# Topology-Aware Stable Multi-Agent Reinforcement Learning Framework for Air–Ground–WANET Inspection

Qingwei Tang, *Student Member, IEEE*, Wei Sun, *Senior Member, IEEE*, Zhi Liu, *Senior Member, IEEE*, Nikos D. Hatziargyriou, *Life Fellow, IEEE*, Yang Xiao, *Fellow, IEEE*, and Xiaohui Yuan, *Senior Member, IEEE*

***Abstract*—Large-scale substations face challenges such as complex inspections, high labor costs, and limited communication, requiring intelligent and efficient solutions. An air-ground system of unmanned aerial vehicles (UAVs) and inspection robots (IRs) offers strong mobility and wide coverage. Wireless Ad Hoc Networks (WANETs), with their scalability, self-organization, and reliability, enable secure, stable, and decentralized communication, making them ideal for supporting substation inspections. The integrated Air-Ground-WANET system features a high-dimensional state space, strong coupling, and challenging multi-objective trade-offs. Dynamic topologies and constraints further complicate optimization. Traditional methods lack real-time adaptability, while conventional reinforcement learning (RL) often suffers from oscillation and instability, limiting practical use. To address these challenges, this paper proposes a multi-agent RL framework with Lyapunov stability guarantees, employing a structurally constrained monotonic neural network controller to ensure convergence and stability during training. A topology-aware graph attention Critic network is designed to effectively capture spatial dependencies and node coordination, improving policy evaluation accuracy. Additionally, a policy optimization mechanism incorporating topology-aware gradient regularization and neighborhood consistency constraints is introduced to enhance coordination and training stability in dynamic communication environments. Extensive simulations demonstrate its effectiveness and superior performance in complex scenarios.**

***Index Terms*—Multi-agent reinforcement learning, substation inspection, wireless ad hoc network, Lyapunov stability, graph attention network.**

## I. Introduction

### A. Background

As THE power grid expands and operating environments grow more complex, substation maintenance faces increasing technical and efficiency challenges [1], [2], [3]. An integrated air-ground inspection system combines ground inspection robots (IRs) with unmanned aerial vehicles (UAVs), leveraging UAVs for efficient, flexible, wide-area coverage and IRs for detailed equipment monitoring. This integration enables comprehensive three-dimensional inspection and maintenance. Real-time data transmission is critical for system operation. However, due to the geographic isolation of substations and strict security requirements, public networks often fail to meet the low-delay and high-security needs of on-site operations. Wireless Ad Hoc Networks (WANETs), with their self-organizing and dynamic topology features, offer an ideal communication infrastructure for air-ground collaborative inspection. In the WANET system, each node can transmit, receive, and relay data, enhancing system flexibility and robustness. WANET technology has shown great potential in various domains, including smart cities, agriculture, and power grids [4], [5], [6].

In a substation, UAVs conduct precise inspections of electrical equipment across wide and complex areas. At the same time, IRs can monitor in detail the infrastructure on the ground, forming a comprehensive air–ground collaborative framework [7]. In such dynamic environments, WANETs ensure stable data transmission, reliably delivering information collected by UAVs and IRs. When nodes are distant from terminals, data is transmitted via multi-hop routing. Lower transmission power increases hop count and delay, while higher power reduces delay, however, it increases interference, contention, and energy use, thereby reducing network efficiency. Due to the continuous movement of inspection devices, WANET topologies must dynamically adapt to changing positions. The integrated UAV-IR-WANET communication system is highly dynamic and uncertain, with a high-dimensional state space and vast solution space, making optimization particularly challenging. Furthermore, although higher transmission power improves communication quality,

Received 1 December 2025; revised 18 March 2026; accepted 26 March 2026. Date of publication 30 March 2026; date of current version 13 April 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 52277087, Grant 62173120, and Grant 52077049 and in part by the Natural Science Foundation of Anhui Province under Grant 2008085UD04. Recommended for acceptance by Prof. Geng Sun. *(Corresponding authors: Wei Sun; Zhi Liu.)*

Qingwei Tang is with the School of Electrical and Automation Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Electrical and Computer Engineering Department, National Technical University of Athens, 15773 Athens, Greece (e-mail: tangqingwei@mail.hfut.edu.cn).

Wei Sun is with the School of Electrical and Automation Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: wsun@hfut.edu.cn).

Zhi Liu is with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: liu@ieee.org).

Nikos D. Hatziargyriou is with the Electrical and Computer Engineering Department, National Technical University of Athens, 15773 Athens, Greece (e-mail: nh@power.ece.ntua.gr).

Yang Xiao is with the Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: yangxiao@ieee.org).

Xiaohui Yuan is with the Department of Science and Engineering, University of North Texas, Denton, TX 76207 USA (e-mail: xiaohui.yuan@unt.edu).

Digital Object Identifier 10.1109/TNSE.2026.3678873

it also increases energy consumption and reduces device endurance [8], [9], reflecting an inherent trade-off among these performance metrics. The system's many interdependent decision nodes further complicate control, a disturbance in one node can trigger cascading effects, leading to performance instability. These globally coupled dynamics require joint optimization of conflicting objectives-such as energy efficiency, delay, and link quality-under stringent real-time and stability requirements. This paper designs a robust real-time control strategy for a UAV-IR-WANET system.

### B. Contribution

(1) This paper formulates the air–ground integrated substation inspection process as a distributed partially observable Markov decision process (Dec-POMDP) with multi-objective coupling, incorporating key states such as delay, energy consumption, and communication quality.

(2) To handle high-dimensional dynamics and multi-node coupling, a Lyapunov-stable multi-agent reinforcement learning (MARL) framework is proposed, featuring a structurally constrained monotonic neural network controller to ensure convergence and stability.

(3) A topology-aware graph attention Critic combining GCN and self-attention is designed to capture spatial dependencies and coordination, while a topology-aware gradient regularization mechanism with neighborhood consistency enhances cooperative training stability.

(4) A 3D substation communication environment is constructed for evaluation. The proposed method is benchmarked against mainstream MARL baselines under routine and specialized inspection tasks.

### C. Paper Organization

This paper is organized as follows: Section II introduces related work; Section III describes the air-ground inspection communication environment and multi-objective optimization model; Section IV outlines the problem formulation and solution framework; Section V presents the stability-guaranteed RL controller and topology-aware graph attention Critic; Section VI provides simulation validation; Section VII concludes the work.

## II. Related Works

### A. Mathematical Optimization Methods

Mathematical modeling and optimization have been widely adopted in wireless communications for resource allocation, topology control, and scheduling due to their rigor and interpretability [3], [10], [11]. Early studies developed cross-layer optimization frameworks to jointly handle routing, power control, and scheduling [12], and further leveraged convex optimization to explicitly capture Physical–MAC coupling for energy-efficient allocation [13]. Energy-efficiency maximization has also been formulated via fractional programming, enabling analytically tractable solutions under QoS constraints [14]. To meet ultra-reliable and low-latency communication (URLLC) requirements, reliability and latency constraints were incorporated into convex cross-layer designs for precise service regulation [15]. Beyond convex settings, several works combined combinatorial and continuous optimization to address multi-hop routing and power allocation under interference cancellation [16], while mixed-integer nonlinear programming has been used for joint subcarrier assignment, scheduling, and power control [17]. With increasing spatial-resource complexity, robust optimization and convex relaxation have been applied to UAV trajectory/beamforming/power design [18], and semidefinite programming has been adopted for phase-shift and beamforming optimization in IRS-assisted systems [19]. Comprehensive overviews of convex and distributed optimization for communication networks can be found in [20]. For energy-efficient network optimization, sequential fractional programming has been used to iteratively convexify and solve efficiency-maximization problems [21], while global optimization frameworks further extend power control and receiver design [22]. Distributed approaches have also been explored, such as congestion-aware routing–scheduling algorithms that improve throughput under fairness and energy constraints [23], and column-generation-based joint scheduling and power control schemes that enhance energy efficiency with near-optimal performance [24].

Despite their theoretical advantages, these methods encounter significant challenges in large-scale, dynamic, or uncertain networks, where constructing accurate models is difficult and computational demands are high, thereby hindering real-time adaptability. Moreover, their strong reliance on precise system modeling restricts flexibility and limits their ability to generalize across diverse scenarios.

### B. Heuristic and Evolutionary Optimization Methods

Heuristic algorithms aim to efficiently obtain near-optimal solutions by leveraging empirical rules and approximation strategies, and have been widely adopted for complex non-convex and NP-hard problems. To improve global exploration and convergence, [25] proposed a hybrid optimizer that integrates an improved Whale Optimization Algorithm with Grey Wolf Optimizer, where dynamic weighting and pseudo-opposition-based learning enhance solution quality in both benchmarks and wireless-network optimization. For IoT service deployment, [26] developed a quantum particle swarm optimization method with quantum encoding and dual-hashing decoding, enabling joint optimization of throughput, energy consumption, delay, and computational load. In [27], a dictionary-based search combined with Pareto-relative dominance achieves energy-efficient IoT service composition while maintaining user satisfaction. Beyond IoT, heuristic search has also been extensively explored in wireless sensor networks (WSNs). [28] introduced a heuristic joint rate and resource allocation strategy to extend network lifetime, while [29] developed an ant colony optimization framework for routing and energy management via pheromone-guided path exploration. A comprehensive survey in [30] summarized particle swarm optimization applications in deployment, localization, clustering, and data aggregation, highlighting the suitability of swarm-intelligence heuristics for WSN

optimization. Bio-inspired heuristics were further applied in [31] to improve intrusion detection using an enhanced V-detector negative-selection mechanism. To address coverage-related objectives, [32] adopted a multi-species evolutionary algorithm for visual sensor networks, and [33] proposed a multifactorial evolutionary framework that simultaneously optimizes coverage, energy, and lifetime via cross-task genetic transfer. In addition, multi-objective evolutionary methods have been investigated for joint deployment and power assignment [34], topology and scheduling co-design [35], and beamforming with power control [36], demonstrating the effectiveness of population-based heuristics in balancing interference mitigation, coverage, and energy efficiency.

Although heuristic and evolutionary algorithms offer strong flexibility and perform well in large-scale, nonlinear, and constrained environments, their effectiveness often depends on empirical design choices. These methods typically lack theoretical convergence guarantees and are susceptible to local optima, particularly under dynamic or uncertain conditions. Moreover, their performance is highly sensitive to parameter tuning, which can reduce robustness and limit their generalizability in real-world applications.

### C. Machine Learning Optimization Methods

In recent years, machine learning (ML) has been extensively applied in wireless and communication networks, providing an effective way to balance system performance and computational complexity [37], [38], [95], [96]. For example, [39] proposes a two-stage self-supervised framework for multi-hop network capacity optimization via structure classification and application identification. Reinforcement learning (RL), which learns decision policies through interactions with the environment, has been widely adopted in robotics and autonomous driving [40], [41], and is gaining increasing attention in topology optimization due to its adaptability and reduced modeling dependency. For instance, [42] develops an RL-based power-control scheme for indoor base stations to maintain target SINR levels. In contrast, [43] introduces a deep RL-assisted local search with permutation-equivariant networks for intelligent neighborhood selection. In [44], deep RL is integrated with graph neural networks (GNNs) for routing, mapping traffic-allocation decisions to element-wise GNN outputs. An RL-based adaptive transmission framework is proposed in [45] to handle nonstationary channels and improve spectral efficiency under Bit Error Rate (BER) constraints. Beyond these, ML/RL techniques have also been applied to scheduling and spectrum utilization [46], large-scale 6G resource management [47], UAV trajectory and communication co-optimization [48], interference-aware device-to-device (D2D) allocation [49]. Joint optimization of power allocation and intelligent reflecting surface (IRS) configurations in OFDM systems [50]. Moreover, [51] investigates DRL-based network slicing for massive MIMO systems under dynamic traffic. Lightweight predictors have been explored for energy-efficient radio access network (RAN) slicing and operational overhead reduction [52], as well as intelligent access-point activation and load balancing [53]. To accelerate training, experience-retention mechanisms are incorporated into DRL for beyond-5G resource allocation [54]. Finally, ML-based optimization frameworks have been summarized for IoT-driven wireless systems [55], wireless sensor networks [56], and massive wearable-device spectrum allocation [57].

It is worth noting, in [93], the authors propose a DRL-based resource scheduling method for UAV-assisted emergency communication networks, while [94] develops a Transformer-based UAV trajectory planning approach for AoI-minimal data collection in UAV-aided IoT networks. In contrast, this paper focuses on distributed online transmit power control for substation inspection Air–Ground–WANETs, jointly optimizing end-to-end delay, endurance, and communication quality under dynamic multi-hop topologies.

ML-based methods have shown strong effectiveness in dynamic and unstructured environments, supporting end-to-end policy learning and autonomous decision-making. However, the inherent stochasticity of their training processes often leads to instability in convergence and limited interpretability.

### D. Stability-Assured Optimization Methods

Recent research has increasingly emphasized optimization frameworks with explicit and theoretically grounded stability guarantees, driven by the demand for reliable performance in highly dynamic wireless and edge computing environments. In [58], a Lyapunov drift-plus-penalty–based edge resource management method is developed to minimize long-term energy consumption while ensuring queue stability. In [59], an energy-efficient goal-oriented communication scheme is proposed, where Lyapunov optimization stabilizes multi-user service queues under dynamic task arrivals. In [60], a stochastic queueing model with per-slot Lyapunov control is applied to Low Earth Orbit (LEO)-assisted mobile edge computing, guaranteeing bounded task queues despite fluctuating demands. In [61], a Lyapunov-guided deep reinforcement learning framework is introduced to achieve provable delay stability for proactive sensing-task offloading at roadside units. In [62], a mobility-aware satellite edge computing strategy is formulated using Lyapunov control to maintain stable offloading and energy–delay performance under uncertain satellite–ground conditions. In [63], a two-timescale hierarchical optimization framework is presented and analytically proven stable through cross-layer stability analysis, enabling robust service deployment and task scheduling. In [64], a robust stability-oriented routing method is developed for LEO satellite networks, addressing unpredictable link and topology variations. In [65], a UAV-assisted computing framework integrates trajectory optimization with a stability-guaranteed offloading mechanism to maintain steady delay queues in dynamic airborne environments. In [66], an industrial IoT resource allocation scheme with built-in stability control is proposed, ensuring reliable operation under fluctuating traffic loads. In [67], a distributed constrained optimization method is designed with formal stability and convergence guarantees for privacy-preserving federated learning over time-varying communication graphs. In [68], a stability-constrained peer-offloading strategy is introduced for satellite edge computing,

TABLE I
SUMMARY OF REPRESENTATIVE COMMUNICATION OPTIMIZATION METHODS

| Studies | Year | Method Categories | Core Mechanism | Limitations |
|---|---|---|---|---|
| Johansson et al.[12] | 2006 | Mathematical Optimization Methods | Nonlinear column-generation | High computational cost; complex model |
| Pollin et al.[13] | 2008 | | Convex optimization capturing coupling | Requires accurate modeling |
| Meshkati et al.[14] | 2009 | | Sequential fractional programming under QoS constraints | Limited to static or slow-varying networks |
| She et al.[15] | 2018 | | Finite blocklength reliability | High computational complexity |
| Xu et al.[19] | 2020 | | Convex relaxation + alternating optimization | Computationally intensive |
| Yang et al.[25] | 2023 | Heuristic and Evolutionary Optimization Methods | Dynamic weighting / pseudo-opposition learning | Parameter-sensitive; potential local minima |
| Bey et al.[26] | 2024 | | Quantum particle encoding / dual hashing decoding | High complexity in decoding and particle control |
| Lin et al.[29] | 2011 | | Pheromone-driven routing & energy exploration | Sensitive to pheromone parameter settings |
| Zhang et al.[32] | 2020 | | Co-evolution across multiple populations | Performance may degrade in highly dynamic contexts |
| Tam et al.[33] | 2021 | | Cross-task genetic transfer for shared optimization | Optimization quality depends on task relatedness |
| Zhang et al.[39] | 2023 | Machine Learning Optimization Methods | Self-supervised machine learning | Limited by pretext-task quality |
| Ding et al.[44] | 2024 | | DRL + GNN Routing | High training cost |
| Qin et al.[48] | 2023 | | DRL for UAV Integrated sensing and communications | Requires accurate simulation for training |
| Yan et al.[51] | 2023 | | DRL for Network Slicing | Potential convergence instability |
| Oliveira et al.[52] | 2023 | | Lightweight ML for radio access network | Limited accuracy in highly dynamic channels |
| Battiloro et al.[58] | 2023 | Stability-Assured Optimization Methods | Drift-plus-penalty for stable edge resource management | Single-agent; limited scalability |
| Ding et al.[60] | 2023 | | Per-slot Lyapunov control for LEO-assisted MEC | Relies on accurate queueing models |
| Zhao et al.[61] | 2024 | | Delay-stable Lyapunov-based reward shaping | High DRL training complexity |
| Tang et al.[63] | 2024 | | Cross-layer two-timescale stability analysis | Not directly applicable to multi-agent coupling |
| Wei et al.[67] | 2025 | | Stability-guaranteed distributed FL over dynamic graphs | Communication overhead in complex distributed settings |
| Ours | 2025 | Machine Learning/Stability-Assured Methods | Lyapunov-stable MARL with topology-aware mechanism | - |

ensuring balanced and stable task execution across heterogeneous nodes. In [69], a reliability- and stability-aware offloading mechanism is developed for integrated satellite–terrestrial networks, providing controllable performance under adversarial or resource disturbances.

However, many existing stability-oriented methods are primarily designed for single-agent or centrally controlled systems and often rely on simplified stability conditions that are not explicitly embedded into the multi-agent policy structure, making them less suitable for systems with coupled dynamics and rapidly varying topologies.

### *E. Communications Enabled by Unmanned Systems*

Recent years have seen a rapid growth of research on unmanned-system–enabled communication networks, in which UAVs and other autonomous platforms function as agile aerial nodes to enhance connectivity, coverage, and service continuity. Beyond performance improvement, an increasing body of work has begun to focus on stability-guaranteed network optimization, aiming to ensure reliable communication under mobility, channel fluctuations, and resource uncertainty. For instance, [70] provides a stability-oriented analysis of UAV-assisted communication, examining the interplay among mobility, channel variation, and resource allocation and summarizing guidelines for maintaining robust connectivity under dynamic flight conditions. [71] offers a survey of UAV–B5G client architectures, highlighting adaptive non-terrestrial access mechanisms that can mitigate link disruptions and throughput oscillations. [72] delivers an analytical overview of UAV-enabled communication models, discussing how coordinated topology management and interference-aware channel allocation improve network reliability. [73] reviews UAV-assisted communication paradigms for 6G networks, emphasizing regulation-aware deployment and robust beamforming as key means to enhance link stability. Building on these insights, [74] proposes a joint trajectory and resource optimization method for UAV-assisted systems that enhances communication stability by stabilizing latency, sensing quality, and data throughput. [75] introduces a multi-service UAV-enabled IoT access mechanism that ensures stable QoS across heterogeneous traffic through dynamic scheduling and interference control. [76] formulates a stability-oriented resource allocation scheme for UAV-enabled wireless-powered mobile edge computing (MEC) networks, leveraging hybrid passive–active communication to guarantee consistent energy supply and data-rate performance under mobility. [77] develops a UAV-assisted maritime communication framework that improves link stability via joint trajectory and power optimization to maintain reliable surveillance in highly volatile maritime channels.

