Multiagent Deep Reinforcement Learning for Joint Movement and User Association of UAV-BS Emergency Indoor User Service

Tae-Yoon Kim, *Member, IEEE*, Jihong Park, Junghwa Kang, *Member, IEEE*, Jaeyeol Lee, *Student Member, IEEE*, Soyi Jung, *Senior Member, IEEE*, and Jae-Hyun Kim, *Member, IEEE*

***Abstract*—This article investigates the joint optimization of uncrewed aerial vehicle-mounted base station (UAV-BS) movement and user association for multiple UAV-BSs providing emergency services to indoor users. Specifically, we focus on optimizing associations between UAV-BSs and indoor users within an outdoor-to-indoor path loss model that accounts for floor penetration. The primary objective is to determine the optimal associations between UAV-BSs and indoor users, while also addressing how multiple UAV-BSs should move to establish these associations quickly. To solve this problem, we propose a novel multiagent reinforcement learning (MARL) architecture featuring three key innovations: a dual-action structure that decouples the complex decision-making process into separate movement and association actions, a multiagent double deep *Q*-network (MADDQN) to learn optimal policies, and prioritized experience replay (PER) to improve learning efficiency. Simulation results demonstrate that the proposed algorithm significantly outperforms baseline methods—including a multiagent deep *Q*-network (MADQN), multiagent independent actor–critic (MAIAC), multiagent deep deterministic policy gradient (MADDPG), and a consensus-based bundle algorithm (CBBA)—across all metrics. Furthermore, a series of rigorous ablation studies systematically validates the contribution of each component. Overall, the simulation results validate the superiority and robustness of our proposed algorithm in dynamic and challenging indoor environments.**

***Index Terms*—Indoor user, joint movement and user association, multiagent double deep *Q*-network (MADDQN), priority experience replay (PER), uncrewed aerial vehicle-mounted base station (UAV-BS).**

Received 3 August 2025; revised 24 September 2025; accepted 1 November 2025. Date of publication 5 November 2025; date of current version 23 December 2025. This work was supported in part by the Institute of Information and Communication Technology Planning and Evaluation (IITP) through the Korea Government [Ministry of Science and ICT (MSIT)], Republic of Korea (Development of Cube Satellites based on Core Technologies in Low Earth Orbit Satellite Communications) under Grant 2024-00396992, (Development of 3D-NET Core Technology for High-Mobility Vehicular Service) Grant 2022-0-00704, and (Development of Ground Station Core Technology for Low Earth Orbit Cluster Satellite Communications) Grant RS-2024-00359235. *(Corresponding authors: Jae-Hyun Kim; Soyi Jung.)*

Tae-Yoon Kim and Jihong Park are with the Department of Artificial Intelligence Convergence Network, Ajou University, Suwon 16499, South Korea (e-mail: xodbsxogjs@ajou.ac.kr; jihong1215@ajou.ac.kr).

Junghwa Kang is with the SW Team (Land), Hanwha Systems, Pangyo 13524, Republic of Korea (e-mail: jh0220.kang@hanwha.com).

Jaeyeol Lee is with the Department of Space Electronics and Information Engineering, Ajou University, Suwon 16499, South Korea (e-mail: jaeyel98@ajou.ac.kr).

Soyi Jung and Jae-Hyun Kim are with the Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, South Korea (e-mail: sjung@ajou.ac.kr; jkim@ajou.ac.kr).

Digital Object Identifier 10.1109/JIOT.2025.3629374

I. Introduction

IN EMERGING and future wireless networks, uncrewed aerial vehicles (UAVs) have attracted interest in many fields, such as the military, civilian, agriculture, and industry, due to their flexibility and cost-effectiveness [1], [2]. In particular, UAVs are increasingly being explored as aerial communication platforms, providing rapid and cost-efficient solutions for wireless networks. Mobile network operators (MNOs), motivated by the success of pilot projects such as AT&T's Flying COW and Nokia's F-cell, are actively investigating UAVs as a means to enhance network coverage and resilience [3]. Specifically, UAV-mounted base stations (UAV-BSs) have been widely utilized to service users when terrestrial networks (TNs) are unavailable due to natural disasters or warfare [4]. The use of UAV-BSs in stable positions can offer a robust framework for next-generation wireless systems, contributing to improved spectral efficiency and quality of service (QoS) in highly dynamic environments [5].

Meanwhile, Huawei estimates that up to 70% of 5G traffic will occur indoors, underscoring the importance of indoor connectivity [6]. This becomes especially critical during emergencies, where many users remain trapped in high-rise buildings, and rescue teams must transmit real-time video, internet of things (IoT) sensor data, and other essential information to ground command centers [7]. UAV-BSs can serve as a rapid and flexible alternative in such scenarios, particularly when conventional networks fail indoors [8]. However, their effectiveness depends heavily on deployment and mobility strategies, as indoor users experience severe signal degradation due to outdoor-to-indoor path loss [9]. Additionally, UAV-BSs have limited onboard energy, unlike TNs with stable power sources, constraining flight time and service capacity. Thus, optimizing both UAV-BS movement and user association is vital to improve energy efficiency and communication performance. Proper optimization ensures reliable indoor connectivity, minimizes service interruptions, and supports effective emergency response.

In recent years, numerous studies have investigated UAV-BS deployment optimization, including trajectory planning, positioning, energy efficiency, and resource allocation [10], [11], [12]. However, most of these works focus on outdoor users with direct line-of-sight (LoS) channels, where signal degradation is minimal. To address indoor coverage, [13] proposed an outdoor-to-indoor path loss and power model

to minimize transmit power while ensuring sufficient indoor coverage. Similarly, Shakhatreh et al. [14] employed a particle swarm optimization (PSO)-based 3-D UAV placement strategy to provide reliable coverage in high-rise buildings during emergencies. While these prior studies provide a foundation for indoor UAV communications, their reliance on window-based propagation models and disregard for vertical attenuation across floors limits their applicability. This assumption is particularly unrealistic in factory-type or windowless structures, especially in emergency scenarios, thereby creating a significant research gap for realistic, floor-aware deployment strategies.

Incorporating realistic floor penetration loss dramatically increases the complexity of joint UAV movement and user association optimization. Even small vertical shifts in UAV altitude can cause abrupt changes in signal quality, creating a highly nonlinear and nonstationary environment. This complexity reveals the limitations of existing learning-based approaches. For example, Ma et al. [15] proposed a machine learning-based UAV-BS deployment and user association mechanism that aims to maximize downlink throughput with minimal computation time. However, their approach decouples user association and UAV-BS positioning and does not consider UAV mobility or energy constraints. Even advanced multiagent deep reinforcement learning (MADRL) frameworks, such as the one proposed in [16], represent a significant step forward by jointly considering UAV mobility and energy constraints to enhance user-level fairness. However, their applicability remains confined to outdoor scenarios with a simplified air-to-ground path loss model. The user association is kept static throughout the mission, and the impact of vertical positioning on signal degradation—particularly floor penetration in indoor environments—is not addressed. As a result, such methods face challenges when applied to realistic emergency scenarios in multifloor buildings, where joint and dynamic control of both movement and association is essential for ensuring reliable communication.

In contrast to prior studies that are constrained to outdoor environments or rely on simplified indoor models, this work presents a novel learning-based framework that realistically addresses emergency indoor communication. Specifically, we propose an MADRL architecture that jointly optimizes UAV-BS movement and user association in a floor-aware indoor environment, reflecting the highly dynamic and nonlinear nature of vertical signal attenuation in multifloor buildings. While a preliminary concept of this framework was introduced in a work-in-progress paper [17], this article presents the first complete and mature version by incorporating novel algorithmic components and comprehensive performance analysis in realistic environments.

The main contributions of this article are as follows.

1) We formulate a realistic indoor channel model by extending the ITU-R outdoor-to-indoor path loss equation to incorporate floor penetration loss, which is especially critical in factory-type or windowless buildings. To the best of our knowledge, such floor-aware propagation characteristics are rarely considered in UAV communication studies [18]. Furthermore, we empirically validate the impact of floor attenuation by conducting real-world signal measurements across multiple floors, confirming the model's fidelity and highlighting its necessity in indoor emergency communication scenarios. Based on this channel model, we compute the signal-to-interference-plus-noise ratio (SINR) and derive the achievable data rate using SINR thresholds defined by the modulation and coding scheme (MCS).
2) We cast the joint UAV movement and user association problem as a partially observable Markov decision process (POMDP), where each UAV agent makes decisions based solely on local observations. To solve this, we propose a state-aware dual-mode policy that dynamically alternates between two operational modes: association mode, activated when a high-quality user connection is feasible; and positioning mode, which guides movement toward strategically advantageous locations when association is suboptimal or unavailable. The policy is trained using a hybrid reward design that blends a common reward—representing system-wide connectivity performance—and an individual reward—reflecting local agent efficiency. This reward design enables agents to balance short-term communication opportunities with long-term deployment strategies. We implement the framework using a multiagent double deep Q-network (MADDQN) with prioritized experience replay (PER), ensuring stable convergence even under sparse and delayed feedback environments [19].
3) Moreover, this article conducts extensive simulations to evaluate the performance of the proposed method in comparison with base algorithms, including the vanilla multiagent deep Q-network (MADQN), a policy gradient-based multiagent independent actor–critic (MAIAC), the multiagent deep deterministic policy gradient (MADDPG) algorithm, and the consensus-based bundle algorithm (CBBA), a representative optimization-based method for multiagent task allocation. The evaluation considers variations in both the minimum SINR threshold for user association and the number of users.

The remainder of this article is organized as follows. Section II reviews related work on UAV-BS systems. Section III describes the system model and formulates the joint optimization problem. Section IV presents the proposed MADDQN algorithm with PER, which jointly optimizes UAV-BS movement and user association for time-efficient indoor emergency services. Section V evaluates the performance of the proposed algorithm through extensive simulations. Finally, Section VI concludes the article.

II. Related Work

A. UAV-BS Operation in Outdoor Environments

In the field of wireless networks, extensive research has been conducted on utilizing UAVs as flying BSs, servers, or relays to provide communication services [20], [21]. In particular, extensive research has been conducted on utilizing UAV-BSs

in disaster scenarios where TNs are unavailable. In such situations, ensuring rapid communication for users is crucial. Additionally, due to the limited battery capacity of UAV-BSs, optimizing their operation for energy-efficient communication is essential [22], [23]. To address these challenges, the proposed method in [24] demonstrates UAV-assisted emergency communication in postdisaster areas using an extended multiarmed bandit (MAB) framework to optimize UAV trajectory for maximizing user coverage under battery constraints. Sambo et al. [25] proposed a genetic algorithm (GA)-based energy-efficient UAV trajectory design for delay-tolerant emergency communication, optimizing flight paths for backhaul connectivity to truck-mounted BSs while considering both straight-and-level and banked-level turns. Furthermore, several studies have focused on optimizing UAV-BS power control and resource allocation to enhance system throughput while maintaining energy-efficient communication [26], [27]. By dynamically adjusting transmission power and efficiently allocating resources, these approaches aim to maximize network performance and user coverage in UAV-assisted communication systems. Hu et al. [26] proposed an uplink throughput optimization scheme for UAV-enabled urban emergency communications, where a UAV acts as a relay using nonorthogonal multiple access (NOMA) to forward data from disconnected ground access points (APs) to a remote BS. Liu et al. [27] propose a joint UAV-BS deployment and power allocation strategy using a loop iterative algorithm to maximize user throughput in maritime emergency communication. However, these studies all assume outdoor environments, while UAV-BS positioning plays a much greater role in signal attenuation in indoor scenarios. As mentioned earlier, most traffic occurs indoors, and ensuring the safety of indoor users is even more critical than outdoor users in disaster scenarios [13], [14].

B. Indoor Wireless Communication With UAV-BSs

Cui et al. [28] proposed a UAV-based decision-making scheme using an indoor–outdoor-iterative optimization approach and a method to estimate outdoor user distribution to optimize bandwidth and power allocation, ensuring fair coverage for emergency indoor and outdoor users. Nevertheless, this work focuses on optimizing UAV-BS placement rather than considering the movement of a single UAV-BS. Guo et al. [29] proposed a joint UAV trajectory and resource allocation optimization algorithm for video streaming in UAV-based emergency indoor–outdoor communication, enhancing uplink throughput and video quality using successive convex approximation and block coordinate descent techniques. In this study, UAV-BS movement is considered, but the limited battery capacity of UAV-BSs is not taken into account. Additionally, the approach focuses solely on a single UAV-BS, lacking consideration for the realistic challenges of multi-UAV deployment. Shakhatreh et al. [30] proposed a UAV-based indoor wireless coverage strategy for high-rise buildings by incorporating an outdoor-to-indoor path loss model and optimizing UAV placement to minimize transmit power. The study formulates single and multiple UAV placement problems, using gradient descent, PSO, and clustering algorithms to enhance coverage efficiency. This is promising in that it considers multiple UAV-BSs, but it still does not account for UAV-BS mobility and user association, making it insufficient for solving problems in more complex and dynamic environments.

C. Multiagent Reinforcement Learning for UAV-BS Control

To address such complex problems, an increasing number of studies are leveraging MADRL-based approaches to tackle challenges that traditional optimization methods fail to solve. Cui et al. [11] proposed an MARL-based resource allocation framework for multi-UAV networks, where UAVs independently optimize user selection, power levels, and subchannel allocation using Q-learning, achieving efficient and scalable decision-making without requiring full network information. However, this approach solely relies on Q-learning, leading to performance degradation as complexity increases, and it only focuses on resource allocation without considering UAV-BS movement, limiting its applicability in dynamic environments. Ding et al. [31] proposed an MADRL-based UAV trajectory design and user access control framework to optimize UAV-BS movement and user association, ensuring fair and high-throughput air–ground coordinated communication while addressing hybrid action space challenges. However, this study does not consider the battery constraints of UAV-BSs, which is a crucial factor for practical deployment and long-term operation. Furthermore, [16] and [32] leveraged MADRL under centralized training with distributed execution (CTDE) to optimize UAV-BS positioning and mobility with energy constraints. However, these studies are confined to outdoor environments and rely on air-to-ground path loss models that assume smooth signal variations. In contrast, indoor multi-floor environments introduce critical learning challenges that undermine the effectiveness of conventional MADRL frameworks. First, conventional Q-learning and actor–critic methods struggle with the highly volatile reward dynamics created by severe floor penetration loss. These algorithms presume a relatively stable state-value landscape and are destabilized by the abrupt, "cliff-like" SINR changes that occur with minor vertical UAV displacements, hindering policy convergence. Second, standard MADRL agents typically learn over a flat action space, which makes it difficult to resolve the ambiguity between positioning errors and association errors, leading to policy divergence. Finally, the intensified reward sparsity of indoor environments is a critical challenge for methods relying on uniform experience replay, as the precise 3-D alignment required for a successful connection makes positive rewards rare.

