

Article

Distributed Decision Making for Electromagnetic Radiation Source Localization Using Multi-Agent Deep Reinforcement Learning

Jiteng Chen ¹, Zehui Zhang ^{1,*}, Dan Fan ², Chaoqun Hou ¹, Yue Zhang ³, Teng Hou ¹, Xiangni Zou ⁴ and Jun Zhao ⁵

¹ School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China; 3120245864@bit.edu.cn (J.C.); houchaoqun@bit.edu.cn (C.H.); 3120245514@bit.edu.cn (T.H.)

² China Satellite Network E-Link Co., Ltd., Beijing 100029, China; xuke19820404@163.com

³ School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; zhangyue@bit.edu.cn

⁴ School of Economics, Jinan University, Guangzhou 510632, China

⁵ The 22th Research Institute of China Electronics Technology Group Corporation, Qingdao 266107, China; zhaojunxinxiang@sohu.com

* Correspondence: zzhui@bit.edu.cn

Abstract: The detection and localization of radiation sources in urban areas present significant challenges in electromagnetic spectrum operations, particularly with the proliferation of small UAVs. To address these challenges, we propose the Multi-UAV Reconnaissance Proximal Policy Optimization (MURPPO) algorithm based on a distributed reinforcement learning framework, which utilizes an independent decision making mechanism and collaborative positioning method with multiple UAVs to achieve high-precision detection and localization of radiation sources. We adopt a dual-branch actor structure for independent decisions in UAV control, which reduces decision complexity and improves learning efficiency. The algorithm incorporates task-specific knowledge into the reward function design to guide UAVs in exploring abnormal radiation sources. Furthermore, we employ a geometry-based three-point localization algorithm that leverages multiple UAVs' spatial distribution for precise abnormal radiation source positioning. Simulations in urban environments demonstrate the effectiveness of the MURPPO algorithm, with the proportion of successfully localized target radiation sources converging to 56.5% in the later stages of training, approaching a 38.5% improvement over a traditional multi-agent proximal policy optimization algorithm. The results indicate that MURPPO effectively addresses the challenges of the intelligent sensing and localization of UAVs in complex urban electromagnetic spectrum operations.

Keywords: deep reinforcement learning; electromagnetic spectrum; localization; multi-UAV



Academic Editor: Oleg Yakimenko

Received: 20 January 2025

Revised: 7 March 2025

Accepted: 9 March 2025

Published: 18 March 2025

Citation: Chen, J.; Zhang, Z.; Fan, D.; Hou, C.; Zhang, Y.; Hou, T.; Zou, X.; Zhao, J. Distributed Decision Making for Electromagnetic Radiation Source Localization Using Multi-Agent Deep Reinforcement Learning. *Drones* **2025**, *9*, 216. <https://doi.org/10.3390/drones9030216>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The electromagnetic spectrum represents a finite natural resource that has taken on an increasingly crucial role in modern urban development [1]. As smart cities continue to evolve and various wireless technologies and communication systems become more widespread [2], electromagnetic spectrum monitoring and control have emerged as vital components of urban infrastructure management [3]. To ensure the smooth operation of smart cities, effective spectrum operations is essential for maintaining reliable communications and efficient city system operations. Such operation must also be capable of

swiftly identifying potential sources of interference [4]. In densely populated urban areas with complex electromagnetic environments, spectrum operations continue to face unique challenges [5].

The rapid proliferation of small UAVs has introduced new challenges to smart city management [6]. These devices, characterized by their accessibility, high mobility, and low visibility [7], pose potential threats to urban airspace management and critical infrastructure security [8]. Consequently, substantial resources are being invested in developing UAV detection and monitoring systems across various urban environments [9]. While current systems incorporate advanced sensing technologies such as signal detection [10], positioning systems, and artificial intelligence [11], developing rapid and robust closed-loop perception and positioning strategies for unauthorized radiation sources (including unregistered UAVs) in complex urban environments remains a crucial task that requires urgent attention. Given these emerging challenges, conventional single-point or static sensing [12] approaches are increasingly inadequate for complex urban environments. We need a more flexible and comprehensive solution to ensure effective spectrum operations and smart city security management.

To tackle the above challenges, we focus on multi-UAV electromagnetic spectrum collaborative reconnaissance and countermeasures for key urban areas. The aim is to achieve a closed-loop process of the detection, localization, and tracking of abnormal radiation sources in complex electromagnetic environments. The contributions of this paper are detailed as follows:

- **Advanced Multi-UAV Framework:** We develop a distributed reinforcement learning framework that enables independent decision-making while facilitating collaborative positioning among multiple UAVs. This framework specifically addresses the challenges of radiation source detection and localization in complex urban environments.
- **Innovative Algorithm Design:** We propose the MURPPO algorithm, an innovative approach that features a dual-branch actor structure to reduce decision complexity and improve learning efficiency. The algorithm incorporates task-specific domain knowledge into its reward function design, enabling effective guidance for UAVs in their exploration tasks. Additionally, we develop a geometry-based three-point localization method that leverages the spatial distribution of multiple UAVs to achieve precise positioning of radiation sources.
- **Performance Validation:** Our extensive simulations demonstrate that MURPPO achieves a 56.5% successful localization rate in later training stages, showing a 38.5% improvement over traditional multi-agent PPO algorithms with robust performance in complex urban electromagnetic environments.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 introduces the system model. Section 4 provides the problem definition and analysis. Section 5 describes the proposed MURPPO algorithm. Performance analysis is presented in Section 6. Finally, Section 7 concludes the paper.

2. Related Work

Current approaches to UAV perception and detection primarily fall into two categories, traditional heuristic algorithms and reinforcement learning algorithms. This section reviews key developments in both approaches and identifies existing research gaps.