However, despite these advances, most existing studies still do not provide explicit stability guarantees for air–ground collaborative communications, as their optimization and control mechanisms are not designed to maintain robust performance under coupled air–ground dynamics and rapidly varying topologies. Table I presents a concise comparison of representative works across the major categories.

## III. PRELIMINARIES

### *A. Problem Description*

Real-time equipment monitoring is crucial for reliable substation operation. We employ an integrated air–ground inspection system utilizing UAVs and IRs. Since public networks are often restricted due to security and geographical constraints [78], WANETs are adopted for their reliable, self-organizing capabilities. As shown in Fig. 1, the UAV and IR collect data along fixed routes and forward it to a terminal via multi-hop relay nodes.

We assume: (a) a common channel with adjustable power; (b) uniform antenna gain and sensitivity; (c) stable transmission; and (d) energy-constrained UAVs and IRs. Recognizing that power settings dictate the range-energy trade-off, this work aims to jointly optimize delay, endurance, and quality via dynamic power control strategies.

Fig. 2 illustrates two communication modes. In direct mode (A), devices transmit to the terminal at maximum power, leading to high energy consumption. Conversely, multi-hop mode (B) utilizes relays to lower power demand and extend endurance. Intelligent agents dynamically adjust power to ensure stable, energy-efficient, and real-time transmission.

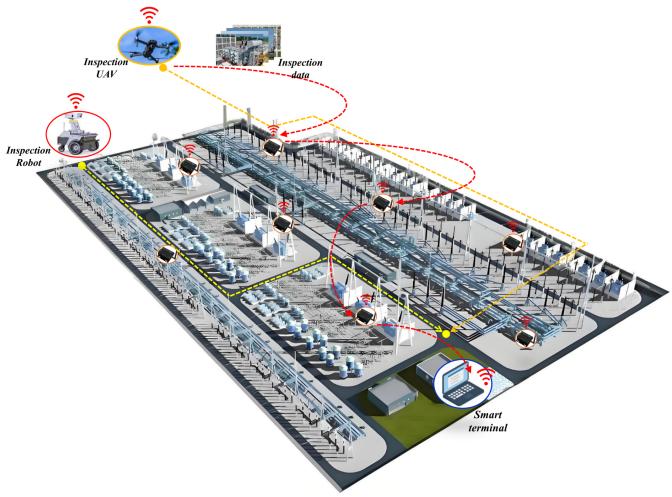


Fig. 1. Air-Ground-WANET substation inspection system.

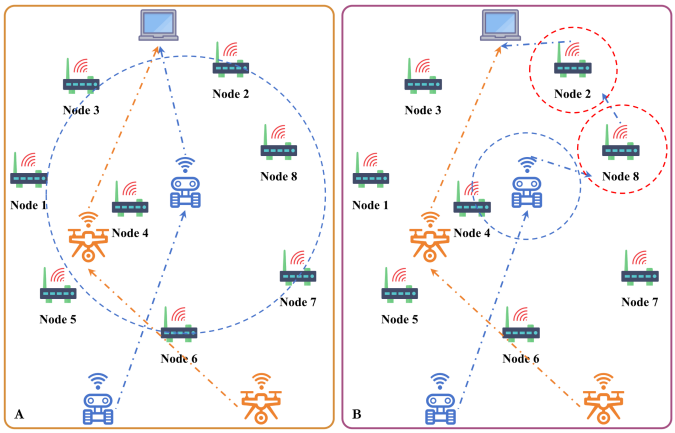


Fig. 2. Two operational modes for frame transmission during inspections.

## B. System Model

*1) Delay Modeling:* In WANET substations, E2E delay is a critical metric for evaluating data transmission efficiency. Delay is primarily affected by factors such as transmission power between nodes, the number of transmission hops, and MAC-layer channel contention. Proper power adjustment optimizes the path, minimizes E2E delay, and ensures reliability. This paper considers system delay as composed of three main parts.

*a) MAC Layer Delay:* The IEEE 802.15.4 protocol uses the CSMA/CA to minimize collisions and improve wireless efficiency. To enhance channel contention, we introduce a custom backoff mechanism. Each node starts at an initial backoff stage with contention window $\omega_0$. If the channel is busy, the stage increases $(i \to i+1)$, and the window expands exponentially: $\omega_i = 2^{\min(i,m)}\omega_0, \quad i \in \{0, 1, \ldots, \mathcal{H}-1\}$, where $m = 5$ is the maximum exponential threshold, and $\mathcal{H}$ is the maximum backoff depth. In each slot, if the channel is idle (probability $1 - \mathbb{P}_\alpha$), the backoff counter decreases; if busy (probability $\mathbb{P}_\alpha$), the node increments the stage and resets the counter. Further details are provided in [79]. The channel busy probability is given in (1).

$$\mathbb{P}_\alpha = 1 - \left(1 - e^{-\lambda \mathfrak{T}_{slot}}\right)^{\mathcal{N}_g^{(node)}} \tag{1}$$

where $\mathbb{P}_g = 1 - e^{-\lambda \mathfrak{T}_{slot}}$ is the probability that a node generates data within a time slot $\mathfrak{T}_{slot}$, and $\mathcal{N}_g^{(node)}$ denotes the number of neighboring nodes. The MAC-layer channel access delay can be derived from the steady-state distribution of a Markov chain model, as shown in [80] (4).

By solving (4) in [80], the average backoff time at stage $i$ can be derived as shown in (2).

$$\mathfrak{T}_b^{(i)} = \sum_{j=0}^{\mathcal{W}_i^{(b)}-1} j\omega_{(i,j)} \mathfrak{t}_s^{(\text{slot})}, i \in [0, \mathcal{H}] \tag{2}$$

where $\mathfrak{T}_b^{(i)}$ is the average duration in the $i$-th backoff stage; $\omega_{(i,j)}$ is the steady-state probability of being in state $(i, j)$; and $\mathfrak{t}_s^{(\text{slot})}$ is the fixed state transition time. More detailed information can be found in [81]. In summary, total MAC delay is the weighted sum of backoff times over all stages, as shown in (3).

$$\mathfrak{T}_{mac}^{(layer)} = \sum_{i=0}^{\mathcal{H}} \left[ \mathfrak{T}_b^{(i)} \sum_{j=0}^{2^{i+3}-1} \omega_{(i,j)} \right] \tag{3}$$

*b) Transmission Delay:* Based on the MAC access process in (a), we model the per-hop transmission delay as an expected value that incorporates potential retransmissions and the associated MAC access overhead. In a WANET, the transmission time $\mathfrak{T}_{tp}$ of a data packet depends on its size $\mathcal{L}_{packet}$ (bits) and the channel bandwidth $\mathfrak{B}_{ch}$, expressed in (4).

$$\mathfrak{T}_{tx}^{(link)} = \frac{\mathcal{L}_{packet}}{\mathfrak{B}_{ch} \log_2(1 + \text{SNR})} \tag{4}$$

In practical communication, data packets may be lost due to channel contention, interference, and noise. Thus, the transmission failure probability $\mathbb{P}_f$ must be considered. It depends on the node's channel contention probability $\mathbb{P}_g$ and the number of neighboring nodes $\mathcal{N}_g^{(node)}$, as shown in (5).

$$\mathbb{P}_f = \mathbb{P}_g \cdot (1 - \mathbb{P}_\alpha)^{\mathcal{N}_g^{(node)}} \tag{5}$$

where $\mathbb{P}_\alpha$ denotes the probability that the channel is busy. If a transmission fails, the data packet must be retransmitted, increasing the total transmission delay. Thus, considering both the channel contention delay $\mathfrak{T}_{mac}^{(layer)}$ and the effective transmission time, the expected total delay is given by (6).

$$\mathfrak{T}_{tp}^{(hop)} = \sum_{k=0}^{\infty} (1 - \mathbb{P}_f)(\mathbb{P}_f)^k \cdot \left( \mathfrak{T}_{tx}^{(link)} + k \cdot \mathfrak{T}_{mac}^{(layer)} \right) \tag{6}$$

*c) Queueing Delay:* Beyond access and transmission latency, relay buffering under stochastic arrivals introduces queueing delay, with the service rate determined by the MAC-layer packet processing time. In the WANET environment, the queueing delay at each node can be modeled as an M/M/1 queue. Packet arrivals follow a Poisson process with rate $\lambda$, and service times are exponentially distributed with rate $\mu = 1/\mathfrak{T}_{mac}^{(layer)}$, where $\mathfrak{T}_{mac}^{(layer)}$ is the MAC layer transmission time per packet. To

maintain stability, the load must satisfy $\rho^{(queue)} = \lambda/\mu < 1$, ensuring arrivals are slower than service. Under steady-state, the average queueing delay $\mathfrak{T}_q^{(node)}$, including both waiting and service time, is given by (7).

$$\mathfrak{T}_q^{(node)} = \frac{1}{\mu - \lambda} = \frac{\mathfrak{T}_{mac}^{(layer)}}{1 - \lambda \mathfrak{T}_{mac}^{(layer)}} \tag{7}$$

According to Little's theorem, the average queue length $\mathcal{A}_q$ and the queueing delay $\mathfrak{T}_q^{(node)}$ relate as shown in (8).

$$\mathcal{A}_q = \lambda \mathfrak{T}_q^{(node)} \tag{8}$$

In summary, the single-hop delay $\mathfrak{T}_{node}$ equals the sum of MAC delay, transmission time, and queueing delay: $\mathfrak{T}_{node} = \mathfrak{T}_{mac}^{(layer)} + \mathfrak{T}_{tx}^{(link)} + \mathfrak{T}_q^{(node)}$. The E2E delay over a multi-hop path is the total of all nodes' single-hop delays, as shown in (9).

$$\mathfrak{T}_{e2e} = \sum_{k \in Path} \mathfrak{T}_{node}^{(k)} \tag{9}$$

where $k$ denotes the $k$-th relay node along the data packet's transmission path from the source node to the terminal node.

*2) Transmission Power Modeling:* In the substation inspection environment studied, modeling and optimizing transmission power is key to balancing communication performance. We relate transmission power to communication distance using the logarithmic path loss model, as shown in (10).

$$\mathbf{\Phi}_t^{(i)} = \mathbf{\Phi}_r^{(i)} + \mathcal{L}_{path}\left(d_0\right) + 10\eta_{path} log_{10}\left(\frac{d}{d_0}\right) + \Xi_\sigma^{(shadow)} \tag{10}$$

The wireless channel model is based on an improved log-distance path loss model, comprising deterministic and stochastic components. In the deterministic part, $\mathbf{\Phi}_r^{(i)}$ denotes the received signal strength (dBm), and $\mathbf{\Phi}_t^{(i)}$ is the transmit power of node $i$. The reference path loss at baseline distance $d_0 = 1\,\mathrm{m}$, denoted $\mathcal{L}_{\text{path}}(d_0)$, depends on antenna characteristics and operating frequency. The distance between transmitter $i$ and receiver $j$ is $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2$, where $\mathbf{p}_i = [x_1, x_2, x_3]^\top$ and $\mathbf{p}_j = [x_1', x_2', x_3']^\top$ are their spatial coordinates. This geometry governs the spatial attenuation of electromagnetic waves. The path loss exponent $\eta_{\text{path}}$, set to 2.5 after calibration, reflects medium attenuation and increases with environmental complexity. The stochastic component $\Xi_\sigma^{(\text{shadow})} \sim \mathcal{N}(0, \sigma_\Xi^2)$ models signal variations due to shadow fading from multipath effects.

*3) Endurance Modeling:* First, a multi-dimensional UAV energy model is developed, capturing flight and communication energy consumption. Equation (11) details the flight energy, which includes blade profile, induced, and fuselage drag power components.

$$\mathcal{E}_i^{fly} = \underbrace{\mathcal{E}^{bla}\left(1 + \frac{3v_i^2}{e_{\text{tip}}^2}\right)}_{\text{blade profile power}} + \underbrace{\mathcal{E}^{\text{ind}}\left(\sqrt{1 + \frac{v_i^4}{4v_0^4}} - \frac{v_i^2}{2v_0^2}\right)^{\frac{1}{2}}}_{\text{induced power}} + \underbrace{\frac{1}{2} d^{fus} \rho^{den} s^{rot} d^{dis} v_i^3}_{\text{fuselage drag power}} \tag{11}$$

where $\mathcal{E}^{\text{bla}}$ and $\mathcal{E}^{\text{ind}}$ denote the blade profile power and induced power, respectively. Specifically, $p^{\text{bla}}$ is the power to overcome aerodynamic drag as the rotor blades spin, while $p^{\text{ind}}$ is the power loss from the downwash airflow generated during lift. Both depend on the UAV's aerodynamic parameters, as given in (12) and (13).

$$\mathcal{E}^{bla} = \frac{c^{pro}}{8} \rho^{den} s^{rot} r^{dis} v_{bla}^3 r_{rad}^3 \tag{12}$$

where, $c^{pro}$ denotes the blade profile drag coefficient; $\rho^{den}$ is the air density; $s^{\text{rot}}$ represents the rotor solidity (the ratio of total blade area to the rotor disc area); $r^{dis}$ is the rotor disc area; $r_{rad}$ denotes the rotor blade radius; and $v_{\text{bla}}$ is the blade angular velocity (rotational speed).

$$\mathcal{E}^{ind} = \frac{\left(1 + f^{inc}\right) w^{\frac{3}{2}}}{\left(2\rho^{den} r^{dis}\right)^{\frac{1}{2}}} \tag{13}$$

where $f^{inc}$ is the induced power increment factor (related to rotor design efficiency), and $w$ denotes the UAV's weight. The energy consumption for data packet transmission and reception is expressed as: $\mathcal{E}_{tc}^{(uav)} = \mathbf{\Phi}_t^{(uav)} \cdot \mathfrak{T}_{tx}^{(uav)} + \mathcal{E}_{circuit}^{(uav)}$, where $\mathbf{\Phi}_t^{(uav)}$ is the UAV's transmission power, $\mathfrak{T}_{tx}^{(uav)}$ is the transmission time, and $\mathcal{E}_{circuit}^{(uav)}$ is the baseline circuit power consumption. Therefore, the total energy consumption of the UAV is: $\mathcal{E}_{total}^{(uav)} = \mathcal{E}_i^{fly} + \mathcal{E}_{tc}^{(uav)}$.

A multi-dimensional energy consumption model is also developed for the IR, focusing on two key aspects: mobility energy consumption and data packet transmission/reception energy consumption. The mobility energy consumption consists of the motor's baseline power and the rolling friction power, as expressed in (14).

$$\mathcal{E}_i^{move} = \underbrace{\mathcal{E}^{base}}_{\text{basic motor power}} + \underbrace{\mathcal{E}^{rol}}_{\text{rolling frictional power}} \tag{14}$$

where $\mathcal{E}^{base}$ represents the baseline energy consumption to keep the motors idling and power the control systems and sensors. It is independent of the robot's speed and is given by: $\mathcal{E}^{base} = (I^{idle} V_{sys}) \cdot t_i$, where $I^{idle}$ is the motor no-load current, $V_{sys}$ the system voltage, and $t_i$ the operation time. $\mathcal{E}^{rol}$ denotes the energy required to overcome rolling friction between the wheels and the ground, proportional to the robot's weight and speed: $\mathcal{E}^{rol} = \frac{f^{rol} m^{rob} g v_i \cdot t_i}{\eta^{mot}}$, where $f^{rol}$ is the rolling resistance coefficient, $m^{rob}$ the robot's mass, $g$ the gravitational acceleration, $v_i$ the speed, and $\eta^{mot}$ the motor efficiency. Similar to the UAV, the IR's energy for data transmission and reception is: $\mathcal{E}_{tc}^{(ir)} = \mathbf{\Phi}_t^{(ir)} \cdot \mathfrak{T}_{tx}^{(ir)} + \mathcal{E}_{circuit}^{(ir)}$. Thus, the IR's total energy consumption is: $\mathcal{E}_{total}^{(ir)} = \mathcal{E}_i^{move} + \mathcal{E}_{tc}^{(ir)}$

*4) Communication Quality Modeling:* Communication quality critically influences data reliability and system stability. This section models WANET communication quality and examines how transmission power affects performance.

In wireless systems, signal quality is evaluated by the Signal-to-Noise Ratio (SNR), which depends on received signal strength. Since transmission power directly impacts the received signal, the received power based on the logarithmic path loss model is expressed in (15).

$$\mathbf{\Phi}_r^{(i)} = \mathbf{\Phi}_t^{(i)} - \mathcal{L}_{path}\left(d_0\right) - 10\eta_{path}\log_{10}\left(\frac{d}{d_0}\right) - \Xi_{\sigma}^{(shadow)} \tag{15}$$

Based on the above equation, the SNR, which serves as a key indicator of signal quality, can be expressed as (16).

$$\Gamma_{ui}^{E2E} = \frac{\mathbf{\Phi}_r^{(i)}\varpi_u^y}{\sum_{q\neq u}^{N}\varrho_{q,y} + \varsigma_{E2E}^2} \tag{16}$$

where $\varpi_u^y$ denotes the channel gain between nodes, accounting for path loss and antenna gain. The term $\sum_{q\neq u}^{N}\varrho_{q,y}$ represents co-channel interference from other inspection devices, while $\varsigma_{E2E}^2$ is the thermal noise power. A higher SNR generally improves data transmission quality, however, excessively high SNR requires increased transmission power, which may shorten node battery life and cause interference to nearby nodes. Therefore, determining the optimal transmission power is crucial when adjusting signal strength.

