Collaborative Charging Optimization for Wireless Rechargeable Sensor Networks via Heterogeneous Mobile Chargers

Jianhang Yao, Hui Kang, Geng Sun, *Senior Member, IEEE*, Jiahui Li, *Member, IEEE*, Hongjuan Li, Jiacheng Wang, Yinqiu Liu, Dusit Niyato *Fellow, IEEE*

***Abstract*—Despite the rapid proliferation of Internet of Things applications driving widespread wireless sensor network (WSN) deployment, traditional WSNs remain fundamentally constrained by persistent energy limitations that severely restrict network lifetime and operational sustainability. Wireless rechargeable sensor networks (WRSNs) integrated with wireless power transfer (WPT) technology emerge as a transformative paradigm, theoretically enabling unlimited operational lifetime. In this paper, we investigate a heterogeneous mobile charging architecture that strategically combines automated aerial vehicles (AAVs) and ground smart vehicles (SVs) in complex terrain scenarios to collaboratively exploit the superior mobility of AAVs and extended endurance of SVs for optimal energy distribution. We formulate a multi-objective optimization problem that simultaneously addresses the dynamic balance of heterogeneous charger advantages, charging efficiency versus mobility energy consumption trade-offs, and real-time adaptive coordination under time-varying network conditions. This problem presents significant computational challenges due to its high-dimensional continuous action space, non-convex optimization landscape, and dynamic environmental constraints. To address these challenges, we propose the improved heterogeneous agent trust region policy optimization (IHATRPO) algorithm that integrates a self-attention mechanism for enhanced complex environmental state processing and employs a Beta sampling strategy to achieve unbiased gradient computation in continuous action spaces. Comprehensive simulation results demonstrate that IHATRPO achieves a 39% performance improvement over the original HATRPO, significantly outperforming state-of-the-art baseline algorithms while substantially increasing sensor node survival rate and charging system efficiency.**

***Index Terms*—Wireless rechargeable sensor network, collaborative charging optimization, heterogeneous mobile chargers, trust region policy optimization**

I. Introduction

With the rapid proliferation of Internet of Things (IoT) applications, wireless sensor networks (WSNs) have become fundamental infrastructures for environmental monitoring, smart cities, industrial automation, and precision agriculture [1], [2]. WSNs are self-organizing wireless networks that monitor physical phenomena such as temperature, sound, vibration, or pollutants [3]. Due to the small size, low power consumption, and autonomous network establishment capabilities of sensor nodes, conventional WSNs offer high flexibility, good adaptability, and low operational costs [4]. However, WSNs face a persistent challenge as the finite energy capacity of sensor nodes severely constrains network lifetime and operational sustainability. Specifically, sensor nodes typically rely on batteries that are difficult or impossible to replace in remote deployments, thereby leading to network degradation and eventual failure as nodes exhaust their energy reserves [5]. Recent research on extending the lifetime of WSNs has concentrated on energy conservation and energy provisioning approaches. While energy conservation techniques [6], [7] can significantly extend network lifetime, those methods cannot guarantee the network stability since batteries will eventually be depleted. Energy provisioning through renewable energy harvesting offers a continuous energy supply, yet is constrained by unpredictable environmental conditions [8], [9].

To address these fundamental limitations, wireless rechargeable sensor networks (WRSNs) have emerged as a transformative paradigm. Specifically, WRSNs integrate wireless power transfer (WPT) technology with conventional sensing capabilities, theoretically providing indefinite operational lifetime [10]. Moreover, WRSNs employ dedicated charging infrastructure that can be categorized into static charging stations and mobile charging platforms. Static charging stations, while providing reliable power delivery, require extensive deployment due to the limited spatial range of WPT technology, resulting in prohibitively high infrastructure costs and reduced deployment flexibility [11]. Conversely, mobile charging platforms offer superior coverage adaptability and dynamic resource allocation capabilities. Among mobile charging solutions, automated aerial vehicles (AAVs) and ground smart vehicles (SVs) represent two complementary approaches with distinct operational characteristics. Specifically, AAVs excel in mobility, rapid deployment, and terrain independence but are constrained by limited energy capacity and weather sensitivity [12], [13], while SVs provide extended operational endurance and robust performance but are restricted by terrain accessibility and mobility limitations [14].

Current WRSN researches focus on single-type charging

Jianhang Yao, Hui Kang, Jiahui Li, and Hongjuan Li are with the College of Computer Science and Technology, Jilin University, Changchun 130012, China (e-mails: yaojx25@mails.jlu.edu.cn; kanghui@jlu.edu.cn; lijiahui@jlu.edu.cn; hongjuan23@mails.jlu.edu.cn).

Geng Sun is with the College of Computer Science and Technology, Jilin University, Changchun 130012, China, and with Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China; he is also affiliated with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: sungeng@jlu.edu.cn).

Jiacheng Wang, Yinqiu Liu, and Dusit Niyato are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mails: jiacheng.wang@ntu.edu.sg; yinqiu001@ntu.edu.sg; dniyato@ntu.edu.sg).

scenarios, which trade-off between mobility and energy efficiency. However, single-type charging approaches, whether AAV-based or SV-based, cannot simultaneously optimize all critical performance metrics due to their individual limitations and the diverse requirements of WRSNs. Such fundamental limitation becomes particularly pronounced in complex deployment environments where sensor nodes exhibit varying energy demands, spatial distributions, and accessibility constraints that exceed the capabilities of any single charging platform. Motivated by these observations, we propose to combine AAV and SV platforms [15] and design a heterogeneous mobile charging architecture to overcome the inherent limitations of homogeneous charging approaches. This strategic coordination between heterogeneous chargers enables adaptive resource allocation that responds to varying sensor node energy demands and environmental constraints, potentially revolutionizing the efficiency and reliability of WRSNs.

However, implementing such heterogeneous mobile charging coordination introduces several significant technical challenges that existing solutions cannot adequately address. *Firstly*, the coordination problem between AAVs and SVs requires sophisticated collaborative decision-making mechanisms that can dynamically balance their respective advantages while accounting for different energy consumption patterns, mobility constraints, and charging capabilities in real-time operational conditions [15]. *Secondly*, the multi-objective optimization nature of the problem involves simultaneously maximizing charging efficiency, minimizing mobility energy consumption, and reducing sensor node mortality, then creating complex trade-offs that traditional optimization approaches cannot effectively resolve due to conflicting objectives and non-convex solution spaces [16]. *Finally*, sensor network conditions exhibit dynamic and time-varying characteristics, including fluctuating energy levels, changing environmental conditions, and evolving communication requirements [17], which necessitate adaptive strategies that can respond to these variations without compromising long-term performance objectives or system stability.

Accordingly, this paper proposes a novel deep reinforcement learning (DRL)-based approach for collaborative charging optimization in WRSNs employing heterogeneous mobile chargers. The main contributions of this paper are summarized as follows:

- *Innovative Heterogeneous Air-Ground Collaborative Charging System Model:* We design a comprehensive system model that strategically integrates the AAV and SV as collaborative charging agents in WRSNs. This architecture is specifically tailored for complex deployment scenarios where single-charger solutions prove inadequate. To the best of our knowledge, this is the first work to systematically investigate the collaborative charging optimization problem for heterogeneous mobile chargers while considering their distinctive mobility characteristics, energy constraints, and charging capabilities.
- *Multi-Objective Optimization Problem with Heterogeneous Charger Interdependencies:* We formulate a multi-objective optimization problem that characterizes the complex interdependencies among charging efficiency maximization, mobility energy minimization, and sensor node mortality minimization in a heterogeneous mobile chargers environment. This formulation enables the identification of fundamental trade-offs inherent in multi-objective optimization, where competing objectives generate a conflicting solution space, thus requiring collaborative coordination mechanisms. Moreover, this problem reveals distinctive coordination dynamics and complementary operational patterns in heterogeneous charger collaboration.
- *DRL Solution with Heterogeneous Trust Region Strategy:* To address the dynamic and multi-objective nature of the optimization challenge, we propose the improved heterogeneous agent trust region policy optimization (IHATRPO) algorithm. This approach incorporates two key innovations. *First*, the self-attention mechanism enables agents to process complex environmental information and inter-agent interactions more effectively. *Second*, the Beta sampling strategy ensures unbiased gradient computation for continuous action spaces with bounded constraints. These enhancements specifically address the challenges of decentralized decision-making in heterogeneous multi-agent environments while ensuring convergence stability.
- *Simulation and Performance Evaluation*: Simulation results demonstrate that the proposed algorithm outperforms various baselines, *e.g.*, PPO, MADDPG, HATRPO. Moreover, the heterogeneous charger coordination approach significantly enhances sensor network survivability while maintaining charging efficiency. In addition, it is also confirmed that collaborative AAV-SV deployment provides adaptive coverage capabilities that effectively respond to dynamic network conditions.

The rest of this paper is organized as follows. Section II reviews the related research activities in WRSNs. Section III presents the system models. Section IV formulates the optimization problem. Section V introduces the proposed IHATRPO algorithm. Section VI provides the comprehensive simulation results and performance analysis, and Section VII concludes the paper with discussions on future research directions.