To the best of our knowledge, no prior work tackles these core learning challenges for indoor UAV-BS control. As summarized in Table I, our approach is the first to integrate a floor-aware indoor propagation model with an MADRL framework featuring a dual-mode control policy that adaptively balances user association and positioning under energy constraints. This holistic design not only unifies and extends the capabilities addressed in prior works but also represents the first MADRL-based solution tailored for emergency indoor scenarios, where realistic floor-aware propagation modeling is

TABLE I
COMPARATIVE ANALYSIS OF UAV-BS-BASED COMMUNICATION WITH EXISTING STUDIES

Reference	UAV-BS Movement	UAV-BS Battery	Indoor	Multi UAV-BS	RL	Floor Loss
J. Cui, *et al.* [13]	✗	✗	✓	✗	✗	✗
H. Shakhatrch, *et al.* [14]	✗	✗	✓	✗	✗	✗
B. Ma, *et al.* [15]	✗	✗	✗	✓	✗	✗
Z. Qin, *et al.* [16]	✓	✓	✗	✓	✓	✗
J. Kang, *et al.* [17]	✓	✗	✓	✗	✗	✓
Y. Lin, *et al.* [23]	✓	✓	✗	✗	✗	✗
Y. A. Sambo, *et al.* [24]	✓	✓	✗	✗	✗	✗
B. Hu, *et al.* [25]	✗	✗	✗	✗	✗	✗
L. Liu, *et al.* [26]	✗	✗	✗	✗	✗	✗
J. Cui, *et al.* [27]	✗	✗	✗	✗	✗	✗
Z. Guo, *et al.* [28]	✓	✗	✓	✗	✗	✗
H. Shakhatrch, *et al.* [29]	✓	✗	✓	✗	✗	✗
J. Cui, *et al.* [30]	✗	✗	✗	✓	✓	✗
R. Ding, *et al.* [11]	✓	✗	✗	✓	✓	✗
J. Kim, *et al.* [32]	✓	✓	✗	✓	✓	✗
Proposed algorithm	✓	✓	✓	✓	✓	✓

Fig. 1. UAV-BSs emergency indoor user service network.

indispensable for enabling stable learning and reliable service delivery.

III. SYSTEM MODEL AND PROBLEM FORMULATION

As shown in Fig. 1, we consider a disaster scenario where TNs are unavailable, and UAV-BSs receive uplink signals from emergency indoor users. The system consists of m multi-UAV-BSs, denoted as $\mathbf{M} = \{1, 2, \ldots, m\}$, and n indoor users in a building, denoted as $\mathbf{N} = \{1, 2, \ldots, n\}$. The coordinates of UAV-BS m and indoor user n at time slot t, where $0 \leq t \leq T$, are expressed by $l_m(t) = [x_m(t), y_m(t), z_m(t)]$ and $l_n(t) = [x_n(t), y_n(t), z_n(t)]$, respectively. The building is assumed to be a factory-type structure, where signal attenuation is significantly affected by UAV-BS positioning. The building

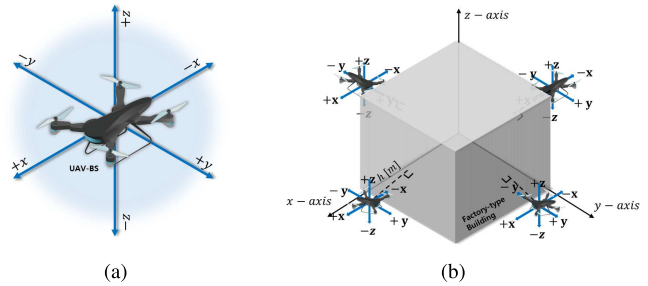


Fig. 2. UAV-BS mobility model and initial deployment. (a) Discrete mobility model of the UAV-BS in the 3-D grid. (b) Initial UAV-BS deployment around a factory-type building.

dimensions are given by $[0, x_b] \times [0, y_b] \times [0, z_b]$. Initially, each UAV-BS is deployed near the north, south, east, and west sides of the building, as illustrated in Fig. 2(b). The figure depicts the 3-D movement constraints of the UAV-BSs around the factory-type building, where each UAV-BS operates within a specific region to ensure efficient coverage of indoor users while minimizing interference. Fig. 2(a) illustrates the 3-D grid-based discrete mobility model. Initially, each UAV-BS is positioned at a horizontal offset of h meters from the building surface. From this initial position, each UAV-BS navigates the 3-D grid by selecting a discrete action at each time step, which corresponds to moving to an adjacent grid cell or hovering, as shown in Fig. 2(b). Discretizing the action space is an effective strategy for managing state–action space complexity and increasing learning stability, thereby making the complex UAV control problem tractable [33], [34], [35], [36]. Meanwhile, indoor users are randomly distributed across all floors of the building, with no guarantee of uniform density. Given these irregular and unpredictable user distributions, UAV-BSs continuously adjust their trajectories to optimize uplink transmission reliability and maximize communication efficiency, particularly under emergency scenarios where users transmit low-rate, survival-critical data in the absence of TN infrastructure.

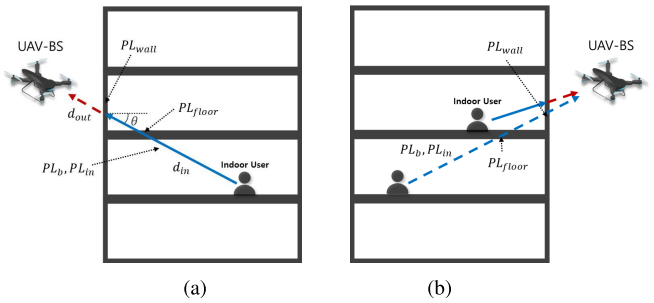


Fig. 3. Path loss modeling scenarios. (a) Outdoor-to-indoor UAV-BS link model with floor penetration loss. (b) Outdoor-to-indoor interference signal.

The considered multi-UAV-BS network operates in a time-slotted manner, where the total service time T is divided into multiple discrete time slots of duration δ_t. This time-slotted structure enables the UAV-BSs to update their positions at regular intervals while maintaining real-time adaptability to indoor user distributions and network conditions. Accordingly, the UAV-BS movement is defined by the following equation:

$$l_m(t+1) - l_m(t) = V\delta_t \tag{1}$$

where $l_m(t)$ represents the position of UAV-BS m at time slot t, and V denotes the velocity of the UAV-BSs.

A. Channel Model

The conventional air-to-ground path loss model [37] is not suitable for indoor environments, as it does not account for user altitude and additional signal losses caused by structural obstructions. Consequently, for indoor scenarios, the International Telecommunication Union-Radiocommunication Sector (ITU-R) outdoor-to-indoor path loss model [9] is commonly used. However, this model primarily assumed window-based signal transmission, making it unsuitable for factory-type buildings that lack windows and exhibit severe floor penetration losses. To address this limitation, our previous study [18] adopted an outdoor-to-indoor path loss model that explicitly incorporates floor penetration loss, as shown in Fig. 3(a). The outdoor-to-indoor path loss under this model, PL_{OI}, is formulated as follows:

$$PL_{\text{OI}} = PL_{\text{b}} + PL_{\text{wall}} + PL_{\text{in}} + PL_{\text{floor}} \tag{2}$$

$$PL_b = 20\log_{10}\left(f_{\text{GHz}}\left(d_{\text{out}} + d_{\text{in}}\right)\right) + 32.4 \tag{3}$$

$$PL_{\text{wall}} = g_1 + g_2\left(1 - \cos\theta\right)^2 \tag{4}$$

$$PL_{\text{floor}} = n_{\text{floor}}\left(g_1 + g_2\left(1 - \sin\theta\right)^2\right) \tag{5}$$

$$PL_{\text{in}} = g_3 d_{\text{in}} \tag{6}$$

where PL_{b} is the free-space path loss, PL_{wall} is the building wall penetration loss, PL_{in} is the indoor loss, PL_{floor} is the building floor penetration loss, g_1 and g_2 are building coefficients determined by wall materials, and g_3 is an in-building constant. To compute the outdoor (d_{out}) and indoor (d_{in}) distances, we first determine the intersection point P_{int}, $P_{\text{int}} = (x_{\text{int}}, y_{\text{int}}, z_{\text{int}})$, where the indoor user uplink signal meets the building wall. The calculation of P_{int} depends on the direction in which the UAV-BS is positioned, as shown in Fig. 2. Depending on whether the x- or y-axis is fixed to the building wall, the intersection coordinates are calculated as follows:

$$(\text{Case 1: } x_{\text{int}} = x_b) \begin{cases} x_{\text{int}} = x_b \\ y_{\text{int}} = y_m + \dfrac{x_b - x_m}{x_n - x_m}(y_n - y_m) \\ z_{\text{int}} = z_m + \dfrac{x_b - x_m}{x_n - x_m}(z_n - z_m) \end{cases}$$

$$(\text{Case 2: } y_{\text{int}} = y_b) \begin{cases} y_{\text{int}} = y_b \\ x_{\text{int}} = x_m + \dfrac{y_b - y_m}{y_n - y_m}(x_n - x_m) \\ z_{\text{int}} = z_m + \dfrac{y_b - y_m}{y_n - y_m}(z_n - z_m). \end{cases} \tag{7}$$

Once the P_{int} is determined, the d_{out} from the UAV-BS m to P_{int} and the d_{in} from P_{int} to the indoor user n are computed as follows:

$$d_{\text{out}} = \sqrt{(x_{\text{int}} - x_m)^2 + (y_{\text{int}} - y_m)^2 + (z_{\text{int}} - z_m)^2} \tag{8}$$

$$d_{\text{in}} = \sqrt{(x_n - x_{\text{int}})^2 + (y_n - y_{\text{int}})^2 + (z_n - z_{\text{int}})^2}. \tag{9}$$

The angle of incidence θ between the UAV-BS m and the indoor user n, which represents the angle at which the signal hits the building wall, is given as follows:

$$\theta = \begin{cases} \arcsin\left(\dfrac{|z_m - z_n|}{d_{\text{out}} + d_{\text{in}}}\right), & \text{if } z_m \neq z_n \\ 0, & \text{if } z_m = z_n. \end{cases} \tag{10}$$

The number of floors, n_{floor}, between the UAV-BS and the indoor user is calculated based on the floor height h_f as follows:

$$n_{\text{floor}} = \left| \left\lfloor \frac{z_n}{h_f} \right\rfloor - \left\lfloor \frac{z_{\text{int}}}{h_f} \right\rfloor \right| \tag{11}$$

where $\lfloor \cdot \rfloor$ represents the floor function, which ensures that the result is an integer corresponding to the number of discrete floors between the UAV-BS m and the indoor user n. This formulation provides the exact number of floors the signal must penetrate.

As illustrated in Fig. 3(b), when an indoor user transmits an uplink signal to a specific UAV-BS, this signal can act as interference to other UAV-BSs. The interference power is computed under the same outdoor-to-indoor propagation conditions as the desired signal, incorporating both floor penetration loss and distance-dependent attenuation. This modeling consistency ensures that both desired and interfering signals are evaluated under realistic indoor communication scenarios.

B. Empirical Validation of the Channel Model

To empirically validate the proposed outdoor-to-indoor path loss model, we conduct a measurement campaign in a multifloor building environment. The primary purpose of this measurement campaign is not to model dynamic effects such as small-scale fading caused by the fine-grained movements of the UAV (e.g., hovering fluctuations), but rather to empirically validate the floor penetration loss, which is the most dominant variable in our multifloor indoor environment. For this reason, the experiment is conducted using fixed nodes to isolate and measure the static path loss component. To this end, the

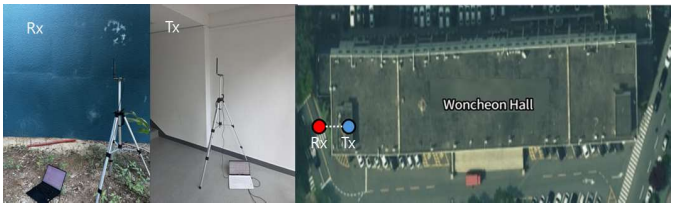


Fig. 4. TX and RX location for channel measurements.

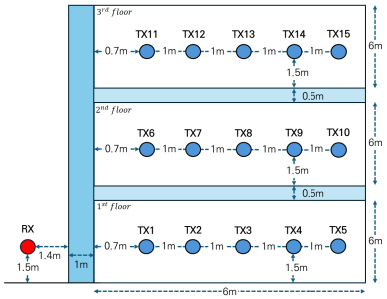


Fig. 5. Layout of TX and RX points.

TABLE II
SPECIFICATION OF THE TRANSCEIVER MODULE FOR CHANNEL MEASUREMENT

Parameter	Specification
Transceiver Model	EFR32FG25 (Silicon Labs)
Transmission Power	16 dBm
Frequency Range	917.1 MHz – 923.3 MHz
Number of Channels	32
Channel Spacing	200 kHz
Modulation Scheme	OFDM Option 4 – MCS 4
Antenna Type	Omnidirectional, 0 dBi gain

experiment utilizes EFR32FG25 Wi-SUN transceivers, with specifications as detailed in Table II. The measurements are performed at Wooncheon Hall on the Ajou University campus, specifically on a windowless facade of the building, as depicted in Fig. 4. As shown in Fig. 5, the receiver (RX) is positioned at a fixed outdoor location 1.4 m from the building surface, while the transmitter (TX) is placed at various indoor locations across the first, second, and third floors, resulting in TX-RX horizontal distances ranging from 3.1 to 7.1 m.