## IV. Problem Formulation

This paper focuses on an air-ground integrated inspection system in substation scenarios. The goal is to dynamically adjust the transmission power of inspection devices (UAVs and IR) and WANET relay nodes to enable efficient multi-hop data transmission, while optimizing E2E communication delay, device endurance, and overall communication quality. As detailed in Section III, this problem is formulated as a multi-objective dynamic optimization problem posed as a vector optimization, as presented in (17)–(23). Specifically, we denote $x(t)$ as the system state and $\Phi(t)$ as the decision variable.

$$\min_{\mathbf{\Phi}_t^{(i)}} \mathbf{F}(\mathbf{\Phi}_t^{(i)}) = [\mathfrak{T}_{e2e}, -\mathcal{E}, -\Gamma_{avg}]^{\top} \tag{17}$$

$$\text{s.t.}\quad C1: \quad \mathbf{\Phi}_t^{(i)\min} \leq \mathbf{\Phi}_t^{(i)}[t] \leq \mathbf{\Phi}_t^{(i)\max} \tag{18}$$

$$C2: \quad \mathbf{\Phi}_t^{(i)} \geq \mathbf{\Phi}_{\min(\text{receive})} + \mathcal{L}_{path}(d_0) + 10\eta_{path}\log_{10}\left(\frac{d_{ij}}{d_0}\right) \tag{19}$$

$$C3: \underline{\mathfrak{T}} \leq \mathfrak{T}_{e2e} \leq \overline{\mathfrak{T}} \tag{20}$$

$$C4: \quad \left\|\Delta_t \mathbf{\Phi}_t^{(i)}[t]\right\| \leq \epsilon \tag{21}$$

$$C5: \quad \mathcal{E}_{Res}^{(uav/ir)}[t] = \mathcal{E}_{Ini}^{(uav/ir)}[0] - \int_0^t \mathcal{E}_{total}^{(uav/ir)}[\tau]d\tau \geq 0 \tag{22}$$

$$C6: \quad \Gamma_{avg} \geq \Gamma_{\min} \tag{23}$$

To ensure that the dynamic power optimization in the substation inspection system meets physical feasibility, communication reliability, and multi-objective coordination, this paper defines the following constraints, C1: Power range constraint. The transmission power of node $i$ at time $t$ is $\mathbf{\Phi}_t^{(i)}[t]$, bounded by the device's minimum and maximum power limits $\mathbf{\Phi}_t^{(i)\min}$ and $\mathbf{\Phi}_t^{(i)\max}$, ensuring power stays within hardware capabilities. C2: Topology connectivity constraint. C3: Constrains the E2E delay $\mathfrak{T}_{e2e}$ within $[\underline{\mathfrak{T}}, \overline{\mathfrak{T}}]$ to ensure timely task execution without violating delay requirements. C4: Power variation rate constraint, limits the instantaneous rate of power change not to exceed a threshold $\epsilon$, preventing communication fluctuations. C5: Endurance constraint, the remaining energy of UAV/IR devices satisfies $\mathcal{E}_{Res}^{(uav/ir)}[t] = \mathcal{E}_{Ini}^{(uav/ir)}[0] - \int_0^t \mathcal{E}_{total}^{(uav/ir)}[t]\,dt \geq 0$, ensuring the residual energy is always non-negative, where $\mathcal{E}_{Ini}^{(uav)}$ denotes the initial Energy and the integral term is the cumulative energy consumption. C6: Lower bound on communication quality. Requires the average communication quality metric to remain above a specified threshold.

As described in Section III, based on the multi-dimensional coupled characteristics of the system state, the impact of power adjustment on delay, energy consumption, and communication quality can be simplified and modeled as a discrete-time linear dynamic system in state-space form, expressed as: $\mathbf{x}(t+1) = \mathbf{H}\mathbf{x}(t) + \mathbf{I}\mathbf{\Phi}(t) + \mathbf{w}(t)$. where, the system state is abstracted as $\mathbf{x}(t)$, simplified as $\mathbf{x}(t) = (\mathfrak{T}_{e2e}, \mathcal{E}_{Res}^{(uav/ir)}, \Gamma_{avg})^{\top}$, and the control input vector is $\mathbf{\Phi}(t) = [\Phi_t^{(1)}(t), \Phi_t^{(2)}(t), \ldots, \Phi_t^{(n)}(t)]^{\top}$. $\mathbf{H} \in \mathbb{R}^{3\times 3}$ is the state transition coefficient matrix. Note that we introduce this simplified linear state-space model for formal expression. The matrix $\mathbf{H} = \text{diag}(\hbar_1, \hbar_2, \hbar_3)$ represents the temporal autocorrelation of each state variable and is approximated as a diagonal structure to simplify theoretical analysis and modeling. Its specific values are implicitly captured during the learning process. $\mathbf{w}$ denotes a stochastic process noise term. Furthermore, the model can be represented as a closed-loop multidimensional control dynamic equation: $\mathbf{x}(t+1) = \mathbf{I}\mathbf{\Phi}(t) + \mathbf{x}^{env}(t)$, where the control input $\mathbf{\Phi}(t)$ is defined as: $\mathbf{\Phi}_t^{(i)}[k+1] = \Phi_t^{(i)}[k] + \Delta T \cdot \pi_i(\mathcal{S}_t)$, with $\mathcal{S}_t$ representing the generalized state input.

## V. Problem Solution

In this section, we reformulate the sequential nonconvex optimization problem as a Dec-POMDP to support real-time, distributed decision-making. Building on this, we develop a stable MATD3 framework [89], as illustrated in Fig. 3. The proposed framework incorporates Lyapunov-based constraints and a topology-aware critic to enforce closed-loop stability and to capture dependencies induced by dynamic network graphs.

### A. Multi-Agent Reinforcement Learning Framework

In the dynamic optimization of the substation air-ground integrated inspection system, we model the power adjustment process as a Dec-POMDP and solve it using Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MATD3). In our multi-agent setup, all agents share a joint environment state $s_t \in \mathcal{S}$. Each agent $i$ selects an action $a_t^i \in \mathcal{A}^i \subseteq \mathbb{R}^{d_i}$, and the joint action is $\mathbf{a}_t = (a_t^1, a_t^2, \ldots, a_t^N) \in \mathcal{A} = \prod_{i=1}^{N}\mathcal{A}^i$.

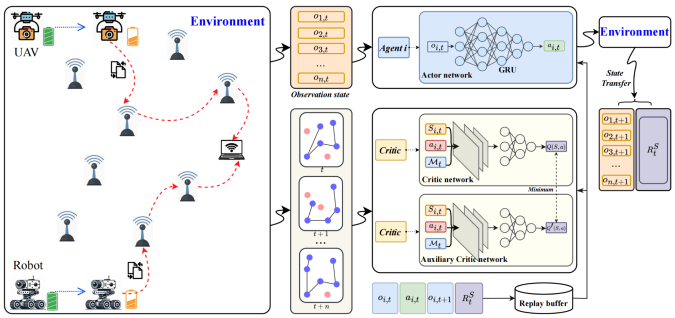


Fig. 3. Overall framework of the method.

The environment evolves according to the transition probability $\mathcal{S}_{t+1} \sim \mathcal{P}(\mathcal{S}_{t+1} \mid \mathcal{S}_t, \mathbf{a}_t)$. Each agent employs an independent policy network $\mu_{\theta_i} : \mathcal{O}^i \to \mathcal{A}^i$, mapping local observations to continuous actions. To improve training stability and reduce Q-value overestimation, each agent also maintains twin Q-networks $\{\mathcal{Q}^{(1)}_{\phi_i^1}(\mathcal{S}, \mathbf{a}), \mathcal{Q}^{(2)}_{\phi_i^2}(\mathcal{S}, \mathbf{a})\}$. After sampling tuples $(\mathcal{S}_t, \mathcal{O}^i_t, \mathbf{a}_t, r^i_t, \mathcal{S}_{t+1}, \mathcal{O}^i_{t+1})$ from the replay buffer, the target Q-value is computed as $y^i_t = r^i_t + \gamma \cdot \min_{k=1,2} \mathcal{Q}^{(k)}_{\phi_i^k}(\mathcal{S}_{t+1}, \mathbf{a}_{t+1})$, where $\mu_{\theta_i}$ and $\mathcal{Q}^{(k)}_{\phi_i^k}$ denote the delayed target networks. The critic networks are updated by minimizing the mean squared error loss, as given in (24).

$$\mathcal{L}_{\phi_i^{(k)}} = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\left(\mathcal{Q}_{\phi_i^{(k)}}(\mathcal{S}_t, \mathbf{a}_t) - y^i_t\right)^2\right], k = 1, 2 \tag{24}$$

The actor network employs a delayed update mechanism, aiming to maximize the state-action value function of agent $i$. The objective function is given by (25).

$$\mathcal{J}_{\theta_i} = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\mathcal{Q}_{\phi_i^{(1)}}\left(\mathcal{S}_t, \mu_{\theta_i}\left(\mathcal{O}^i_i\right), \mathbf{a}^{-i}_t\right)\right] \tag{25}$$

where $\mathbf{a}^{-i}_t$ denotes the actions currently taken by all agents except agent $i$. The parameter updates are performed via gradient ascent as described in (26).

$$\nabla_{\theta_i}\mathcal{J}_{\theta_i} \approx \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\nabla_{\theta_i}\mu_{\theta_i}\left(\mathcal{O}^i_i\right) \cdot \nabla_{a^i}\mathcal{Q}_{\phi_i^{(1)}}\left(\mathcal{S}_t, a^i, \mathbf{a}^{-i}_t\right)\big|_{a^i=\mu_{\theta_i}(\mathcal{O}^i_i)}\right] \tag{26}$$

### B. Markov Decision Process

To transform the dynamic power adjustment problem in the substation inspection scenario into an RL task, we model it as a Dec-POMDP. This process is defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{N}$ is the set of agents (UAV, IR, and WANET relay nodes); $\mathcal{S}$ denotes the joint state space; $\mathcal{A}$ represents the continuous action space; $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ is the multi-objective reward function; $\gamma \in [0, 1)$ is the discount factor.

*1) State Space:* The system state space $\mathbf{S}$ is defined as the Cartesian product of delay, transmission power, energy, and communication quality states: $\mathbf{S} = \mathbf{T} \times \mathbf{\Phi} \times \mathbf{E} \times \mathbf{\Gamma}$, where $\mathbf{T}$ represents the system delay state. The total system delay in the WANET is expressed as $\mathbf{T} = \{\sum_{i=1}^{n} \mathfrak{T}^{(i)}_{node} : \mathfrak{T}^{(i)}_{node} \in (0, \mathbf{d})\}$, where $\mathfrak{T}^{(i)}_{node}$ denotes the single-hop delay at node $i$. $\mathbf{\Phi}$ denotes the transmission power state, including the real-time transmit power of UAV/IR devices and all WANET relay nodes: $\mathbf{\Phi} = \{\Phi^{(i)}_t \mid \Phi^{(i)}_t \in [\Phi^{(i)\,\min}_t, \Phi^{(i)\,\max}_t]\}$. $\mathbf{E}$ quantifies the energy state of inspection devices, representing the residual battery levels normalized between 0 and 1: $\mathbf{E} = \{\mathcal{E}^{(uav)}_{Res}, \mathcal{E}^{(ir)}_{Res}\} \subseteq [0, 1]^2$. $\mathbf{\Gamma}$ evaluates link stability and interference levels, defined as: $\mathbf{\Gamma} = \{\Gamma_{avg}, \min_{k\in\mathcal{P}_{active}} \Gamma^{(k)}_{SNR}\}$, where $\min_{k\in\mathcal{P}_{active}} \Gamma^{(k)}_{SNR}$ represents the minimum SNR among all nodes on the critical communication path $\mathcal{P}_{active}$, which determines communication reliability.

*2) Observation Space:* In the distributed MARL architecture, each agent-including UAVs, IRs, and wireless ad hoc network nodes-has a local observation space designed to encompass four key state dimensions: 1. E2E system total delay $\mathfrak{T}_{e2e}$, representing task execution efficiency; 2. Node's real-time transmission power $\Phi^{(i)}_t$, reflecting communication energy consumption; 3. Residual energy of mobile inspection devices $\{\mathcal{E}^{(uav)}_{Res}, \mathcal{E}^{(ir)}_{Res}\}$, indicating the sustainable operation time of the devices; 4. Composite communication quality metrics $\{\Gamma_{avg}, \min_{k\in\mathcal{P}_{active}} \Gamma^{(k)}_{SNR}\}$. This observation space, by capturing multidimensional state information, provides a comprehensive environmental feature representation for the agents' cooperative decision-making. The observation space serves as the actual input for each agent to perform real-time decisions. For example, for Agent 1, the real-time observation at time $t$ is expressed as $\mathbf{O}_{1,t} = \{\mathfrak{T}_{e2e}, \Phi^{(1)}_t, \mathcal{E}^{(uav)}_{Res}, \mathcal{E}^{(ir)}_{Res}, \Gamma_{avg}, \min_{k\in\mathcal{P}_{active}} \Gamma^{(k)}_{SNR}\}$.

*3) Action:* To achieve multi-node collaborative power optimization in the substation inspection scenario, this paper constructs a distributed agent architecture based on RL. For the heterogeneous network system composed of WANET nodes and mobile inspection devices, a discrete-time dynamic control equation is defined as: $\mathbf{\Phi}^{(i)}_t[k+1] = \mathbf{\Phi}^{(i)}_t[k] + \Delta T \cdot \pi_i(\mathcal{S}^{(i)}_t)$, where $\Delta T$ is the control period. Each agent generates a power adjustment value constrained by hardware limits through its policy network $\boldsymbol{\pi}_\theta$. Specifically, each agent optimizes the power increment in real time via the policy $\pi_\theta : \mathcal{S} \to \mathcal{A}$, where the action space of agent $i$ is $\mathcal{A}_i = \{\Delta a \in \mathbb{R}^+ \mid -c \le \Delta a \le c\}$, and $c$ denotes the maximum transmit power increment the agent can generate. The power update rule is defined as: $a^{k+1}_i = a^k_i + \Delta a$.

*4) Reward:* In dynamic power optimization for air-ground integrated substation inspection, balancing E2E delay, device endurance, and communication quality is vital for efficient and stable performance. To enable agents to handle multi-objective trade-offs, we design a dynamically weighted composite reward function that adapts to real-time conditions and the time-varying complexity of the scenario, as defined in (27).

$$\mathcal{R}(\mathcal{S}, \mathcal{A}) = \omega_d \cdot \exp\left(-\frac{\mathfrak{T}_{\text{ref}} - \mathfrak{T}_{\text{e2e}}}{\mathfrak{T}_{\text{ref}}}\right) + \omega_\Phi \cdot \left(\frac{\Delta\Phi^{(\text{uav/ir})}_t}{1 + \Delta\Phi^{(\text{uav/ir})}_t}\right)$$

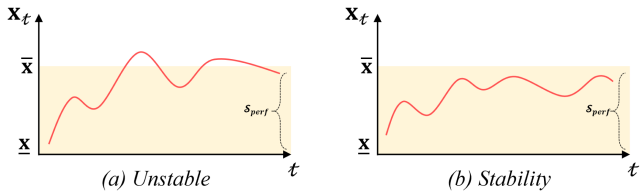


Fig. 4. Definition of system stability.

$$+\omega_e \cdot \ln\left(1+\frac{1}{\mathcal{E}_{\text{total}}+\epsilon}\right)+\omega_q \cdot \left(\frac{\Gamma_{\text{avg}}-\Gamma_{\min}}{\Gamma_{\max}-\Gamma_{\min}}\right) \tag{27}$$

The reward function $\mathcal{R}(\mathcal{S},\mathcal{A})$ integrates key performance metrics-communication delay, energy consumption. The first term, $e^{-(\frac{\mathfrak{T}_{\text{ref}}-\mathfrak{T}_{\text{e2e}}}{\mathfrak{T}_{\text{ref}}})}$, captures E2E delay using exponential decay to encourage low delay. The second term, $\frac{\Delta\Phi_t^{(uav/ir)}}{1+\Delta\Phi_t^{(uav/ir)}}$, measures transmission power variation, penalizing sudden changes to reduce energy spikes and prolong operation. The third term, $\ln(1+\frac{1}{\mathcal{E}_{total}+\epsilon})$, evaluates total energy usage, where lower consumption yields higher reward; $\epsilon$ is a numerical stability constant, and the log form prevents gradient explosion. The fourth term, $\frac{\Gamma_{\text{avg}}-\Gamma_{\min}}{\Gamma_{\max}-\Gamma_{\min}}$, normalizes average channel quality to promote reliable communication. The reward weights $\omega_d,\omega_\Phi,\omega_e,\omega_q$ are empirically tuned through multiple trials to balance the multi-objective optimization goals and ensure stable training performance.

### C. Joint Inspection System Stability Conditions

Based on the hybrid systems theoretical framework, we propose a stability-guaranteed DRL architecture to address multi-objective optimization and control problems in substation collaborative inspection tasks. Communication stability is defined as the maintenance of key performance indicators within acceptable bounds, as detailed in Definition 1. The core of the proposed method lies in constraining the policy search space within the feasible region defined by Lyapunov stability analysis and establishing convergence criteria based on the generalized LaSalle's invariance principle. A set of necessary and sufficient conditions is then derived to ensure the stable operation of the closed-loop system. This theoretical framework introduces a state-dependent constraint mechanism that confines the RL exploration process within a positively invariant region of attraction, thereby ensuring that system responses during policy optimization consistently converge to the neighborhood of the equilibrium point. Fig. 4 shows the stable and unstable states of the system.

*Definition 1. (Multi-Metric Stabilization Set):* A closed-loop air-ground integrated multi-hop communication system is said to be multi-metric stable if, for any initial system state $(\mathfrak{T}_{\text{e2e}}(0),\mathcal{E}(0),\Gamma_{\text{avg}}(0))$, its trajectory $(\mathfrak{T}_{\text{e2e}}(t),\mathcal{E}(t),\Gamma_{\text{avg}}(t))$ eventually converges to a feasible performance set $\mathcal{S}_{\text{perf}}\subset\mathbb{R}^3$, which is defined as: $\mathcal{S}_{\text{perf}}=\{(\mathfrak{T}_{\text{e2e}},\mathcal{E},\Gamma_{\text{avg}})\in\mathbb{R}^3\,|\,\underline{\mathfrak{T}}\le\mathfrak{T}_{\text{e2e}}\le\overline{\mathfrak{T}},\ \mathcal{E}\ge\underline{\mathcal{E}},\ \Gamma_{\text{avg}}\ge\underline{\Gamma}\}$. If the system satisfies the following convergence condition: $\lim_{t\to\infty}\text{dist}\,((\mathfrak{T}_{\text{e2e}}(t),\mathcal{E}(t),\Gamma_{\text{avg}}(t)),\ \mathcal{S}_{\text{perf}})=0$, then the system is considered multi-metric stable. where, the distance function is defined as: $\text{dist}(\mathbf{x},\mathcal{S}_{\text{perf}})=\min_{\mathbf{x}'\in\mathcal{S}_{\text{perf}}}\|\mathbf{x}-\mathbf{x}'\|$, where $\mathbf{x}=(\mathfrak{T}_{\text{e2e}},\mathcal{E},\Gamma_{\text{avg}})$.