II. Related Work

In this work, we aim to propose a collaborative charging optimization framework in WRSNs by using heterogeneous mobile chargers. This topic involves the charging system architecture in WRSNs, optimization objectives in WRSN charging systems, and optimization methods for WRSN charging. Thus, we briefly introduce the related works of these areas as follows.

A. Charging System Architectures in WRSNs

Various charging system architectures have been designed to prolong the network lifetime in WRSNs. Traditional ground-based charging strategies have been extensively investigated, where mobile charging vehicles traverse the network to replenish sensor nodes. For example, the authors in [18] proposed a

periodic charging and scheduling scheme aimed at optimizing the charging time and sensor selection of charging vehicles. Moreover, the authors in [19] proposed an on-demand charging strategy that incorporates spatial, temporal, and event domain characteristics of nodes, while utilizing an improved K-means algorithm for network partitioning with terrestrial wireless charging vehicles. Further building upon this ground-based mobile charger architecture, the authors in [20] focused on optimizing for network tasks by jointly selecting sensors and allocating energy.

With the advancement of AAV technology, aerial charging systems have emerged as promising alternatives for WRSN energy replenishment. For example, the authors in [21] proposed a joint scheduling and trajectory optimization problem for single-AAV based charging scenarios, thus improving charging efficiency by reducing repeated charging nodes while minimizing hovering points and flight distance. Furthermore, the authors in [22] investigated a multi-AAV deployment optimization problem and proposed an improved firefly algorithm to optimize charging efficiency, motion energy consumption, and sensor coverage. In [23], the authors proposed a cooperative air-ground architecture where one AAV charges sensors, and a ground-based vehicle provides battery replacement for the AAV, using a Deep Q-Network to optimize the strategy.

However, these works treat ground-based and aerial charging systems as independent solutions, thus overlooking the potential collaborative benefits of air-ground cooperative charging. Different from these methods, we design a heterogeneous charging system that simultaneously coordinates both the AAV and SV to achieve complementary operational advantages and compensate for individual limitations.

B. Optimization Objectives in WRSN Charging Systems

The optimization objectives in WRSN charging systems have been primarily focused on network lifetime maximization and node survival rate enhancement. For instance, the authors in [24] proposed a hybrid approach targeting network longevity through optimized charging scheduling, where inner rings adopt single-node charging with flat topology while outer rings employ multi-node charging with cluster topology. Moreover, the authors in [25] proposed an energy-efficient adaptive directional charging algorithm that focuses on maximizing sensor node survival rates by adaptively selecting single-node or multi-node charging based on sensor node density.

Energy consumption optimization of mobile chargers represents another critical research direction. The authors in [26] proposed a DRL-based mobile safety policy intervention algorithm specifically targeting single mobile charger energy efficiency in an uncertain environment with mobile obstacles. Moreover, the authors in [27] combined SV deployment with recovery operations, jointly optimizing charging and recovery scheduling to minimize overall system energy consumption while handling increased charging requests.

Charging efficiency has also received considerable attention in recent studies. Specifically, the authors in [28] proposed efficient algorithms for increasing energy efficiency in WRSNs for cyber-physical systems through intelligent scheduling and sensor node prioritization without requiring prior knowledge of energy levels. Furthermore, trajectory optimization has emerged as a key goal for enhancing charging efficiency, where researchers focus on minimizing travel distances and optimizing charging paths to improve overall system performance.

However, these works predominantly optimize individual objectives in isolation without considering trade-offs between competing goals. Different from these approaches, we adopt a multi-objective optimization framework that jointly considers the mortality of sensor nodes, energy consumption of chargers, and charging efficiency.

C. Optimization Methods for WRSN Charging

Conventional optimization methods have been widely employed for WRSN charging problems. For example, the graph-based optimization approaches have been extensively used, where the authors in [29] proposed comprehensive frameworks by using hexagonal decomposition and boustrophedon path planning for energy-aware coordination of one AAV in WRSN, thus addressing simultaneous period-area coverage, charging scheduling, and resource allocation challenges. Moreover, evolutionary computation methods have also demonstrated effectiveness, as shown in [5], which proposed an improved firefly and NSGA-II-based solution for many-objective charging optimization in WRSNs. Additionally, heuristic optimization techniques have been applied in some works, where researchers employ greedy algorithms and local search methods to solve charging scheduling problems with polynomial time complexity.

Recent advances in DRL have introduced intelligent decision-making capabilities to WRSN charging systems. For instance, the authors in [30] proposed a novel DRL approach with a hybrid action space for mobile charging, specifically employing the deep deterministic policy gradient (DDPG) algorithm to determine optimal charging time allocation and achieve improved network lifetime through continuous action space control. Furthermore, the authors in [31] introduced an asynchronous and scalable multi-agent proximal policy optimization (ASM-PPO) algorithm for cooperative charging, thus demonstrating enhanced charging coordination through distributed policy optimization with improved scalability for large-scale scenarios.

However, these DRL-based works primarily focus on homogeneous multi-agent systems without considering the coordination challenges inherent in heterogeneous agent environments. Current approaches lack the collaborative mechanisms required to handle heterogeneous agent coordination between the AAV and SV with fundamentally different operational characteristics. These limitations motivate us to propose a specialized multi-agent DRL algorithm capable of managing heterogeneous agent interactions.

D. Motivation and Contributions of This Work

Different from these words, we consider a heterogeneous air-ground cooperative charging system by using both the

AAV and SV. Moreover, we formulate a multi-objective optimization problem that jointly considers the mortality of sensor nodes, energy consumption of chargers, and charging efficiency. To solve it, we propose an innovative heterogeneous multi-agent DRL method specifically designed for coordinating agents with diverse operational characteristics and capabilities. In the following section, therefore, we present a detailed description of the system model under consideration.

III. System Models and Preliminaries

In this section, we introduce the models of the considered heterogeneous air-ground collaborative charging system (HAGCCS), including the network model, wireless charging model, and energy consumption models of the AAV and SV.

A. Network Model

The HAGCCS under consideration is illustrated in Fig. 1, and it comprises the following elements:

- A set of sensor nodes $\mathcal{S} = \{1, 2, \ldots, N_{\mathcal{S}}\}$. These sensor nodes are stationary and randomly distributed throughout the network, primarily tasked with data collection. Note that each sensor node can transmit data to a remote base station or receive commands from it [15]. Moreover, each sensor node is equipped with an energy harvesting unit and an energy storage unit, which means that it can receive and store wireless energy transferred by mobile chargers [32].
- A pair of heterogeneous mobile chargers. Specifically, the heterogeneous mobile chargers consist of an AAV and an SV. Note that both the AAV and SV are capable of processing data from sensor nodes, remote base stations, and other mobile chargers [33]. Moreover, the AAV and SV can travel freely within the network area to provide charging service for the sensor nodes within a specified radius [34], and their batteries power both of them.
- A remote base (BS) station that acts as a data fusion center. This BS is located at the edge of the region for data collection, and without loss of generality, we consider that the BS has no energy constraint since it has a sufficient energy supply [35].

In HAGCCS, the energy consumption of sensor nodes typically follows certain protocols and cycles to ensure efficient network operation and prolong network lifetime. In this case, we consider a discrete-time system evolving over the timeline $\mathcal{T} = \{t|1, 2, ..., T\}$. Specifically, each time slot t consists of two main phases that are the sensing phase and charging phase, as illustrated in Fig. 2. In the sensing phase, sensor nodes perform data collection, data processing, and data transmission. In the charging phase, the AAV and SV provide wireless energy transfer to the sensor nodes.

Based on this, we consider that all the sensor nodes and SV are located within the same two-dimensional plane, while the AAV maintains a constant altitude when flying or hovering. As such, the locations of the i-th sensor node, AAV, SV, are denoted as $(x_i^{\mathcal{S}}, y_i^{\mathcal{S}}, 0)$, (x^{AAV}, y^{AAV}, h), $(x^{SV}, y^{SV}, 0)$, respectively.

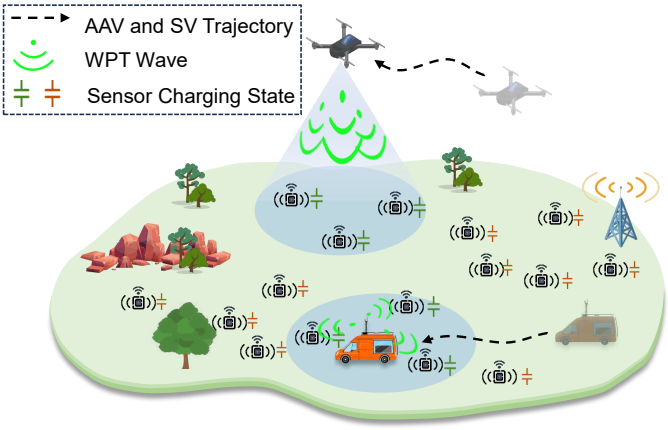

Fig. 1. Architecture diagram of the HAGCCS for the WRSN. The AAV and SV travel within the WRSN to collaboratively provide energy for sensors through WPT waves.