The experimental results are presented in Fig. 6, which compares the measured RSSI values against the calculated values from both our proposed floor-aware model and the standard model. For the cofloor scenario (TX: first, RX: first floor), where floor penetration is not a factor, both models show excellent fidelity, with an average error of only 1.3% compared to the measured data. However, a significant discrepancy emerges when cross-floor attenuation is introduced. For the second-floor transmission, the standard model deviates from the empirical data by an average of 26.4%, whereas our proposed model remains highly accurate with an average error of just 1.2%. This trend is even more pronounced for the third-floor transmission, where the predictions of the standard model have a substantial average error of 33.3%, while our model maintains its accuracy with a 2.1% error. Furthermore, despite some instability in the measurements (±5 dB), our

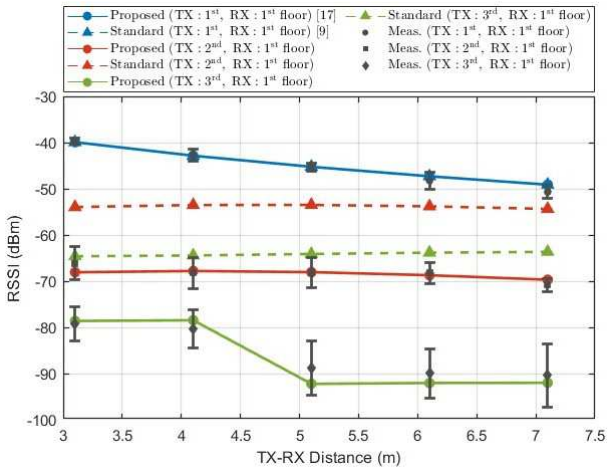


Fig. 6. Comparison of measured RSSI with the proposed and standard models.

proposed model successfully captures the nonlinear trend in the third-floor data.

These results quantitatively and qualitatively validate the superior accuracy and necessity of the proposed channel model for realistic indoor scenarios.

C. Association and Interference-Aware Signal Modeling

In this article, we assume a communication model where each UAV-BS operates on the same channel, leading to uplink interference for indoor users at time t due to other indoor users transmitting to a UAV-BS at the same time. We define an association indicator function $b_m^n(t)$ as follows:

$$b_m^n(t) = \begin{cases} 1, & \text{if } n \text{ is associated with } m \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

At each time t, we assume that all UAV-BSs associate with an indoor user if there exists at least one connectable indoor user. However, if any UAV-BS has no connectable indoor user, all UAV-BSs will move without user association. The UAV-BS and indoor user association at time t, Φ_t, is denoted as follows:

$$\Phi_t = \bigcup_{m=1}^{M} \left\{ n \mid b_m^n(t) = 1 \right\}. \tag{13}$$

Whether a UAV-BS can associate with an indoor user is determined by the SINR, which depends on the set of indoor users associated with each UAV-BS.

In this study, we assume that all indoor users transmit with the same transmission power as follows:

$$P_n = P_I \quad \forall n \in 1, 2, \ldots, N \tag{14}$$

where P_n represents the transmission power of indoor user n, and P_I is the fixed uplink transmission power for all indoor users.

The received power at the UAV-BS can be calculated based on the uplink transmission power of the indoor user as follows:

$$P_{r,nm} = P_I \cdot 10^{-\frac{PL_{nm}(t)}{10}} \tag{15}$$

TABLE III
UPLINK SINR-TO-MCS LOOKUP TABLE [38]

SINR Range (dB)	MCS Index	Modulation	Code Rate
(∼, -6.4]	N/A	N/A	N/A
(-6.4, -4.4]	0	QPSK	120/1024
(-4.4, -2.4]	1	QPSK	193/1024
(-2.4, -0.1]	2	QPSK	308/1024
(-0.1, 1.6]	3	QPSK	449/1024
(1.6, 4.2]	4	QPSK	602/1024
(4.2, 5.3]	5	16QAM	378/1024
(5.3, 6.4]	6	16QAM	434/1024
(6.4, 7.2]	7	16QAM	490/1024
(7.2, 8.1]	8	16QAM	553/1024
(8.1, 8.7]	9	16QAM	616/1024
(8.7, 10]	10	16QAM	658/1024
(10, 11.1]	11	64QAM	466/1024
(11.1, 12]	12	64QAM	517/1024
(12, 13]	13	64QAM	567/1024
(13, 13.7]	14	64QAM	616/1024
(13.7, 14.7]	15	64QAM	666/1024
(14.7, 15.5]	16	64QAM	719/1024
(15.5, 16.5]	17	64QAM	772/1024
(16.5, 17.6]	18	64QAM	822/1024
(17.6, 19.7]	19	64QAM	873/1024

where $P^n_{r,m}$ represents the received power at UAV-BS m from indoor user n, and $PL_{nm}(t)$ denotes an outdoor-to-indoor path loss between indoor user n and UAV-BS m at time t.

However, this uplink signal can act as interference to other UAV-BSs. Therefore, the interference signal received by another indoor user can be calculated as follows:

$$I_{jm} = P_I \cdot 10^{-\frac{PL_{jm}(t)}{10}} \quad \forall i \in \mathbf{N}, \quad m \in \mathbf{M}; i \neq j \tag{16}$$

where I_{jm} represents the interference power received by UAV-BS m due to the uplink transmission from the indoor user j, and $PL_{jm}(t)$ denotes an indoor-to-indoor path loss between indoor users j and UAV-BS m at time t.

The SINR is determined by the set of indoor users associated with each UAV-BS and is calculated as follows:

$$\gamma^n_m(t) = \frac{b^n_m(t) P_{r,nm}}{\sum_{k\in\mathbf{M},k\neq m} \sum_{j\in\mathbf{N},j\neq n} b^j_k(t) I_{jm}(t) + \sigma^2} \tag{17}$$

where $\sum_{k=1,k\neq m}^{M} \sum_{j=1,j\neq n}^{N} b^j_k(t) I_{jm}(t)$ represents the total interference power received by UAV-BS m from other indoor users communicating with different UAV-BSs. This equation accounts for all interference signals received by UAV-BS m from other indoor users communicating with different UAV-BSs, ensuring that the SINR calculation properly reflects the impact of multiuser interference in an indoor wireless environment.

In this study, we assume that all UAV-BSs can associate with an indoor user if the SINR for all UAV-BS and indoor user associations exceeds a predefined threshold at each time t. This SINR threshold is determined based on the MCS index, which plays a crucial role in adapting transmission parameters to varying channel conditions [39]. Therefore, the SINR value, computed based on (18), is used to determine the corresponding MCS index from Table III, which subsequently defines the achievable data rate.

Our physical layer model enables adaptive communication by considering the trade-off between reliability and efficiency. A lower MCS index, such as QPSK, is more resilient to noise and interference, ensuring reliable connections in poor channel conditions at the cost of lower data rates [40]. Conversely, a higher MCS index, such as 64-QAM, can be utilized to achieve greater data rates when channel conditions are favorable. To concretely model this trade-off, our simulation does not consider all possible MCS indices, but instead focuses on three representative scenarios: index 1 (QPSK), index 5 (16-QAM), and index 11 (64-QAM). Table III serves as a reference, providing the standard-based SINR thresholds for these specific indices to ensure that our simulation parameters are well-justified. By incorporating these MCS-based SINR thresholds into the association process, our model ensures that only users with sufficiently strong wireless links are connected, preventing link failures and enabling UAV-BSs to establish stable and efficient connections.

D. Energy Consumption Model for UAV-BS Operations

Efficient energy management is essential for UAV-BSs, which are energy-constrained and rely on battery-powered electric propulsion [42]. Unlike conventional aircraft, UAV-BSs' energy consumption is driven by aerodynamic power calculations. To ensure reliable communication, each UAV-BS must monitor its energy levels to prevent depletion, as a fully discharged UAV-BS cannot serve users. While energy is consumed for both aviation and communication, this study focuses on aviation-related energy consumption, as it is the dominant factor [43].

1) *UAV-BS Hovering Energy Consumption:* The hovering energy consumption of UAV-BS m can be expressed as follows [44]:

$$e^h_m(t) = \underbrace{\frac{\delta}{8}\rho s A \Omega^3 R^3}_{\text{blade profile}} + \underbrace{(1+k)\frac{W^{3/2}}{\sqrt{2\rho A}}}_{\text{induced}} \tag{18}$$

where $e^h_m(t)$ denotes the total hovering energy consumption of UAV-BS m, which is given by the sum of two components: the blade profile and the induced power. The blade profile represents the power required to turn the rotors' blades, where δ, ρ, s, A, Ω, and R are the profile drag coefficient, air density, rotor solidity, rotor disk area, blade angular velocity, and rotor radius, respectively. The induced power is the energy required to overcome the induced drag during lift generation, where k is an incremental correction factor, and W represents the aircraft weight, which includes the battery and propellers.

2) *UAV-BS Propulsion Energy Consumption:* The propulsion energy consumption of UAV-BS m can be expressed as follows [45]:

$$e^p_m(V(t)) = \underbrace{\frac{\delta}{8}\rho s A \Omega^3 R^3 \left(1 + \frac{3V(t)^2}{V^2_{\text{tip}}}\right)}_{\text{blade profile}}$$

$$+\underbrace{(1+k)\frac{W^{3/2}}{\sqrt{2\rho A}}\left(\sqrt{1+\frac{V(t)^4}{4V_i^4}}-\frac{V(t)^2}{2V_i^2}\right)^{1/2}}_{\text{induced}}+\underbrace{\frac{1}{2}d_0\rho sAV(t)^3}_{\text{parasite}} \tag{19}$$

where $e_m^p(t)$ denotes the total propulsion energy consumption of UAV-BS m, which is the sum of three components: the blade profile, the induced power, and the parasite power. The tip speed V_{tip}, the mean rotor induced velocity V_i, and the fuselage drag ratio d_0 are key parameters that contribute to the overall energy consumption. With the flying speed $V(t)$, the power consumption function is convex, meaning it increases with respect to the blade profile and parasite power, while decreasing with the induced power.

As a result, based on (19) and (20), the remaining energy of the UAV-BS at time t can be expressed as follows:

$$E_m(t+1)=\begin{cases}\max\left\{0,E_m(t)-e_m^h(t)\right\}, & \text{if hovering}\\ \max\left\{0,E_m(t)-e_m^p(V(t))\right\}, & \text{if moving}\end{cases} \tag{20}$$

where $E_m(t+1)$ represents the remaining energy of UAV-BS m at time $t+1$. It depends on whether the UAV-BS is hovering or moving at time t. In this article, we utilize the specifications of a real UAV-BS, as presented in Table IV.

E. Problem Formulation

In this article, we aim to minimize $S_m^n(t)$, the time required for UAV-BS m to establish uplink communication with indoor user n, by using the defined association method between UAV-BSs and indoor users, along with the SINR formula. $S_m^n(t)$ is composed of two parts: UAV-BS movement time and indoor user uplink transmission time. If a UAV-BS does not have an indoor user that satisfies the SINR threshold, the UAV-BS will move, and this is defined as UAV-BS movement time. The indoor user uplink transmission time refers to the time required for all UAV-BSs to complete the uplink service after establishing associations with indoor users at a specific time t. Therefore, to minimize $S_m^n(t)$, the UAV-BS must be associated with indoor users who exceed the SINR threshold while minimizing movement, and simultaneously considering the limited battery capacity of the UAV-BS.

In this study, we aim to jointly optimize the variables $x_m(t)$, $y_m(t)$, $z_m(t)$, $b_m^n(t)$, and $E_m(t)$ at each time interval t, with the goal of minimizing the total uplink communication time with indoor users, as formulated below

$$\mathbf{P1:}\ \min_{\mathbf{x_m(t)},\mathbf{y_m(t)}\mathbf{z_m(t)},\mathbf{b_m^n(t)},\mathbf{E_m(t)}}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}S_m^n(t), \tag{21}$$

$$\text{s.t.}\ \ \mathbf{C1:}\ \ b_m^n(t)=\{0,1\}\quad\forall m\quad\forall n\quad\forall t \tag{22}$$

$$\mathbf{C2:}\ \ \sum_m b_m^n(t)=1\quad\forall m\quad\forall t \tag{23}$$

$$\mathbf{C3:}\ \ \sum_m\sum_n b_m^n(t)=M,\quad\left\{\gamma_m^n(t)\geq\Gamma\right\}\quad\forall m,n\in\mathbf{b}_t \tag{24}$$

TABLE IV
SPECIFICATION OF UAV-BS [41]

Parameter	Value
Profile drag coefficient, δ	0.012
Air density, ρ	1.225 kg/m^3
Rotor solidity, s	0.05
Rotor disc area, A	0.503 m^2
Blade angular velocity, Ω	300 radius/s
Rotor radius, R	0.4 m
Tip speed of the rotor blade, V_{tip}	120
Fuselage drag ratio, d_0	0.6
Mean rotor-induced velocity in hovering, V_i	4.03
Incremental correction factor to induced power, k	0.1
Flight speed, V	20 m/s
Average maximum flight time	30 min
Capacity of flight battery	5,870 mAh
Aircraft weight, including battery and propellers, W	1,375 g

$$\sum_m\sum_n b_m^n(t)=0,\quad\left\{\gamma_m^n(t)<\Gamma\right\}\quad\forall m,\quad n\in\mathbf{b}_t$$

$$\mathbf{C4:}\ \ E_m(t)>0,\quad\forall m\quad\forall t \tag{25}$$

$$\mathbf{C5:}\ \ \|x_m(t)\|\leq x_{\max},\quad\|y_m(t)\|\leq y_{\max}\quad\|z_m(t)\|\leq z_{\max}\quad\forall m\quad\forall t \tag{26}$$

$$\mathbf{C6:}\ \ \|l_m(t+1)-l_m(t)\|=V\delta_t\quad\forall m\quad\forall t. \tag{27}$$

In **P1**, the goal is to minimize the total indoor user service time $S_m^n(t)$, which tightly couples UAV positioning with real-time communication performance. This time-oriented formulation contrasts with conventional throughput-maximization or energy-minimization objectives and is more aligned with the latency-critical nature of indoor emergency communications. In this problem formulation, the association between UAV-BSs and indoor users is governed by several constraints. **C1** and **C2** ensure that $b_m^n(t)$ represents a Boolean variable, and that each indoor user is associated with exactly one UAV-BS. **C3** guarantees that the association can only occur if the SINR for all UAV-BS and indoor user associations at time t exceeds the specified SINR threshold; otherwise, the association will not be established. Unlike conventional models that rely on simplified outdoor or window-based indoor propagation assumptions, our problem formulation is grounded in a floor-aware channel model that explicitly incorporates vertical signal attenuation through building floors. This makes the SINR constraint in **C3** significantly more dynamic and nonlinear, as even small UAV movements can lead to drastic SINR fluctuations, especially in high-rise or windowless industrial buildings. **C4** ensures that the remaining battery capacity of all UAV-BSs is always positive, preventing them from running out of energy during operation. **C5** guarantees that the UAV-BSs remain within their authorized operational regions, while **C6** limits their maximum movement distance.