2.1. Traditional Heuristic Algorithms

Traditional heuristic algorithms have demonstrated notable efficacy in specific, well-defined environments [12]. These algorithms use specific problem information and heuristic rules to find solutions. They incorporate concepts such as greedy strategies, local search,

and simulations of natural processes to efficiently identify acceptable solutions for complex problems. While these approaches are straightforward and efficient, they tend to become stuck in local optima when dealing with complex problems and require extensive fine-tuning.

In the field of radiation source localization, significant contributions have been made using various heuristic approaches. Abdelhakim [13] explored heuristic-based maximum likelihood estimation for radioactive source localization. This is achieved through the implementation of various optimization algorithms (FFA, PSO, ACO, ABC), k-sigma detection method, and parameter estimation using sensor network readings, significantly reducing computation time while maintaining localization accuracy. Yu et al. proposed in [14] an Improved Multi-UAV Collaborative Search Algorithm based on Binary Search Algorithm (IMUCS-BSAE), which supports multi-UAV collaborative search under UAV energy consumption constraints, ensuring higher efficiency and stronger scalability of the algorithm. Additionally, a Motion-encoded Particle Swarm Optimization algorithm (MPSO) is proposed in [15], where UAV-based target search is converted into a probability-based cost function optimization using Bayesian theory, significantly improving the search efficiency for mobile targets.

2.2. Reinforcement Learning Algorithms

Recent years have seen major breakthroughs in artificial intelligence across fields like reinforcement learning [16], natural language processing [17], and computer vision [18]. Meanwhile, significant progress has also been made in resource allocation, including multi-UAV cooperative computing [19], intelligent decision-making [20], and reinforcement learning-based optimization [21]. Additionally, artificial intelligence algorithms have been successfully applied to radiation source sensing and localization [22]. As a key branch of artificial intelligence, reinforcement learning has made significant strides in target perception and detection, showing great promise in electromagnetic spectrum warfare with several notable achievements [23].

In the context of single-agent reinforcement learning, Wu et al. [24] proposed an adaptive transition speed Q-learning algorithm, enabling UAVs to autonomously navigate and plan paths for multiple search and rescue tasks in unknown environments. This is achieved through a two-stage strategy (rapid task search and optimal path planning), enhanced state and action space design, a composite reward function, and Q-table initialization. To address the limitations of traditional Q-learning due to the size of the Q-table, deep Q-learning can be employed for improvements. In [25], a multi-UAV control method based on deep Q-learning is proposed, achieving high-precision, low-cost tracking of multiple emergency responders in complex environments. This approach effectively addresses challenges such as obstacles, occlusions, and measurement noise. The actor–critic reinforcement learning algorithm [26] addresses the challenges of mathematical modeling and high-dimensional value functions in UAV area coverage. By optimizing UAV action timing, it achieves more energy-efficient coverage compared to traditional methods, providing a new solution for efficient area coverage in complex environments.

In the field of multi-agent systems, Hou et al. [27] proposed a distributed collaborative technique based on multi-agent deep deterministic policy gradient for UAV swarm search. This approach combines scene decomposition with centralized training and distributed execution, achieving high scalability and efficient collaboration in large-scale scenarios. It enhances search coverage and task efficiency while significantly reducing computational and communication resource consumption. Alagha et al. [28] developed a multi-agent deep reinforcement learning framework for target localization under uncertainty. This is achieved through PPO-based optimization for multi-agent decision making, a three-dimensional action space design (search, detection, and reachability assessment), and transfer learning

mechanisms for handling unreachable targets. Liu et al. [29] proposed a three-body cooperative active defense guidance law for air combat scenarios. This is achieved through a small speed ratio perspective analysis, zero-effort-miss and zero-effort-velocity concepts, and a sliding mode control method with overload constraints. These successful applications demonstrate the inherent flexibility and adaptability of reinforcement learning in addressing complex problem domains.

2.3. Research Gaps and Our Approach

Compared to heuristic algorithms, reinforcement learning offers significant advantages such as strong adaptability, consideration of long-term rewards, and model-free characteristics. Reinforcement learning can automatically adjust its strategy through continuous interaction with the environment, adapting to dynamic changes. It focuses on long-term rewards, enabling it to find optimal solutions in UAV perception and detection problems. Additionally, this method does not rely on an explicit environmental model, making it suitable for learning in unknown or complex situations.

A detailed analysis of existing reinforcement learning approaches reveals several specific limitations in current research. The work in [24] proposed adaptive transition speed Q-learning, but it is constrained by Q-table size, which limits its scalability in complex environments. While [25,28] introduced the multi-UAV control framework and multi-agent DRL framework, respectively, both rely solely on single-point position estimation or single UAV readings rather than utilizing multi-UAV triangulation for improved accuracy. The approach in [26] handles high-dimensional value functions for energy-efficient area coverage but is restricted to coverage tasks without closing the perception-localization loop. Although [27] implemented centralized training with distributed execution, it primarily focuses on reducing search overlap without leveraging multi-UAV cooperation for precise localization. The three-body cooperative defense guidance proposed in [29] lacks the incorporation of advanced intelligent optimization algorithms. These methods are summarized in Table 1. Our proposed MURPPO algorithm addresses these identified limitations through the following:

- It establishes a distributed multi-UAV collaborative perception and localization architecture to facilitate efficient cooperative operations.
- It integrates a branched actor network architecture for enhanced computational efficiency while utilizing trilateration methods to exploit the spatial distribution of UAVs for precise localization.
- It uses sophisticated reward mechanisms to ensure robust closed-loop perception and detection.

Table 1. Comparison of different reinforcement learning approaches.