TABLE I
Summary of Main Notations

Notation	Description
μ	Charging efficiency
ρ	Air density
θ_i	Heading angle of agent i
γ	Discount factor
α, β	Shape parameters of Beta distribution
$d_{\max}$	Maximum charging radius of AAV/SV
d_i	Travel distance of agent i
$\mathcal{D}$	Trajectory buffer of AAV/SV
D_{KL}	KL divergence
f_1, f_2, f_3	Charging efficiency, travel distance, node mortality
G_s, G_r	Antenna gain of transmitter and receiver
h	Flight altitude of AAV
k_1, k_2, k_3	Control parameters for SV motor
$\mathcal{N}$	Agent set
P_0	Transmit power of AAV/SV
P_i	Received power at sensor node i
$P_{AAV}(v)$	Motion energy consumption of AAV
$P_{SV}(v)$	Motion energy consumption of SV
q_i^t	Energy level of sensor node i at time t
$\mathcal{S}$	Set of sensor nodes
$\mathcal{T}$	Set of time slots
v	Travel/flight speed
X_{max}, Y_{max}	Maximum range of WRSN area
$\boldsymbol{Z}_t$	Decision variables at time slot t

As such, during each time slot, the AAV and SV travel freely within the sensor network to charge nearby sensor nodes, which aims to improve the charging efficiency and extend the network lifetime. In the following, we model the wireless charging model and energy consumption model of the AAV and SV to identify the key decision variables for optimizing wireless energy transfer and its transmission efficiency.

B. Wireless Charging Model

In WRSNs, WPT enables the transmission of electrical energy wirelessly from the transmitter to the receiver across the air gap. We consider a radio-frequency (RF) based omnidirectional WPT model [36], which utilizes RF waves at a

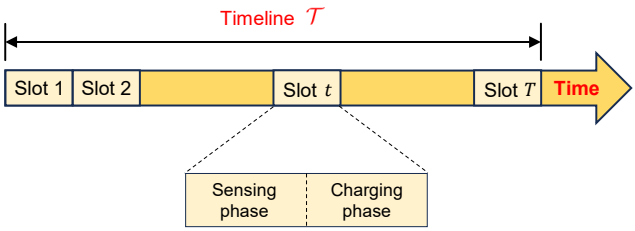

Fig. 2. The time slot division model in HAGCCS.

specific frequency for energy transmission, thereby allowing energy to propagate in all directions.

As such, the charging efficiency μ of the AAV or SV for sensor nodes can be defined as follows:

$$\mu = \frac{G_s G_r \eta}{L_p} \left(\frac{\lambda}{4\pi(d+\beta)} \right)^2, \tag{1}$$

where G_s denotes the antenna gain of the AAV or SV, G_r represents the antenna gain of the sensor nodes as the receiver, λ is the wavelength of the RF signal, η is the rectifier efficiency, L_p is the polarization loss, β is a tunable parameter in the Friis free-space equation, and d is the distance between the AAV or SV and the sensor node.

Since in Eq. (1), all parameters except for d and β are constant values in a specific WRSN, the calculation for the charging efficiency μ can be simplified as $\mu = \alpha/(d+\beta)^2$, where α is a constant that encompasses the parameter values of G_s, G_r, η, L_p, λ, and others from Eq. (1). Let P_0 represent the transmit power of the AAV or SV. Then, the received power P_i at the i-th sensor node S_i can be given by $P_i = \mu_i P_0$.

From Eq. (1), it can be observed that the received power at the sensor node primarily depends on the distance between the AAV or SV and the sensor node, as all parameters except for d can be considered constants. As such, we set the max charging distance d_{max} to assess the impact of distance on the received power. Specifically, when the distance between the AAV (or SV) and sensor node exceeds d_{max}, the received power at the sensor node becomes too low for energy rectification, thus preventing effective charging. Therefore, d_{max} can be regarded as the effective charging radius. The received power P_i can then be further expressed as follows:

$$P_i = \begin{cases} \frac{\alpha P_0}{(d_i+\beta)^2} & d_i \le d_{max} \\ 0 & d_i > d_{max} \end{cases}. \tag{2}$$

C. Energy Consumption Model of AAV and SV

The total energy consumption of the AAV and SV consists of two main components. The first part is the energy consumed by the AAV and SV for charging sensor nodes. The second part is the energy consumed during the movement of the AAV and SV, including propulsion and hovering for the AAV, as well as the travel of the SV. Moreover, the energy consumption caused by communication among sensor nodes, mobile chargers, and BS is negligible compared to the movement energy consumption. Therefore, we focus on the wireless charging energy in Section III-B and motion energy consumption in this section. Based on this, we consider the use of rotary-wing AAV and SV equipped with DC motors, with their respective motion energy consumption models as follows:

For a rotary-wing AAV with a flight speed of v, its motion energy consumption [37] can be given by

$$\begin{aligned} P_{AAV}(v) =& P_B \left(1 + \frac{3v^2}{v_{tip}^2}\right) + \\ & P_I \left(\sqrt{1 + \frac{v^4}{4v_0^4}} - \frac{v^2}{2v_0^2} \right)^{1/2} + \frac{1}{2} d_0 \rho s A v^3, \end{aligned} \tag{3}$$

where P_B and P_I represent the blade power and induced power of the AAV in a hovering state, respectively. Moreover, v_{tip} denotes the tip speed of the rotor blades, while v_0 represents the average induced rotor speed of the AAV in the hovering state. Additionally, d_0 and ρ are the body drag coefficient and air density, respectively. Meanwhile, s and A represent the solidity and area of the rotor of the AAV, respectively.

For an SV with a travel speed of v and using a permanent magnet direct current (PMDC) motor model, its motion energy consumption [36] can be given by

$$P_{SV}(v) = k_1 v^2 + k_2 v + k_3, \tag{4}$$

where k_1, k_2, and k_3 are the respective control parameters.

Without loss of generality, we disregard the additional increase or decrease in energy consumption of the AAV and SV due to acceleration or deceleration during motion, as these account for only a small fraction of their total operating time.

IV. Problem Formulation and Analyses

In this section, we analyze the collaborative charging problem of HAGCCS. *First*, we analyze several key factors involved in the charging phase. *Second*, we formulate and analyze the collaborative charging problem.

A. Problem Statement

In this work, we focus on three optimization objectives, *i.e.*, improving the charging efficiency of the AAV and SV, reducing the travel distance of the AAV and SV, and minimizing the mortality of the sensor nodes. These three optimization objectives involve inherent trade-offs. Specifically, if the AAV and SV are positioned closer to the sensor nodes, a larger number of nodes will fall within the charging range, thereby improving the charging efficiency. The location of the AAV and SV is directly related to their energy consumption, which means that if the positioning results in more frequent or longer travel of the AAV and SV, the energy consumption will increase accordingly. Moreover, improper positioning may lead to inadequate coverage of sensor nodes, thereby preventing some nodes from receiving sufficient charging support, which means that the node mortality increases.

As such, the corresponding decision variables are represented as $\boldsymbol{Z}_t = \{x_t^{SV}, y_t^{SV}, x_t^{AAV}, y_t^{AAV}, h_t^{AAV}\}$, whose variables correspond to the coordinates of the AAV and SV.

In HAGCCS, we aim to enhance the charging efficiency of the AAV and SV to supply more energy to the sensor nodes, thereby extending the lifetime of WRSN. According to Eq. (2), the AAV or SV can charge all sensor nodes within the effective charging radius d_{max}. Therefore, the charging efficiency of the AAV or SV, which is the first optimization objective f_1, can be expressed as follows:

$$f_1 = \sum_{i=1}^{N_S} P_i. \tag{5}$$

By reducing the travel distance of the AAV and SV, the energy consumption caused by their travel distance is minimized. Therefore, more energy can be allocated for charging the sensor nodes, thereby effectively improving their energy utilization efficiency. Let $(x_{\text{init}}, y_{\text{init}}, z_{\text{init}})$ and $(x_{\text{target}}, y_{\text{target}}, z_{\text{target}})$ represent the initial and target positions of the AAV or SV, respectively, in a single movement, then the travel distance of the AAV or SV, which is the second optimization objective f_2, can then be expressed as follows:

$$f_2 = \sqrt{(x_{\text{target}} - x_{\text{init}})^2 + (y_{\text{target}} - y_{\text{init}})^2 + (z_{\text{target}} - z_{\text{init}})^2}. \tag{6}$$

The mortality of sensor nodes is a key indicator for evaluating the performance and efficiency of WRSNs. Specifically, an increase in sensor node mortality leads to deterioration in WRSN stability and reliability, while also reducing the integrity of collected data. As such, we consider minimizing the mortality of sensor nodes in this network as the third optimization objective. Specifically, the third objective f_3, *i.e.*, the mortality of sensor nodes,

$$f_3 = \frac{\sum_{i=1}^{N_S} b_i}{N_{\mathcal{S}}}, \tag{7}$$

where b_i is a binary variable defined as follows:

$$b_i = \begin{cases} 1, & \text{if sensor node } i \text{ is alive} \\ 0, & \text{if sensor node } i \text{ is dead} \end{cases}. \tag{8}$$

To improve the charging efficiency, the AAV and SV need to move frequently between sensor nodes that need to be charged, which results in an increase in their travel distance. However, as the travel distances of the AAV and SV increase, their energy consumption also rises, which means that they cannot charge more sensors. As a result, the mortality of sensor nodes will increase. Therefore, three optimization objectives have a conflicting relationship. Thus, we formulate this problem by using multi-objective optimization theory.