The formulated problem **P1** is inherently nonconvex and high-dimensional, involving discrete association decisions, continuous movement variables, and dynamic SINR constraints. Moreover, the time-varying network state and interagent dependencies make it challenging to apply conventional convex optimization methods, which typically require full system observability, centralized coordination, and static

environments. To overcome these limitations, we adopt an MADRL approach. This framework enables each UAV-BS to independently learn an optimal policy using only local observations, while collectively achieving global objectives through decentralized cooperation.

IV. Proposed MADRL Framework

In this section, we propose an MADDQN approach integrated with PER to optimize UAV-BS indoor user service and to minimize the indoor user service time. To begin, we formulate the problem as a POMDP, where each UAV-BS operates independently and cannot access information from other UAV-BSs. POMDP enables decision-making without complete state information by utilizing a belief state. This allows the agent to select the optimal action based on incomplete observations.

A. POMDP Design

The POMDP of the proposed indoor user service networks with M UAV-BSs can be modeled as $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, \gamma >$. Similar to a fully observable MDP, $s \in \mathcal{S}$ represents the global true system state of the environment, while $a_m \in \mathcal{A}_m$ denotes the set of possible actions for UAV-BS m. The state transition probability function, $\mathcal{P}$, is defined as $P(s'|s, a)$, describing the probability of transitioning to a new state s' given the current state s and action a. Each UAV-BS agent receives a reward r_m, governed by the reward function $\mathcal{R}$, based on the action taken while only observing partial information $o_m \in \mathcal{O}_m$. The observation probability function, Ω, represents the likelihood of receiving an observation in a given state and is generated according to the probability distribution $o \sim \mathcal{O}(s)$, derived from the underlying system state. Finally, γ is the discount factor, which determines the relative importance of future rewards, where $0 \le \gamma \le 1$.

In Section IV-A, we formalize the states, observations, actions, and rewards for the proposed UAV-BS emergency indoor user service network.

1) *State and Observation:* At each time slot, the agent develops an optimal policy based on its observation of the environment. The global state $s(t)$ represents the complete system information and is defined as follows:

$$s(t) = \left(\bigcup_{m=1}^{M} \{l_m(t), \Phi_m(t), E_m(t)\}, \bigcup_{m=1}^{M} \bigcup_{n=1}^{N} \{D_m^n(t)\}, \bigcup_{\Phi(t)} \gamma_m^n(t) \right) \tag{28}$$

where $l_m(t)$ denotes the coordinates of UAV-BS m, $\Phi_m(t)$ denotes the indoor user associated with UAV-BS m, $E_m(t)$ indicates the remaining battery of UAV-BS m, and $D_m^n(t)$ represents the uplink signal-to-noise ratio (SNR) of indoor user n with UAV-BS m. The global state thereby provides comprehensive information on UAV-BS locations, user associations, and communication conditions required for calculating the SINR, $\gamma_m^n(t)$.

However, in our realistic POMDP formulation, each agent makes decisions based on local observations and does not have access to the global state. To model practical sensing constraints, we limit an agent's observation to a subset of candidate users, denoted as $\mathcal{N}_m(t)$. This set includes only users from whom agent m can receive a signal stronger than a predefined SNR threshold, which effectively constrains the observation to communicatively relevant users. Accordingly, the local observation $o_m(t)$ for agent m consists of its own state and the SNR values from this candidate set

$$o_m(t) = \left(l_m(t), \Phi_m(t), E_m(t), \bigcup_{n \in \mathcal{N}_m(t)} \{D_m^n(t)\} \right). \tag{29}$$

Thus, each UAV-BS selects its action based on its current location, $l_m(t)$, the users it has previously been associated with, $\Phi_m(t)$, its remaining battery level, $E_m(t)$, and the SNR values of the currently observable users, $\bigcup_{n \in \mathcal{N}_m(t)} \{D_m^n(t)\}$.

2) *Action:* In this study, rather than employing a single, flat action space, we introduce a state-dependent, dual-mode control policy to structure the agent's decision-making process. This approach decomposes the complex joint optimization problem into two distinct operational modes—positioning and association—allowing the agent to focus on the most relevant objective at each time step based on its current observation. When the agent's observation indicates that no high-quality user connection is currently possible, it operates in the positioning mode. Here, the action space is restricted to movement actions, and the agent's objective is to navigate to a more strategically advantageous location to improve future association opportunities. Conversely, if one or more suitable association candidates are available, the agent switches to the association mode. The action space is then restricted to user association actions, and the agent's task is to select the optimal user to serve from the set of available candidates.
Therefore, the set of actions available to a UAV-BS m at time t can be expressed as follows:

$$a_m(t) = \left(\pm V_x \delta_t, \pm V_y \delta_t, \pm V_z \delta_t, \bigcup_{n=1}^{N} b_m^n(t) \right) \tag{30}$$

where $V_x\delta_t, V_y\delta_t$, and $V_z\delta_t$ represent the movements in the x-, y-, and z-directions, respectively, and $b_m^n(t)$ represents the binary association variables. The agent's current mode, determined by the state, dictates which subset of these actions is available at any given time. Additionally, as shown in Fig. 2, each agent is positioned on one of the four sides of the building and moves in 3-D within its designated region.

3) *Reward:* To effectively train the state-dependent, dual-mode policy, a multicomponent reward structure is designed to provide targeted learning signals for both positioning and association while fostering multiagent cooperation. The total reward for each agent is a

weighted sum of individual, common, and penalty components. The principal mechanism guiding this dual-mode behavior is the individual reward, $r_m(t)$, which is structured to provide a distinct incentive for each operational mode

$$r_m(t) = \sum_{n=1}^{N} b_m^n(t) D_m^n(t) + \begin{cases} \max\limits_{n=1,\ldots,N} \left\{D_m^n(t)\right\}, & \text{if } \sum\limits_{n=1}^{N} b_m^n(t) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

when in association mode, the agent receives a reward proportional to the resulting link quality, given by $\sum_{n=1}^{N} b_m^n(t) D_m^n(t)$, which encourages the formation of high-quality connections. Conversely, when in positioning mode, the agent is guided by a potential-based shaping reward, $\max_{n=1,\ldots,N}\{D_m^n(t)\}$. This component encourages the agent to navigate toward unassociated users with the highest potential SNR, thereby improving its strategic position for future association attempts.

To ensure that locally optimal decisions align with the global system objective, we introduce a common reward, $r_c(t)$

$$r_c(t) = \sum_{m=1}^{M} \sum_{n=1}^{N} b_m^n(t). \tag{32}$$

This common reward is distributed to all agents only upon the formation of a valid joint association that satisfies the system-wide SINR threshold. This mechanism compels agents to learn cooperative policies that inherently avoid mutual interference.

Finally, the total reward $r_{\text{tot},m}(t)$ is formulated to integrate the individual and common rewards with several crucial penalties that guide the agent toward a practical and efficient policy. The weights ω_i are introduced as normalization factors to ensure that each component contributes meaningfully to the learning signal. The total reward is given by

$$r_{\text{tot},m}(t) = \omega_1 r_m(t) + \omega_2 r_c(t) - \omega_3 \zeta_m^1 - \omega_4 \zeta_m^2, \quad -\omega_5 p_m^e(t) \tag{33}$$

where ω_{1-5} are normalization weights that scale each term to a comparable magnitude. $p_m^e(t)$ is the energy consumption penalty, which is proportional to the energy consumed for the action taken at time t. ζ_m^1 is a penalty for moving outside the designated operational area. ζ_m^2 is the penalty for inefficient mode selection. This is applied when an agent either chooses a movement action when a valid user association is possible, or attempts an association action when it should be repositioning. This penalty encourages the agent to learn the correct state-dependent switching for its dual-mode policy.

B. Proposed MADDQN Algorithm

In this study, we propose a PER-based MADDQN algorithm to optimize the UAV-BS association and movement strategy for indoor user service networks. The proposed algorithm builds upon traditional DQN and MADQN frameworks, introducing PER for more efficient learning and DDQN to mitigate overestimation bias.

The proposed MADDQN algorithm is designed to train each UAV-BS agent to learn an optimal policy π^*, enabling it to make user association and movement decisions based on observations in a partially observable environment. This approach is particularly suited for indoor UAV-BS networks, where each agent has only local observations and must make autonomous decisions while coordinating with other UAV-BSs. As illustrated in Fig. 7, our work adopts the CTDE paradigm, implemented via the MADDQN with PER framework, to address several inherent challenges of MARL in POMDP environments. In this paradigm, agents are trained in a centralized manner that has access to the joint observations and actions of all agents. The key advantage of the CTDE structure is its ability to directly mitigate the nonstationarity issue inherent in multiagent learning. This is achieved by providing a stable training environment where each agent's update is conditioned on the actions of all others. This centralized process facilitates cooperative learning and stable value estimation.

It is important to clarify the role of communication during the execution phase. We assume that a standard, low-level communication link is available for essential flight operations such as localization and basic collision avoidance. The key advantage of our CTDE approach, however, is that it obviates the need for any additional communication overhead for the high-level decision-making task of cooperative user service. Each agent operates in a fully decentralized fashion based solely on its local observations. This design offers a significant practical advantage by allowing the limited wireless bandwidth to be fully dedicated to the primary mission: transmitting data to users in need.

To achieve efficient and robust training, the algorithm integrates several key components.

1) Proposed Double Deep Q-Networks: In the UAV-BS indoor user association and movement optimization problem, the agent must learn an optimal policy to dynamically adapt to a partially observable environment. Traditional Q-learning estimates the action-value function using the Bellman equation, where the target Q-value is computed as follows:

$$y_m^t = r_m^t + \gamma \max_{a_{t+1}} Q\left(o_{t+1}, a_{t+1}; \theta_m\right). \tag{34}$$

However, this approach suffers from overestimation bias because the same network is used to both select the best action and estimate its value. Over time, this can lead to unstable learning, causing the UAV-BSs to make suboptimal movement or association decisions, ultimately degrading network performance.

To mitigate this issue, DDQN introduces a decoupled action selection and evaluation mechanism by incorporating a target network. The updated target Q-value in DDQN is computed as follows:

$$y_m^t = r_m^t + \gamma Q\left(o_{t+1}, \arg\max_{a_{t+1}} Q\left(o_{t+1}, a_{t+1}; \theta_m\right), m; \hat{\theta}_m\right). \tag{35}$$

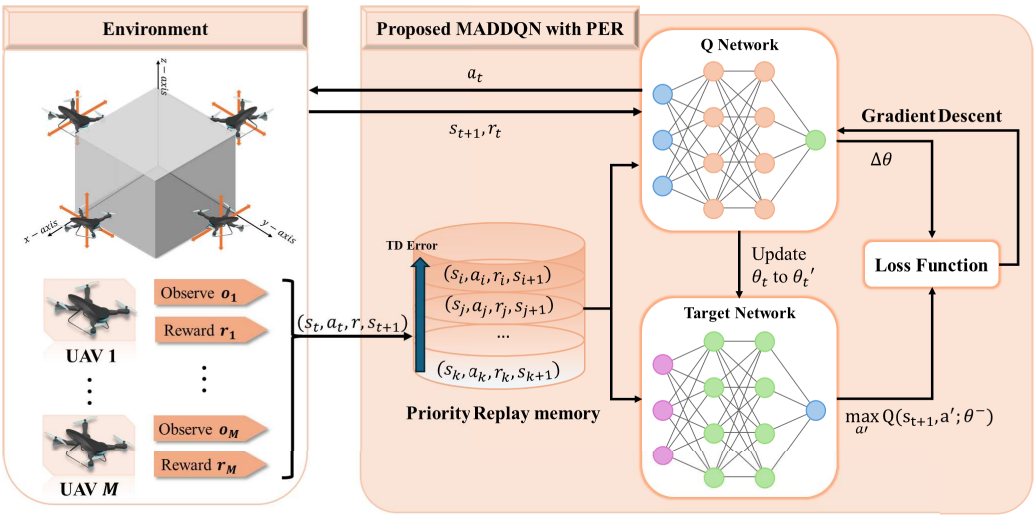


Fig. 7. Overview of the proposed MADDQN algorithm with PER framework.

In this formulation, the main Q-network is responsible for selecting the best action

$$a^* = \arg\max_{a_{t+1}} Q\left(o_{t+1}, a_{t+1}; \theta_m\right) \tag{36}$$

while the target network is used to evaluate the value of the selected action. This separation prevents overestimation bias, resulting in more accurate Q-value estimations and enhancing the stability of learning.

To update the Q-network, the mean squared error (MSE) loss function is used to minimize the difference between the predicted Q-value and the target Q-value

$$L_m^t = \left(y_m^t - Q\left(o_t, a_t, m; \theta_m\right)\right)^2 \tag{37}$$

where y_m^t is the target Q-value obtained from the target network, $Q(o_t, a_t, m; \theta_m)$ is the estimated Q- value from the current Q-network, and the objective is to adjust θ_m such that the temporal difference (TD) error (difference between predicted and target Q-value) is minimized. To update the network parameters, gradient descent is applied

$$\theta_m \leftarrow \theta_m - \eta \nabla_{\theta_m} L_m^t \tag{38}$$

where η is the learning rate.

By minimizing this loss function, the Q-network gradually improves its policy, leading to better user association and UAV-BS movement optimization.

To maintain stable Q-value updates, a target network is used. Instead of instantly updating the target network, a soft update mechanism is applied

$$\hat{\theta}_m \leftarrow \tau\theta_m + (1-\tau)\,\hat{\theta}_m \tag{39}$$

where $\tau \ll 1$ (e.g., 0.01) controls the update rate of the target network. The soft update mechanism prevents sudden Q-value fluctuations, allowing UAV-BSs to stabilize their learning process, which is critical for ensuring smooth policy improvements over time.