Reference	Main Contributions	Key Limitations	Proposed Solutions
[24]	- Adaptive transition speed Q-learning - Two-stage strategy: rapid search and optimal planning	- Limited by Q-table size - Only for single UAV scenarios - Limited state-action space representation	- Deep neural network to handle large state spaces - Multi-UAV cooperative framework
[25]	Multi-UAV control framework	Relies solely on single-point position estimation	Multi-UAV triangulation for improved accuracy
[26]	Handles high dimensional value functions	The perception and localization loop has not been closed	Closed-loop control with real-time feedback

Table 1. *Cont.*

Reference	Main Contributions	Key Limitations	Proposed Solutions
[27]	- Centralized training, distributed execution - Enhanced search coverage and efficiency	The focus is on reducing search overlap and coverage, without using multiple UAVs for precise triangulation	- Cooperative triangulation strategy - Optimized multi-UAV positioning
[28]	- Multi-Agent DRL framework - Transfer learning for unreachable targets	Relies on single UAV readings instead of multi-UAV triangulation for better accuracy	- Collaborative sensing approach - Enhanced measurement fusion algorithm
[29]	Three-body cooperative defense guidance	Lacks intelligent optimization algorithms	Adaptive DRL optimization strategy

3. System Model

This section constructs environmental and UAV models. It establishes the environmental settings for critical urban areas, defines the UAV motion model, and sets the minimum safe distances between UAVs, as well as between UAVs and obstacles to ensure operational safety. For better readability, Table 2 provides definitions of the key mathematical symbols and parameters used in this paper, including essential components related to UAV movement, radiation detection, and the MURPPO algorithm framework.

Table 2. Principal mathematical notations and parameters in multi-UAV radiation source detection.

Symbol	Definition
PT	Position information table
ST	Sensing information table
RT	Radiation propagation table
P_r	Received radiation signal power
P_t	Transmitted radiation signal power
B_i	Directional sensing matrix of UAV i
ST_i	Sensing information table for UAV i
$d_i(t)$	Distance from UAV i to radiation source at time t
$p_i(t)$	Position coordinates of UAV i at time t
$\phi_i(t)$	Direction angle of UAV i at time t
$v_i(t)$	Speed of UAV i at time t
$\psi_i(t)$	Sensor heading angle of UAV i at time t
M	Set of valid positions within map boundaries
$VM(t)$	Set of UAVs with valid measurements at time t
$I(t)$	Indicator function for successful localization at time t
$A_{t,i}^{\text{norm}}$	Normalized advantage function of UAV i at time t
$L_{\text{move}}^{\text{CLIP}}(\theta_i)$	Clipped policy loss for movement actions of UAV i
$L_{\text{scan}}^{\text{CLIP}}(\theta_i)$	Clipped policy loss for scanning actions of UAV i
$S[\pi_{\theta_i, \text{move}}]$	Entropy of movement policy of UAV i
$S[\pi_{\theta_i, \text{scan}}]$	Entropy of scanning policy of UAV i
$L^{\text{MURPPO}}(\theta_i)$	Overall MURPPO objective function of UAV i

3.1. Environment Model

In the urban environment depicted in Figure 1, multiple UAVs engage in collaborative search operations to locate anomalous radiation source targets. We construct an intelligent

spectrum game platform for multi-UAV cooperation, integrating it with reinforcement learning framework to train and evaluate multi-UAV models. The target search area is characterized as a rectangular two-dimensional plane with dimensions $X \times Y$, further divided into grids. Let $\mathcal{T} = \{0, 1, \dots, T\}$ denote the discrete time horizon. In the defensive reconnaissance process, the red team has multiple UAVs with sensing and localization capabilities, while the blue team has one anomalous radiation source. Each UAV is assigned a unique identifier. The red team's UAVs have movement and scan rotation capabilities, equipped with both sensing and localization abilities. Initially, the red team's UAVs are distributed across the map at a fixed altitude H . The blue team's radiation source is randomly placed in certain critical areas of the urban map, hovering to gather intelligence, also at a fixed altitude H . The red team's UAVs perform sensing and detection operations in the defense area to locate the blue team's radiation source. When three or more red team's UAVs successfully detect the target simultaneously, a three-point localization process is triggered. The UAVs then continue to track and localize the target.

The urban electromagnetic spectrum model employs a layered information table design, including position information table PT , sensing information table ST , and radiation propagation table RT . The position table PT records the positions of UAVs and environmental information, where the positions of UAVs are marked with their unique identifiers (positive integers), obstacles are marked as -1 , and actionable areas are marked as 0 . The sensing table ST records detection states, with 0 for undetected areas, 1 for detected areas without radiation source, and 5 for areas with detected radiation signals. The radiation propagation table RT is derived from the received power [30].

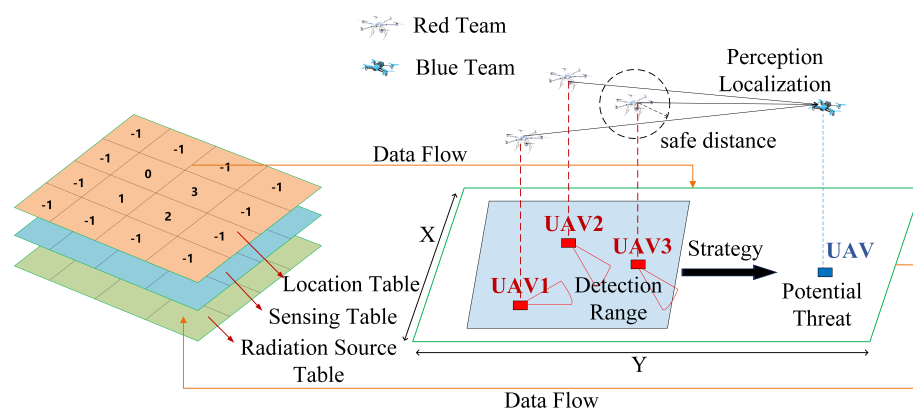


Figure 1. Urban electromagnetic spectrum model, consisting of UAV teams, multi-layered information tables, and electromagnetic sensing systems.

To model the radiation signal propagation and detection in our system, we first establish the relationship between transmitted and received power based on the Friis transmission equation. The received power (P_r) equals the transmitted power (P_t) minus the free-space path loss (L):

$$P_r = P_t - [20 \log_{10}(d) + 20 \log_{10}(f) + 20 \log_{10}(4\pi/c)], \quad (1)$$

where d represents the propagation distance, c is the speed of light (3×10^8 m/s), and f denotes the signal frequency (2.4 GHz). This equation enables us to calculate the expected signal strength at given distance from the radiation source, which is crucial for the detection and localization process.