According to the three optimization sub-objectives above, our optimization problem can be formulated as follows:

$$\begin{aligned}
(\text{P1}): \quad & \max_{\boldsymbol{Z}_t} \sum_{t=1}^{\mathcal{T}} (f_1, -f_2, -f_3), && \text{(9a)} \\
\text{s.t.} \quad & 0 \le x_t^{AAV} \le X_{max}, \quad \forall t \in \mathcal{T} && \text{(9b)} \\
& 0 \le y_t^{AAV} \le Y_{max}, \quad \forall t \in \mathcal{T} && \text{(9c)} \\
& 0 \le x_t^{SV} \le X_{max}, \quad \forall t \in \mathcal{T} && \text{(9d)} \\
& 0 \le y_t^{SV} \le Y_{max}, \quad \forall t \in \mathcal{T} && \text{(9e)}
\end{aligned}$$

where X_{max} and Y_{max} represent the maximum ranges of the WRSN area along the x-axis and y-axis, respectively. Moreover, the boundary constraints (9b)-(9c) and (9d)-(9e) ensure that both the AAV and SV operate within the WRSN boundaries, respectively.

B. Problem Analyses

Based on the HAGCCS and the optimization sub-objectives, problem (P1) exhibits the following characteristics. *Firstly*, problem (P1) exhibits strong dynamic and stochastic characteristics. Specifically, the energy consumption magnitude of sensors varies randomly, thereby making the current overall network sensor energy consumption level unpredictable, thus demonstrating strong dynamic properties. Moreover, the uncertainty in travel distances of the AAV and SV results in their energy consumption being stochastic as well. *Secondly*, this problem involves both long-term and short-term optimization objectives. Specifically, the long-term objective is to maximize the WRSN lifetime, while the short-term objective is to minimize the energy consumption of the AAV and SV within each time slot. Therefore, during the optimization process, we should consider both the current and long-term interests. *Finally*, since the AAV is an energy-sensitive system that requires real-time decision-making during flight operation, the solution used to solve this problem should satisfy real-time computational requirements.

Accordingly, the problem (P1) exhibits dynamic characteristics, long-term slot properties, and real-time decision-making requirements. Thus, conventional optimization methods or evolutionary computation algorithms are unsuitable for this problem. Specifically, conventional optimization methods typically rely on a known and fixed environment model [16]. Even if heuristic or evolutionary algorithms are used, they are often predefined or require a considerable amount of time to run, which prevents real-time adjustments in practical applications [38]. Moreover, these methods generally focus on immediate optimization and struggle to balance both current and long-term benefits. Though conventional methods may maximize short-term gains, they overlook the sustainability of long-term network performance and stability.

Accordingly, we adopt the advantageous DRL to address the considered problem. Specifically, DRL enables adaptive decision-making in dynamic environments and optimizes long-term network performance by learning from real-time feedback, thereby making it well-suited for problem (P1) in HAGCCS.

V. Heterogeneous Trust Region Strategy Optimization-Based Decentralized Solution

In this section, we propose a decentralized solution to address the collaborative charging problem (P1) in the WRSN. *Firstly*, we formulate the optimization problem as a Markov game [39] involving the AAV and SV agents. *Secondly*, we introduce the IHATRPO algorithm that integrates a self-attention mechanism and Beta sampling to enhance multi-agent coordination. *Finally*, we analyze the computational and space complexity of the proposed algorithm.

A. Markov Game Formulation

We first model Problem (P1) as a Markov game. Specifically, MG can be formally represented by the tuple $\langle \mathcal{N}, \{\mathcal{S}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{P}, \{\boldsymbol{R}_i\}_{i\in\mathcal{N}}, \gamma\rangle$. The key elements of MG are given as follows:

1) Agent Set: The HAGCCS employs two agents that are assigned to control the AAV and SV, respectively, *i.e.*,

$$\mathcal{N} = \{A^{AAV}, A^{SV}\}. \tag{10}$$

At each time slot t, both agents independently observe the environmental state and execute actions, so as to maximize their respective expected total rewards.

2) State Space: Both agents share the same global state space, which can ensure complete environmental observability for decision-making. As such, the state space is defined by the positions of the sensor nodes, their energy levels, as well as the positions of the AAV and SV. Specifically, the sensor-related information can be obtained through the communication protocol, while the positions of the AAV and SV can be acquired via global positioning system (GPS). Thus, the state space is defined as follows:

$$\mathcal{S} = \{s_t | s_t = (\mathcal{S}_t, AAV_t, SV_t), \forall t \in \mathcal{T}\}, \tag{11}$$

where $\mathcal{S}_t = \{x_t^1, x_t^2, ..., x_t^{Ns}, y_t^1, y_t^2, ..., y_t^{Ns}, q_t^1, q_t^2, ..., q_t^{Ns}\}$ represents the set of coordinates, and current energy levels of each sensor node at the beginning of time slot t. Meanwhile, $AAV_t = \{x_t^{AAV}, y_t^{AAV}, h^{AAV}\}$ and $SV_t = \{x_t^{SV}, y_t^{SV}\}$ denote the coordinates of the AAV and SV, respectively, at the start of time slot t.

3) Action Space: Each agent operates within its own action space, representing distinct decision variables for controlling vehicle motion parameters. Both the AAV and SV agents follow the same mathematical formulation while maintaining independent control over their respective vehicles. Based on environmental observations, each agent governs two critical motion parameters, which are the heading angle θ and travel distance d. Note that these two parameters can be corresponded to the decision variable $\boldsymbol{Z}_t$ of problem (P1). Consequently, the action space for each agent is defined as:

$$\mathcal{A}_i = \{a_t^i | a_t^i = (\theta_t^i, d_t^i), \forall t \in \mathcal{T}, i \in \mathcal{N}\}. \tag{12}$$

4) Reward Function: The reward mechanism is designed to motivate both agents to optimize their respective contributions to the HAGCCS performance. Each agent receives individual rewards based on its performance, with both agents sharing the same mathematical reward structure to ensure consistency and fairness in the learning process. According to the optimization objectives in problem (P1), the reward function incorporates three key performance indicators: charging efficiency, energy consumption (represented by travel distance), and network sustainability (measured by node mortality). The reward function is defined as follows:

$$\mathcal{R}_i = \{r_t^i | r_t^i = \lambda_1 f_{1,t}^i - \lambda_2 f_{2,t}^i - \lambda_3 f_{3,t}, \forall t \in \mathcal{T}, i \in \mathcal{N}\}, \tag{13}$$

where $f_{1,t}^i$, $f_{2,t}^i$, and $f_{3,t}$ are corresponding to Eq. 5, Eq. 6, and Eq. 7 during time slot t. The weighting coefficients λ_1, λ_2, and λ_3 serve as balancing factors that ensure appropriate emphasis on the relative importance of each reward component in the overall system performance.

Algorithm 1: IHATRPO

Input: Number of heterogeneous agents n, Max training episodes $max\_episodes$, max time slots max_time_slots
Output: Optimized policy network parameters $\{\theta_i\}_{i=1}^n$

/* Initialization: */
for *agent* $i \in [1, n]$ **do**
  Initialize Actor network parameters θ_i and Critic network parameters ω_i
end
for *episode* = 1 *to* $max\_episodes$ **do**
  Reset sensor nodes, initialize power levels for AAV/SV, initialize trajectory buffer $\mathcal{D}$
  for t = 1 *to* max_time_slots **do**
    for *agent* $i \in [1, n]$ **do**
      Agent i constructs Beta distribution from state s_t, samples action a_i^t
      Execute action a_i^t, receive reward r_i^t
    end
    Update environment state $s_t \to s_{t+1}$
    for *agent* $i \in [1, n]$ **do**
      Store transition $(s_t, a_i^t, s_{t+1}, r_i^t)$ in trajectory buffer $\mathcal{D}$
    end
    if $E_{AAV} \leq 0$ *and* $E_{SV} \leq 0$ **then**
      break
    end
  end
  /* Policy Update: */
  for *agent* $i \in [1, n]$ **do**
    Compute GAE advantages $\hat{A}^{\pi_{\theta_i}}$ from $\mathcal{D}$ and normalize
    Update ω_i
    Update θ_i via TRPO using Eq. 17
  end
end
Return $\theta = \{\theta_1, \ldots, \theta_n\}$.

B. IHATRPO Algorithm

In the section, we handle the MG through the IHATRPO algorithm, where the AAV and SV are each treated as an agent. In the following, we first introduce the conventional HATRPO. Subsequently, we present two improvement measures, namely a self-attention mechanism and Beta sampling, to enhance the ability of HATRPO to handle the MG.

1) Preliminaries of HATRPO: HATRPO integrates the multi-agent framework with trust region policy optimization to enhance MADRL, thus achieving monotonic improvement.