2) Prioritized Experience Replay: One of the key challenges in reinforcement learning, particularly in UAV-BS indoor communication environments, is the sparse reward problem [46], [47]. In such environments, rewards are often provided only when significant events occur, making it difficult for the agent to learn effective policies. Traditional experience replay in DQN stores past transitions in a buffer and samples experiences uniformly to break temporal correlations between consecutive states. However, uniform sampling treats all experiences equally, which can slow down learning, as rare but critical experiences (e.g., high TD error transitions) may not be sampled frequently enough. PER addresses this issue by assigning a priority to each experience, ensuring that more informative transitions are sampled more frequently. PER is introduced to address this issue by ensuring that high-impact experiences are sampled more frequently, thus improving learning efficiency and convergence speed. Each UAV-BS agent maintains a replay buffer, where experiences (o, a, r, o') are stored along with their TD error, which determines their priority as follows:

$$\delta_m^t = \left|r_m^t + \gamma \max_{a_{t+1}} Q\left(o_{t+1}, a_{t+1}, m; \hat{\theta}_m\right) - Q\left(o_t, a_t, m; \theta_m\right)\right| \tag{40}$$

where δ_m^t is the TD error of UAV-BS at t.

The priority score for a given experience i is defined as follows:

$$P(i) = \frac{|\delta_i|^\alpha}{\sum_j |\delta_j|^\alpha} \tag{41}$$

where δ_i represents the TD error of experience i, indicating how surprising or important the transition is, α controls the degree of prioritization, with $\alpha = 0$ reducing PER to uniform sampling, and $\sum_j |\delta_j|^\alpha$ normalizes the priorities across all stored experiences by summing over all experiences j in the replay buffer.

To ensure stability and prevent overfitting to high-priority samples, importance sampling weights are introduced

$$w(i) = \left(\frac{1}{N} \cdot \frac{1}{P(i)}\right)^{\beta} \tag{42}$$

where N is the total number of experiences in the replay buffer, and β controls the correction of sampling bias (β is annealed from a small value to 1 over time). These importance weights adjust the learning process to counteract bias introduced by prioritized sampling, ensuring that updates remain balanced.

C. Training Workflow of the Proposed MADDQN Algorithm With PER

As shown in Algorithm 1, the training process of UAV-BS agents using the proposed MADDQN algorithm with PER begins with initializing the environment, where UAV-BSs and indoor users are randomly placed within the designated area. UAV-BSs are deployed near the north, south, east, and west sides of the building, while indoor users are randomly distributed across different floors. At this stage, a PER buffer is also initialized to efficiently store and manage past experiences.

During each time step t, UAV-BSs select their actions $a_m(t)$ based on an ϵ-greedy policy, balancing exploration and exploitation. The probability of selecting an exploratory action decays over time, allowing UAV-BSs to gradually transition from exploring new actions to exploiting learned policies. After executing their selected actions, each UAV-BS records its local observation transition (o_m, a_m, r_m, o'_m). These individual experiences are then aggregated to construct the global transition $(\mathbf{o}, \mathbf{a}, \mathbf{r}, \mathbf{o}')$, which is stored in the PER buffer. In this process, priority is assigned based on the TD error, ensuring that more informative experiences are sampled more frequently, thereby improving learning efficiency. Once a sufficient number of experiences are stored, a mini-batch of transitions is sampled from the PER buffer based on priority scores, and the importance sampling weight w_i is computed to correct for sampling bias. The target Q-value is then computed using DDQN to mitigate overestimation bias, and the loss function is updated using gradient descent with importance weighting. After updating the Q-network, the TD error is recomputed, and the priority values in the PER buffer are updated accordingly to reflect new learning progress. Finally, the target network is updated using a soft update mechanism to ensure training stability, while the exploration rate ϵ decays over time, shifting UAV-BSs from exploration to exploitation. This training cycle is repeated iteratively until the predefined number of episodes is completed, allowing UAV-BSs to continuously refine their policies. Throughout the training process, UAV-BSs learn to efficiently associate with indoor users and determine movement strategies that minimize service time while maintaining energy efficiency. Through the integration of PER for efficient sampling, DDQN for stable Q-value estimation, and target networks for controlled updates, the proposed algorithm ensures robust learning and optimal UAV-BS coordination in dynamic indoor environments.

Algorithm 1 PER-MADDQN Algorithm for UAV-BS Control

Initialization: Initialize action-value function Q with random weights θ, target action-value function $\hat{Q}$ with weights $\theta^- = \theta$ for each UAV-BS agent. Initialize prioritized experience replay buffer (B) with capacity C, minibatch size F, total episodes E.
Deploy each UAV-BS agent randomly near the north, south, east, and west sides of the building.
Deploy indoor users randomly across each floor of the building.
Set initial exploration rate $\epsilon = \epsilon_{\text{init}}$.
for *episode = 1* ***to*** E **do**
  Initialize the UAV-BS environment.
  for $t = 1$ ***to*** T **do**
    for $m = 1 : M$ **do**
      ▷ UAV-BS m selects an action $a_m(t)$ using an ϵ-greedy policy:

$$a_m(t) = \begin{cases} \arg\max_{\pi} Q(o_m(t), a, m; \theta_m), & 1-\epsilon, \\ \text{random action}, & \epsilon. \end{cases}$$

    end
    ▷ Observe next state o_{t+1} and receive reward r_t.
    ▷ Compute TD error before storing experience by (40):
    ▷ Compute priority score by (41):
    ▷ Store experience tuple $(\mathbf{o}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{o}_{t+1})$ into prioritized replay buffer B with priority P_t.
  end
  if B *has sufficient samples for training* **then**
    ▷ Sample a minibatch F from the PER buffer B based on priority scores.
    for $m = 1 : M$ **do**
      for $i = 1$ ***to*** F **do**
        ▷ Compute TD target:

$$y_i = r_i + \gamma Q(o'_i, \arg\max_{a'} Q(o'_i, a', m; \theta_m), m; \hat{\theta}_m)$$

        ▷ Compute importance sampling weight by (42):
        ▷ Perform gradient descent step to minimize the loss:

$$L_m^t = w_i \cdot (y_i - Q(o_i, a_i, m; \theta_m))^2$$

      end
      for i ***in*** B **do**
        ▷ Update priority scores in B by (41):
      end
      ▷ Soft update the target network:

$$\hat{\theta}_m \leftarrow \tau\theta_m + (1-\tau)\hat{\theta}_m$$

    end
  end
  ▷ Decay exploration rate ϵ:

$$\epsilon \leftarrow \max(\epsilon_{\min}, \epsilon \cdot \epsilon_{\text{decay}})$$

end

V. Performance Evaluation

A. Simulation Environments

For the performance evaluation of the proposed approach, we consider the simulation settings summarized in Table V. The UAV-BS model is based on the specifications of the DJI

TABLE V
SIMULATION PARAMETERS

Parameters	Value
Number of UAV-BSs, M	4
Number of Indoor users, N	$20 \sim 40$
Carrier frequency, f_c	2 GHz
Bandwidth	20 MHz
Indoor User Tx power, P_I	20 dBm
Noise power, σ^2	-120 dBm
Building size, x_b, y_b, z_b	40m, 40m, 50m
Building floor,	10
Building, UAV-BS distance, h,	10m
Building coefficient, g_1, g_2, g_3	14, 15, 0.5
Learning rate	0.0001
Discount Factor, γ	0.99
Buffer size and capacity, B,C	1,000, 50,000
Episodes, E	4,000
Mini-batch size, F	256
Time step, T	60
Importance of priority, α	0.6
Strength of importance sampling, β	0.4

Phantom 4 Pro v2.0 [42]. UAV-BSs are randomly deployed 10 m away from each of the four sides of the building, while indoor users are randomly distributed across all floors of the building. Each episode consists of 60 time steps, corresponding to a total duration of 30 min, with each time step representing 30 s.

The neural network in the proposed algorithm consists of six layers: an input layer, an output layer, and four hidden layers. Each of the four hidden layers comprises 1024 neurons, leveraging fully connected layers to capture complex feature representations. The rectified linear unit (ReLU) activation function is applied to each hidden layer to introduce nonlinearity while mitigating the vanishing gradient problem, ensuring efficient gradient propagation. To train the network, we employ the Adam optimizer, which adapts the learning rate dynamically for each parameter update, leading to faster convergence and improved stability compared to traditional stochastic gradient descent (SGD). The target network is updated periodically using a soft update method with a smoothing factor of $\tau = 0.01$, preventing drastic changes in policy updates and promoting stable convergence. Additionally, to enhance the exploration–exploitation trade-off, we implement an ϵ-greedy policy where the exploration rate ϵ is initially set to 0.7 and gradually decreases by 0.00035 per time step following $\epsilon = \max(\epsilon_{\min}, \epsilon - \epsilon_{\text{decay}})$.

In this article, the effectiveness of the proposed PER-based MADDQN is evaluated by comparing it with these existing algorithms, as detailed below.

1) *Multiagent Deep Q-Network [48]:* This method serves as the baseline reinforcement learning approach, where each UAV-BS trains its policy using a standard MADQN framework. It utilizes a single-network Q-value update mechanism and employs a uniform experience replay buffer, where samples are randomly selected without prioritization.
2) *Multiagent Independent Actor–Critic [49]:* MAIAC represents a decentralized policy-based multiagent reinforcement learning approach, where each UAV-BS independently learns its own actor and critic without parameter sharing or centralized training. As a representative algorithm of the policy gradient family, the actor is trained using policy gradients based on locally estimated advantages, while the critic learns to approximate the state-value function using temporal difference learning.
3) *Multiagent Deep Deterministic Policy Gradient [50]:* MADDPG is a model-free, off-policy actor–critic algorithm that extends DDPG to multiagent settings by learning deterministic policies through CTDE. As MADDPG was originally developed for continuous action spaces, we adapted it to our discrete problem using the Gumbel-Softmax relaxation technique, which enables differentiable sampling from categorical distributions. This discrete adaptation ensures a fair and consistent comparison, as our proposed method also operates within a multiagent, off-policy learning framework.
4) *Consensus-Based Bundle Algorithm [51]:* CBBA is a decentralized, auction-based optimization for solving multiagent task assignment problems. The algorithm iterates between a bundle construction phase, where each agent greedily builds a sequence of tasks to maximize its local score, and a conflict resolution phase. In this study, each agent competes to associate with the user that yields the highest SINR. If the resulting SINR with the selected user is lower than a predefined threshold, the agent is designed to relocate to the position where the achievable SINR is maximized. To facilitate this decision-making process, we assume that each agent has a comprehensive local observation, including the potential channel conditions for all users relative to itself.
5) *Ablation 1 (PER-MADDQN Without Dual-Action Space):* This algorithm serves to isolate the contribution of the dual-action mechanism. It uses MADDQN with PER but operates on a combined, larger action space.
6) *Ablation 2 (Vanilla MADDQN):* This is the foundational baseline, using a standard MADDQN framework without either PER or the dual-action structure. It employs a uniform experience replay buffer and a combined action space. This comparison allows us to quantify the combined performance improvement from both PER and the dual-action design.

To ensure a fair and rigorous evaluation, all DRL algorithms, including our proposed method and the baselines, are implemented with an identical optimizer and hyperparameter settings as described above. Furthermore, to maintain consistency across the decision-making process, MADQN, MADDPG, MAIAC, and CBBA algorithms are adapted to employ the same dual-action space and multicomponent reward structure. For the actor–critic-based baselines (MAIAC and MADDPG), the critic network is configured separately: it consists of four layers, with the hidden layers containing 512 neurons, and is trained with a learning rate of 0.001.

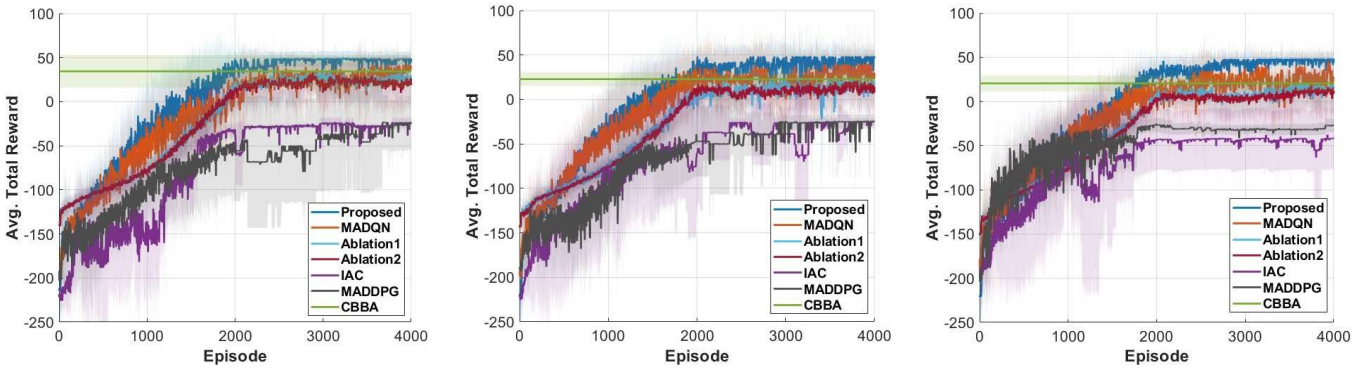


Fig. 8. Comparison average total reward of each method according to the SINR threshold. (a) SINR threshold = −4.4 dB (QPSK). (b) SINR threshold = 4.2 dB (16-QAM). (c) SINR threshold = 10 dB (64-QAM).

TABLE VI
CONVERGENCE AVERAGE TOTAL REWARD VALUE OF THE PROPOSED AND BASELINE ALGORITHMS

Algorithm	-4.4 dB	4.2 dB	10 dB
Proposed	**49.0**	**48.7**	**48.1**
MADQN	35.2	29.9	25.4
Ablation 1	24.8	15.3	14.8
Ablation 2	18.7	13.9	10.1
MAIAC	−25.3	−25.2	−41.1
MADDPG	−23.8	−24.9	−27.8
CBBA	34.5	22.8	20.3

B. Simulation Results

In this section, we present a comparative analysis of the simulation results between the proposed algorithm and the baseline approaches. The performance evaluation is conducted based on the following key metrics: 1) reward convergence; 2) user connectivity; 3) user service quality; 4) trained behaviors of agent UAV-BSs across episodes; and 5) computational cost.

1) Reward Convergence: In this section, we evaluate the average total reward convergence of the proposed algorithm against the baselines and two ablated versions of our method. Fig. 8 presents the average total reward per episode under three different SINR threshold levels for user scenarios ranging from 20 to 40, and Table VI outlines the final average reward achieved by each algorithm.