Building upon the signal propagation model in Equation (1), we define the sensing mechanism for each UAV. In the sensing process, each UAV i has a directional sensing matrix B_i . The sensing result ST_i of UAV i at position (x_i, y_i) is defined as follows:

$$ST_i = B_i \times RT, \quad (2)$$

where B_i represents the sensor's directional detection pattern (weights in different directions). Their multiplication models how effectively the sensor detects radiation from different directions at its current position. The sensing table ST is constructed by concatenating all individual sensing results ST_i from each UAV. When a UAV receives signal strength exceeding the predetermined threshold within its sensing range, the area is marked as potentially containing a radiation source. To ensure localization accuracy, the system employs a multi-UAV cooperative verification mechanism: the three-point localization process is triggered only when three or more UAVs simultaneously detect suspected radiation sources within their respective sensing ranges. This design ensures detection reliability while avoiding false judgments from individual UAVs.

3.2. UAV Kinematic Model

This section provides a detailed description of the UAV model's movement, with particular focus on changes in speed and direction. We simultaneously establish a safe distance model to ensure safe UAV operation in various environments, while exploring methods for locating target radiation sources. To simplify, we use a 2D model instead of 3D, and make it discrete to fit a map.

We begin by defining the position formula, followed by the introduction of speed and direction sets based on real-world scenarios, which are used to update positions. The UAV position is given by the following equations:

$$p_i(t) = (x_i(t), y_i(t)), \quad (3)$$

where $p_i(t)$ represents the UAV i position coordinates at time t . Based on Equation (3), the UAV position update is given by the following equations:

$$p_i(t+1) = p_i(t) + v_i(t) [\lceil \cos \phi_i(t) \rceil, \lceil \sin \phi_i(t) \rceil]^T, \quad (4)$$

where $v_i(t)$ is the speed, and $\phi_i(t)$ is the direction angle of UAV i at time t . This update equation captures the discretized motion of UAVs in the 2D plane. To model realistic UAV capabilities, the speed $v_i(t)$ is defined as follows:

$$v_i(t) \in \{0, 1\}, \quad (5)$$

where 0 represents that the UAV is hovering in place, and 1 indicates moving horizontally. Each UAV has two types of physical constraints. For the onboard radiation detection system, the sensing direction constraint is defined as follows:

$$|\phi_i(t+1) - \phi_i(t)| \leq \Delta\phi_{\max}, \quad (6)$$

where $\Delta\phi_{\max}$ is the maximum allowable change in angle per step. As shown in Equation (6), this constraint ensures that the radiation detector's orientation change remains within its physical capabilities. For the UAV's movement, the acceleration constraint is defined as follow:

$$|v_i(t+1) - v_i(t)| \leq \Delta v_{\max}, \quad (7)$$

where $\Delta v_{\max}$ is the maximum allowable change in speed per step, ensuring smooth transitions in UAV motion control.

To ensure the safe operation of the red team's UAVs, it is essential to maintain a certain safe distance to protect it from external interference and potential hazards. This approach effectively prevents damage from environmental factors and ensures normal and stable operation. The safety constraints are formulated through three key equations. Based on the position vectors defined in Equation (3), the physical-safe distance constraint is formulated as follows:

$$|p_i(t) - p_j(t)| \geq D, \quad (8)$$

where D is the minimum physical-safe distance between any two UAVs, and the constraint applies for all pairs of different UAVs ($i \neq j$) at all time steps t . This safety constraint is essential for collision avoidance in multi-UAV systems. Each UAV must remain within the map boundaries:

$$p_i(t) \in M, \quad (9)$$

where

$$M = \{(x, y) \mid 0 \leq x \leq X_{\max}, 0 \leq y \leq Y_{\max}\}. \quad (10)$$

This boundary constraint requires all UAV positions to stay within the valid region at all times, restricting their movement when approaching these limits. Together, Equations (8)–(10) establish comprehensive safety constraints that ensure both collision avoidance between UAVs and containment within the operational area. This approach establishes an appropriate safety margin for the red team's UAVs, ensuring their optimal operational effectiveness.

4. Problem Definition and Analysis

In Section 4, we focus on problem formulation for the multi-UAV collaborative sensing and localization of anomalous radiation sources in urban environments and develop a Markov Decision Process (MDP) model for this scenario. This section is divided into two parts. First, it presents the problem formulation for sensing and localization along with its associated constraints; then, it proceeds to MDP modeling.

4.1. Problem Formulation for Perception

The red team UAVs achieve continuous localization of the anomalous radiation source through cooperation. To perform perception and detection of the anomalous radiation source, certain requirements must be met. These include UAV motion constraints and collision avoidance considerations to ensure safe operation. For localizing the anomalous radiation source, successful perception by three or more red team UAVs is required to achieve localization and detection. Therefore, a three-point localization algorithm is employed.

Each red team UAV i perceives the anomalous radiation source through Received Signal Strength Indication (RSSI). The distance can be derived from the received RSSI values [30]:

$$d_i(t) = 10^{\frac{\text{RSSI}_i(t) - P_0}{-10 \cdot n}}, \quad (11)$$

where $d_i(t)$ represents the distance from UAV i to the radiation source at time t , $\text{RSSI}_i(t)$ is the measured signal strength, P_0 is the RSSI value at the reference distance of 1 m, and n is the path loss exponent. As shown in Equation (11), the exponential relationship between RSSI and distance allows us to estimate the UAV's distance from the radiation source. Due to the limited sensing range of UAVs, the measurement is valid only when

$$d_i(t) < d_{\max}, \quad (12)$$

where d_{max} is the maximum detection distance. Building upon the distance estimation in Equation (11) and the range constraint in Equation (12), the set of UAVs with valid measurements at time t is defined as follows:

$$VM(t) = \{i \mid RSSI_i(t) > RSSI_{th}\}, \quad (13)$$

where $RSSI_{th}$ is the threshold value for valid RSSI measurements. The three-point localization method requires at least three spatially distributed measurement points to uniquely determine the position of a radiation source through geometric triangulation. Therefore, successful localization is only possible when $|VM(t)| \geq 3$, meaning three or more UAVs have obtained valid RSSI measurements simultaneously. To simplify the model, each UAV is equipped with a directional sensor that can be oriented in eight discrete directions. The sensor heading angle $\psi_i(t)$ of UAV i at time t can only take values from a discrete set of angles separated by $\pi/4$:

$$\psi_i(t) \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\}. \quad (14)$$