To ensure system stability during constrained RL, we use Lyapunov stability theory to limit the policy search to controllers satisfying stability conditions. Specifically, we derive and analyze stability based on the Lyapunov direct method combined with a generalized LaSalle's invariance principle.

*Proposition 1:* (LaSalle's Theorem for Discrete-Time Systems [82]). Consider a discrete-time dynamical system $x(t+1)=f(x(t))$, and suppose there exists a continuously differentiable function $\mathcal{V}:\mathbb{R}^n\to\mathbb{R}$ such that: 1. Non-negativity: $\mathcal{V}(x)\ge 0$ for all $x\in\mathbb{R}^n$; 2. Non-increasing property: $\mathcal{V}(f(x))-\mathcal{V}(x)\le 0$ for all $x\in\mathbb{R}^n$. Define the set $\mathcal{E}:=\{x\in\mathbb{R}^n\mid\mathcal{V}(f(x))-\mathcal{V}(x)=0\}$, and let $\mathcal{M}\subseteq\mathcal{E}$ be the largest invariant set contained in $\mathcal{E}$ (i.e., all trajectories that start in $\mathcal{M}$ remain in $\mathcal{M}$ for all time). If there exists a constant $a\in\mathbb{R}^+$ such that the sublevel set $\mathcal{L}_a:=\{x\in\mathbb{R}^n\mid\mathcal{V}(x)\le a\}$ is bounded, then for any initial state $x(0)\in\mathcal{L}_a$, the trajectory $x(t)$ asymptotically converges to $\mathcal{M}$, that is: $\lim_{t\to\infty}\text{dist}(x(t),\mathcal{M})=0$. Furthermore, if $\mathcal{V}(x)\to\infty$ as $\|x\|\to\infty$ (i.e., $\mathcal{V}$ is radially unbounded), then the above result holds for any initial state $x(0)\in\mathbb{R}^n$.

To ensure the stability of the proposed control algorithm, it is crucial to identify a control policy $\boldsymbol{\pi}=-\psi_\theta(\mathcal{S}_t)$ and a Lyapunov function $\mathcal{V}$ that jointly satisfy the stability conditions stated in Proposition 1. Since the control policy depends on the system state $\mathcal{S}_t$, the system dynamics can be approximated as $\mathbf{x}(t+1)=\mathbf{x}(t)-\psi_\theta(\mathcal{S}_t):=\mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}(t))$. To facilitate the stability analysis, we construct a quadratic Lyapunov function as shown in (28), where $A$ is a positive definite identity matrix.

$$\mathcal{V}(\mathbf{x})=(\mathbf{x}-\mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}))^\top A^{-1}(\mathbf{x}-\mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x})) \tag{28}$$

According to the LaSalle's invariance principle stated in Proposition 1, if the Lyapunov function $\mathcal{V}(\cdot)$ satisfies the condition $\mathcal{V}(\mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}))-\mathcal{V}(\mathbf{x})\le 0$ for the system state update mapping $\mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x})$, and the equality holds if and only if $\mathbf{x}\in\mathcal{S}_{\text{perf}}$, i.e., $\mathcal{V}(\mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}))-\mathcal{V}(\mathbf{x})=0\Leftrightarrow\mathbf{x}\in\mathcal{S}_{\text{perf}}$, then it follows that, for any initial state $\mathbf{x}(0)\in\mathbb{R}^n$, the system trajectory $\mathbf{x}(t)$ will asymptotically converge to the largest invariant subset contained in $\mathcal{S}_{\text{perf}}$. Under the control policy $\boldsymbol{\pi}$, the system exhibits asymptotic stability and will eventually remain within the set of states that satisfy the performance-stability conditions.

Our primary objective is to construct a control policy of the form $\boldsymbol{\pi}=-\psi_\theta(\mathcal{S}_t)$ such that the resulting closed-loop system satisfies the requirement of asymptotic stability. To achieve this, we introduce a Lyapunov function $\mathcal{V}(\cdot)$ and impose the following two conditions: 1. For any state $\mathbf{x}\notin\mathcal{S}_{\text{perf}}$, $\mathcal{V}(\mathfrak{f}_\pi(\mathbf{x}))-\mathcal{V}(\mathbf{x})<0$, indicating that the system state strictly decreases in this region; 2. For all $\mathbf{x}\in\mathcal{S}_{\text{perf}}$, $\mathcal{V}(\mathfrak{f}_\pi(\mathbf{x}))-\mathcal{V}(\mathbf{x})=0$, implying that the Lyapunov function remains invariant over the performance-optimal subset. In Theorem 1, we propose a sufficient structural condition to ensure the above properties hold, thereby guaranteeing the stability of the system.

*Theorem 1. (Inspection Stability Condition):* Suppose that for any controller $i$, its corresponding control function $\psi_{\theta_i}(\cdot)$ is continuously differentiable over its domain. When the system

state $\mathbf{x}_i$ lies within the interval $[\mathbf{x}_i, \overline{\mathbf{x}}_i] \triangleq \mathcal{S}_{\text{perf}}$, the control policy satisfies $\boldsymbol{\pi}_i = -\psi_{\theta_i}(\mathcal{S}_t) = \mathbf{0}$. Moreover, when the system state $\mathbf{x}_i \notin \mathcal{S}_{\text{perf}}$, the Jacobian of the controller with respect to the state satisfies the following stability inequality: $-\frac{2}{\Delta\mathbb{T}}A^{-1} \preceq \frac{\partial \boldsymbol{\pi}}{\partial x} \preceq 0$, where $A \succ 0$ is a positive definite matrix representing system stability, and $\Delta\mathbb{T}$ denotes the control update interval. This ensures that when the system state deviates from the stable region, the controller can provide timely feedback to drive the state back, thus maintaining system stability. (Proof is provided in Appendix 1.)

Based on the Lyapunov function $\mathcal{V}(\mathbf{x})$ constructed in (28), to ensure system stability under the control policy, we need to further derive that it satisfies the monotonic non-increasing condition. Specifically, $\mathfrak{f}_\pi(\mathbf{x})$ denotes the target state trajectory obtained by mapping the current system state $\mathbf{x}$ through the policy $\pi$. A smaller value of the Lyapunov function $\mathcal{V}(\mathbf{x})$ indicates that the system is closer to the desired stable equilibrium. Therefore, we expect $\mathcal{V}(\mathbf{x})$ to be non-increasing during the system evolution. This property can be indirectly ensured by constraining the partial derivative of the policy function with respect to the state variables, i.e., requiring $\frac{\partial \boldsymbol{\pi}}{\partial x} \preceq \mathbf{0}$, which guarantees that the state feedback induces a convergent behavior in the system dynamics, thereby contributing to the stability of the system.

Furthermore, as the sampling period $\Delta T$ increases, $\frac{\partial \boldsymbol{\pi}}{\partial x}$ must satisfy a bounded constraint with a lower bound of $-\frac{2}{\Delta T}$ and an upper bound of 0. Considering that contemporary wireless nodes typically operate at very high sampling frequencies, in most scenarios the left-hand condition $-\frac{2}{\Delta T}A^{-1} \preceq \frac{\partial \boldsymbol{\pi}}{\partial x}$ is naturally satisfied. Therefore, our primary focus is on the upper bound constraint. The decentralized nature of MARL implies that $\frac{\partial \boldsymbol{\pi}}{\partial x}$ is a diagonal matrix, as shown in (29). To meet the stability requirements, we design each element $\frac{\partial \psi_{\theta_i}}{\partial x_i} > 0$ to be strictly monotonically increasing.

$$\frac{\partial \boldsymbol{\pi}}{\partial x} = -\operatorname{diag}\left(\frac{\partial \psi_{\theta_1}}{\partial x_1}, \frac{\partial \psi_{\theta_2}}{\partial x_2}, \ldots, \frac{\partial \psi_{\theta_i}}{\partial x_i}\right) \tag{29}$$

### D. Design of Stability Controllers

Combining the structural constraints on stable controllers given in Theorem 1, this work designs an RL control strategy with guaranteed stability. According to the stability criterion in Theorem 1, to ensure the system state ultimately converges to the equilibrium point, the policy function must strictly satisfy a monotonic increasing property. To achieve this, we incorporate a monotonicity constraint structure into the policy network and model the policy function using neural networks with explicit monotonic features. Previous studies [83], [84], [85] have proposed several feasible methods for constructing monotonic neural networks.

A multivariate function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}$ is said to be monotonically non-decreasing with respect to the variable $x_i$ if and only if:

$$x_i^0 > x_i^1 \Rightarrow \mathbf{f}\left(x_1, \ldots, x_i^0, \ldots, x_n\right) \geq \mathbf{f}\left(x_1, \ldots, x_i^1, \ldots, x_n\right) \tag{30}$$

Additionally, a monotonically increasing indicator vector $\mathbf{t}$ is introduced, such that when $\mathbf{t}_i = 1$, the corresponding partial derivative is required to satisfy the enforced constraint $\frac{\partial \mathbf{f}(x_i)}{\partial x_i} \geq 0$.

*Definition 2. (Monotonicity-Constrained Weight Transformation):* For a weight matrix $\mathbf{W} \in \mathbb{R}^{n\times m}$, the constrained operation $|\cdot|_t$ is defined as shown in (31).

$$|\mathbf{W}|_t = \begin{bmatrix} |\omega_{11}| & \cdots & |\omega_{1m}| \\ \vdots & \ddots & \vdots \\ |\omega_{n1}| & \cdots & |\omega_{nm}| \end{bmatrix} \circ \begin{bmatrix} \mathbf{t} \\ \vdots \\ \mathbf{t} \end{bmatrix} 1^T \tag{31}$$

where $\circ$ denotes the Hadamard product, and $\mathbf{1}$ is an all-ones vector of dimension $m$.

*Lemma 1. (Monotonicity-Preserving Neural Network via ReLU Composition):* Consider a function $\phi : \mathbb{R}^d \to \mathbb{R}$ implemented by a piecewise-constructed neural network. When modeled in the following form, its output components $\phi^{(+)}(\mathbf{x})$ and $\phi^{(-)}(\mathbf{x})$ satisfy strict monotonicity and origin-centered controllability constraints, respectively. The positive component is defined as shown in (32).

$$\phi^{(+)}(\mathbf{x}) = \sum_{k=1}^{M} \mathbf{W}_k \rho\left(\mathbf{1}\mathbf{x} + b_k\right) \tag{32}$$

where the activation function $\rho(\cdot)$ is the element-wise ReLU defined as $\rho(\mathbf{x}) = \max(0, \mathbf{x})$, and satisfies the constraints in (33)-(34).

$$\sum_{k=1}^{\ell} \mathbf{W}_i \succeq \mathbf{0}, \quad \forall \ell = 1, 2, \ldots, M \tag{33}$$

$$b_1 = 0, b_\ell \leq b_{\ell-1}, \forall \ell = 2, 3, \ldots, M \tag{34}$$

where $\rho(\cdot) = \max(x, 0)$ is the ReLU activation function; $\mathbf{W}_i = [\omega_i^1, \omega_i^2, \ldots, \omega_i^m]$ is the weight vector corresponding to Definition 2; and $\mathbf{b}_i = [b_i^1, b_i^2, \ldots, b_i^m]^T$ is the bias vector. This construction must satisfy the conditions of (35).

$$\phi^{(-)}(\mathbf{x}) = \sum_{k=1}^{M} |\mathbf{W}_i|_t \, \rho\left(-1\mathbf{x} + c_k\right) \tag{35}$$

The weights and biases are required to satisfy the monotonicity conditions of (36).

$$\sum_{k=1}^{\ell} \mathbf{W}_i \preceq \mathbf{0}, \quad \forall \ell = 1, 2, \ldots, M \tag{36}$$

$$c_1 = 0, \quad c_\ell \leq c_{\ell-1}, \quad \forall \ell = 2, 3, \ldots, M \tag{37}$$

The following Theorem 2 formally establishes that the function forms constructed in (32) and (35) can approximate any target function satisfying the specified conditions.

*Theorem 2:* (Monotone Function Approximation Theorem) Let $\mathbf{r}(\mathbf{x})$ be a continuous, Lipschitz continuous, bounded, and monotonic function that passes through the origin, with bounded derivatives, defined on a compact set $\mathbf{X}$ and mapping to the real numbers $\mathbb{R}$. For any $\boldsymbol{\varepsilon} > \mathbf{0}$, there exists a function $\phi(\mathbf{x}) = \phi^{(+)}(\mathbf{x}) + \phi^{(-)}(\mathbf{x})$, constructed as in (32) and (35), such that for all $\mathbf{x} \in \mathbf{X}$, $|\mathbf{r}(\mathbf{x}) - \phi(\mathbf{x})| < \boldsymbol{\varepsilon}$.

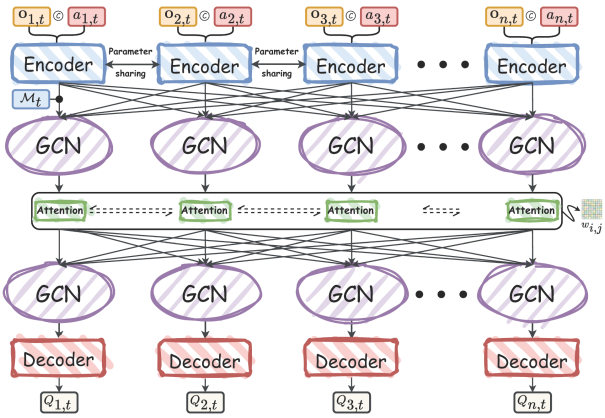


Fig. 5. Topology-aware graph attention collaborative critic network.

### E. Topology-Aware Graph Attention Collaborative Critic Network

As described in Section V.C, after the stability controller generates control actions, a topology-aware graph attention Critic network evaluates the system's state-action value in real time. As shown in Fig. 5, the architecture includes three modules: a multi-source feature embedding layer that fuses topology, node states, and actions; a hierarchical topology coupling module with a cascaded GCN–Self-Attention–GCN structure to capture spatial and global dependencies; and a distributed value output layer that balances local accuracy with global consistency. At the input stage, all agents share a unified encoder module. Each agent's input includes its local observation $o_{i,t}$, current action $a_{i,t}$, and the system's real-time adjacency matrix $\mathcal{M}_t$. The encoder, composed of fully connected layers, concatenates $o_{i,t}$ and $a_{i,t}$ into a feature vector $\mathbf{h}_i^{(0)}$. This feature, together with $\mathcal{M}_t$, is then fed into the first GCN layer to perform graph convolution as defined in (38).

$$\mathbf{h}_i^{(1)} = \sigma\left(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{h}_i^{(0)}\mathbf{W}^{(1)}\right) \tag{38}$$

where $\mathbf{h}_i^{(1)} \in \mathbb{R}^{N\times d_l}$ is the node feature matrix of the first GCN layer; $\mathbf{W}^{(1)} \in \mathbb{R}^{d_l\times d_{l+1}}$ is the trainable weight matrix; $\sigma(\cdot)$ denotes the ReLU function; $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix $\mathbf{A}$ with added self-loops (identity matrix); $\widetilde{\mathbf{D}}$ is the degree matrix of $\widetilde{\mathbf{A}}$; and $\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}$ represents the symmetrically normalized adjacency matrix.

After obtaining the topology embedding $\mathbf{h}_i^{(1)}$, a parameter-shared Self-Attention module is introduced to capture global interactions among all nodes. This module adopts a unified weight-sharing mechanism to ensure consistent and generalizable feature interactions. Let the output of the first GCN layer be $\mathbf{H}^{(1)} = [\mathbf{h}_1^{(1)}, \ldots, \mathbf{h}_n^{(1)}] \in \mathbb{R}^{n\times d}$. The computation of the Self-Attention module is described in (39).

$$k\left(\mathbf{h}_i^{(1)}\right) = v\left(\frac{\mathbf{h}_i^{(1)}\mathbf{W}_q \cdot \left(\mathbf{h}_i^{(1)}\mathbf{W}_k\right)^\top}{\sqrt{d_k}}\right) \cdot \left(\mathbf{h}_i^{(1)}\mathbf{W}_v\right) \tag{39}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d_k}$ are the query, key, and value projection matrices shared by all nodes, ensuring that the model captures interactions among all nodes using only a single set of parameters. The output features are denoted as $\mathbf{H}^{(2)} = [\mathbf{h}_1^{(2)}, \ldots, \mathbf{h}_n^{(2)}]$, which serve as the input to the subsequent GCN layer. The features fused through the attention mechanism are then fed into the second GCN layer to enhance global context awareness. The features after the second GCN layer are denoted as $\mathbf{H}^{(3)} = [\mathbf{h}_1^{(3)}, \ldots, \mathbf{h}_n^{(3)}]$. The final $\mathbf{h}_i^{(3)}$ represents the high-dimensional embedding after two rounds of graph convolution and one attention interaction, fully integrating topology structure, node states, and system collaborative semantics.

Finally, the feature $\mathbf{h}_i^{(3)}$ is passed to the corresponding Decoder module to estimate the action-state value for each agent at the current time step. This module uses a parameter-independent multilayer perceptron (MLP), enabling each node to independently output its local action value $Q_{i,t}$ while preserving collaborative interaction information for policy training. Notably, when constructing the topology-aware collaborative policy objective, we introduce a set of MLPs to nonlinearly map the attention mask $\mathbf{A} = (\mathbf{h}_i^{(1)}\mathbf{W}_q \cdot (\mathbf{h}_i^{(1)}\mathbf{W}_k)^\top)/\sqrt{d_k}$, generating weighting coefficients $\omega_{ij}^{(n)}$ corresponding to the number of agents ( (41)). These weights are adaptively learned from feature similarity and topology, reflecting neighbors' influence on the agent's policy adjustment. Note that $\omega_{ij}$ is used only during policy training for regularization and does not affect policy execution or action generation.

### F. Learning and Optimization

This section presents the optimization and update mechanisms of the Actor–Critic networks. To address the high-dimensional and non-convex inspection control problem, we employ the MATD3 algorithm with a topology-aware graph attention Critic, enabling coordinated agent training to balance delay, energy consumption, and communication quality.