Fig. 8(a) illustrates the reward convergence under a relaxed SINR threshold of −4.4 dB (QPSK). The proposed algorithm clearly outperforms all other methods, converging to the highest reward of 49.0, while the MADQN achieves a reward of 35.2. This performance gap demonstrates the significant advantage of using a DDQN structure combined with PER, which effectively mitigates overestimation bias and focuses learning on critical experiences. Our ablation studies further dissect this performance gain. Ablation 1, which removes the dual-action structure, shows a dramatic performance drop to 24.8, confirming the critical role of the dual-action mechanism in making the complex action space tractable. Furthermore, Ablation 2, which also removes PER, drops to 18.7, clearly showing the additional, valuable contribution of PER in enhancing learning stability. This trend of superior performance is maintained and even amplified as the channel quality requirements become more stringent, as shown in Fig. 8(b) (4.2 dB, 16-QAM) and (c) (10 dB, 64-QAM). While the reward of the proposed algorithm remains consistently high (48.7 and 48.1, respectively), the performance of all baseline and ablated methods degrades significantly under these tougher conditions. This demonstrates the robustness of our approach, confirming its ability to find stable, high-quality policies even when successful associations are significantly more challenging to achieve.

In contrast, the other baselines exhibit clear limitations that are amplified as the SINR threshold increases. Under the relaxed −4.4 dB condition, MADDPG and MAIAC converge to significantly lower rewards, near −23.8 and −25.3, respectively. As the threshold increases to 4.2 dB, their performance remains poor at −24.9 (MADDPG) and −25.2 (MAIAC). Under the most stringent 10 dB threshold, the reward of MADDPG drops further to −27.8, while the performance of MAIAC deteriorates significantly to −41.1.

The CBBA, a nonlearning heuristic, performs respectably under the easiest condition with a reward of 34.5. However, because it lacks the long-term foresight of a learned policy, its performance progressively degrades to 22.8 (at 4.2 dB) and 20.3 (at 10 dB) as the greedy SINR-based choices become less effective in more challenging scenarios.

Overall, the robust and superior performance of the proposed algorithm is systematically validated by our ablation studies. These studies confirm that the dual-action mechanism is the most critical innovation, providing a substantial performance gain by making the high-dimensional action space tractable. Building on this, the DDQN structure and PER provide additional stability and efficiency, effectively addressing the overestimation bias and sparse reward challenges that limit the baseline MADQN. In contrast, the external baselines falter, revealing their limitations in a discrete and volatile environment. The challenges of applying actor–critic methods are particularly evident; MADDPG, originally designed for continuous control, struggles with the necessary adaptations for a discrete action space, while MAIAC suffers from the high variance inherent to policy gradient methods, which is exacerbated by the abrupt reward changes. The nonlearning CBBA lacks the necessary foresight for long-term

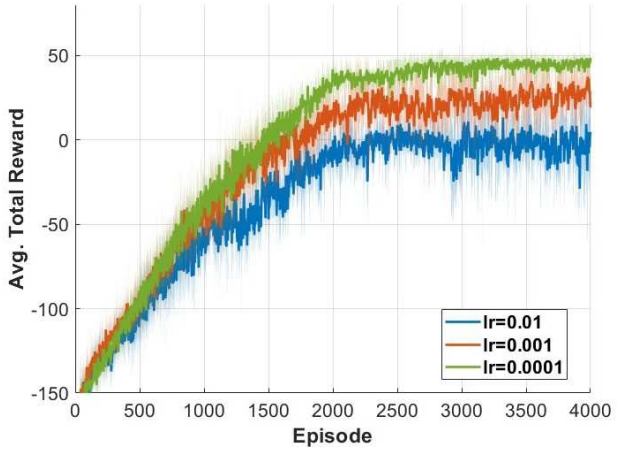


Fig. 9. Proposed algorithm average total reward with different learning rate.

optimization. Therefore, our proposed algorithm's consistent success across all scenarios confirms the synergistic benefits of its core components.

To investigate the sensitivity of the proposed MADDQN algorithm to the learning rate, we conducted training with three different values: 0.01, 0.001, and 0.0001. Fig. 9 shows the average total reward over 4000 episodes, where each curve represents the performance averaged across all SINR threshold conditions and user scenarios ranging from 20 to 40 users. This experiment aims to assess how the learning rate affects convergence speed, reward stability, and overall performance in a dynamic and discrete UAV-BS control environment.

The learning rate of 0.01 (blue curve) facilitates fast early-stage learning but introduces significant fluctuations in later phases, resulting in unstable convergence and lower final rewards. Reducing the learning rate to 0.001 (orange curve) yields a more stable trajectory, yet the final reward remains suboptimal. Notably, 0.0001 (green curve) demonstrates the most stable learning trajectory, converging steadily to the highest average reward with minimal variance across episodes. A high learning rate (e.g., 0.01 and 0.001) causes the optimizer to take excessively large steps, leading to an overshooting effect where the policy parameters repeatedly jump past the optimal region. This results in the high-variance oscillations and poor final performance seen in the blue and orange curves, as the agent fails to settle into a stable, high-reward policy.

While a smaller learning rate such as 0.0001 enhances stability and prevents abrupt parameter updates, it also presents a potential risk of overfitting, particularly in discrete and low-dimensional environments. Excessively slow updates may cause overspecialization to training scenarios, reducing adaptability to unseen conditions. Therefore, the choice of 0.0001 reflects a balanced trade-off among stability, final performance, and generalization capability. It showed consistent superiority across various SINR conditions and user densities, making it a practical and robust choice for the subsequent simulations under the tested conditions.

2) User Connectivity: Fig. 10 presents the performance results across different user densities and SINR thresholds. Specifically, panels (a) through (f) display the results for networks with 20, 24, 28, 32, 36, and 40 users, respectively. In each panel, the three graphs from left to right correspond to SINR thresholds of −4.4, 4.2, and 10 dB. Table VII complements these graphical results by summarizing the average connection percentage achieved by each algorithm across the three SINR thresholds.

TABLE VII
AVERAGE CONNECTION PERCENTAGE BY SINR THRESHOLD

Algorithm	-4.4 dB	4.2 dB	10 dB
Proposed	100.0%	100.0%	100.0%
MADQN	82.22%	77.77%	71.11%
Ablation 1	77.77%	71.11%	55.55%
Ablation 2	62.22%	57.77%	51.11%
MAIAC	20.00%	13.33%	13.33%
MADDPG	28.88%	22.22%	17.78%
CBBA	84.44%	66.67%	60.00%

As shown in Fig. 10(c), which corresponds to the scenario with 28 users, the proposed algorithm successfully connects with all users, achieving a 100% association rate across all SINR thresholds. In this less dense environment, MADQN and our two ablation studies, Ablation 1 and Ablation 2, also demonstrate strong performance by reaching full association in all cases, though it requires slightly more steps than the proposed method. This indicates that their core learning structures are effective for problems of moderate complexity. In contrast, the heuristic-based CBBA fails to achieve full connectivity, associating with 16 users (57%) at the lower thresholds and degrading to 12 users (42%) at 10 dB. The other learning-based methods, MADDPG and MAIAC, show clear limitations, consistently associating with only 8 (28%) and 4 (14%) users, respectively.

This performance gap becomes more pronounced under higher user densities, as shown in Fig. 10(f) for the 40-user case. The proposed algorithm is the only method that maintains a 100% success rate, demonstrating excellent scalability. The ablation studies starkly illustrate the breakdown of the incomplete models. The performance of Ablation 1 collapses, associating with only 16 users (40%) at −4.4 dB and dropping to a mere four users (10%) at 10 dB. This confirms that the dual-action structure is critical for managing the expanded action space in dense scenarios. Ablation 2 also shows poor scalability, four users (10%). The performance of the standard MADQN also drops sharply, associating with only 20 users (50%) at −4.4 dB and four users (10%) at the 10 dB threshold, as its standard Q-learning approach is ineffective under these sparse reward and high-complexity conditions. The scalability of CBBA, MADDPG, and MAIAC is similarly poor, with all failing to connect a majority of users as conditions become stricter. These results indicate a severe lack of generalization for most baseline and ablated methods under dense and challenging network conditions.

Finally, as summarized in Table VII, the average connection ratio across all user counts and SINR conditions further highlights the proposed algorithm's superior performance. It achieves 100.0% success consistently across all SINR thresholds. In contrast, all baseline and ablated methods failed to

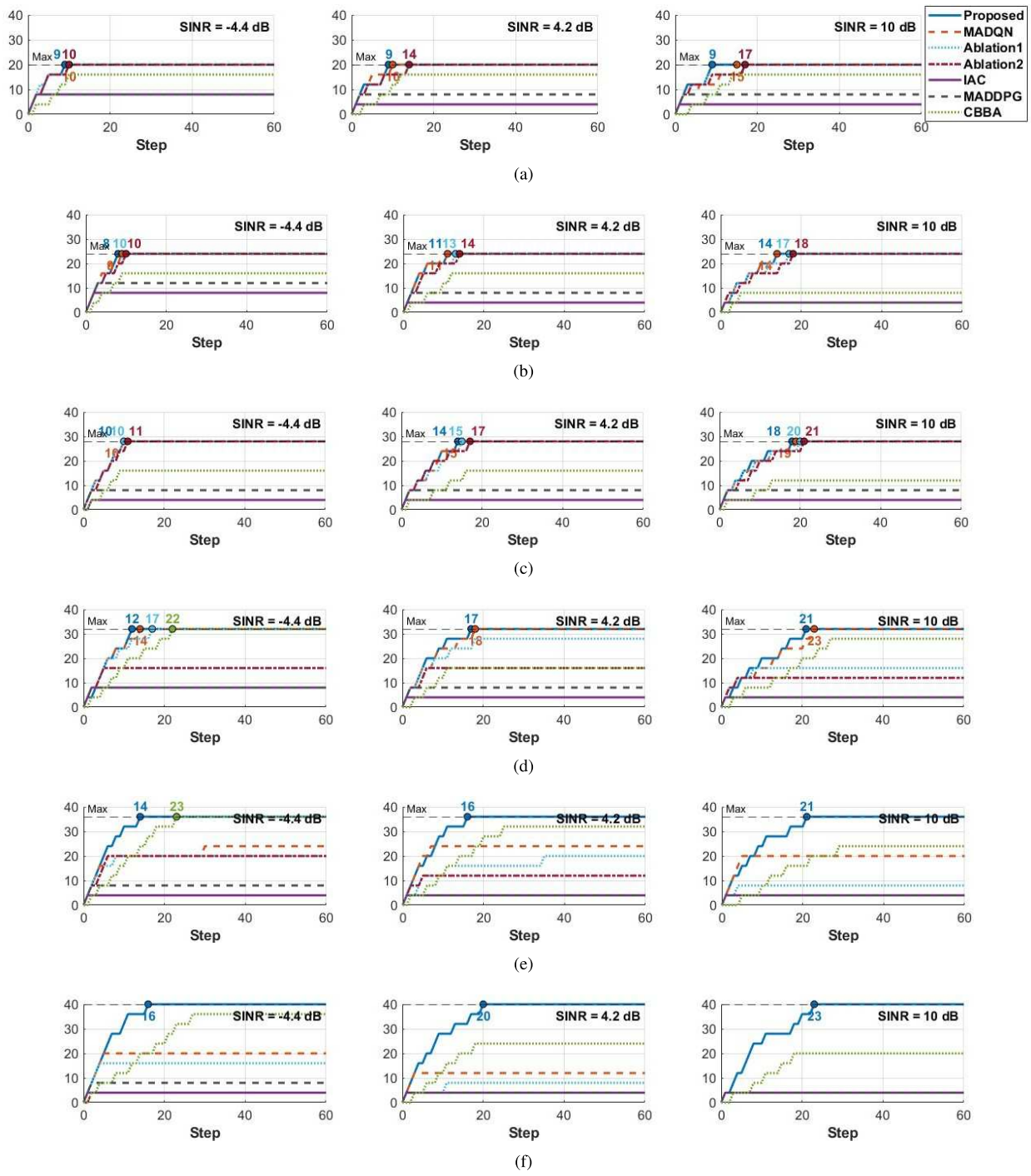
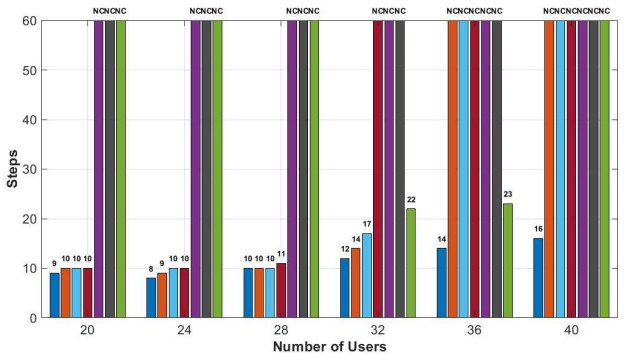


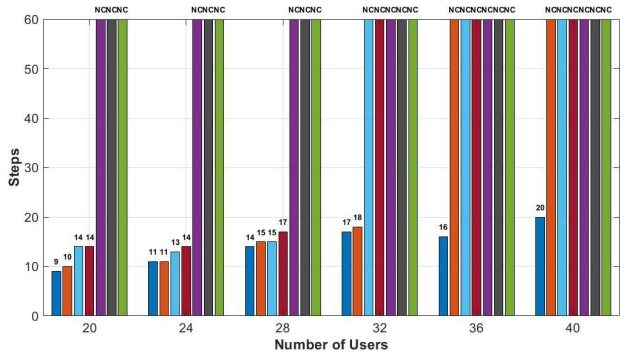
Fig. 10. Association performance by algorithm: Stepwise cumulative user association progress under varying SINR thresholds and user loads. (a) Number of users = 20. (b) Number of users = 24. (c) Number of users = 28. (d) Number of users = 32. (e) Number of users = 36. (f) Number of users = 40.

achieve full connectivity, exhibiting significant performance degradation as conditions became stricter. While CBBA and MADQN showed moderate success under the relaxed −4.4 dB threshold with connection ratios of 84.44% and 82.22%, respectively, their performance consistently dropped at higher SINR thresholds. Our ablation studies quantitatively confirm the importance of our proposed components. Ablation 1 and Ablation 2 achieved initial success rates of 77.77% and 62.22% in the relaxed setting, but their performance also degraded significantly to 55.55% and 51.11% in the most stringent environment. This validates that both the dual-action structure and PER are critical for robust performance. The other learning-based algorithms, MADDPG and MAIAC, struggled significantly, with their peak performance reaching only 28.88% and 20.00%, respectively. These results quantitatively validate the superior adaptability and robustness of the proposed method in handling dynamic, dense, and noisy UAV-BS environments.