In an N-agent MG, the joint policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ represents collective decision-making of agents. Specifically, at time slot t, given state s^t, each agent takes an action a_i^t according to its policy. Subsequently, the environment computes the reward $\boldsymbol{r}^t = (r_1^t, \ldots, t_N^t)$ based on the joint action $\boldsymbol{a}^t = (a_1^t, \ldots, a_N^t)$ and updates the state to s^{t+1}. The optimization goal is to maximize expected cumulative reward by updating policy parameters from θ_i to θ_i', where the

objective function difference from the policy update is given by

$$J(\theta_i') - J(\theta_i) = \mathbb{E}_{\tau\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_{\theta_i}}(s_i^t, a_i^t)\right], \tag{14}$$

where τ is the trajectory, $\gamma \in (0,1)$ is the discount factor, and $A^{\pi_{\theta_i}}$ is the advantage function under policy π_{θ_i}. However, since the updated policy $\pi_{\theta_i'}$ cannot be computed directly, we approximate the objective function using the state distribution of the pre-update policy π_{θ_i} and apply importance sampling to correct the action distribution. The objective is then given by

$$L(\theta_i'|\theta_i) = \mathbb{E}_{s_i\sim\nu^{\boldsymbol{\pi}}}\mathbb{E}_{a_i\sim\pi_{\theta_i}(\cdot|s_i)}\left[\frac{\pi_{\theta_i'}(a_i|s_i)}{\pi_{\theta_i}(a_i|s_i)}A^{\pi_{\theta_i}}(s_i,a_i)\right]. \tag{15}$$

To maintain proximity between the updated and original policies, we adopt the Kullback-Leibler (KL) divergence within the trust region policy optimization framework [40]. Specifically, the divergence between the pre-update policy π_{θ_i} and post-update policy $\pi_{\theta_i'}$ is denoted by $D_{KL}(\pi_\theta||\pi_{\theta_i'})$. By setting δ as the update step size threshold, we formulate the optimization problem as:

$$\begin{aligned} &\max_{\theta_i'} \quad L(\theta_i'|\theta_i) \\ &\text{s.t.} \quad \mathbb{E}_{s_i\sim\nu^{\boldsymbol{\pi}}}\left[D_{KL}(\pi_{\theta_i}||\pi_{\theta_i'})\right] \le \delta. \end{aligned} \tag{16}$$

To simplify the computation, we apply linear and quadratic approximations to the objective function and KL constraint, respectively, thereby yielding the closed-form update as follows:

$$\theta_i^{k+1} = \theta_i^k + \alpha^j\sqrt{\frac{2\delta}{(g_i)^T(H_i)^{-1}g_i}}{H_i}^{-1}g_i, \tag{17}$$

where θ_i^k represents the policy parameters after the k-th iteration of the i-th agent, $\alpha^j \in (0,1)$ is the backtracking line search coefficient, which ensures that θ_i' is superior to θ_i^k and satisfies the KL divergence constraint. Moreover, $g_i = \nabla_{\theta_i'}\mathbb{E}_{s_i\sim\nu^{\boldsymbol{\pi}}}\mathbb{E}_{a_i\sim\pi_{\theta_i^k}(\cdot|s_i)}[\pi_{\theta_i'}(a_i|s_i)/\pi_{\theta_i^k}(a_i|s_i)A^{\pi_{\theta_i^k}}(s_i,a_i)]$ is the gradient of the optimization objective, and $H_i = \mathcal{H}\left[\mathbb{E}_{s_i\sim\nu^{\boldsymbol{\pi}}}\left[D_{KL}\left(\pi_{\theta_i}||\pi_{\theta_i'}\right)\right]\right]$ represents the Hessian matrix derived from the KL divergence.

2) Self-Attention Mechanism: In HAGCCS, heterogeneous charging agents must simultaneously process multi-dimensional state information, including their own states, distributed sensor node conditions, and inter-agent coordination requirements within a non-stationary environment. Conventional MADRL approaches treat all state information equally through conventional feature extraction, thereby failing to capture varying component importance and dynamic relationships between agents and sensor nodes, which leads to suboptimal decision-making.

The self-attention mechanism addresses these limitations by dynamically assigning importance weights to input elements based on contextual relevance. Different from traditional approaches, the self-attention mechanism captures complex dependencies through parallel processing while adaptively focusing on critical information for decision-making. In our IHATRPO, we integrate the self-attention mechanism into the actor-critic networks of both AAV and SV agents. Specifically, the self-attention mechanism [41] computes context-aware representations by measuring similarity between input elements using Query (Q), Key (K), and Value (V) vectors, which can be given by

$$A(Q,K,V) = \sigma\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{18}$$

where d_k represents the dimension of K. By using self-attention integration, heterogeneous agents dynamically prioritize relevant information based on context and achieve a deep understanding of state interdependencies for informed decision-making.

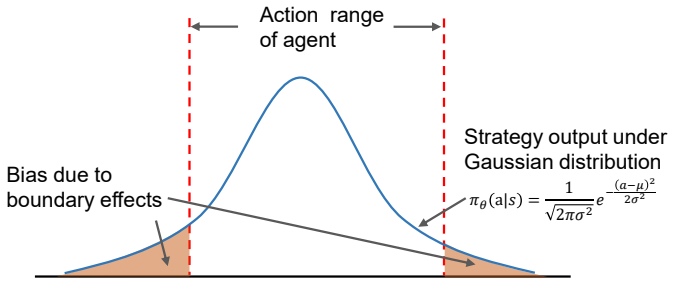

Fig. 3. Boundary effects on Gaussian distribution bias. The shaded areas represent probability mass falling outside the valid action range, which must be truncated during sampling.

3) Beta Sampling: The HAGCCS requires continuous action control for the AAV and SV travel within bounded action spaces constrained by the finite distribution range of the WRSN. However, conventional continuous control methods utilize Gaussian distributions for action sampling, whose unbounded nature conflicts with the bounded action spaces, thereby resulting in boundary effects and distributional bias that compromise gradient computation accuracy, as demonstrated in Fig 3.

In this case, the Beta distribution addresses these limitations through its inherent bounded property on $[0,1]$, thereby ensuring all sampled actions remain within valid ranges without truncation. Unlike Gaussian distributions that require clipping or rescaling, Beta distributions naturally maintain unbiased gradient computation while respecting action space constraints [42]. The probability density function of the Beta distribution is given by

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{19}$$

where $\Gamma(\cdot)$ is the Gamma function, and α and β serve as shape parameters that collectively determine the distribution shape. We adopt $\pi_\theta(a|s) = f(c\cdot a;\alpha,\beta)$ to characterize the stochastic policy, which is referred to as the Beta sampling strategy. The parameters $\alpha = \alpha_\theta(s)$ and $\beta = \beta_\theta(s)$ are modeled by a neural network parameterized by θ. The parameter c is determined based on the value ranges of travel direction and distance for the AAV or SV in the action space, thereby ensuring that action outputs satisfy their respective action space constraints.

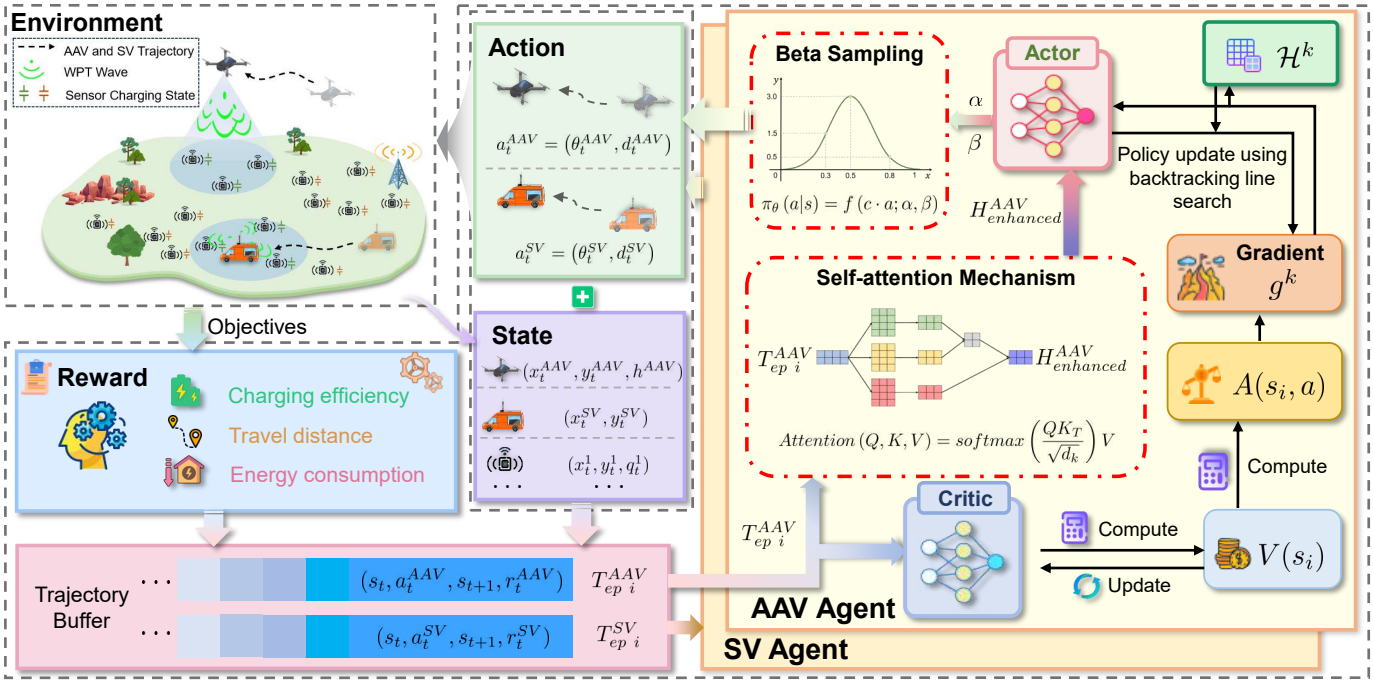

Fig. 4. Framework of IHATRPO for heterogeneous air-ground collaborative charging in the WRSN. The algorithm integrates Beta distribution-based action sampling, self-attention mechanism enhanced state processing, and heterogeneous actor-critic structures for the AAV and SV to optimize multi-objective charging strategies.