During each training iteration, agents sample state transitions from the replay buffer to update their twin Critic networks $\mathcal{Q}\psi_1^i$ and $\mathcal{Q}\psi_2^i$. The target Q-value is computed using the delayed target networks as $\mathcal{Y}_t^i = \mathcal{R}_t^i + \gamma \min_{k=1,2} \mathcal{Q}_{\overline{\psi_\theta}^i}^{(k)}(\mathcal{O}_{t+1}, \overline{\psi_\theta}^i(\mathcal{O}_{t+1}))$, where $\overline{\psi_\theta}^i$ and $\mathcal{Q}\overline{\psi\theta}^{i(k)}$ denote the target networks. To improve stability and generalization of the Critic under multi-agent collaboration, a topology-aware gradient regularization term is introduced, incorporating action sensitivity control and neighborhood consistency constraints, as defined in (40).

$$\mathcal{L}_{\psi_i^{(k)}} = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\left(\mathcal{Q}_{\psi_i^{(k)}}(\mathcal{S}_t, \mathbf{a}_t) - y_t^i\right)^2\right] + \left\|\nabla_{\mathbf{a}_t^i}\mathcal{Q}_\psi^i(\mathcal{O}_t, \mathbf{a}_t)\right\|^2 + \sum_{j\in\mathcal{N}(i)}\left\|\nabla_{\mathbf{a}_t^j}\mathcal{Q}_\psi^i - \nabla_{\mathbf{a}_t^j}\mathcal{Q}_\psi^j\right\|^2 \tag{40}$$

where Term 1 is the TD error that minimizes the mean-squared difference between the Critic's predicted and target Q-values. Term 2 penalizes the squared norm of the Critic's action gradient $\nabla_{\mathbf{a}_t^i}Q$, limiting sensitivity to the agent's action and promoting

function smoothness. Term 3 enforces neighborhood action-gradient consistency by minimizing the squared differences between the action gradients of agent $i$ and its neighbors $j \in \mathcal{N}(i)$, fostering coordinated policies and stable multi-agent training.

There exist complex topological couplings among agents, where each agent's action influences both its own performance and that of its neighbors. To improve coordination and overall system performance, a topology-aware collaborative policy objective is introduced. Specifically, the policy objective of agent $i$ is defined in (41).

$$\mathcal{J}(\theta_i) = \mathbb{E}_{\mathcal{O}_t}\left[\mathcal{Q}_\psi^i\left(\mathcal{O}_t, \psi_\theta^i(\mathcal{O}_t)\right)\right] - \omega_{ij}^{[i]}\mathbb{E}_{j\in\mathcal{N}(i)}\left[\left\|\psi_\theta^i(\mathcal{O}_t) - \psi_\theta^j(\mathcal{O}_t)\right\|^2\right] \tag{41}$$

where the first term represents the value estimate of the current action by the Critic network corresponding to the policy, aiming to maximize the agent's long-term return; the second term is the neighborhood action consistency regularization, which characterizes the policy correlation between agent $i$ and its neighbors. By applying a weighted penalty on the differences in policy outputs, it encourages neighboring agents to maintain consistent action updates, thereby enhancing coordination among policies and overall system stability. During the training of the policy network, the parameters $\theta^i$ are updated by maximizing the objective function $\mathcal{J}(\theta^i)$, with the corresponding gradient given in (42).

$$\nabla_{\theta^i}\mathcal{J}_{\theta^i} = \mathbb{E}_{\delta\sim\mathcal{O}_t}\left[\nabla_{\mathbf{a}_t^i} Q_{\psi^i}(\mathcal{O}_t, \mathbf{a}_t)\cdot\nabla_{\theta^i}\psi_\theta^i(\mathcal{O}_t)\right] - 2\omega_{ij}\sum_{j\in\mathcal{N}(i)}\left(\psi_\theta^i(\mathcal{O}_t) - \psi_\theta^j(\mathcal{O}_t)\right)\cdot\nabla_{\theta^i}\psi_\theta^i(\mathcal{O}_t) \tag{42}$$

This work employs a soft update mechanism to maintain the target network parameters. Specifically, the Critic's target parameters $\overline{\psi_\theta}^i$ are updated from the main network $\psi_\theta^i$ using $\overline{\psi}^i \leftarrow \tau\psi^i + (1-\tau)\overline{\psi}^i$, where $\tau \in (0,1)$ is the soft update coefficient controlling the update rate for smooth parameter changes. The Actor's target parameters are updated in the same manner, synchronized with the Critic. The overall procedure is summarized in Algorithm 1.

**Algorithm 1:** MARL Algorithm for Air-Ground-WANET Substation Inspection Problem.

Initialize policy/target policy network parameters $\theta, \theta'$, Q-network/target Q-network parameters $\psi_1, \psi_2, \psi_1', \psi_2'$, and replay buffer $\mathcal{D}$
**for** agent $i = 1, 2, \ldots, N$ **do**
  Initialize actor network $\theta$ with random parameters
  Initialize target networks $\theta' \leftarrow \theta$
**end for**
Set global time step: $T \leftarrow 0$
**for** $episode = 1$ to $T$ **do**
  **for** time step $t = 1, 2, \ldots, T$ **do**
    Interact with the environment and obtain a transition $(s_t, o_t^i, a_t, r_t^i, s_{t+1}, o_{t+1}^i)$, and store it in $\mathcal{D}$
    Sample a mini-batch of $N$ transitions $(s_t, o_t^i, a_t, r_t^i, s_{t+1}, o_{t+1}^i)$ from $\mathcal{D}$
  **end for**
  Update global time step: $T \leftarrow T + 1$
  **for** agent $i = 1, 2, \ldots, N$ **do**
    Select action with exploration noise $a \sim \pi_\psi(s) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$
    Observe reward $r$ and new state $s_{t+1}$, then store transition tuple $(s_t, o_t^i, a_t, r_t^i, s_{t+1}, o_{t+1}^i)$ in $\mathcal{D}$
    **for** $j = 1$ to $N_c$ **do**
      For each transition $\tau \sim (s_t, o_t^i, a_t, r_t^i, s_{t+1}, o_{t+1}^i)$ in the batch, calculate the target Q-value $y_i = r + \gamma \min_{i=1,2} Q_{\psi_i}^-(s_{t+1}, \mu_\theta^-(s_{t+1}))$
      Calculate the online Q-values $Q_{\psi_1}(s_t, \mu_\theta(s_t))$ and $Q_{\psi_2}(s_t, \mu_\theta(s_t))$
      Update the parameters $\psi_1, \psi_2$ of the critic network
    **end for**
    **for** $j = 1$ to $N_a$ **do**
      Prevent overestimation by using the smaller Q-values $\min Q_{\psi_i}(s_t, \mu(s_t;\theta))$
      Compute the action sensitivity term $\|\nabla_{\mathbf{a}_t^i}\mathcal{Q}_\psi^i(\mathcal{O}_t, \mathbf{a}_t)\|^2$
      Compute the regularization term $\sum_{j\in\mathcal{N}(i)}\|\nabla_{\mathbf{a}_t^j}\mathcal{Q}_\psi^i - \nabla_{\mathbf{a}_t^j}\mathcal{Q}_\psi^j\|^2$
      Calculate the critic network loss using (40)
      Compute the action consistency regularization $\omega_{ij}^{[i]}\mathbb{E}_{j\in\mathcal{N}(i)}[\|\psi_\theta^i(\mathcal{O}_t) - \psi_\theta^j(\mathcal{O}_t)\|^2]$
      Calculate the actor network loss using (41)
    **end for**
    Perform soft updates on the target critic and actor networks
  **end for**
**end for**

## VI. Case Study

### A. Experimental Setup

To verify the effectiveness of the proposed approach, we implemented the scenario in Fig. 1 within a simulation environment. A 3D space of 50 m × 50 m × 10 m was constructed for node deployment, as illustrated in Fig. 6. The system includes 11 nodes, with node 1 serving as the terminal receiver. The algorithm is developed and implemented in PyTorch. All simulations are performed on a workstation equipped with an NVIDIA GeForce RTX 4080 GPU (16GB) and an Intel Core$^{\text{TM}}$i7-14700KF processor. Algorithm parameters, environment settings, and device endurance parameters are listed in Tables II–IV. Additional environment and energy consumption model configurations are detailed in [79], [86], [87], [88].

TABLE II
Algorithm Parameter Settings

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Actor Hidden Size | 256 | Maximum Steps per Episode | 200 |
| Critic Hidden Size | 128 | Maximum Episodes | 1000 |
| Actor Learning Rate | 1e-4 | Action Noise Standard Deviation | 0.3 |
| Critic Learning Rate | 1e-3 | Replay Buffer Capacity | 1e5 |
| Discount Factor ($\gamma$) | 0.99 | Batch Size | 256 |
| Target Update Rate ($\tau$) | 0.005 | Random Seeds | 100, 200, 300 |
| Optimizer Type | Adam | Computational Cost per Agent | 9266 Params (4.3e4 FLOPs) |

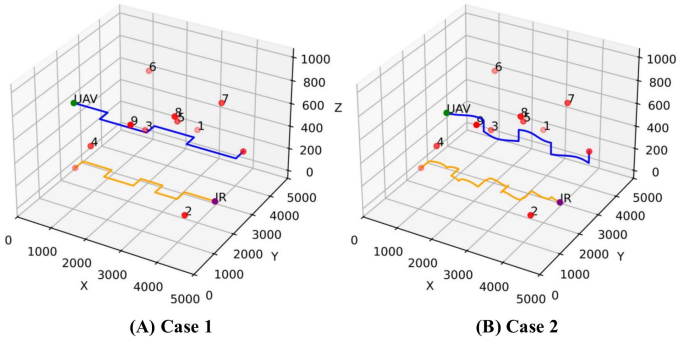


Fig. 6. Representative air-ground WANET inspection scenarios.

TABLE III
ENVIRONMENTAL PARAMETER SETTINGS

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $\omega_0$ | 8 | SIFS | 28 $\mu$s | $m$ | 5 |
| DIFS | 128 $\mu$s | $\mathcal{H}$ | 5 | ACK | 50 $\mu$s |
| $\mathfrak{T}_{slot}$ | 50 ms | $\mathcal{L}_{path}(d_0)$ | 40 dB | $\mathbb{P}_\alpha$ | 0.1 |
| $\eta_{path}$ | 2 | $\mathrm{t}_s^{(\mathrm{slot})}$ | 320 $\mu$s | $\sigma^2_\Xi$ | 6 dB |
| $\mathcal{L}_{packet}$ | 256 bits | $d_0$ | 1 m | $\mathfrak{B}_{ch}$ | 2.4 MHz |
| $\lambda$ | 1 | $\varrho_{q,y}$ | -80 dBm | $\varsigma^2_{E2E}$ | -100 dBm |

TABLE IV
DEVICE ENDURANCE PARAMETER SETTINGS

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $c^{pro}$ | 0.15 | $r_{rad}$ | 0.2 m | $\rho^{den}$ | 1.225 kg/m$^3$ |
| $r^{dis}$ | 0.1 m | $s^{\mathrm{rot}}$ | 0.1 | $v_{\mathrm{bla}}$ | 600 |
| $f^{inc}$ | 0.1 | $\mathcal{E}^{(uav)}_{circuit}$ | 0.3 W | $w$ | 1.5 kg |
| $I^{idle}$ | 0.3 A | $V_{sys}$ | 12 V | $m^{rob}$ | 6 kg |
| $f^{rol}$ | 0.01 | $v_i$ | 0.6 m/s | $\omega_d$ | 0.5 |
| $\omega_\Phi$ | 0.5 | $\omega_e$ | 0.4 | $\omega_q$ | 0.6 |

To verify the adaptability and effectiveness of the proposed communication-aware air-ground collaborative inspection strategy in real-world settings, two representative scenarios are designed, as shown in Fig. 6. The figure depicts two typical air-ground WANET inspection modes: (A) Case 1 represents a routine inspection with evenly distributed targets, where the UAV and IR follow predefined paths to inspect aerial and ground equipment, enabling clear task division and efficient planning; (B) Case 2 reflects special tasks involving concentrated equipment or temporary faults, requiring manual adjustments to the UAV's altitude and the IR's route to detect the task area. To verify the superiority of the proposed algorithm, we reproduced several mainstream MARL algorithms in the same environment, including MATD3 [89], MAPPO [90], and QMIX [91].

### B. Comparison of Rewards for the Two Cases

Fig. 7 illustrates the reward trends of different algorithms in Case 1 and Case 2. The proposed method consistently outperforms the baselines in both scenarios. In Case 1, it rapidly increases and stabilizes the reward at 55–60, notably higher than MATD3 (30–35), MAPPO (40), and QMIX (45), with smaller fluctuations and narrower shaded regions, indicating higher stability. In Case 2, although task complexity reduces overall rewards, the proposed method still achieves the best performance, stabilizing around 40–45 compared to 30–38 for other algorithms. It converges faster with lower variance, demonstrating strong adaptability and stable effectiveness across both cases.

To evaluate performance in practical scenarios, we tested the proposed method alongside three baselines in a unified environment. To simulate dynamics, the inspection device node update rate was increased tenfold, while other parameters remained unchanged. Fig. 8 shows the reward curves for Case 1 and Case 2. The proposed method consistently outperforms the baselines, exhibiting higher stability and smaller fluctuations. In Case 1, its reward stabilizes around 60, surpassing QMIX (50), MAPPO, and MATD3 (40–45). In Case 2, despite increased complexity, it maintains about 45, outperforming QMIX (37), MATD3, and MAPPO (30–35). These results indicate that the proposed method is more stable and adapts better to changing conditions.

### C. Comparison of Delay for the Two Cases

Fig. 9 illustrates the average delay variations during training for different algorithms under Case 1 and Case 2, with error bars representing standard deviations. All algorithms achieve delay reduction, indicating policy optimization. However, the proposed method consistently attains the lowest delay and smallest variance, showing strong stability and fast convergence. In Case 1, delay reduction is smooth across algorithms, while in Case 2, delay increases due to higher environmental complexity. Even so, our method maintains superior generalization, robustness, and minimal fluctuations. MATD3 performs well but slightly lags behind, whereas MAPPO and QMIX exhibit notable volatility in Case 2, reflecting greater sensitivity to environmental changes.

Fig. 10 compares the delay performance of different algorithms in two scenarios. In Case 1, all algorithms exhibit small delay fluctuations (0.28–0.36), with the proposed method achieving the lowest average delay, indicating superior scheduling efficiency and stability. In Case 2, higher task complexity slightly increases delays, especially for MATD3 and MAPPO, showing degraded scheduling performance. In contrast, the proposed method maintains low delay and stable convergence, demonstrating better generalization and robustness.

### D. Comparison of Transmission Power for the Two Cases

Fig. 11 illustrates the dynamic evolution of UAV and IR transmit power (dBm) across four algorithms in two inspection scenarios. All algorithms eventually converge, demonstrating learning capability. However, the proposed method achieves faster convergence, better power stability, and smoother control, maintaining lower and more consistent power levels. In contrast, MATD3 and MAPPO exhibit strong mid-training oscillations, while QMIX stabilizes later but fluctuates significantly in the early stages.

Fig. 12 shows the variations in UAV and IR transmit power over training episodes for different algorithms under two cases. In Case 1, our method maintains the lowest and most stable power levels around 16 dBm, outperforming MATD3, MAPPO, and QMIX. In Case 2, although environmental complexity

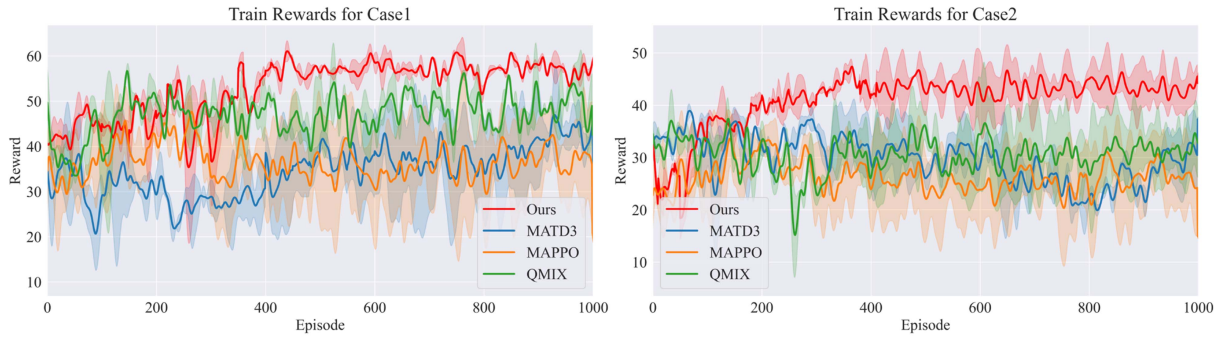


Fig. 7. Training reward for different algorithms.

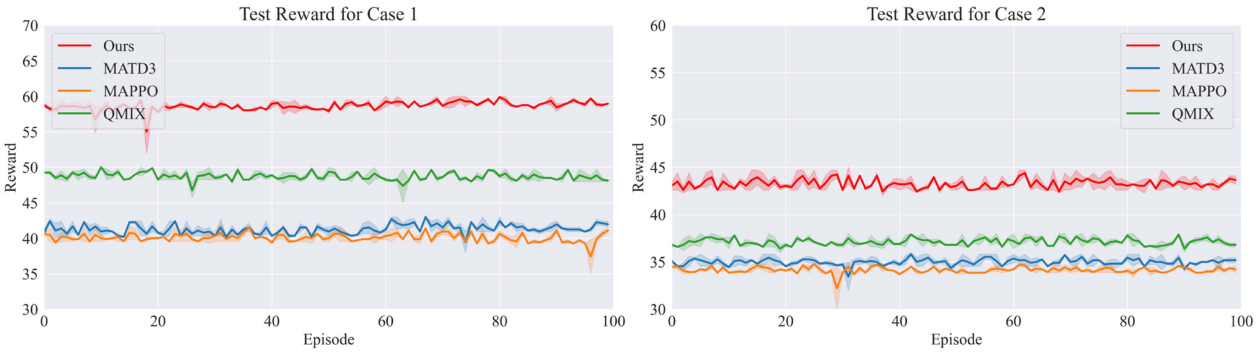


Fig. 8. Testing reward for different algorithms.

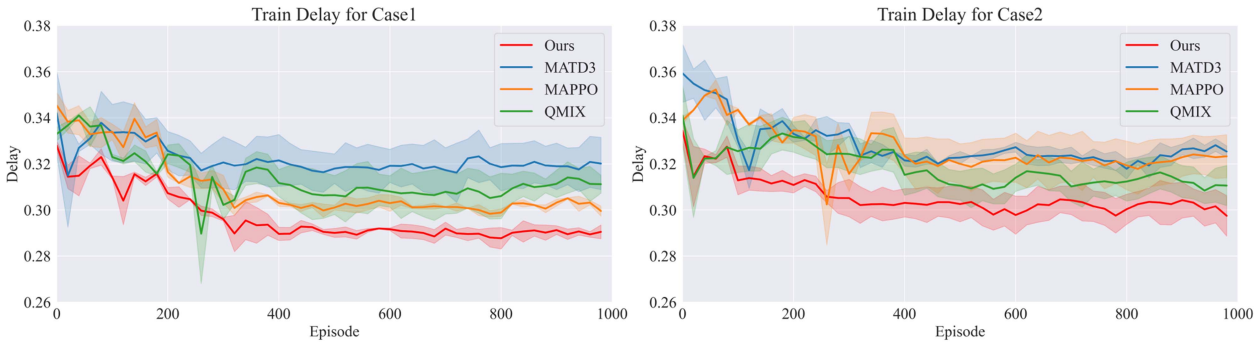


Fig. 9. Training delay for different algorithms.