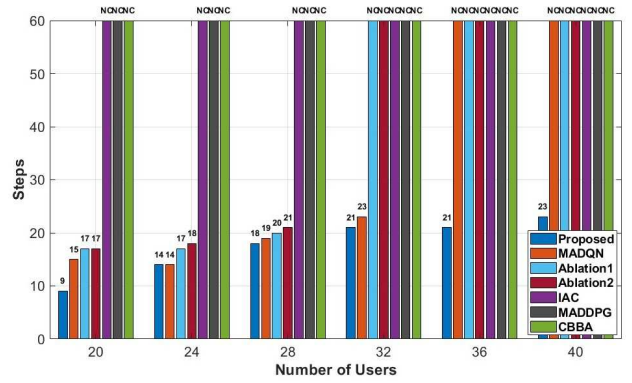
3) User Service Quality: The results in Fig. 11 compare the number of steps required for full user association under



(a)



(b)



(c)

Fig. 11. Algorithmwise comparison of steps required for full association. (a) SINR threshold = −4.4 dB (QPSK). (b) SINR threshold = 4.2 dB (16-QAM). (c) SINR threshold = 10 dB (64-QAM).

varying SINR thresholds and user densities. Each subplot presents the performance of all four algorithms under a specific SINR condition. The proposed method consistently achieves full association across all scenarios while requiring significantly fewer steps than the baseline methods.

For instance, as shown in Fig. 11(a), under an SINR threshold of −4.4 dB, the proposed algorithm demonstrates high efficiency by achieving full association in just nine steps for 20 users and 16 steps for 40 users. Even as the SINR threshold becomes more restrictive at 4.2 and 10 dB [as illustrated in Fig. 11(b) and (c)], the required steps for 40 users increase only modestly to 20 and 23, respectively. These results confirm that the proposed method finds a solution rapidly and reliably, completing full user association well within the 60-step episode limit even under the most challenging conditions.

In contrast, as shown in Fig. 11(a), a clear scalability hierarchy emerges at 32 users and beyond. Ablation 2 is the first to fail among the ablations, unable to achieve full association from the 32-user mark onward. Ablation 1 and MADQN manage to succeed at 32 users (requiring 17 and 14 steps, respectively) but consistently fail at higher densities. This performance gap becomes even more pronounced under the most stringent 10 dB threshold, as shown in Fig. 11(c). The scalability of the other methods collapses much earlier. The limit of the MADQN is reached at 32 users (requiring 23 steps before failing), while the ablation studies fail to scale beyond 28 users (where Ablation 1 and 2 required 20 and 21 steps, respectively). This confirms that only the complete proposed architecture provides the necessary efficiency and robustness to scale effectively in challenging, high-density environments.

The heuristic-based CBBA showed very limited success; it only achieved full association for 32 and 36 users under the −4.4 dB threshold, requiring a high number of steps (22 and 23, respectively). In all other user density and SINR conditions, CBBA failed to achieve full association. The other learning-based methods, MADDPG and MAIAC, demonstrated more severe limitations, consistently failing to achieve full association across every tested scenario, revealing an inability to generalize in these complex conditions.

Fig. 12 illustrates the average power consumption of both the initial and the proposed trained policies under varying SINR thresholds and user densities. Across all scenarios, the trained policy consistently consumes significantly less power than the initial baseline. This improvement becomes more prominent as user density increases or the SINR threshold becomes more stringent—conditions under which efficient resource management becomes increasingly critical. For instance, under an SINR threshold of −4.4 dB [Fig. 12(a)], the trained policy reduces power consumption from 69 to 17.1 W when serving 20 users, corresponding to a 75.22% reduction. Even with 40 users, it achieves a 49.27% reduction, consuming only 35 W compared to the constant 69 W baseline. Similar power-saving patterns are observed under stricter SINR conditions in Fig. 12(b) and (c), where the trained policy still maintains over 36% savings even in high-density scenarios. These results demonstrate the robustness and adaptability of the proposed trained policy. Importantly, this efficiency gain is not at the cost of service quality. The proposed method maintains full user association and requires fewer steps to do so, as discussed earlier. In summary, the proposed algorithm achieves dual optimization in both time and energy dimensions. It ensures rapid user association while minimizing power consumption, underscoring its practical viability in real-time, energy-constrained UAV-BS environments.

4) Trained Behaviors of UAV-BSs Across Episodes: The UAV-BS trajectories illustrated in Fig. 13 [subplots (a)–(f)] compare two deployment scenarios: the left column represents Scenario 1: Open Floor, where no internal obstructions are present, and the right column represents Scenario 2: With Walls, where interior walls introduce signal blockage. These trajectories provide key insights into how the proposed

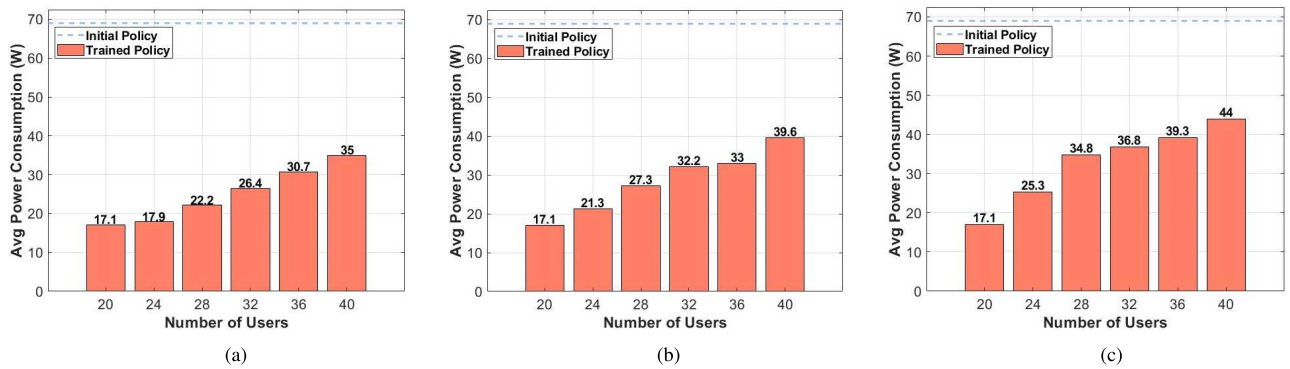


Fig. 12. Comparison of average power consumption before and after training of the proposed algorithm under varying SINR thresholds. (a) SINR threshold = −4.4dB (QPSK). (b) SINR threshold = 4.2 dB (16-QAM). (c) SINR threshold = 10 dB (64-QAM).

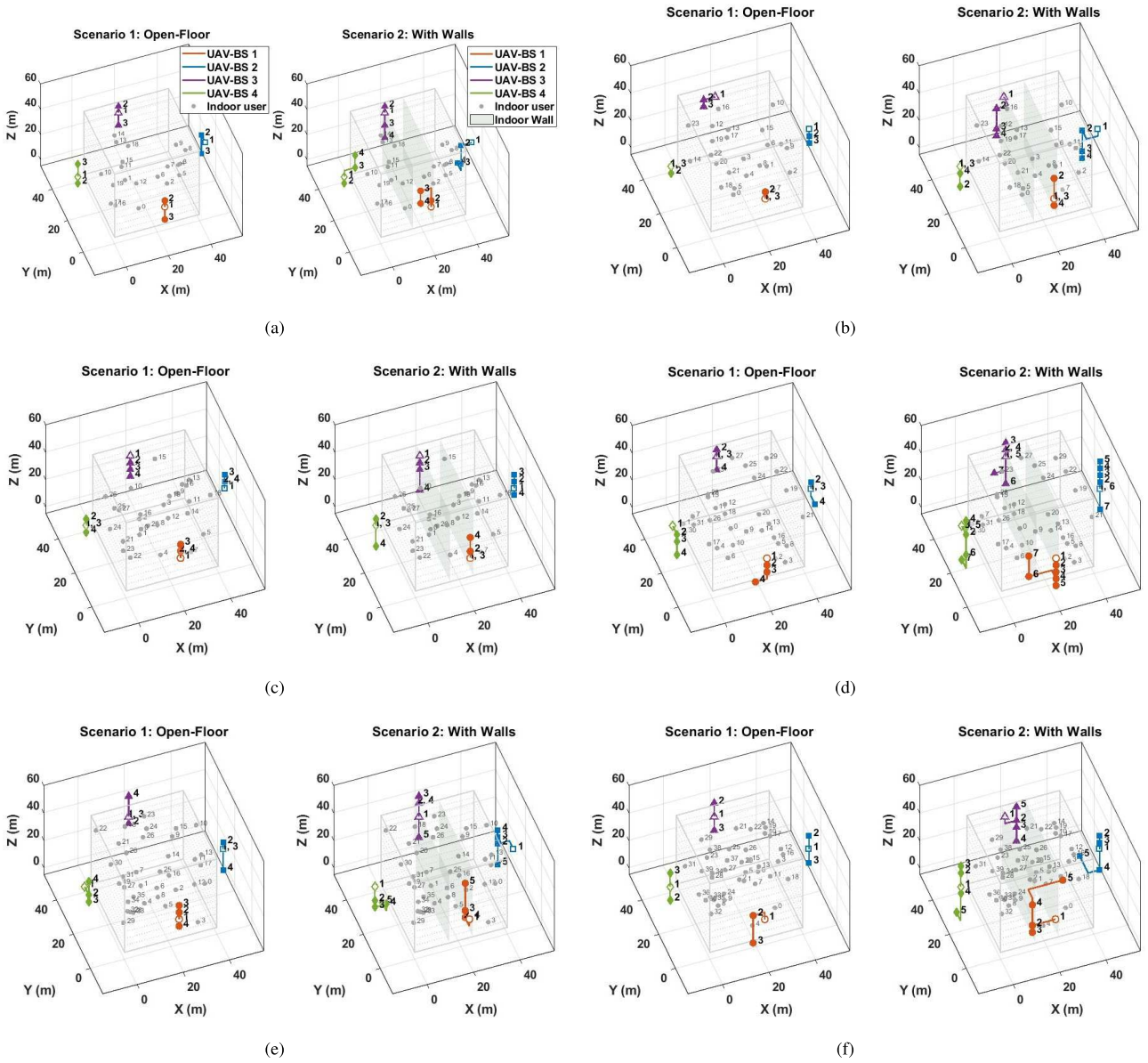


Fig. 13. UAV trajectory under SINR threshold of −4.4 using the proposed algorithm. (a) Number of users = 20. (b) Number of users = 24. (c) Number of users = 28. (d) Number of users = 32. (e) Number of users = 36. (f) Number of users = 40.

algorithm dynamically adjusts UAV movements to maintain full user association under an SINR threshold of −4.4 dB. As the number of indoor users increases from 20 to 40, a clear pattern emerges in both scenarios: UAV-BSs exhibit progressively more complex and adaptive 3-D flight paths. This growing mobility reflects the increased challenge of

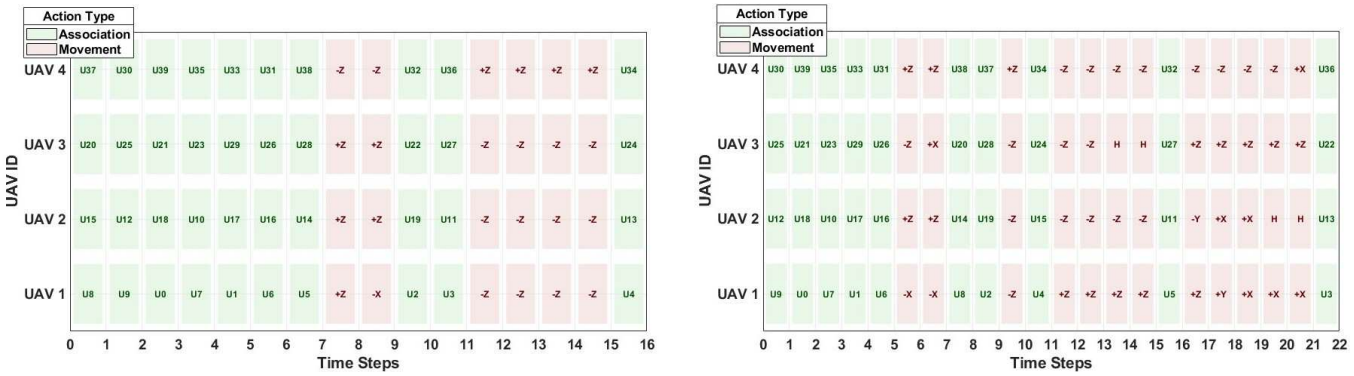


Fig. 14. Time-slot action sequences of UAVs for user association and 3-D movement under a 40-user scenario with a SINR threshold of −4.4 dB. (a) Scenario 1: open floor. (b) Scenario 2: with walls.

TABLE VIII
COMPARISON OF COMPUTATIONAL COST WITH PROPOSED AND BASELINE ALGORITHMS

Metric	Computational Cost [FLOPS]
Proposed	1,138,688
MADQN	1,097,728
MAIAC	1,303,040
MADDPG	1,249,536

serving users distributed across multiple floors and spatial regions.

Notably, Scenario 2 results in sharper trajectory variations, particularly in the vertical and lateral dimensions. This is due to the presence of internal obstacles, which force the UAVs to adjust their positions more frequently in order to circumvent signal blockages. Further evidence of this is provided in Fig. 14, which presents the detailed time-slot action sequences of all UAV-BSs for both scenarios. In Fig. 14(a), 28 users are associated immediately at the UAV-BSs' initial positions, and full association is completed within only 16 steps. In contrast, Fig. 14(b) starts with just 20 users connected and requires 22 steps to achieve complete association. This increased delay is primarily due to the movement constraints imposed by walls. For instance, UAV 1 in Scenario 2 must perform multiple directional movements to reach User 3, who remains isolated behind an interior wall. As a result, UAVs expend more time navigating the obstructed environment to maintain connectivity and satisfy the SINR threshold constraints.

Overall, this dual-scenario comparison underscores the robustness of the proposed approach in both idealized and realistic indoor environments, where static or heuristic-based methods would likely fail to maintain full association under such constraints.

