulted in a higher number of collected rewards.

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

Fig. 8(a) depicts the optimal policy π∗SARSA for an optimized 3-D UAV trajectory with the 26-DOF SARSA method.
The optimal policy for the optimized 3-D UAV trajectory with the 26-DOF-based Q-learning method is shown in Fig. 8(b). Similar to the optimized trajectory for a fixed-altitude UAV, the optimal policies resulting from the SARSA and Q-learning methods in identical environments were different. A different optimal policy can generate different sum of rewards results.

Fig. 9 shows the energy efficiency of a UAV under a 3-D trajectory with two randomly distributed sensors. “Energy efficiency” refers to average sum rewards of (27), where higher rewards indicate a better performance. Moreover, “iterations” indicates the training iteration steps that optimized the action Q-values (Section V-C). In this study, the iteration steps were set to 10000. As shown in the figure, the extended directional space with 26-DOFs achieved a higher energy efficiency than that with benchmark model 6-DOFs ([26], [37]). Moreover, Q-learning converged faster than SARSA in this scenario because of the limited $T$ of the UAV. Specifically, during the learning process, SARSA required more actions than Q-learning, as shown in Algorithm 1 for Q-learning and Algorithm 2 for SARSA. This occurred because SARSA learns the optimal policy based on the current policy, which requires more time, whereas Q-learning learns based on the best action performed in a greedy manner. Consequently, the UAV fails to reach the optimal result within a finite flight time $T$.

Fig. 10 shows the comparison of $R$ values between proposed 26-DOFs with related studies [26], [37] on 6-DOFs under Q-learning and SARSA for four sensor scenarios. The figure shows similar trend results for the two sensors shown in Fig. 9. The figure clearly shows that the UAV with 26-DOFs outperformed the 6-DOF UAV in maximizing energy efficiency. Moreover, the proposed model under Q-learning achieved a better performance by converging faster than SARSA. Furthermore, the additional two sensors resulted in a higher energy efficiency of the UAV because a higher amount of data could be collected.

Fig. 11 depicts the $R$ values of proposed 26-DOF and 6-DOF [26], [37] models under Q-learning and SARSA. As shown in the figure, 26-DOF with Q-learning achieved higher rewards than the 26-DOF SARSA and 6-DOF models, which was a similar trend to the previous results (Fig. 9 and 10). Moreover, the result indicated a higher number of collected rewards, which were directly affected by the number of sensors.
VII. CONCLUSION

This article proposed a multidimensional search space that employs higher DOFs to optimize the trajectory of UAV. Moreover, the trajectory was optimized utilizing the RL method, specifically Q-learning, to maximize the energy efficiency of UAV-aided IoT networks. In addition, the energy consumption model of a UAV was designed by considering practical environments. The performance of the proposed model was compared with those using limited DOFs and the SARSA method. The proposed extended DOFs exhibited a higher energy efficiency than the limited DOFs and SARSA. Moreover, Q-learning achieved faster iteration convergence than SARSA. The following points can be considered in future studies.

1) Intersensor interference from other sensors can be considered for a data transmissions scheme.
2) A deep neural network RL scheme (called DRL), i.e., [32], [33] can be used to further improve energy efficiency of UAV-aided IoT networks.
3) For a better performance, a cooperative multiagent RL (MARL) scheme can also be applied to this scenario, such as in [19], [21], and [37].