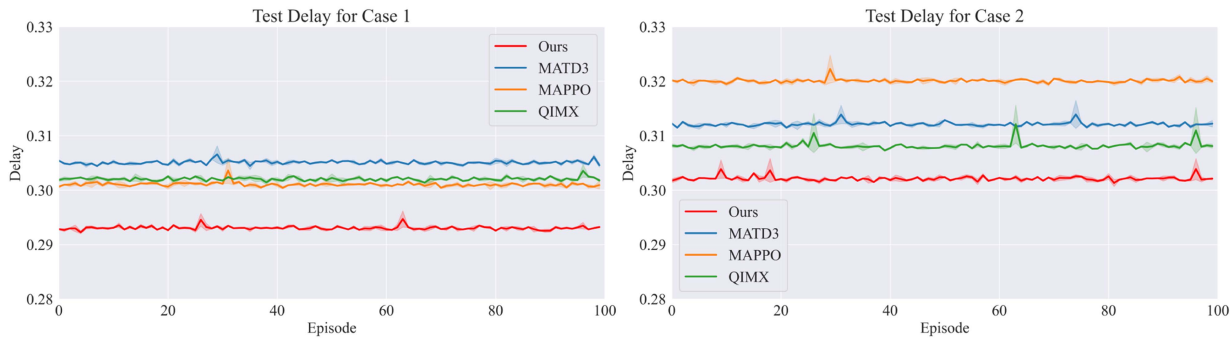


Fig. 10. Testing delay for different algorithms.

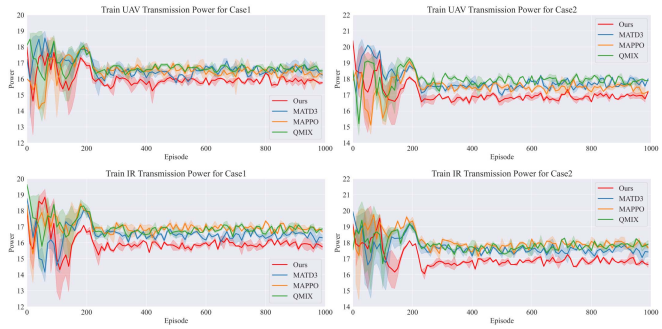


Fig. 11. Training transmission power for different algorithms.

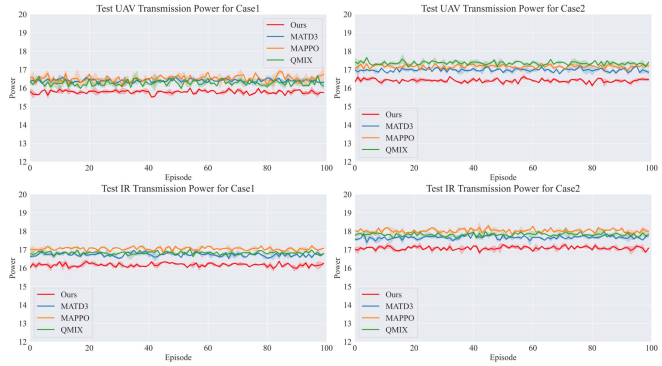


Fig. 12. Testing transmission power for different algorithms.

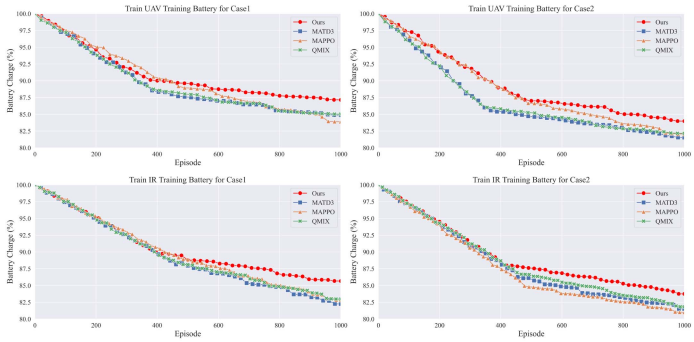


Fig. 13. UAV and IR battery levels during training in Case 1/Case 2.

increases and baseline algorithms exhibit higher fluctuations, our method remains stable around 16.5 dBm, demonstrating superior robustness and energy efficiency. Overall, the proposed approach achieves low, stable transmit power across varying scenarios.

### *E. Comparison of Device Endurance for the Two Cases*

Fig. 13 illustrates the UAV and IR battery trends during training in Case 1 and Case 2. As training progresses, battery levels decrease for all algorithms, but at different rates. The proposed method achieves the slowest depletion and highest final stability in both cases, indicating superior endurance. In Case 1, the UAV battery stabilizes after about 400 episodes with 12.5% consumption, maintaining 87.5% remaining power—higher than MATD3 (85%), MAPPO (83%), and QMIX (85%). In Case 2, it retains about 85% battery, while others drop to 80–82.5%. The IR shows similar results, maintaining about 86% in Case 1 and 84% in Case 2, 2–4 points higher than other methods. The proposed method effectively reduces UAV and IR energy consumption via adaptive power control, enhancing endurance and energy efficiency in dynamic environments.

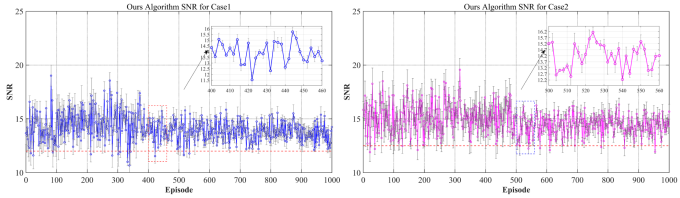


Fig. 14. The SNR performance of our algorithm.

### *F. Comparison of SNR for the Two Cases*

Fig. 14 illustrates the SNR (in dB) variation during training for the proposed method under two inspection scenarios, Case 1 and Case 2. In both cases, the average SNR remains above the predefined threshold, confirming that the proposed power control strategy maintains communication quality while optimizing other metrics. In Case 1, the SNR stabilizes around 14 dB with a narrow fluctuation range (13–15.5 dB), indicating stable communication. In Case 2, despite higher task complexity, the SNR stays around 14.5 dB with a slightly wider range (13–16.5 dB), showing that the method effectively adapts to complex conditions and ensures reliable link quality.

### *G. Sensitivity Analysis on the Number of Communication Nodes*

To evaluate the scalability of the proposed method across network sizes, we performed a node-scale sensitivity study. Holding the task scenario and all other settings constant, we increased the number of communication nodes from 11 to 25, randomly placing the newly added nodes, and fixed the random seed at 100. The results are presented in Fig. 15. As the network grows, multi-agent coordination becomes more challenging, which degrades the performance of the baseline methods. In Case 1, our method maintains a stable training reward of approximately 56 across all network sizes, yielding nearly a 40% gain over the weakest baseline, MAPPO ($\approx 40$), and preserving a clear advantage over the stronger baseline, QMIX ($\approx 50$). In Case 2, when the number of nodes reaches 25, MAPPO's delay increases to above 0.33 s. By contrast, aided by an efficient communication mechanism, our algorithm consistently maintains a low delay of approximately 0.30 s with minimal fluctuations.

### *H. Effect of Interfering Factors on Performance*

*Node Failure:* To evaluate the robustness of the proposed method to communication node failures, we inject a node-fault disturbance in both Case 1 and Case 2. Specifically, at episode

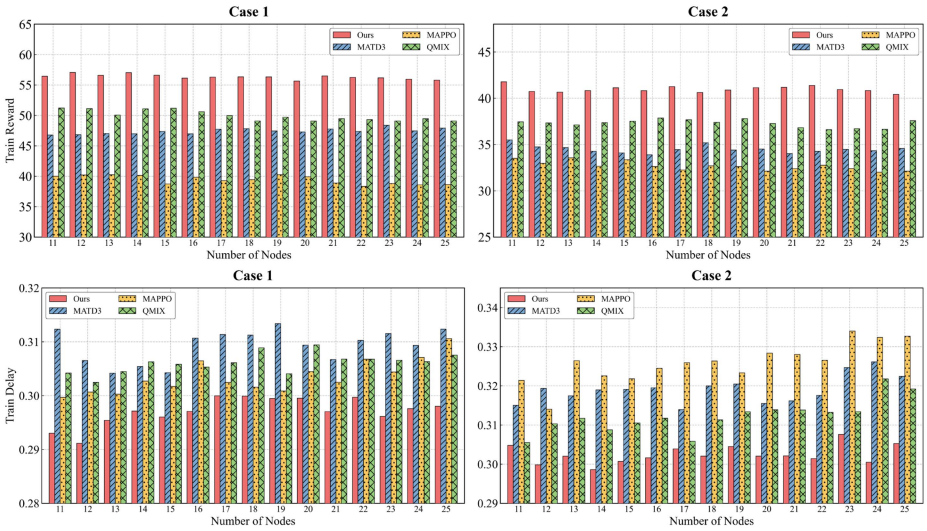


Fig. 15. Sensitivity analysis of algorithm performance under different network scales.

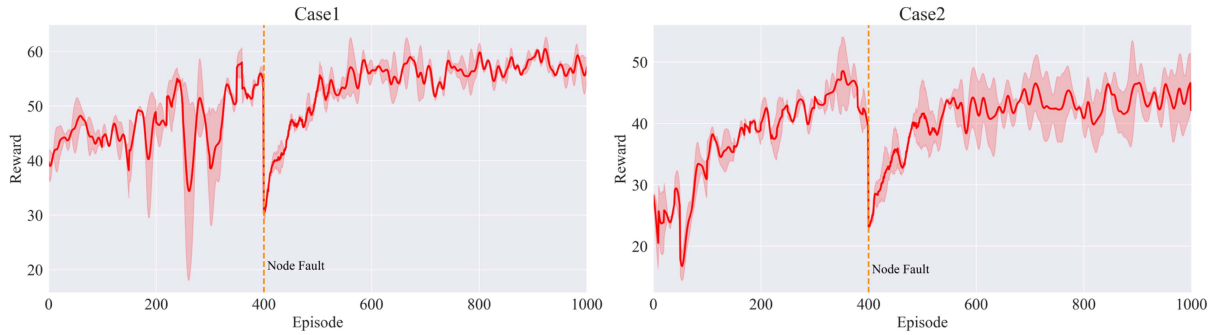


Fig. 16. Reward under a node-failure scenario in Case 1 and Case 2.

400, relay node 6 is forced to fail, emulating typical WANET disruptions caused by hardware malfunction or temporary disconnection due to battery depletion. Fig. 16 shows the reward trajectories before and after the fault. The node failure causes a short-term performance drop; however, the system gradually recovers and re-converges to a stable operating regime. In Case 1, the reward decreases from approximately 55–57 to 31–33 and then recovers to 56–60; in Case 2, it drops from roughly 46–48 to 24–27 and subsequently rebounds, stabilizing at about 42–46. These results suggest that the proposed method can quickly recover from disturbances and remains stable when the network topology changes abruptly.

*Signal Interference:* To evaluate system robustness under complex interference, we model external disturbances as reduced link reliability, quantified by a lower packet delivery ratio (PDR). Specifically, we examine three PDR levels—90%, 80%, and 70%—corresponding to weak, moderate, and strong interference, respectively; the no-interference setting serves as the baseline. As shown in Fig. 17, once interference is introduced at Episode 20, all disturbed cases exhibit a marked increase in end-to-end delay and larger oscillations, indicating that reduced reliability triggers retransmissions and route changes, thereby degrading short-term performance. A consistent pattern emerges: lower PDR leads to higher delay peaks and a longer recovery period before the system reaches a stable regime. Nevertheless, after several episodes of policy self-adaptation, the delay decreases and converges to a steady state. Experimental results demonstrate that the proposed algorithm can tolerate abrupt disturbances and achieve self-recovery under external interference.

### *I. Effect of Stability-Constrained Controllers on Performance*

To evaluate the effect of the proposed monotonic structural constraint on multi-agent policy learning, we ablate the monotonicity constraint while keeping all other experimental settings unchanged. Specifically, we replace the monotonic network with a standard (non-monotonic) multilayer perceptron (MLP). Fig. 18 presents the training return curves for Case 1 and Case 2. The monotonic constraint accelerates convergence and improves the final return in both scenarios. In contrast, removing the constraint leads to larger training fluctuations and consistently lower returns in the later training stages. Experimental results indicate that the monotonic constraint reduces training oscillations and improves learning robustness.


Fig. 17. E2E delay under different PDR levels with interference injection.

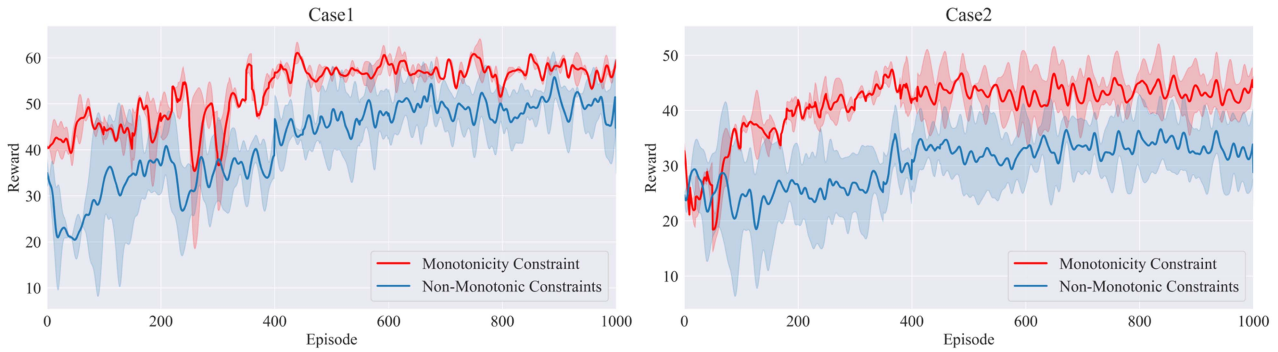


Fig. 18. Effect of monotonicity constraint on reward.

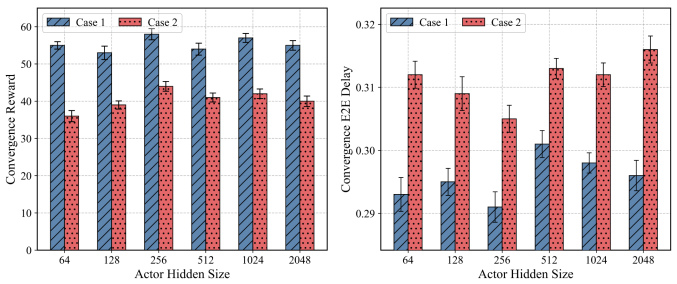


Fig. 19. Sensitivity of reward and delay to actor hidden size under Case 1 and Case 2.

## J. Ablation Studies

*Model size:* As shown in Fig. 19, we vary the Actor hidden-layer width over a broader range (64, 128, 256, 512, 1024, and 2048) to further assess the effectiveness of the lightweight Actor design. Overall, Case 1 is only mildly sensitive to network width: the converged return remains stable across configurations, and reducing the Actor width from 256 to 64 results in only an approximately 5% performance drop, indicating limited reliance on large model capacity under relatively stable conditions. In the more challenging Case 2, increasing the width to a moderate scale yields a more pronounced improvement (about 22% from 64 to 256), suggesting higher representational demands under strong perturbations. Beyond 256, however, further widening (512-2048) provides no additional gains and may slightly degrade performance, reflecting potential over-parameterization. Accordingly, a width of 128 offers a balanced choice for edge devices, whereas 256 is preferable for performance-oriented deployment on resource-rich nodes. The observed delay trend is consistent with the reward-based analysis, further supporting a favorable balance between performance and efficiency.

*Each module:* To quantify the contribution of each major component in the proposed framework, we conduct an ablation study under Case 1 and Case 2, comparing the full model (Full) with three degraded variants: w/o Stable (removing the stability-structure constraint), w/o Topo-Critic (removing the topology-aware critic), and w/o Topo-Reg (removing the topology-consistency regularization). As shown in Fig. 20, Full consistently yields higher rewards and lower E2E delays across both cases, while also achieving better SNR and energy-saving ratios, demonstrating the complementary benefits of these modules. Notably, w/o Stable suffers the most severe degradation, with a marked reward drop and increased delay, indicating that the stability structure is critical to convergence quality and training stability. Removing Topo-Critic or Topo-Reg also degrades performance, with the impact becoming more pronounced in the more challenging Case 2. Experiments demonstrate that topology-aware evaluation and consistency regularization are critical to robust collaborative decision-making, protecting against deterioration in delay, SNR, and energy metrics.

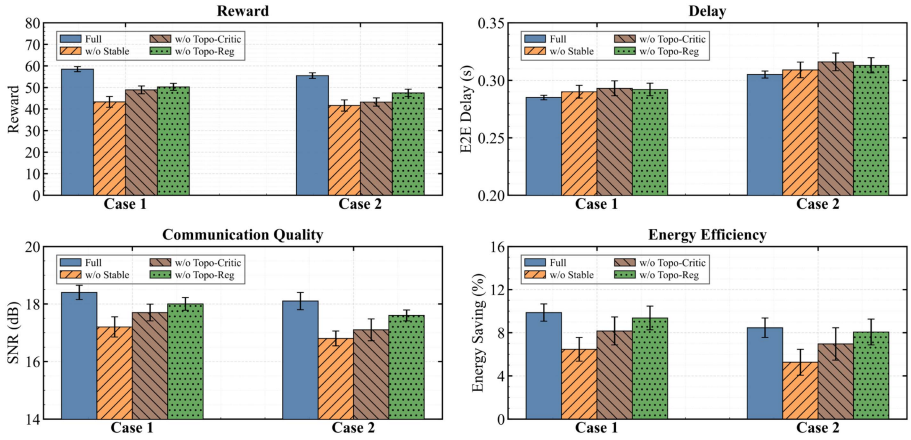


Fig. 20. Ablation results under two cases.