5) Computational Cost: In addition to performance and energy efficiency, the computational complexity of each algorithm is evaluated in terms of the number of floating-point operations (FLOPS) required for a single inference. As summarized in Table VIII, the proposed algorithm requires approximately 1.14 million FLOPS, which represents a slight 3.59% increase compared to MADQN (1.10 million FLOPS), but is still 12.6% lower than MAIAC, which incurs the cost at 1.30 million FLOPS. Furthermore, MADDPG exhibits computational load at 1.25 million FLOPS, which is 8.87% higher than that of the proposed algorithm. This modest increase over MADQN is well-justified by the significantly improved performance achieved by the proposed method in terms of association success rate, convergence speed, and energy efficiency. In contrast, MAIAC and MADDPG not only incur substantially higher computational costs but also fail to ensure reliable user association across scenarios, resulting in an inefficient trade-off between complexity and effectiveness. These results clearly indicate that the proposed algorithm maintains strong computational efficiency while achieving superior learning performance.

Overall, the algorithm's ability to ensure real-time responsiveness and full association success at relatively low computational cost highlights its practicality for real-world UAV-BS deployment, especially in resource-constrained environments.

VI. CONCLUSION

In this article, we proposed a PER-based MADDQN algorithm to jointly optimize the movement and user association of multiple UAV-BSs for emergency indoor user service. To realistically model indoor communication constraints, we formulated the problem as a POMDP, incorporating a floor-penetrating path loss model and limited UAV energy resources. Also, we validated this channel model against realistic deployment conditions to ensure accurate modeling of signal attenuation in complex indoor environments. The proposed algorithm introduced a dual-mode policy structure that enables each UAV-BS to alternately select either a movement or an association action at each time step. Additionally, a hybrid reward function, composed of both individual and common components, guides learning toward both local efficiency and global coordination. This design allows agents to minimize total service time while satisfying SINR constraints and avoiding mutual interference. Extensive simulations demonstrated that the proposed algorithm achieves complete user association under all tested SINR thresholds and user densities, while consuming fewer association steps and

significantly less energy than baseline methods. Furthermore, the proposed algorithm maintained a low computational cost despite its superior performance. These results confirm that the proposed algorithm achieves robust, scalable, and energy-efficient user service while maintaining high computational efficiency, making it well-suited for real-time UAV-BS operations in complex indoor emergency environments.

REFERENCES

[1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[2] M. Zhang, S. Fu, and Q. Fan, "Joint 3D deployment and power allocation for UAV-BS: A deep reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2309–2312, Oct. 2021.

[3] B. Kirubakaran, O. Vikhrova, S. Andreev, and J. Hosek, "UAV-BS integration with urban infrastructure: An energy efficiency perspective," *IEEE Commun. Mag.*, vol. 63, no. 3, pp. 1–7, Mar. 2025.

[4] A. Fotouhi et al., "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart., 2019.

[5] S. Chandrasekharan et al., "Designing and implementing future aerial communication networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 26–34, May 2016.

[6] O. Y. Kolawole and M. Hunukumbure, "UAV based 5G indoor localization for emergency services," in *Proc. 5th Int. ACM Mobicom Workshop Drone Assist. Wireless Commun. 5G Beyond*, New York, NY, USA, Oct. 2022, pp. 43–48.

[7] J. Cui, B. Hu, and S. Chen, "Resource allocation and location decision of a UAV-relay for reliable emergency indoor communication," *Comput. Commun.*, vol. 159, pp. 15–25, Jun. 2020.

[8] X. Liu et al., "Deployment of UAV-BSs for on-demand full communication coverage," *Ad Hoc Netw.*, vol. 140, Mar. 2023, Art. no. 103047.

[9] *Guidelines for Evaluation of Radio Interface Technologies for IMT-advanced*, document 638, 2009.

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

[11] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[12] S. Lim, H. Yu, and H. Lee, "Optimal tethered-UAV deployment in A2G communication networks: Multi-agent Q-learning approach," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18539–18549, Oct. 2022.

[13] J. Cui, H. Shakhatreh, B. Hu, S. Chen, and C. Wang, "Power-efficient deployment of a UAV for emergency indoor wireless coverage," *IEEE Access*, vol. 6, pp. 73200–73209, 2018.

[14] H. Shakhatreh, A. Khreishah, A. Alsarhan, I. Khalil, A. Sawalmeh, and N. S. Othman, "Efficient 3D placement of a UAV using particle swarm optimization," in *Proc. 8th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2017, pp. 258–263.

[15] B. Ma, J. Zhang, Z. Zhang, and J. Zhang, "Time-efficient joint UAV-BS deployment and user association based on machine learning," *IEEE Internet Things J.*, vol. 10, no. 14, pp. 13077–13094, Jul. 2023.

[16] Z. Qin, Z. Liu, G. Han, C. Lin, L. Guo, and L. Xie, "Distributed UAV-BSs trajectory optimization for user-level fair communication service with multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12290–12301, Dec. 2021.

[17] T.-Y. Kim, Y. Ahn, J. Lee, and J.-H. Kim, "Joint movement and user association of UAV-BS for indoor user service: A multi-agent deep reinforcement learning approach," in *Proc. IEEE 22nd Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2025, pp. 1–4.

[18] J. Kang, K. Kim, H. Lee, and J.-H. Kim, "Lyapunov optimization-based online positioning in UAV-assisted emergency communications," *IEEE Access*, vol. 11, pp. 60835–60843, 2023.

[19] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2015, *arXiv:1511.05952*.

[20] X. Gu and G. Zhang, "A survey on UAV-assisted wireless communications: Recent advances and future trends," *Comput. Commun.*, vol. 208, pp. 44–78, Aug. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366423001743

[21] C. T. Cicek, H. Gultekin, B. Tavli, and H. Yanikomeroglu, "UAV base station location optimization for next generation wireless networks: Overview and future research directions," in *Proc. 1st Int. Conf. Unmanned Vehicle Systems-Oman (UVS)*, Feb. 2019, pp. 1–6.

[22] A. Carreras-Coch, J. Navarro, C. Sans, and A. Zaballos, "Communication technologies in emergency situations," *Electronics*, vol. 11, no. 7, p. 1155, Apr. 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/7/1155

[23] 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.

[24] Y. Lin, T. Wang, and S. Wang, "UAV-assisted emergency communications: An extended multi-armed bandit perspective," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 938–941, May 2019.

[25] Y. A. Sambo, P. V. Klaine, J. P. B. Nadas, and M. A. Imran, "Energy minimization UAV trajectory design for delay-tolerant emergency communication," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2019, pp. 1–6.

[26] B. Hu, L. Wang, S. Chen, J. Cui, and L. Chen, "An uplink throughput optimization scheme for UAV-enabled urban emergency communications," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4291–4302, Mar. 2022.

[27] L. Liu, B. Lin, and Y. Che, "Joint UAV-BS deployment and power allocation for maritime emergency communication system," in *Proc. 13th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2021, pp. 1–5.

[28] J. Cui, B. Hu, and S. Chen, "A decision-making scheme for UAV maximizes coverage of emergency indoor and outdoor users," *Ad Hoc Netw.*, vol. 112, Mar. 2021, Art. no. 102391.

[29] Z. Guo, B. Hu, S. Chen, B. Zhang, and L. Wang, "Joint resource and trajectory optimization for video streaming in UAV-based emergency indoor-outdoor communication," *Telecommun. Syst.*, vol. 87, no. 1, pp. 199–211, Sep. 2024.

[30] H. Shakhatreh, A. Khreishah, and I. Khalil, "Indoor mobile coverage problem using UAVs," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3837–3848, Dec. 2018.

[31] R. Ding, Y. Xu, F. Gao, and X. Shen, "Trajectory design and access control for air–ground coordinated communications system with multiagent deep reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5785–5798, Apr. 2022.

[32] J. Kim, S. Park, S. Jung, and C. Cordeiro, "Cooperative multi-UAV positioning for aerial Internet service management: A multi-agent deep reinforcement learning approach," *IEEE Trans. Netw. Service Manage.*, vol. 21, no. 4, pp. 3797–3812, Aug. 2024.

[33] E. Eldeeb, J. M. D. S. Sant'Ana, D. E. Pérez, M. Shehab, N. H. Mahmood, and H. Alves, "Multi-UAV path learning for age and power optimization in IoT with UAV battery recharge," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5356–5360, Apr. 2023.

[34] J. Tang, Y. Liang, and K. Li, "Dynamic scene path planning of UAVs based on deep reinforcement learning," *Drones*, vol. 8, no. 2, p. 60, Feb. 2024. [Online]. Available: https://www.mdpi.com/2504-446X/8/2/60

[35] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5994–6006, Sep. 2021.

[36] Y.-J. Chen and D.-Y. Huang, "Joint trajectory design and BS association for cellular-connected UAV: An imitation-augmented deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 9, no. 4, pp. 2843–2858, Feb. 2022.

[37] 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.

[38] L. Li, Q. Jiang, and W. Luo, "A unified Non-CQI-based AMC scheme for 5G NR downlink and uplink transmissions," in *Proc. IEEE 6th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Apr. 2021, pp. 881–886.

[39] T.-Y. Kim, J.-K. Kim, W.-J. Lee, S. Jung, and J.-H. Kim, "Energy-efficient full-duplex MAC protocol design for air-terrestrial communication," *J. Commun. Netw.*, vol. 25, no. 3, pp. 333–343, Jun. 2023.

[40] B. S. K. Reddy and B. Lakshmi, "Adaptive modulation and coding with channel state information in OFDM for WiMAX," *Int. J. Image, Graph. Signal Process.*, vol. 7, no. 1, pp. 61–69, Dec. 2014.

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

[42] S. Jung, W. J. Yun, M. Shin, J. Kim, and J.-H. Kim, "Orchestrated scheduling and multi-agent deep reinforcement learning for cloud-assisted multi-UAV charging systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5362–5377, Jun. 2021.
[43] A. R. S. Bramwell, D. Balmford, and G. Done, *Bramwell's Helicopter Dynamics*. Amsterdam, The Netherlands: Elsevier, 2001.
[44] S. Jung, J. Kim, and J.-H. Kim, "Joint message-passing and convex optimization framework for energy-efficient surveillance UAV scheduling," *Electronics*, vol. 9, no. 9, p. 1475, Sep. 2020.
[45] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.
[46] J. Hare, "Dealing with sparse rewards in reinforcement learning," 2019, *arXiv:1910.09281*.
[47] R. Devidze, P. Kamalaruban, and A. Singla, "Exploration-guided reward shaping for reinforcement learning under sparse rewards," in *Proc. 36th Conf. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 5829–5842.
[48] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[49] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[50] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
[51] H.-L. Choi, L. Brunet, and J. P. How, "Consensus-based decentralized auctions for robust task allocation," *IEEE Trans. Robot.*, vol. 25, no. 4, pp. 912–926, Aug. 2009.

Tae-Yoon Kim (Member, IEEE) received the B.S. degree from the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in artificial intelligence convergence network.

His current research interests include nonterrestrial network communication, medium access control protocol, and reinforcement learning-based UAV communication.

Jihong Park received the B.S. degree in avionics engineering from Hanseo University, Seosan, South Korea, in 2024. He is currently pursuing the M.S. degree in artificial intelligence convergence network with Ajou University, Suwon, South Korea.

His current research interests include medium access control protocol and reinforcement learning-based UAV communication.

Junghwa Kang (Member, IEEE) received the B.S. degree in electrical and computer engineering and the M.S. degree in artificial intelligence convergence network from Ajou University, Suwon, South Korea, in 2021 and 2023, respectively.

She has been an Engineer at SW Team (Land), Hanwha Systems, Pangyo, Republic of Korea, since February 2023. Her current research interests include reinforcement learning for target allocation.

Jaeyeol Lee (Student Member, IEEE) received the B.S. degree from the Department of Smart Automobile, Soonchunhyang University, Cheonan, South Korea, in 2023. He is currently pursuing the Ph.D. degree in space electronics and information engineering with Ajou University, Suwon, South Korea.

His current research interests include reinforcement learning, resource management, and NTN-TN integrated networks.

Soyi Jung (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Ajou University, Suwon, South Korea, in 2013, 2015, and 2021, respectively.

She has been an Assistant Professor at the Department of Electrical and Computer Engineering, Ajou University, since September 2022. Before joining Ajou University, she was an Assistant Professor at Hallym University, Chuncheon, Republic of Korea, from 2021 to 2022; a Visiting Scholar at Donald Bren School of Information and Computer Sciences, University of California at Irvine, Irvine, CA, USA, from 2021 to 2022; a Research Professor at Korea University, Seoul, Republic of Korea, in 2021; and a Researcher at Korea Testing and Research (KTR) Institute, Gwacheon, Republic of Korea, from 2015 to 2016.

Dr. Jung was a recipient of the Best Paper Award by KICS in 2015, the Young Women Researcher Award by WISET and KICS in 2015, the Bronze Paper Award from IEEE Seoul Section Student Paper Contest in 2018, and the IEEE ICOIN Best Paper Award in 2021.

Jae-Hyun Kim (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Hanyang University, Ansan, South Korea, in 1991, 1993, and 1996, respectively.

In 1996, he was with the Communication Research Laboratory, Tokyo, Japan, as a Visiting Scholar. From April 1997 to October 1998, he was a Post-Doctoral Fellow with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA, USA. From November 1998 to February 2003, he was a member of Technical Staff with the Performance Modeling and QoS Management Department, Bell Laboratories, Lucent Technologies, Holmdel, NJ, USA. Since 2003, he has been with the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea, as a Professor. He has been the Dean of College of ICT, since 2022. His research interests include medium access control protocols, QoS issues, and cross-layer optimization for wireless communication and satellite communication.

Dr. Kim is a member of KICS, the Institute of Electronics and Information Engineers (IEIE), and the Korea Information Scientists and Engineers (KIISE). He is the Center Chief of the Cooperative Multi-Orbit Radio Platform for Advanced CubeSat System-Radio Research Center (COMPASS-RRC) sponsored by the Institute for Information and Communications Technology Promotion, South Korea. He has been the Chairperson of the Digital space network committee of 6G Forum and International affairs and Service of Satellite communication Forum in South Korea, since 2018 and 2021, respectively. He also served as a Chair of IEEE VTS Seoul chapter from 2020 to 2021.