Combining all the measurement and sensing constraints defined in Equations (11)–(14), our objective is to maximize the time duration $I(t)$ when the radiation source can be effectively localized. Therefore, we formulate the following optimization problem to coordinate the UAVs' motion and sensing directions:

$$\max_{\{\mathbf{v}_i(t), \phi_i(t), \psi_i(t)\}_{t=1}^T, \forall i} \sum_{t=1}^T I(t) \quad (15)$$

$$\text{s.t. } (4)–(10), (12), (14), \quad (15a)$$

$$I(t) = \begin{cases} 1, & \text{if } |VM(t)| \geq 3 \\ 0, & \text{otherwise} \end{cases}. \quad (15b)$$

This optimization problem aims to find the optimal velocity $\mathbf{v}_i(t)$, heading angle $\phi_i(t)$, and sensor orientation $\psi_i(t)$ for each UAV i over the time horizon $\mathcal{T}$, subject to the motion and sensing constraints. At time t , the indicator function $I(t)$ evaluates to 1 only when three or more UAVs have valid measurements simultaneously, enabling three-point localization. This condition may not be satisfied at all times as UAVs may temporarily lose detection due to various factors such as distance, orientation, or environmental interference. By maximizing the sum of $I(t)$, we seek to maximize the duration during which the radiation source can be effectively localized.

4.2. Markov Decision Process Modeling

This section discusses the construction of a MDP model for electromagnetic spectrum operations in complex urban environments and critical areas. The discussion is based on an analysis of various factors and constraints that red team UAVs must consider when perceiving and localizing anomalous radiation sources. The MDP framework is particularly suitable for capturing the sequential decision-making nature and environmental uncertainties inherent in this problem. The interaction is illustrated in Figure 2.

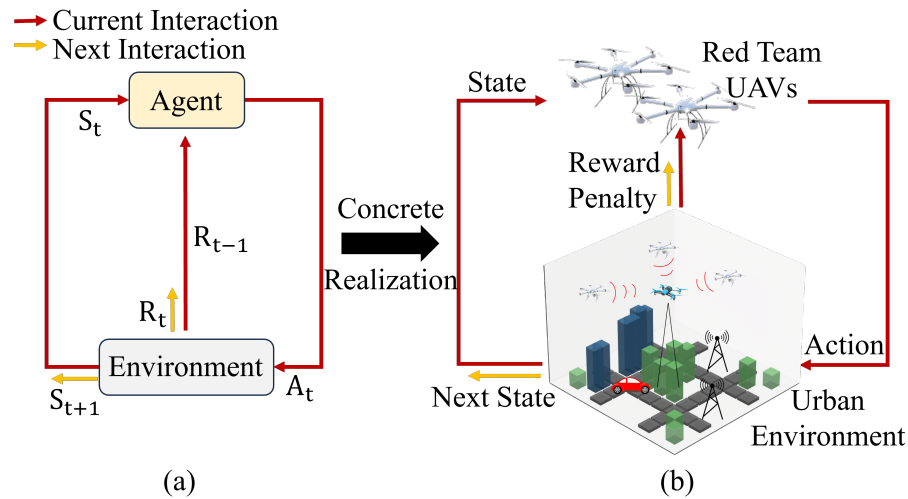


Figure 2. The MDP framework and its implementation: (a) standard MDP interaction framework; (b) implementation in urban electromagnetic environment with red team UAVs as agents.

As shown in Figure 2, the framework follows a MDP where state S_t represents the agent's observation of the environment at time step t , which transitions to the next state S_{t+1} after the agent takes an action. Action A_t defines the agent's possible operations. The reward signals include both immediate reward R_t and the previous time step's reward R_{t-1} , providing feedback on the agent's performance in both current and past states. Through this temporal interaction loop, the agent learns to optimize its behavior based on both current observations and historical experience. This MDP framework is implemented in our urban electromagnetic environment where UAVs act as agents to perform sensing and detection tasks. The key components of our MDP framework are detailed as follows:

- **State:** The state primarily focuses on the red team UAV's current position (x, y) , distance d , detection direction dir , current time step t , the number of successfully detected UAVs n , the anomalous radiation source position of the most recent effective scan (x_v, y_v) , and its distance d_v . The state space is constructed as a set $\langle x, y, d, dir, t, n, x_v, y_v, d_v \rangle$. Different UAVs have varying attributes and represent different systems. Based on their specific characteristics, the state transitions for movement and detection are configured accordingly;
- **Action:** The red team UAV's actions are primarily divided into movement direction m_d and detection direction d_d , where m_d is determined by velocity $\mathbf{v}_i(t)$ and heading angle $\phi_i(t)$, while d_d is determined by sensor orientation $\psi_i(t)$. For movement, the red team UAV has nine options: moving in cardinal and diagonal directions or staying. For detection, the red team UAV also has nine options: scanning or not scanning in 45-degree increments on the current plane. The action space is constructed as a set $\langle m_d, d_d \rangle$. The actual radiation detection depends on both sensor orientation and RSSI measurements. When three red team UAVs simultaneously detect the anomalous radiation source (i.e., $|VM(t)| \geq 3$), it triggers the three-point localization function;
- **Reward:** The environment generates a scalar signal as reward feedback based on the state and actions taken by the red team UAV. In this paper, we design rewards from multiple aspects, providing rewards for individual UAV successful detection, first successful detection, and successful three-point localization. This multi-faceted reward design balances individual performance and team collaboration, while addressing both task speed and accuracy. The individual detection reward keeps all red team UAVs motivated, the first successful detection reward speeds up task completion through competition, and the three-point localization reward encourages coopera-