## VII. Conclusion

We propose a multi-objective optimization framework for integrated aerial–ground substation inspection based on multi-agent graph RL. By jointly controlling the transmit power of UAVs, IRs, and WANET relay nodes, the framework balances E2E delay, energy efficiency, and communication quality. We model the system as a Dec-POMDP and develop a stability-constrained algorithm that incorporates a topology-aware graph-attention critic and gradient regularization to improve training stability and multi-agent coordination. The proposed method outperforms MATD3, MAPPO, and QMIX in stability, adaptability, and energy efficiency, demonstrating superior robustness and generalization in dynamic tasks.

Despite its strong stability performance in simulation, the proposed framework has several limitations. First, the simulated substation WANET abstracts channel dynamics and interference patterns. Second, the topology-aware graph-attention critic and gradient regularization may impose substantial computational overhead, particularly in large-scale networks. Future work will incorporate more realistic channel models, perform hardware-in-the-loop validation to improve scalability and deployment efficiency.

## Appendix A
## Proof of Theorem 1

*Proof:* Recall the closed-loop delay dynamics $\mathbf{x}(t+1) = \mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}(t))$ with $\boldsymbol{\pi} = -\psi_\theta(\mathcal{S}_t)$. Define $y(\mathbf{x}(t)) = \mathbf{x}(t) - \mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}(t))$. The Lyapunov function is defined as shown in (43).

$$\mathcal{V}(\mathbf{x}(t)) = y(\mathbf{x}(t))^\top A^{-1} y(\mathbf{x}(t)) \tag{43}$$

We further express $y(\mathbf{x}(t+1))$ in terms of $y(\mathbf{x}(t))$ as shown in (44).

$$y(\mathbf{x}(t+1)) = y(\mathbf{x}(t)) + \int_0^1 \frac{\partial y}{\partial \mathbf{x}} (\mathbf{x}(t) + x \cdot \Delta\mathbf{x}(t)) \cdot \Delta\mathbf{x}(t)\, dx \tag{44}$$

By Kowalewski's Mean Value Theorem (Theorem 1 in [92]), we have:

$$y(\mathbf{x}(t+1)) = y(\mathbf{x}(t)) + J_h \left(\mathbf{x}(t+1) - \mathbf{x}(t)\right) \tag{45}$$

where,

$$J_h = \sum_{i=1}^{n} \lambda_i \frac{\partial y}{\partial \mathbf{x}} \left(\mathbf{x}(t) + k_i(\mathbf{x}(t+1) - \mathbf{x}(t))\right) \tag{46}$$

with $k_i \in [0,1]$, $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. Note that:

$$\mathbf{x}(t+1) - \mathbf{x}(t) = \mathfrak{f}_{\boldsymbol{\pi}}(\mathbf{x}(t)) - \mathbf{x}(t) = -y(\mathbf{x}(t)) \tag{47}$$

so we get (48).

$$y(\mathbf{x}(t+1)) = (I - J_h)\, y(\mathbf{x}(t)) \tag{48}$$

Then the Lyapunov function at the next step becomes:

$$\begin{aligned} \mathcal{V}(\mathbf{x}(t+1)) &= y(\mathbf{x}(t+1))^\top A^{-1} y(\mathbf{x}(t+1)) \\ &= y(\mathbf{x}(t))^\top (I - J_h)^\top A^{-1} (I - J_h) y(\mathbf{x}(t)) \end{aligned} \tag{49}$$

Let $\mathcal{G}(\mathbf{x}, \theta) = \frac{\partial \mathfrak{f}_{\boldsymbol{\pi}}}{\partial \mathbf{x}} + \frac{\partial \mathfrak{f}_{\boldsymbol{\pi}}}{\partial \boldsymbol{\pi}} \frac{\partial \boldsymbol{\pi}}{\partial \mathbf{x}}$ be the Jacobian of the closed-loop dynamics, and define:

$$J_G = \sum_{i=1}^{n} \lambda_i \mathcal{G}\left(\mathbf{x}(t) + k_i(\mathbf{x}(t+1) - \mathbf{x}(t)), \theta\right) \tag{50}$$

From the definition, we have $J_G = I - J_h$. Therefore:

$$\mathcal{V}(\mathbf{x}(t+1)) - \mathcal{V}(\mathbf{x}(t)) = y(\mathbf{x}(t))^\top \left(J_G^\top A^{-1} J_G - A^{-1}\right) y(\mathbf{x}(t)) \tag{51}$$

By Jensen's inequality, for any $x \in \mathbb{R}^n$:

$$\begin{aligned} x^\top J_G^\top A^{-1} J_G x &= \left\| A^{-1/2} J_G x \right\|^2 \\ &= \left\| \sum_{i=1}^{n} \lambda_i A^{-1/2} \mathcal{G}\left(\mathbf{x}(t) + k_i(\mathbf{x}(t+1) - \mathbf{x}(t)), \theta\right) x \right\|^2 \\ &\leq \sum_{i=1}^{n} \lambda_i \left\| A^{-1/2} \mathcal{G}\left(\mathbf{x}(t) + k_i(\mathbf{x}(t+1) - \mathbf{x}(t)), \theta\right) x \right\|^2 \end{aligned}$$

$$= \sum_{i=1}^{n} \lambda_i x^\top \mathcal{G}^\top A^{-1} \mathcal{G} x \tag{52}$$

where $\mathcal{G}$ is evaluated at $\mathbf{x}(t) + k_i(\mathbf{x}(t+1) - \mathbf{x}(t))$. Therefore, if $\mathcal{G}^\top A^{-1}\mathcal{G} - A^{-1} \prec 0$ for all $\mathbf{x} \in \mathcal{X}$, then $\mathcal{V}(\mathbf{x}(t+1)) - \mathcal{V}(\mathbf{x}(t)) < 0$ whenever $y(\mathbf{x}(t)) \neq 0$, i.e., the Lyapunov function is strictly decreasing along system trajectories.

Finally, since $\psi_{\theta_i}(\mathbf{x}) = 0$ for $\mathbf{x} \in [\underline{\mathbf{x}}, \overline{\mathbf{x}}]$, we have $\mathcal{V}(\mathbf{x}(t+1)) - \mathcal{V}(\mathbf{x}(t)) = 0$ implies $\mathbf{x}(t) \in S_{\mathbf{x}}$. Given that $\mathcal{G}(\mathbf{x}, \theta) = I + I_{\Delta T} A \frac{\partial \boldsymbol{\pi}}{\partial \mathbf{x}}$, the condition becomes:

$$\left(I + I_{\Delta T} A \frac{\partial \boldsymbol{\pi}}{\partial \mathbf{x}}\right)^\top A^{-1} \left(I + I_{\Delta T} A \frac{\partial \boldsymbol{\pi}}{\partial \mathbf{x}}\right) - A^{-1} \prec 0 \tag{53}$$

Due to the decentralized structure, $\frac{\partial \boldsymbol{\pi}}{\partial \mathbf{x}}$ is diagonal. Expanding the matrix product yields:

$$-\frac{2}{\Delta T} A^{-1} \prec \frac{\partial \boldsymbol{\pi}}{\partial \mathbf{x}} \prec 0 \tag{54}$$

By LaSalle's Invariance Principle and the condition $\lim_{\|\mathbf{x}\| \to \infty} \|\psi_\theta(\mathbf{x})\| = \infty$, the stability condition is thus established as summarized in Theorem 1. Proof completed. ■

## Appendix B
## Proof of Lemma 1

*Proof:* Constraint (34) specifies a decreasing structure for the bias terms $b_i^l$, that is, $b_i^1 \geq b_i^2 \geq \cdots \geq b_i^M$, which imposes a monotonicity condition across the layers $l = 1, 2, \ldots, M$.

$$b_i^1 = 0, b_i^2 \leq b_i^1, b_i^3 \leq b_i^2, \ldots, b_i^m \leq b_i^{m-1} \tag{55}$$

This ensures that each sub-function $\psi_i^l(\mathbf{x}_i) = \mathbf{W}_i^l \cdot \sigma(\mathbf{1}\mathbf{x}_i + b_i^l)$ is sequentially activated as the input $\mathbf{x}_i$ increases. Specifically, when $\mathbf{x}_i \in [-b_i^2, -b_i^1)$, only the first unit $\psi_i^1$ is activated; when $\mathbf{x}_i \in [-b_i^3, -b_i^2)$, both $\psi_i^1$ and $\psi_i^2$ are activated simultaneously; and so on. Therefore, the overall function (56) is piecewise linear.

$$\phi_i^{(+)}(\mathbf{x}_i) = \sum_{l=1}^{m} \psi_i^l(\mathbf{x}_i) \tag{56}$$

where the slope of each segment corresponds to the cumulative sum of the weights associated with the currently activated units. Let the $k$-th segment correspond to the interval $[\mathbf{x}_k, \mathbf{x}_{k+1}]$, then its slope is given by $\mathbb{C}_k = \sum_{j=1}^{k} \mathbf{W}_i^j$. According to Constraint (33), the cumulative weights for each segment satisfy (57).

$$\sum_{j=1}^{k} \mathbf{W}_i^j \geq 0, \forall k = 1, 2, \ldots, m \tag{57}$$

Thus, the slope of each segment is non-negative, and the overall function remains monotonically increasing within its domain. Consequently, for any $x_a < x_b$, we have:

$$\phi_i^{(+)}(x_b) - \phi_i^{(+)}(x_a) = \int_{x_a}^{x_b} \frac{d\phi_i^{(+)}}{dx} dx \geq 0 \tag{58}$$

Moreover, the function near the origin satisfies: when $\mathbf{x}_i = 0$, since $b_i^1 = 0$ and for all $l \geq 2$, $b_i^l \leq 0$, we have $\rho(b_i^l) = \max(0, b_i^l) = 0$. Therefore, $\phi^{(+)}(0) = \sum_{l=1}^{m} \mathbf{W}_i^l \cdot \rho(b_i^l) = 0$. When $\mathbf{x}_i < 0$, because all $b_i^l \leq 0$, it follows that $\mathbf{x}_i + b_i^l < 0$, so $\rho(\mathbf{x}_i + b_i^l) = 0$ for all $l$, which implies $\phi_i^{(+)}(\mathbf{x}_i) = 0$. Similarly, according to constraints 36-37, it can be proven that the negative term construction $\phi_i^{(-)}(\mathbf{x}_i)$ is zero for $\mathbf{x}_i \geq 0$, and is a piecewise linear, monotonically increasing function for $\mathbf{x}_i < 0$. Proof completed. ■

## Appendix C
## Proof of Theorem 2

*Proof:* Let the first derivative of the function $r(x)$ be bounded by a constant $\alpha$ on the domain $\mathbb{X}$. Consider dividing $\mathbb{X}$ into an equidistant grid with interval $\beta = \frac{1}{n}$. For each interval $[k\beta, (k+1)\beta]$, we define the linear interpolation function as shown in (59).

$$y(x) = r(k\beta) + \frac{r((k+1)\beta) - r(k\beta)}{\beta}(x - k\beta) \tag{59}$$

The function satisfies $y(k\beta) = r(k\beta)$ and $y((k+1)\beta) = r((k+1)\beta)$ at the two endpoints. Since $r(x)$ is a monotonic function, it is easy to conclude that for any $x \in [k\beta, (k+1)\beta]$, the following holds:

$$r(k\beta) \leq r(x) \leq r((k+1)\beta), r(k\beta) \leq y(x) \leq r((k+1)\beta) \tag{60}$$

Thus, the upper bound of the interpolation error is shown in (61).

$$|y(x) - r(x)| \leq |r((k+1)\beta) - r(k\beta)|. \tag{61}$$

Next, by applying the mean value theorem, we have:

$$r((k+1)\beta) - r(k\beta) = \beta \cdot \frac{\partial r(c)}{\partial x} \tag{62}$$

where the point $c$ lies within the interval $(k\beta, (k+1)\beta)$. Since the derivative is bounded by $\alpha$, we further obtain:

$$|y(x) - r(x)| \leq \beta\alpha \tag{63}$$

Next, we consider constructing a class of piecewise linear functions as shown in (64).

$$y(x) = r(k\beta) + \frac{r((k+1)\beta) - r(k\beta)}{\beta}(x - k\beta) \tag{64}$$

We will show that this class of functions can be constructed by (32) and (35). For simplicity, let $y(x)$ be the non-negative part approximated by the function $\phi^{(+)}(\mathbf{x})$. Define the initial parameters as $b_1^1 = 0$ and $\mathbf{W}_1^1 = r(\beta)$, and then set:

$$b_k^* = (k-1)\beta, \sum_{j=1}^{k} \mathbf{W}_j^i = \frac{r(k\beta) - r((k-1)\beta)}{\beta},$$
$$k = 2, 3, \ldots, n. \tag{65}$$

The function $f^+(x)$ constructed in this way matches the behavior of $y(x)$ on each interval. Therefore, $|f(x) - r(x)| \leq \beta\alpha$. By choosing $\beta < \frac{\varepsilon}{\alpha}$, the error can be controlled within any given precision $\varepsilon$. Proof completed. ■

## References

[1] Y. Xue and W. Chen, "Efficient deceptive path planning for UAVs via attention-based reinforcement learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 13, pp. 539–551, 2026.

[2] S. Chaturvedi, Z. Liu, V. A. Bohara, A. Srivastava, and P. Xiao, "Resource allocation in SCMA-empowered multi-UAV transmission system," *IEEE Trans. Netw. Sci. Eng.*, vol. 13, pp. 815–827, 2026.

[3] N. C. Luong et al., "Incentive mechanisms for data relay and scene graph transmission in UAV-assisted networks with image fidelity awareness," *IEEE Trans. Commun.*, vol. 73, no. 12, pp. 15264–15278, Dec. 2025, doi: 10.1109/TCOMM.2025.3594777.

[4] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of Things for the future of smart agriculture: A comprehensive survey of emerging technologies," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 4, pp. 718–752, Apr. 2021.

[5] A. Taik, B. Nour, and S. Cherkaoui, "Empowering prosumer communities in smart grid with wireless communications and federated edge learning," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 26–33, Dec. 2021.

[6] Q. Tang, W. Sun, Z. Liu, Q. Li, and X. Yuan, "Multi-agent reinforcement learning based dynamic self-coordinated topology optimization for wireless mesh networks," *J. Netw. Comput. Appl.*, vol. 239, 2025, Art. no. 104177.

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

[8] M. Dai, C. Dou, Y. Wu, L. Qian, R. Lu, and T. Q. Quek, "Multi-UAV aided multi-access edge computing in marine communication networks: A joint system-welfare and energy-efficient design," *IEEE Trans. Commun.*, vol. 72, no. 9, pp. 5517–5531, Sep. 2024.

[9] N. N. Ei, M. Alsenwi, Y. K. Tun, Z. Han, and C. S. Hong, "Energy-efficient resource allocation in multi-UAV-assisted two-stage edge computing for beyond 5g networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16421–16432, Sep. 2022.

[10] Q.-V. Pham, S. Mirjalili, N. Kumar, M. Alazab, and W.-J. Hwang, "Whale optimization algorithm with applications to resource allocation in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4285–4297, Apr. 2020.

[11] M. M. Alam, M. Y. Arafat, S. Moh, and J. Shen, "Topology control algorithms in multi-unmanned aerial vehicle networks: An extensive survey," *J. Netw. Comput. Appl.*, vol. 207, 2022, Art. no. 103495.

[12] M. Johansson and L. Xiao, "Cross-layer optimization of wireless networks using nonlinear column generation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 435–445, Feb. 2006.

[13] S. Pollin et al., "MEERA: Cross-layer methodology for energy efficient resource allocation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 617–628, Feb. 2007.

[14] F. Meshkati, H. V. Poor, S. C. Schwartz, and R. V. Balan, "Energy-efficient resource allocation in wireless networks with quality-of-service constraints," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3406–3414, Nov. 2009.

[15] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[16] C. Jiang et al., "Cross-layer optimization for multi-hop wireless networks with successive interference cancellation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5819–5831, 2016.

[17] D. W. K. Ng and R. Schober, "Resource allocation and scheduling in multi-cell OFDMA systems with decode-and-forward relaying," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2246–2258, Jul. 2011.

[18] D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Multiuser MISO UAV communications in uncertain environments with no-fly zones: Robust trajectory and resource allocation design," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3153–3172, May 2020.

[19] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for IRS-assisted full-duplex cognitive radio systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7376–7394, Dec. 2020.

[20] G. Tychogiorgos and K. K. Leung, "Optimization-based resource allocation in communication networks," *Comput. Netw.*, vol. 66, pp. 32–45, Jun. 2014.

[21] A. Zappone and E. A. Jorswieck, "Energy-efficient resource allocation in future wireless networks by sequential fractional programming," *Digit. Signal Process.*, vol. 60, pp. 324–337, Jan. 2017.

[22] A. Zappone et al., "Globally optimal energy-efficient power control and receiver design in wireless networks," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2844–2859, Jun. 2017.

[23] V. A. Kumar, M. V. Marathe, S. Parthasarathy, and A. Srinivasan, "Algorithmic aspects of capacity in wireless networks," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, 2005, pp. 133–144.

[24] L. Liu, Y. Cheng, X. Cao, S. Zhou, Z. Niu, and P. Wang, "Joint optimization of scheduling and power control in wireless networks: Multi-dimensional modeling and decomposition," *IEEE Trans. Mobile Comput.*, vol. 18, no. 7, pp. 1585–1600, Jul. 2019.

[25] Q. Yang, J. Liu, Z. Wu, and S. He, "A fusion algorithm based on whale and grey wolf optimization algorithm for solving real-world optimization problems," *Appl. Soft Comput.*, vol. 146, 2023, Art. no. 110701.

[26] M. Bey, P. Kuila, B. B. Naik, and S. Ghosh, "Quantum-inspired particle swarm optimization for efficient IoT service placement in edge computing systems," *Expert Syst. Appl.*, vol. 236, 2024, Art. no. 121270.

[27] M. E. Khanouche, Y. Amirat, A. Chibani, M. Kerkar, and A. Yachir, "Energy-centered and QoS-aware services selection for Internet of Things," *IEEE Trans. Automat. Sci. Eng.*, vol. 13, no. 3, pp. 1256–1269, Jul. 2016.

[28] Y. Hou, Y. Shi, and H. D. Sherali, "Rate allocation and network lifetime problems for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 722–733, Apr. 2006.

[29] Y. Lin, J. Zhang, H. S.-H. Chung, W. H. Ip, Y. Li, and Y.-H. Shi, "An ant colony optimization approach for maximizing the lifetime of heterogeneous wireless sensor networks," *IEEE Trans. Syst., Man, Cybern., Part C, (Appl. Rev.)*, vol. 42, no. 3, pp. 408–420, May 2012.

[30] R. V. Kulkarni and G. K. Venayagamoorthy, "Particle swarm optimization in wireless-sensor networks: A brief survey," *IEEE Trans. Syst., Man, Cybern., Part C, (Appl. Rev.)*, vol. 41, no. 2, pp. 262–267, Mar. 2011.