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),” J. Intell. Robot. Syst., vol. 54, pp. 61–78, Mar. 2009.
[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. Surveys Tuts., vol. 21, no. 3, pp. 2334–2360, 3rd Quart, 2019.
[4] S. ur Rahman, G.-H. 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. 2018.
[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] D. Gupta, R. Jain, and G. Vaszkun, “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, C. Jiang, Z. Han, Y. Ren, R. G. Maunder, and L. Hanzo, “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. 5, 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.-C. 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 Internet Things J., 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 Internet Things J., vol. 6, no. 4, pp. 7025–7037, Aug. 2019.
[18] X. Lin et al., “The sky is not the limit: LTE for unmanned aerial vehicles,” IEEE Commun. Mag., vol. 56, no. 4, pp. 204–210, Apr. 2018.
[19] J. Zhang, J. F. Campbell, D. C. Sweeney II, and A. C. Huffman, “Energy consumption models for delivery drones: A comparison and assessment,” Transport. Res. Part D: Transp. Environ., vol. 90, Jan. 2021, Art. no. 102668.
[20] S. Shakoor, Z. Kaleem, M. I. Baig, O. Chughtai, T. Q. Duong, and L. D. Nguyen, “Role of UAVs in public safety communications: Energy efficiency perspective, IEEE Access, vol. 7, pp. 140665–140679, 2019.
[21] J. Y. Wang, R. Ruby, Q.-W. Pham, and K. Wu, “Aiding a disaster spot via multi-UAV-based IoT networks: Energy and mission completion time-aware trajectory optimization,” IEEE Internet Things J., vol. 9, no. 8, pp. 5853–5867, Apr. 2020.
[22] C. Zhan and H. Lai, “Energy minimization in Internet-of-Things system based on rotary-wing UAV,” IEEE Wireless Commun. Lett., vol. 8, no. 5, pp. 1341–1344, Oct. 2019.
[23] 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, Sep. 2021.
[24] 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.
[25] T. Liu, M. Cui, G. Zhang, Q. Wu, X. Chu, and J. Zhang, “3D trajectory and transmit power optimization for UAV-enabled multi-link relay systems,” IEEE Trans. Commun. Netw., vol. 5, no. 1, pp. 392–405, Mar. 2021.
[26] 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.
[27] 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, Sep. 2020.
[28] 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, Jun. 2019.
[29] 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.

[30] 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.

[31] 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 Things J.*, vol. 8, no. 24, pp. 17275–17289, Dec. 2021.

[32] 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 Access*, vol. 7, pp. 149503–149514, 2019.

[33] 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 Things J.*, vol. 8, no. 24, pp. 17275–17289, Dec. 2021.

[34] 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 Access*, vol. 7, pp. 149503–149514, 2019.

[35] N. 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, Mar. 2020.

[36] N. Zhao et al., “UAV-assisted emergency networks in disasters,” *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 45–51, Feb. 2019.

[37] N. Zhao et al., “Joint trajectory and precoding optimization for UAV-assisted NOMA networks,” *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3723–3735, May 2019.

[38] S. Shakoor, Z. Kaleem, D.-T. Do, O. A. Dobre, and A. Jamalipour, “Joint optimization of UAV 3-D placement and path-loss factor for energy-efficient maximal coverage,” *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9776–9786, Jun. 2021.

[39] X. Liu, Y. Liu, and Y. Chen, “Reinforcement learning in multiple-UAV networks: Deployment and movement design,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.

[40] S. Fu et al., “Energy-efficient UAV-enabled data collection via wireless charging: A reinforcement learning approach,” *IEEE Internet Things J.*, vol. 8, no. 12, pp. 10209–10219, Jun. 2021.

Silvirianti received the bachelor’s degree in electrical engineering from Telkom University, Bandung, Indonesia, in 2017, and the master’s degree from the Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea, in 2020, where she is currently pursuing the Ph.D. degree.

Her major research interests include UAV communications, nonorthogonal multiple access, energy-efficient communication, robotics, and autonomous vehicle.

Soo Young Shin (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1999, 2001, and 2006, respectively.

He was a Visiting Scholar with the FUN Laboratory, University of Washington, Seattle, WA, USA, from July 2006 to June 2007. After three years working with the WiMAX Design Laboratory, Samsung Electronics, Suwon, South Korea, he has been an Associate Professor with the School of Electronics, Kumoh National Institute of Technology, Gumi, South Korea, since September 2010. His research interests include wireless LAN, WPAN, WBAN, wireless mesh networks, sensor networks, coexistence among wireless networks, industrial and military networks, cognitive radio networks, and next-generation mobile wireless broadband networks.