Through Beta sampling implementation, the agents maintain unbiased gradient computation within bounded action spaces, eliminate boundary effects that degrade policy performance, and ensure natural action space compliance without additional constraints or post-processing steps.

C. Complexity Analysis of IHATRPO

The computational and space complexity of IHATRPO during training and execution phases are analyzed as follows [43].

The computational complexity of IHATRPO during the training phase is $\mathcal{O}(N_A(|\boldsymbol{\theta}| + |\boldsymbol{\omega}| + N_E(T(1+V) + |\boldsymbol{\omega}| + N_T(3+N_K+N_M) + |\boldsymbol{\theta}|(3+N_K+N_M)))$, which can be summarized as follows:

- *Network Initialization:* This phase involves the initialization of network parameters of the AAV and SV. Specifically, the computational complexity is expressed as $\mathcal{O}(N_A(|\boldsymbol{\theta}|+|\boldsymbol{\omega}|))$, where N_A is the number of agents, the $|\cdot|$ operation represents the number of parameters in the networks.
- *Action Selection:* This phase entails selecting actions according to the output scores of the actor network, and its complexity is $\mathcal{O}(N_A N_E T)$. Here, N_E denotes the number of training episodes, and T is the number of steps per episode.
- *Reward Calculation and State Transitions:* The computational complexity of reward calculation and state transitions is $\mathcal{O}(N_A N_E T V)$, where V represents the complexity of interacting with the environment.
- *Network Update:* The updating phase consists of two main parts that are the updates of the critic networks, as well as the updates of the actor networks. *First*, the advantage function is calculated, and the critic network parameters are updated subsequently. This part has the complexity of $\mathcal{O}(N_A N_E(|\boldsymbol{\omega}| + N_T))$, where N_T is the length of the sampled training data. *Second*, the actor network is updated by calculating the target value of the surrogate function, calculating the conjugate gradient, and linearly searching for parameters that meet the conditions. Therefore, the corresponding complexity is $\mathcal{O}(N_A N_E(N_T(2+N_K+N_M)+|\boldsymbol{\theta}|(3+N_K+N_M)))$, where N_K is the number of iterations of the conjugate gradient and N_M is the number of iterations of the linear search. Thus, the complexity of this phase is calculated as $\mathcal{O}(N_A N_E(|\boldsymbol{\omega}|+N_T(3+N_K+N_M)+|\boldsymbol{\theta}|(3+N_K+N_M)))$.

Besides, the space complexity of IHATRPO during the training phase is $\mathcal{O}(N_A(|\boldsymbol{\theta}|+|\boldsymbol{\omega}|)+|\mathcal{D}|(2|\boldsymbol{s}|+\boldsymbol{a}+1)$, where $|\mathcal{D}|$ denotes the size of the trajectory buffer. As such, the space complexity is mainly for storing neural network parameters and sampled trajectories.

During the evaluation phase, the computational complexity of IHATRPO is $\mathcal{O}(N_E N_A)$, which can be attributed to action selection and transition according to the current state using the feature and actor network. Moreover, the space complexity during the execution phase is $N_A|\boldsymbol{\theta}|$ since the feature and actor network parameters need to be stored in memory for action selection.

VI. Simulations and Analyses

In this section, we first introduce the simulation setting and baselines. Subsequently, we present the optimization results,

the comparison analyses with state-of-the-art baselines, and the analysis of agent spatial movement patterns.

A. Simulation Setups

1) Scenario and Algorithm Setups: In the simulations, we consider the scenario that the AAV and SV provide wireless charging to a sensor network. The primary parameters are shown in Table II. Additionally, following the methodology in [44], we set the charging efficiency parameters α and β in Eq. 2 to 36 and 30, respectively. The energy consumption rate of sensor nodes per round is randomly generated within the range of 0.025 J to 0.04 J.

In the proposed IHATRPO, the algorithm parameters are shown in Table II. Both the policy network and value network are configured with two hidden layers, each containing 256 neurons. Additionally, we set the number of training iterations to 6.5×10^5 and employ the Adam optimizer for neural network updates. Note that these algorithm parameters are determined by careful tuning to ensure performance and convergence. We consider the heterogeneity between the AAV and SV by assigning different reward weight coefficients. Therefore, in the reward function, we assign a higher λ_2 for the SV and a higher λ_3 for the AAV.

TABLE II
SIMULATION SETTINGS

Parameters	Values
Network area	$100 \times 100\ m^2$
Number of sensor nodes	100
Transmit power of AAV and SV	3 W [45]
Reception threshold of the sensor node	5 mW
The max energy of the sensor node	2 J
The charging radius of AAV and SV	6 m
Learning rate of neural network	5×10^{-5}
KL threshold	5×10^{-5}
Linear search step	0.5
GAE scaling factor λ	0.98
Entropy coefficient	0.01
Discount factor	0.96
Time step of each episode	200

2) Baselines: To demonstrate the superiority of the proposed IHATRPO, we introduce the following comparative baselines. Note that these baselines adopt the same parameters as mentioned above and integrate the schedule policy of the AAV and SV.

- *PPO*: PPO is a policy gradient method that improves training stability through clipped surrogate objectives [46]. As a single-agent baseline, PPO treats the multi-agent environment as a stationary single-agent MDP by training each agent independently, therefore ignoring the non-stationary nature caused by other learning agents.
- *DDPG*: DDPG is an actor-critic method designed for continuous control tasks that combines policy gradient methods with Q-learning [47]. When applied to multi-agent settings, each agent is trained independently using DDPG and treats other agents as part of the environment dynamics without explicit coordination mechanisms.
- *MADDPG*: MADDPG is DDPG-based classical MADRL approach based on the CTDE architecture [48]. This baseline allows agents to access global information during training while maintaining individual policies and shows effectiveness in multi-agent continuous control tasks.
- *HAPPO*: HAPPO adapts PPO for heterogeneous multi-agent settings where agents have different observation and action spaces [49]. This method serves as a baseline given its capability to handle heterogeneous AAV-SV coordination.
- *HATRPO*: HATRPO extends TRPO to a heterogeneous multi-agent environment by maintaining individual trust regions for each agent [49]. Furthermore, the details of this approach are elaborated in Section V-B1. The implementation ensures stable policy updates through KL divergence constraints across diverse agents.

As such, the comparisons with PPO and DDPG demonstrate the necessity of multi-agent coordination mechanisms, the comparison with MADDPG shows the effectiveness of handling different types of agents, the comparison with HAPPO illustrates the superiority of the HATRPO-based framework in handling heterogeneous multi-agent scenarios, and the comparison with HATRPO can assess the effectiveness of two improvement measures of IHATRPO. In the following analyses, we first present the performance of multiple optimization sub-objectives under the IHATRPO, and then conduct a comparative analysis of convergence performance and total reward feedback between these baselines and IHATRPO, and the following analysis of agent trajectories.

B. Performance Evaluation

1) Optimziation Results: As can be seen in Fig. 5, Fig. 5(a) shows the respective cumulative reward of the AAV and SV, Fig. 5(b), Fig. 5(c), and Fig. 5(d) illustrate optimization of objectives in terms of the charging efficiency (f_1), travel distance (f_2) of the AAV and SV, and the mortality (f_3) of sensor nodes. As can be seen, the AAV and SV agents exhibit similar convergence trends and converge after approximately 200k iterations in Fig. 5(a), which demonstrates that IHATRPO applied heterogeneous optimization framework achieves excellent optimization effects for heterogeneous agents. Moreover, each objective achieves good optimization results with increasing training episodes in Fig. 5(a), (b), and (c), which demonstrates that the designed reward function in Eq. 13 can better balance the relationship between the AAV and SV. Moreover, it is noteworthy that a significant reduction in sensor node mortality from an initial rate exceeding 90% to below 10% in Fig. 5(d), which indicates that through the scheduling of the AAV and SV, the sensor node mortality can be reduced and HAGCCS achieves better energy efficiency.

2) Comparison Results: Fig. 6 illustrates the cumulative rewards for each episode of IHATRPO in comparison to other benchmark algorithms. As can be seen, IHATRPO achieves faster convergence speed and the highest reward. This can be explained by several factors. *First*, the self-attention mechanism enables IHATRPO to dynamically prioritize relevant information and capture complex dependencies among multi-

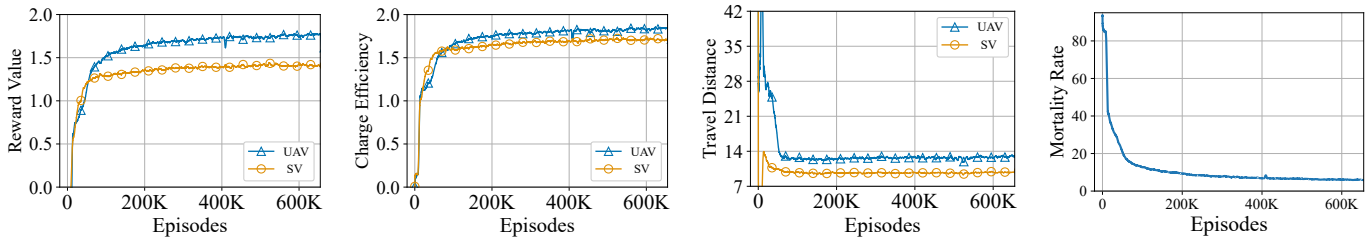

Fig. 5. Visualization results obtained by IHATRPO. (a) The total reward of the AAV and SV. (b) The charging efficiency of the AAV and SV. (c) The travel distance of the AAV and SV. (d) The mortality of sensor nodes.