[31] Z. Sun, Y. Xu, G. Liang, and Z. Zhou, "An intrusion detection model for wireless sensor networks with an improved V-detector algorithm," *IEEE Sensors J.*, vol. 18, no. 5, pp. 1971–1984, Mar. 2018.

[32] D. Zhang and J. Zhang, "Multi-species evolutionary algorithm for wireless visual sensor networks coverage optimization with changeable field of views," *Appl. Soft Comput.*, vol. 96, 2020, Art. no. 106680.

[33] N. T. Tam et al., "Multifactorial evolutionary optimization to maximize lifetime of wireless sensor network," *Inf. Sci.*, vol. 576, pp. 355–373, 2021.

[34] A. Konstantinidis et al., "A multi-objective evolutionary algorithm for the deployment and power assignment problem in wireless sensor networks," *Comput. Netw.*, vol. 54, no. 6, pp. 960–976, 2010.

[35] K. P. Ferentinos and T. A. Tsiligiridis, "A memetic algorithm for optimal dynamic design of wireless sensor networks," *Comput. Commun.*, vol. 30, no. 13–14, pp. 2753–2764, Sep. 2007.

[36] M. Khodier and G. Saleh, "Beamforming and power control for interference reduction in wireless communications using particle swarm optimization," *AEU–Int. J. Electron. Commun.*, vol. 64, no. 6, pp. 489–502, Jun. 2010.

[37] R. Zhang et al., "Generative AI-enabled vehicular networks: Fundamentals, framework, and case study," *IEEE Netw.*, vol. 38, no. 4, pp. 259–267, Jul. 2024.

[38] R. Zhang et al., "Generative AI for space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 10–20, Dec. 2024.

[39] S. Zhang, O. T. Ajayi, and Y. Cheng, "A self-supervised learning approach for accelerating wireless network optimization," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8074–8087, Jun. 2023.

[40] M. Hua et al., "Multi-agent reinforcement learning for connected and automated vehicles control: Recent advancements and future prospects," *IEEE Trans. Automat. Sci. Eng.*, vol. 22, pp. 16266–16286, 2025.

[41] P. Liu, H. Bou-Ammar, J. Peters, and D. Tateo, "Safe reinforcement learning on the constraint manifold: Theory and applications," *IEEE Trans. Robot.*, vol. 41, pp. 3442–3461, 2025.

[42] O. Alhussein and W. Zhuang, "Dynamic topology design of NFV-enabled services using deep reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1228–1238, Jun. 2022.

[43] O. Franek, "Phasor alternatives to Friis' transmission equation," *IEEE Antennas Wireless Propag. Lett.*, vol. 17, no. 1, pp. 90–93, Jan. 2018.

[44] M. Ding, Y. Guo, Z. Huang, B. Lin, and H. Luo, "Grom: A generalized routing optimization method with graph neural network and deep reinforcement learning," *J. Netw. Comput. Appl.*, vol. 229, 2024, Art. no. 103927.

[45] X. Lin et al., "Intelligent adaptive MIMO transmission for nonstationary communication environment: A deep reinforcement learning approach," *IEEE Trans. Commun.*, vol. 73, no. 8, pp. 5965–5979, Aug. 2025.

[46] D. G. S. Pivoto, F. A. P. d. Figueiredo, C. Cavdar, G. R. d. L. Tejerina, and L. L. Mendes, "A comprehensive survey of machine learning applied to resource allocation in wireless communications," *IEEE Commun. Surv. Tut.*, vol. 28, pp. 1986–2053, 2026.

[47] Y. Shi et al., "Machine learning for large-scale optimization in 6G wireless networks," *IEEE Commun. Surv. Tut.*, vol. 25, no. 4, pp. 2088–2132, Fourthquarter 2023.
[48] Y. Qin et al., "Deep reinforcement learning based resource allocation and trajectory planning in integrated sensing and communications UAV network," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8158–8169, Nov. 2023.
[49] P.-G. Ye, Y.-G. Wang, and W. Tang, "S-MFRL: Spiking mean field reinforcement learning for dynamic resource allocation of D2D networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1032–1047, Jan. 2023.
[50] W. Wu, F.-C. Yang, F.-H. Zhou, Q. Wu, and R. Q. Hu, "Intelligent resource allocation for IRS-Enhanced OFDM communication systems: A hybrid deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 4028–4042, Jun. 2023.
[51] D.-D. Yan, B. K. Ng, W. Ke, and C. T. Lam, "Deep reinforcement learning based resource allocation for network slicing with massive MIMO," *IEEE Access*, vol. 11, pp. 75899–75911, 2023.
[52] A. Oliveira and T. Vazão, "Towards green machine learning for resource allocation in beyond 5G RAN slicing," *Comput. Netw.*, vol. 233, 2023, Art. no. 109877.
[53] L. R. Frank, A. Galletta, L. Carnevale, A. B. Vieira, and E. F. Silva, "Intelligent resource allocation in wireless networks: Predictive models for efficient access point management," *Comput. Netw.*, vol. 254, 2024, Art. no. 110762.
[54] Á. G. Andrade and A. Anzaldo, "Accelerated resource allocation based on experience retention for B5G networks," *J. Netw. Comput. Appl.*, vol. 213, 2023, Art. no. 103593.
[55] M. Liaq, S. Sharif, S. Zeadally, and W. Ejaz, "Utilization of machine learning in future wireless networks for resource optimization: A survey," *Ad Hoc Netw.*, vol. 178, 2025, Art. no. 103983.
[56] G. S. Kori et al., "Wireless sensor networks and machine learning centric resource management schemes: A survey," *Ad Hoc Netw.*, vol. 167, 2025, Art. no. 103698.
[57] J. Huang et al., "Deep reinforcement learning-based spectrum resource allocation for the web of healthcare things with massive integrating wearable gadgets," *Digit. Commun. Netw.*, vol. 11, no. 3, pp. 671–680, 2025.
[58] C. Battiloro, P. Di Lorenzo, M. Merluzzi, and S. Barbarossa, "Lyapunov-based optimization of edge resources for energy-efficient adaptive federated learning," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 1, pp. 265–280, Mar. 2023.
[59] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Multi-user goal-oriented communications with energy-efficient edge resource management," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 4, pp. 1709–1724, Dec. 2023.
[60] C. Ding, J.-B. Wang, M. Cheng, M. Lin, and J. Cheng, "Dynamic transmission and computation resource optimization for dense LEO satellite assisted mobile-edge computing," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 3087–3102, May 2023.
[61] W. Zhao et al., "DRL connects Lyapunov in delay and stability optimization for offloading proactive sensing tasks of RSUs," *IEEE Trans. Mobile Comput.*, vol. 23, no. 7, pp. 7969–7982, Jul. 2024.
[62] J. Zhou, Q. Yang, L. Zhao, H. Dai, and F. Xiao, "Mobility-aware computation offloading in satellite edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 10, pp. 9135–9149, Oct. 2024.
[63] Q. Tang et al., "Joint service deployment and task scheduling for satellite edge computing: A two-timescale hierarchical approach," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 5, pp. 1063–1079, May 2024.
[64] Z. Huang, Z.-P. Jiang, Z. Han, and Y. Liu, "Robust Lyapunov optimization for LEO satellite networks routing control," *IEEE Trans. Mobile Comput.*, vol. 24, no. 12, pp. 13293–13308, Dec. 2025.
[65] J. Wang, L. Wang, K. Zhu, and P. Dai, "Lyapunov-based joint flight trajectory and computation offloading optimization for UAV-assisted vehicular networks," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 22243–22256, Jun. 2024.
[66] J. Zhang, Y. Zhai, Z. Liu, and Y. Wang, "A Lyapunov-based resource allocation method for edge-assisted industrial Internet of Things," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 39464–39472, Dec. 2024.
[67] M. Wei, W. Yu, D. Chen, M. Kang, and G. Cheng, "Privacy distributed constrained optimization over time-varying unbalanced networks and its application in federated learning," *IEEE/CAA J. Automatica Sinica*, vol. 12, no. 2, pp. 335–346, Feb. 2025.
[68] X. Zhang et al., "Energy-efficient computation peer offloading in satellite edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 4, pp. 3077–3091, Apr. 2024.
[69] W. Lanet al., "Security-sensitive task offloading in integrated satellite-terrestrial networks," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 2220–2233, Mar. 2025, doi: 10.1109/TMC.2024.3489619.
[70] 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.
[71] M. M. H. Qazzaz et al., "Non-terrestrial UAV clients for beyond 5G networks: A comprehensive survey," *Ad Hoc Netw.*, vol. 157, 2024, Art. no. 103440.
[72] F. Wanet al., "Advancements and challenges in UAV-based communication networks: A comprehensive scholarly analysis," *Results Eng.*, vol. 24, 2024, Art. no. 103271.
[73] M. Mahbub et al., "UAV-assisted wireless communications in the 6G-and-beyond era: An extensive survey on characteristics, standardization and regulations, enabling technologies, challenges, and future directions," *Veh. Commun.*, vol. 56, 2025, Art. no. 100977.
[74] Y. Zhou, X. Liu, X. Zhai, Q. Zhu, and T. S. Durrani, "UAV-enabled integrated sensing, computing, and communication for Internet of Things: Joint resource allocation and trajectory design," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 12717–12727, Apr. 2024.
[75] H. Hellaoui, M. Bagaa, A. Chelli, T. Taleb, and B. Yang, "On supporting multiservices in UAV-enabled aerial communication for Internet of Things," *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13754–13768, Aug. 2023.
[76] Q. Li, L. Shi, Z. Zhang, and G. Zheng, "Resource allocation in UAV-enabled wireless powered MEC networks with hybrid passive and active communications," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2574–2588, Feb. 2023.
[77] L. Wu et al., "UAV-assisted maritime legitimate surveillance: Joint trajectory design and power allocation," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13701–13705, Oct. 2023.
[78] J. Gaspar, T. Cruz, C.-T. Lam, and P. Simões, "Smart substation communications and cybersecurity: A comprehensive survey," *IEEE Commun. Surv. Tut.*, vol. 25, no. 4, pp. 2456–2493, Fourth Quarter 2023.
[79] C. Zhao, W. Sun, Z. Fang, J. Wang, Q. Li, and H. Zhang, "End-to-end delay optimisation for ieee 802.11 string topology multi-hop wireless networks in overhead transmission line system," *IET Commun.*, vol. 15, no. 3, pp. 487–495, 2021.
[80] W. Sun, L. Zhang, Q. Lv, Z. Liu, W. Li, and Q. Li, "Dynamic collaborative optimization of end-to-end delay and power consumption in wireless sensor networks for smart distribution grids," *Comput. Commun.*, vol. 202, pp. 87–96, 2023.
[81] W. Sun et al., "Multi-agent reinforcement learning for dynamic topology optimization of mesh wireless networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 10501–10513, Sep. 2024.
[82] W. Zhang, X. Lin, and B.-S. Chen, "LaSalle-type theorem and its applications to infinite horizon optimal control of discrete-time nonlinear stochastic systems," *IEEE Trans. Autom. Control*, vol. 62, no. 1, pp. 250–261, Jan. 2017.
[83] X. Liu, X. Han, N. Zhang, and Q. Liu, "Certified monotonic neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 15427–15438.
[84] J. Feng, Y. Shi, G. Qu, S. H. Low, A. Anandkumar, and A. Wierman, "Stability constrained reinforcement learning for decentralized real-time voltage control," *IEEE Trans. Control Netw. Syst.*, vol. 11, no. 3, pp. 1370–1381, Sep. 2024.
[85] R. Zhou, W. Pu, M.-Y. You, and Q. Shi, "Harnessing monotonic neural networks for performance prediction and threshold determination in multichannel detection," *IEEE Trans. Signal Process.*, vol. 73, pp. 2154–2169, 2025.
[86] Q. Tang et al., "Multi-agent reinforcement learning based delay and power optimization for UAV-WMN substation inspection," *IEEE Trans. Netw. Service Manag.*, vol. 22, no. 4, pp. 3060–3076, Aug. 2025.
[87] A. I. Ameur, O. S. Oubbati, A. Lakas, A. Rachedi, and M. B. Yagoubi, "Efficient vehicular data sharing using aerial p2p backbone," *IEEE Trans. Intell. Veh.*, vol. 10, no. 1, pp. 413–426, Jan. 2025.
[88] X. Kong, C. Ni, G. Duan, G. Shen, Y. Yang, and S. K. Das, "Energy consumption optimization of UAV-assisted traffic monitoring scheme with tiny reinforcement learning," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 21135–21145, Jun. 2024.
[89] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[90] C. Yu et al., "The surprising effectiveness of PPO in cooperative multi-agent games," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 24611–24624.
[91] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *J. Mach. Learn. Res.*, vol. 21, no. 178, pp. 1–51, 2020.
[92] S. Janković and M. Merkle, "A mean value theorem for systems of integrals," *J. Math. Anal. Appl.*, vol. 342, no. 1, pp. 334–339, 2008.
[93] C. Wang et al., "Resource scheduling based on deep reinforcement learning in UAV assisted emergency communication networks," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3834–3848, Jun. 2022.
[94] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, and Z. Gao, "UAV trajectory planning for AoI-minimal data collection in UAV-aided IoT networks by transformer," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1343–1358, Feb. 2023.
[95] R. Zhang, K. Xiong, Y. Lu, D. W. K. Ng, P. Fan, and K. B. Letaief, "SWIPT-enabled cell-free massive MIMO-NOMA networks: A machine learning-based approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6701–6718, Jul. 2024.
[96] R. Zhang et al., "Toward edge general intelligence with agentic AI and agentification: Concepts, technologies, and future directions," *IEEE Commun. Surv. Tut.*, vol. 28, pp. 4285–4318, 2026.

**Qingwei Tang** (Student Member, IEEE) is currently working toward the Ph.D. degree with the School of Electrical and Automation Engineering, Hefei University of Technology, Hefei, China. His research interests include multi-agent reinforcement learning and its applications in complex systems, particularly wireless communication networks, modern power systems, improving the intelligence, sustainability, and resilience of critical infrastructures through advanced learning, and optimization methods.

**Wei Sun** (Senior Member, IEEE) received the B.E. degree in automation, the M.S. degree in detection technology and automatic equipment, and the Ph.D. degree in electrical engineering from the Hefei University of Technology, Hefei, China, in 2004, 2007, and 2012, respectively. He is currently a Professor with the Hefei University of Technology. His research interests include wireless networks, networked control systems, and microgrids.

**Zhi Liu** (Senior Member, IEEE) received the Ph.D. degree in informatics from the National Institute of Informatics, Tokyo, Japan. He is currently an Associate Professor with The University of Electro-Communications, Chofu, Japan. His research interests include video network transmission and mobile edge computing. He is the Editorial Board member of Springer *Wireless Networks* and IEEE TRANSACTIONS ON MULTIMEDIA.

**Nikos D. Hatziargyriou** (Life Fellow, IEEE) has been with the National Technical University of Athens (NTUA), Athens, Greece, since 1984. He became Professor of power systems in 1995 and was named Professor Emeritus with NTUA in 2022. He is also a Part-time Professor with the University of Vaasa, Vaasa, Finland. He has more than 10 years of industrial experience, having served as the Chair and CEO of Hellenic Distribution Network Operator and Executive Vice-Chair and Deputy CEO of Public Power Corporation, with responsibility for the Transmission and Distribution Divisions. He has authored or coauthored more than 300 journal papers and 600 conference papers. He has participated in more than 60 R&D projects funded by the European Commission, electric utilities, and industry, covering both fundamental research and practical applications. He served as the Chair and Vice-Chair of ETIP-SNET. He was the past Editor-in-Chief of IEEE TRANSACTIONS ON POWER SYSTEMS. He is currently the Editor-in-Chief-at-Large of IEEE Power and Energy Society Transactions. He was included in the Thomson Reuters lists of the top 1% most cited researchers in 2016, 2017, and 2019, respectively. His honors include the 2020 Globe Energy Prize, 2017 IEEE/PES Prabha S. Kundur Power System Dynamics and Control Award, and 2023 IEEE Herman Halperin Electric Transmission and Distribution Award.

**Yang Xiao** (Fellow, IEEE) received the B.S. and M.S. degrees in computational mathematics from Jilin University, Changchun, China, in 1989 and 1991, respectively, and the M.S. (second) and Ph.D. degrees in computer science and engineering from Wright State University, Dayton, OH, USA, in 2000 and 2001, respectively. He is currently a Full Professor with the Department of Computer Science, The University of Alabama, Tuscaloosa, AL, USA. Dr. Xiao directed more than 20 doctoral dissertations and supervised more than 20M.S. theses/projects. He has authored or coauthored more than 300 Science Citation Index (SCI)-indexed journal papers (including more than 70 IEEE/ACM Transactions) and 300 Engineering Index (EI)-indexed refereed conference papers and book chapters related to his research interests which include cyber-physical systems, Internet of Things, security, wireless networks, smart grids, and telemedicine. Dr. Xiao was a Voting Member of IEEE 802.11 Working Group from 2001 to 2004, involving IEEE 802.11 (Wi-Fi) standardization work. He is a fellow of IET, AAIA, AIIA, and ACIS. Dr. Xiao was the Guest Editor 37 times of different international journals, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS during 2022–2023, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING in 2021, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING in 2021, IEEE NETWORK in 2007, IEEE WIRELESS COMMUNICATIONS in 2006 and 2021, respectively, *IEEE Communications Standards Magazine* in 2021, and *Mobile Networks and Applications* (MONET)(ACM/Springer) in 2008. He is also the Editor-in-Chief of *Cyber-Physical Systems Journal*, *International Journal of Sensor Networks*, and *International Journal of Security and Networks*. Dr. Xiao has been the Editorial Board member or Associate Editor for 20 international journals, including IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING since 2022, IEEE TRANSACTIONS ON CYBERNETICS since 2020, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS from 2014 to 2015, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2009, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2007 to 2014. He is/was a member of Technical Program Committee for more than 300 conferences. He was the recipient of the IEEE TNSE Excellent Editor Award in 2022 and 2023, respectively.

**Xiaohui Yuan** (Senior Member, IEEE) is currently an Associate Professor and Director of Computer Vision and Intelligent Systems Lab, University of North Texas, Denton, TX, USA. His research interests include artificial intelligence and machine learning. Dr. Yuan was the recipient of the Ralph E. Powe Professor Award in 2008 and U.S. Air Force Visiting Professor Award in 2011, 2012, and 2013, respectively. He serves as an Associate Editor, Editorial Board member, and Guest Editor for several journals and Organizing member of many international conferences.