tion, enhancing localization accuracy. To effectively guide the UAVs' behavior, we design a distance-based reward function that provides continuous feedback during the detection process:

$$R(d, r_p) = r_p \cdot \left(1 - e^{-3 \cdot \left(1 - \frac{d - d_{\min}^m}{d_{\max}^m - d_{\min}^m} \right)} \right), \quad (16)$$

where r_p is the preset detection reward value, $d_{\max}^m$ is the maximum cross-range distance of the key protection area (or the diagonal length if the area is square), $d_{\min}^m$ is the minimum distance, and d is the distance calculated from RSSI using Equation (11). By using this method, we link the successful detection reward for an individual UAV to the distance, creating a dynamic reward mechanism. This guides the UAV to approach the anomalous radiation source for better detection results. The total reward for a corresponding action is then calculated. For the first successful detection and successful three-point localization, additional rewards are assigned to each UAV.

- **Penalty:** Due to the sparsity of rewards caused by the large spatial range, we implement penalties to encourage the red team UAVs to perform mobile detection behaviors. A penalty is applied at each step when a UAV fails to detect the anomalous radiation source. Additionally, a penalty is imposed if the UAV loses track of the detection target in the time step immediately following a successful detection, and when collisions occur between UAVs or with obstacles.

MDPs provide a powerful and flexible framework for modeling decision-making problems. In this study, we have applied an MDP-based framework to the challenge of detecting and localizing anomalous radiation sources. Building on this framework, we showcase in later chapters the implementation of highly adaptive reinforcement learning techniques. These techniques allow us to derive optimal strategies in environments that are either unknown or only partially known.

5. Multi-UAV Reconnaissance Proximal Policy Optimization Algorithm

In Section 5, we first introduce commonly used reinforcement learning algorithms; then, we propose the MURPPO algorithm and describe its network structure and algorithmic process. The proposed algorithm incorporates innovative improvements based on the original algorithm. We analyze how these modifications might be beneficial for sensing and localizing anomalous radiation sources within the environment described in this paper.

5.1. Reinforcement Learning Algorithm

Reinforcement learning leverages the framework of MDP, yet it does not require prior knowledge of a complete model [31]. In our multi-UAV radiation source search scenario, the objective is to find an optimal policy that guides UAVs to efficiently locate the radiation source. To address the complex challenges in multi-UAV radiation source search, we propose MURPPO, a novel variant of Proximal Policy Optimization (PPO) algorithm. The PPO framework is particularly suitable for our scenario as it provides stable policy updates while handling the uncertainties in radiation measurements and complex multi-UAV coordination.

To train the value network effectively, we need a loss function that measures the accuracy of our value predictions. The mean squared temporal difference (TD) error serves as an ideal loss function as it quantifies the discrepancy between predicted and actual value estimates. The value network loss function is as follows [32]:

$$\mathcal{L}(\omega) = \frac{1}{2}(\delta_t)^2. \quad (17)$$

The core component of this loss function is the TD error δ_t , which measures the difference between our current value estimate and a more accurate target value. The TD error is calculated by comparing the current value estimate with the sum of the immediate reward and the discounted value of the next state [33]:

$$\delta_t = r + \gamma V(s_{t+1}) - V(s_t), \quad (18)$$

where r represents the immediate reward, γ is the discount factor, $V(s_t)$ is the value estimate of the current state s_t , and $V(s_{t+1})$ is the value estimate of the next state s_{t+1} . This TD error enables the UAV to learn from immediate radiation measurements while accounting for potential future detections.

During the policy optimization process, large policy updates can lead to performance collapse, especially in scenarios with noisy measurements like radiation detection. To address this challenge, the PPO algorithm [34] introduces a trust region-based approach that limits the policy change in each update step. This stability is achieved through a clipped objective function:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t)], \quad (19)$$

where $r_t(\theta) = \frac{\pi_{\theta}^{new}(a_t|s_t)}{\pi_{\theta}^{old}(a_t|s_t)}$ represents the probability ratio of the new policy to the old policy, A_t is the advantage function evaluating the relative benefit of actions under the current policy, ϵ is a small constant controlling the magnitude of policy updates, and $\text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)$ constrains the ratio $r_t(\theta)$ to the interval $[1 - \epsilon, 1 + \epsilon]$ to prevent excessive policy changes. This clipping mechanism helps stabilize the training process by preventing large policy updates when the UAV receives varying radiation measurements.

The advantage function A_t represents the relative superiority of an action in a given state. It is calculated as the difference between the TD target and the current value estimate. In the PPO algorithm, temporal difference methods are used to estimate the advantage function, which then guides policy optimization. The formula is as follows [35]:

$$A_t \approx \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t). \quad (20)$$

This advantage estimation helps the UAV evaluate whether its current actions contribute to effective source detection compared to the estimated state value.

The overall objective of the PPO algorithm is to maximize the weighted sum of policy loss, value function loss, and entropy bonus. The total objective function can be expressed as follows [34]:

$$L^{PPO}(\theta) = L^{CLIP}(\theta) - c_1 L^{VF}(\omega) + c_2 S[\pi_{\theta}](\theta), \quad (21)$$

where c_1 and c_2 are weight coefficients, $S[\pi_{\theta}](s_t)$ represents the policy entropy that encourages exploration, and $L^{VF}(\omega)$ is the value function loss that minimizes the error between estimated and actual returns. This objective function balances stable policy updates with effective exploration, enabling the UAV to maintain consistent search patterns while discovering optimal detection strategies.