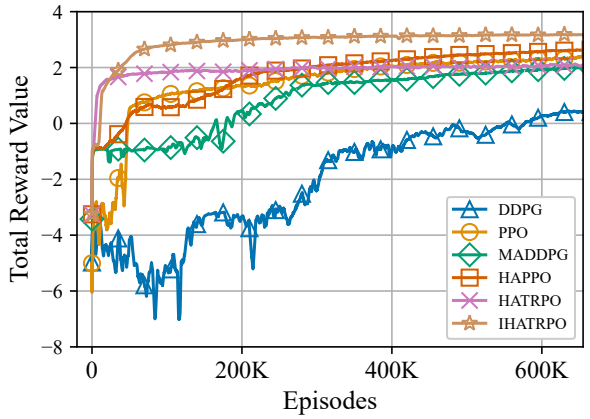

Fig. 6. Convergence performance comparison of PPO, DDPG, MADDPG, HAPPO, HATRPO, IHATRPO.

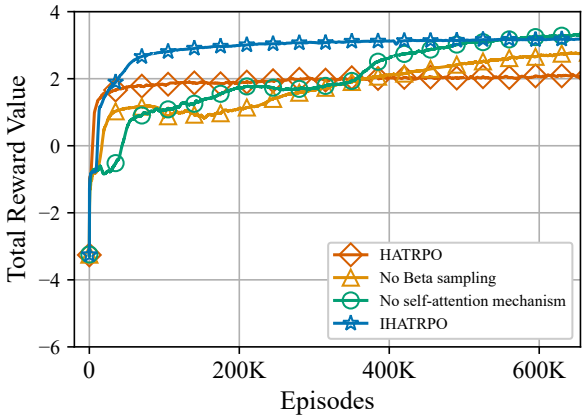

Fig. 7. Effectiveness of different techniques. (Self-attention mechanism and Beta sampling strategy)

dimensional states. *Second*, the Beta sampling provides naturally bounded action sampling complying with the HAGCCS characteristics for IHATRPO. While HATRPO demonstrates the fastest convergence performance in the initial phase, this algorithm achieves lower cumulative rewards after convergence due to its limited capability in processing abundant information and coordination between heterogeneous agents. Among the single-agent baselines, PPO shows better convergence than DDPG, but both struggle with the multi-agent coordination challenges in the convergence phase. MADDPG fails to account for the heterogeneity between the AAV and SV, thereby resulting in slower convergence and lower reward. HAPPO and HATRPO exhibit slower convergence due to action boundary violations caused by Gaussian sampling.

The performance improvement over the original HATRPO particularly validates that the integration of the self-attention mechanism and Beta sampling strategy effectively enhances policy optimization capability, thereby achieving superior learning performance in the heterogeneous multi-agent collaborative charging scenario.

3) Ablation Analysis: Fig. 7 presents the contribution of each proposed component in IHATRPO. Specifically, we examine the effects of removing the self-attention mechanism and Beta sampling strategy, respectively. As can be seen, the complete IHATRPO achieves the highest total reward value and demonstrates stable convergence, thereby highlighting the synergistic effect of its components. When the self-attention mechanism is removed, slower convergence suggests that self-attention enhances the capability to extract and integrate critical state information from the complex environment. Similarly, the removal of the Beta sampling causes a noticeable decline of the final reward value, which indicates that the Beta sampling strategy supports the policy in exploring the bounded action space more effectively.

Quantitatively, the integration of the self-attention mechanism and Beta samplilng strategy yields an overall performance improvement of approximately 39% compared with the original HATRPO algorithm. This result in Fig. 7 confirms that both components contribute significantly to enhance the learning performance and overall reward of IHATRPO.

4) Spatial Movement Patterns Analysis: Fig. 8 shows the trajectory patterns and spatial distribution of the AAV and SV in the WRSN obtained through IHATRPO optimization. As observed in the trajectory visualization, the AAV primarily operates in the lower region of the sensor network, while the SV predominantly covers the upper region. The middle area demonstrates overlapping coverage where sensor nodes may be served by either agent, with the actual charging responsibility determined dynamically based on real-time charging requirements and spatial proximity.

This territorial division emerges naturally from the embedded coordination mechanism in IHATRPO. The self-attention

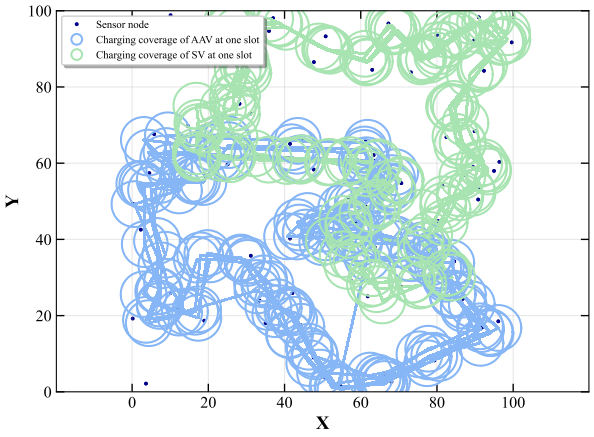

Fig. 8. The trajectory of the AAV and SV obtained by IHATRPO.

mechanism enables each agent to dynamically assess charging priorities and spatial distribution based on the current WRSN conditions, which results in an efficient labor division that minimizes redundant coverage. Moreover, the Beta sampling enables agents to discover optimal territorial boundaries that balance workload distribution and service efficiency. This territorial coordination demonstrates the effectiveness of IHATRPO in achieving intelligent spatial resource allocation without explicit territorial assignment protocols or centralized coordination mechanisms.

VII. Conclusion

This paper has investigated a collaborative charging optimization problem for WRSNs using heterogeneous mobile chargers in complex terrain scenarios. Following this, we have formulated a multi-objective optimization problem to simultaneously maximize charging efficiency, minimize mobility energy consumption, and reduce sensor node mortality by coordinating the AAV and SV. The problem has proved highly challenging due to its dynamic nature with real-time adaptation requirements and complex trade-offs between competing objectives in heterogeneous multi-agent environments. To address these challenges, we have proposed the novel IHATRPO algorithm that incorporates the self-attention mechanism for enhanced environmental processing and the Beta sampling strategy for unbiased gradient computation in continuous action spaces. Simulation results have demonstrated that the proposed IHATRPO algorithm achieves faster convergence and superior performance compared to baselines, with sensor node mortality dramatically reduced from over 90% to below 10%. Spatial movement patterns analysis shows that the AAV and SV naturally develop complementary coverage patterns through the embedded coordination mechanism, with each agent specializing in different network regions to achieve efficient spatial division of labor. Future work will focus on extending the framework to larger-scale networks and multiple heterogeneous charging agents, while investigating the scalability limits of the proposed coordination mechanism. Additionally, integrating energy harvesting techniques may potentially yield even better performance by reducing charging demands and enabling more efficient resource allocation strategies.

References

[1] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, "Applications of wireless sensor networks: An up-to-date survey," *Applied System Innovation*, vol. 3, no. 1, 2020.

[2] W. W. Greenwood, J. P. Lynch, and D. Zekkos, "Applications of UAVs in civil infrastructure," *Journal of Infrastructure Systems*, vol. 25, no. 2, p. 04019002, 2019.

[3] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[4] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, "Applications of wireless sensor networks: An up-to-date survey," *Applied System Innovation*, vol. 3, no. 1, 2020.

[5] J. Li, G. Sun, A. Wang, M. Lei, S. Liang, H. Kang, and Y. Liu, "A many-objective optimization charging scheme for wireless rechargeable sensor networks via mobile charging vehicles," *Comput. Networks*, vol. 215, p. 109196, 2022.

[6] D. Dhabliya, R. Soundararajan, P. Selvarasu, M. S. Balasubramaniam, A. S. Rajawat, S. B. Goyal, M. S. Raboaca, T. C. Mihaltan, C. Verma, and G. Suciu, "Energy-efficient network protocols and resilient data transmission schemes for wireless sensor networks—an experimental survey," *Energies*, vol. 15, no. 23, 2022.

[7] M. Y. A. Khan, M. Hussain, M. Halim, S. Ibrahim, and A. Haque, "A comprehensive review on techniques and challenges of energy harvesting from distributed renewable energy sources for wireless sensor networks," *Control Systems and Optimization Letters*, vol. 2, pp. 15–22, 01 2024.