5.2. MURPPO Based on Branch Training

Red team UAVs need to efficiently perceive the reconnaissance environment using multiple devices simultaneously. In this scenario, multiple UAVs interact with the environment concurrently, known as multi-UAV reinforcement learning. System state transitions are determined by the collective actions of UAVs, while rewards depend on their joint actions. Multi-UAV systems can be categorized into centralized and distributed control. Centralized control involves a central entity for information gathering and decision-making,

with all UAVs following its instructions. This approach enables global optimization but suffers from poor scalability. In contrast, distributed control allows each UAV to make decisions independently, offering better scalability and fault tolerance.

This paper employs a MURPPO algorithm with distributed control and branched training to learn complex decision-making tasks. Red team UAVs' behavior is controlled by the algorithm. Each UAV uses an independent algorithm model, and cooperation in perception and localization is achieved through reward design. When three UAVs simultaneously detect an anomalous radiation source, it triggers a three-point localization function (i.e., $|VM(t)| \geq 3$), enabling the reconnaissance and localization of the anomalous radiation source.

To implement our distributed reinforcement learning framework effectively, we design the policy network with a dual-branch structure to reduce decision complexity and improve learning efficiency. The policy network is improved through branched training, enabling a single network to output both movement direction and detection suppression direction actions. The branched structure enables separate optimization of movement and detection tasks, where the first branch outputs the movement direction action in the movement action space. The second branch outputs the detection direction action in the detection action space. Since these two actions serve different purposes and have different gradient directions, separate network parameters are required for each branch. The policy network structure is illustrated in Figure 3.

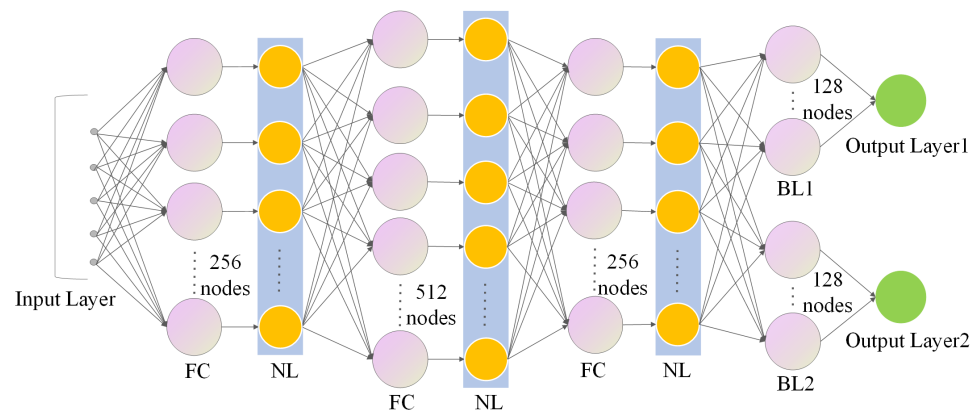


Figure 3. Network structure of the actor.

Additionally, a value network is designed that takes state information as input and outputs a value to evaluate the policy network. The structure of this network is shown in Figure 4. Based on Equation (17) and TD error in Equation (18), the value loss $L^{VF}(\omega_i)$ is designed for each UAV to optimize its value estimation. This network architecture, working in conjunction with our task-specific reward design, forms the foundation of our MURPPO algorithm's capability in handling complex radiation source detection tasks.

In MURPPO, to enhance training stability, we normalize the advantage function from Equation (20) for the i -th UAV by subtracting the mean and dividing by the standard deviation. The formula for the normalized advantage function is defined as follows:

$$A_{t,i}^{\text{norm}} = \frac{A_{t,i} - \mu_{A,i}}{\sigma_{A,i} + 1 \times 10^{-8}}, \quad (22)$$

where $\mu_{A,i}$ represents the mean of all advantage values, calculated as $\frac{1}{N} \sum_{t=1}^N A_{t,i}$, while $\sigma_{A,i}$ is the standard deviation, computed as $\sqrt{\frac{1}{N} \sum_{t=1}^N (A_{t,i} - \mu_{A,i})^2}$. By subtracting the mean and dividing by the standard deviation, the advantages are scaled to have zero mean and unit variance across the batch of experiences. The small constant 1×10^{-8} added to

the denominator prevents potential division by zero, ensuring numerical stability. This normalization is particularly important in radiation detection tasks as it helps stabilize the training process.

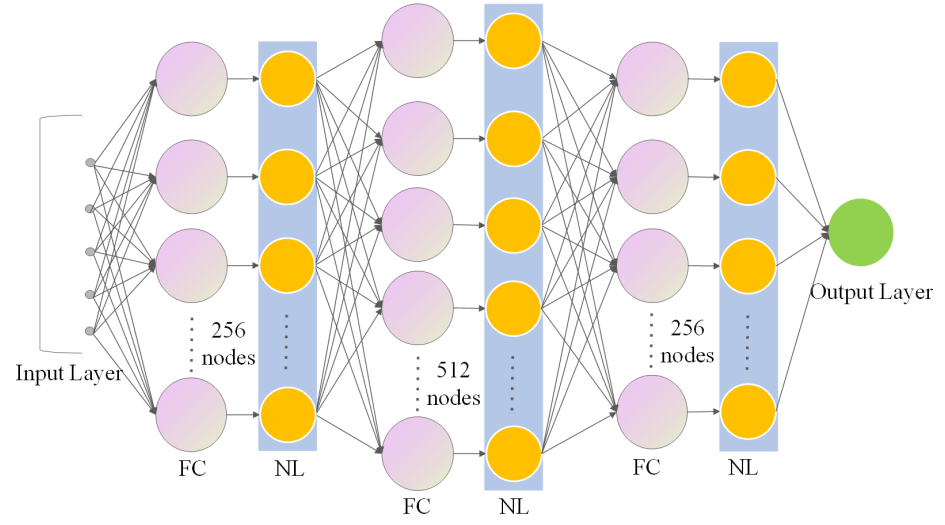


Figure 4. Network structure of the critic.