[8] M. N. Hussain, M. A. Halim, M. Y. A. Khan, S. Ibrahim, and A. Haque, "A comprehensive review on techniques and challenges of energy harvesting from distributed renewable energy sources for wireless sensor networks," *Control Systems and Optimization Letters*, vol. 2, no. 1, pp. 15–22, 2024.

[9] B. Y. León Ávila, C. A. García Vázquez, O. Pérez Baluja, D. T. Cotfas, and P. A. Cotfas, "Energy harvesting techniques for wireless sensor networks: A systematic literature review," *Energy Strategy Reviews*, vol. 57, p. 101617, 2025.

[10] B. Qureshi, S. A. Aziz, X. Wang, A. Hawbani, S. H. Alsamhi, T. Qureshi, and A. Naji, "A state-of-the-art survey on wireless rechargeable sensor networks: Perspectives and challenges," *Wirel. Networks*, vol. 28, no. 7, pp. 3019–3043, 2022.

[11] A. Kaswan, P. K. Jana, and S. K. Das, "A survey on mobile charging techniques in wireless rechargeable sensor networks," *IEEE Commun. Surv. Tutorials*, vol. 24, no. 3, pp. 1750–1779, 2022.

[12] G. Sun, L. Zhang, J. Li, J. Wu, J. Wang, Z. Sun, C. Zhao, and V. C. M. Leung, "Age of information optimization in laser-charged uav-assisted iot networks: A multi-agent deep reinforcement learning method," 2025.

[13] X. Mou, D. Gladwin, J. Jiang, K. Li, and Z. Yang, "Near-field wireless power transfer technology for unmanned aerial vehicles: A systematical review," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 4, no. 1, pp. 147–158, 2023.

[14] C. Lin, F. Gao, H. Dai, J. Ren, L. Wang, and G. Wu, "Maximizing charging utility with obstacles through fresnel diffraction model," in *Proc. IEEE INFOCOM*, 2020, pp. 2046–2055.

[15] N. Liu, C. Luo, J. Cao, Y. Hong, and Z. Chen, "Trajectory optimization of laser-charged UAVs for charging wireless rechargeable sensor networks," *Sensors*, vol. 22, no. 23, p. 9215, 2022.

[16] S. He, J. Chen, F. Jiang, D. K. Y. Yau, G. Xing, and Y. Sun, "Energy provisioning in wireless rechargeable sensor networks," *IEEE Trans. Mob. Comput.*, vol. 12, no. 10, pp. 1931–1942, 2013.

[17] C. Lin, Z. Wang, D. Han, Y. Wu, C. Yu, and G. Wu, "TADP: enabling temporal and distantial priority scheduling for on-demand charging architecture in wireless rechargeable sensor networks," *J. Syst. Archit.*, vol. 70, pp. 26–38, 2016.

[18] H. Dai, Q. Ma, X. Wu, G. Chen, D. K. Y. Yau, S. Tang, X. Li, and C. Tian, "CHASE: Charging and scheduling scheme for stochastic event capture in wireless rechargeable sensor networks," *IEEE Trans. Mob. Comput.*, vol. 19, no. 1, pp. 44–59, 2020.

[19] Y. Dong, G. Bao, Y. Liu, M. Wei, Y. Huo, Z. Lou, Y. Wang, and C. Wang, "Instant on-demand charging strategy with multiple chargers in wireless rechargeable sensor networks," *Ad Hoc Networks*, vol. 136, p. 102964, 2022.

[20] T. Wu, P. Yang, H. Dai, C. Xiang, X. Rao, J. Huang, and T. Ma, "Joint sensor selection and energy allocation for tasks-driven mobile charging in wireless rechargeable sensor networks," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 505–11 523, 2020.

[21] Y. Liu, H. Pan, G. Sun, A. Wang, J. Li, and S. Liang, "Joint scheduling and trajectory optimization of charging UAV in wireless rechargeable sensor networks," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 11 796–11 813, 2022.

[22] S. Liang, Z. Fang, G. Sun, C. Lin, J. Li, S. Li, and A. Wang, "Charging UAV deployment for improving charging performance of wireless rechargeable sensor networks via joint optimization approach," *Comput. Networks*, vol. 201, p. 108573, 2021.

[23] N. Liu, J. Zhang, C. Luo, J. Cao, Y. Hong, Z. Chen, and T. Chen, "Dynamic charging strategy optimization for uav-assisted wireless rechargeable sensor networks based on deep q-network," *IEEE Internet of Things Journal*, vol. 11, no. 12, pp. 21 125–21 134, 2024.

[24] Y. Yang, X. Liu, K. Tang, W. Che, and Q. Xue, "Multi-type charging scheduling based on area requirement difference for wireless rechargeable sensor networks," *IEEE Trans. Sustain. Comput.*, vol. 9, no. 2, pp. 182–196, 2024.

[25] D. Lee, C. Lee, G. Jang, W. Na, and S. Cho, "Energy-efficient directional charging strategy for wireless rechargeable sensor networks," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 19 034–19 048, 2022.

[26] X. Zhang, R. Jia, Q. Yin, Z. Zheng, and M. Li, "Intelligent trajectory design and charging scheduling in wireless rechargeable sensor networks with obstacles," *IEEE Trans. Mob. Comput.*, vol. 23, no. 9, pp. 8664–8679, 2024.

[27] L. Li, Y. Feng, N. Liu, Y. Li, and J. Zhang, "Deep reinforcement learning-based dynamic charging–recycling scheme for wireless rechargeable sensor networks," *IEEE Sensors Journal*, vol. 24, no. 9, pp. 15 457–15 471, 2024.

[28] E. F. Orumwense and K. Abo-Al-Ez, "On increasing the energy efficiency of wireless rechargeable sensor networks for cyber-physical systems," *Energies*, vol. 15, no. 3, 2022.

[29] C. Lin, S. Hao, W. Yang, P. Wang, L. Wang, G. Wu, and Q. Zhang, "Maximizing energy efficiency of period-area coverage with a UAV for wireless rechargeable sensor networks," *IEEE/ACM Trans. Netw.*, vol. 31, no. 4, pp. 1657–1673, 2023.

[30] C. Jiang, W. Chen, X. Chen, S. Zhang, and W. Xiao, "Deep reinforcement learning approach with hybrid action space for mobile charging in wireless rechargeable sensor networks," *Expert Syst. Appl.*, vol. 249, p. 123752, 2024.

[31] Y. Liang, H. Wu, and H. Wang, "ASM-PPO: Asynchronous and scalable multi-agent PPO for cooperative charging," in *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, P. Faliszewski, V. Mascardi, C. Pelachaud, and M. E. Taylor, Eds. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022, pp. 798–806.

[32] Z. Ning, H. Ji, X. Wang, E. C. H. Ngai, L. Guo, and J. Liu, "Joint optimization of data acquisition and trajectory planning for UAV-assisted wireless powered internet of things," *IEEE Trans. Mob. Comput.*, vol. 24, no. 2, pp. 1016–1030, 2025.

[33] T. Chen, J. Chen, X. Gao, and T. Chen, "Mobile charging strategy for wireless rechargeable sensor networks," *Sensors*, vol. 22, no. 1, p. 359, 2022.

[34] L. Xie, Y. Shi, Y. T. Hou, and H. D. Sherali, "Making sensor networks immortal: An energy-renewal approach with wireless power transfer," *IEEE/ACM Trans. Netw.*, vol. 20, no. 6, pp. 1748–1761, 2012.

[35] J. Yi and I. Yoon, "Efficient energy supply using mobile charger for solar-powered wireless sensor networks," *Sensors*, vol. 19, no. 12, p. 2679, 2019.

[36] Y. Mei, Y. Lu, Y. C. Hu, and C. S. G. Lee, "Energy-efficient motion planning for mobile robots," in *Proceedings of the 2004 IEEE International Conference on Robotics and Automation, ICRA 2004, April 26 - May 1, 2004, New Orleans, LA, USA*, 2004, pp. 4344–4349.

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

[38] Y. Shu, H. Yousefi, P. Cheng, J. Chen, Y. J. Gu, T. He, and K. G. Shin, "Near-optimal velocity control for mobile charging in wireless rechargeable sensor networks," *IEEE Trans. Mob. Comput.*, vol. 15, no. 7, pp. 1699–1713, 2016.

[39] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: A survey," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 895–943, 2022.

[40] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 1889–1897.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.

[42] P.-W. Chou, "The beta policy for continuous control reinforcement learning," Master's thesis, Carnegie Mellon University, Pittsburgh PA, June 2017.

[43] W. Xie, G. Sun, B. Liu, J. Li, J. Wang, H. Du, D. Niyato, and D. I. Kim, "Joint optimization of UAV-carried IRS for urban low altitude mmwave communications with deep reinforcement learning," *CoRR*, vol. abs/2501.02787, 2025.

[44] L. Fu, P. Cheng, Y. Gu, J. Chen, and T. He, "Optimal charging in wireless rechargeable sensor networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 278–291, 2016.

[45] C. Hou and Q. Huang, "Energy supply control of wireless powered piecewise linear neural network," *IEEE Trans Autom. Sci. Eng.*, vol. 21, no. 4, pp. 6892–6907, 2024.

[46] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[47] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[48] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6379–6390.

[49] J. G. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang, "Trust region policy optimisation in multi-agent reinforcement learning," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.