Based on Equation (19), we modify the policy loss in MURPPO to account for the dual-branch structure. As defined in Equations (23) and (24), the policy loss is calculated separately for movement and scanning actions:

$$L_{\text{move}}^{\text{CLIP}}(\theta_i) = \mathbb{E}_t[\min(r_{t,\text{move}}(\theta_i)A_{t,i}^{\text{norm}}, \text{clip}(r_{t,\text{move}}(\theta_i), 1 - \epsilon, 1 + \epsilon)A_{t,i}^{\text{norm}})], \quad (23)$$

$$L_{\text{scan}}^{\text{CLIP}}(\theta_i) = \mathbb{E}_t[\min(r_{t,\text{scan}}(\theta_i)A_{t,i}^{\text{norm}}, \text{clip}(r_{t,\text{scan}}(\theta_i), 1 - \epsilon, 1 + \epsilon)A_{t,i}^{\text{norm}})], \quad (24)$$

where $r_{t,\text{action}}(\theta_i) = \frac{\pi_{\theta_i}^{\text{new}}(a_{t,\text{action}}|s_t)}{\pi_{\theta_i}^{\text{old}}(a_{t,\text{action}}|s_t)}$ represents the policy ratio for the i -th UAV, with action representing either movement (move) or scanning (scan) actions. The clipping parameter ϵ in policy loss function limits the magnitude of policy updates. When the probability ratio between new and old policies exceeds $1 + \epsilon$ or falls below $1 - \epsilon$, the clipping mechanism prevents excessive changes. This mechanism is applied to both movement and scanning policy branches independently, ensuring stable updates across both action spaces. The final policy loss for the i -th UAV combines Equations (23) and (24):

$$L_{\text{total}}^{\text{CLIP}}(\theta_i) = L_{\text{move}}^{\text{CLIP}}(\theta_i) + L_{\text{scan}}^{\text{CLIP}}(\theta_i). \quad (25)$$

To maximize immediate rewards, a greedy policy often selects the currently known best action, but it can become stuck in local optima. Therefore, this paper adopts a decaying ϵ -greedy policy [31] that primarily selects the best action based on the current policy, while with a small probability ϵ , it chooses a random action to promote exploration. Initially, ϵ is set to 1 and decays over time, allowing for extensive early-stage exploration while gradually focusing on exploitation as the policy improves. As shown in Equation (26), the action selection process follows a probability-based mechanism:

$$\text{Action} = \begin{cases} \text{output from the Actor,} & \text{with probability } 1 - \epsilon, \\ \text{output a random action,} & \text{with probability } \epsilon. \end{cases} \quad (26)$$

Incorporating task-specific knowledge, when a target is detected, the UAV maintains its position and scanning direction to avoid losing the target. If no target is detected, the UAV follows the action determined by the Actor network. With probability ε , a random action is selected, regardless of whether a target was detected. This design ensures sufficient exploration during early training while optimizing performance in later stages by leveraging both learned strategies and domain knowledge.

In MURPPO, to ensure effective exploration of both movement and scanning actions, we calculate the entropy loss separately for each action branch of the i -th UAV. Equations (27) and (28) define the entropy losses for movement and scanning policies, respectively:

$$S[\pi_{\theta_i, \text{move}}](s_t) = \mathbb{E}_t[\mathcal{H}(\pi_{\theta_i, \text{move}}(\cdot|s_t))], \quad (27)$$

$$S[\pi_{\theta_i, \text{scan}}](s_t) = \mathbb{E}_t[\mathcal{H}(\pi_{\theta_i, \text{scan}}(\cdot|s_t))], \quad (28)$$

where $\mathcal{H}$ is the entropy of the policy. The entropy loss encourages exploration by preventing the policy from becoming overly deterministic. For our radiation detection scenario, this is particularly important as it helps discover diverse movement patterns and scanning strategies. As shown in Equation (29), the total entropy loss combines both branches:

$$S[\pi_{\theta_i}](s_t) = S[\pi_{\theta_i, \text{move}}](s_t) + S[\pi_{\theta_i, \text{scan}}](s_t). \quad (29)$$

To balance the learning progress between branches, MURPPO employs advantage function normalization using Equation (22) to ensure both branches optimize in a comparable way, while summing the entropy losses of both branches using Equation (29) unifies the exploration pressure, preventing one branch from converging prematurely while the other still requires extensive exploration.

Building upon the PPO objective function in Equation (21), MURPPO's objective function incorporates the specialized components discussed above. The complete objective function for the i -th UAV in MURPPO is given by Equation (30):

$$L^{MURPPO}(\theta_i) = L_{total}^{CLIP}(\theta_i) - L^{VF}(\omega_i) + 0.01 \cdot S[\pi_{\theta_i}](s_t), \quad (30)$$

where $c_1 = 1.0$ and $c_2 = 0.01$. The value of c_1 ensures balanced learning between the value function and policy, providing reliable advantage estimates for both movement and scanning decisions. The small value of c_2 maintains sufficient exploration while preventing it from overshadowing the primary learning objectives in the UAV target search task.

The MURPPO framework is implemented through multiple UAVs executing this optimization process simultaneously in a distributed manner, enabling effective radiation source localization when three or more UAVs have valid measurements (i.e., $|VM(t)| \geq 3$). The detailed algorithmic procedure of MURPPO, incorporating the aforementioned components, is presented in Algorithms 1 and 2. Due to its decentralized architecture, the algorithm maintains computational efficiency with constant complexity $O(1)$ for each UAV regardless of team size, ensuring effectiveness across different scales of UAV teams.

Algorithm 1 MURPPO Algorithm Update Process

```

1: Input: Trajectories  $Tr$ , mini-batch size  $M$ , policy network  $\pi_{\theta_i}$ , value network  $V_{\omega_i}$ 
2: Output: Updated parameters  $\theta_i$  and  $\omega_i$  for the  $i$ -th UAV
3: Initialize actor learning rate  $\alpha$ , critic learning rate  $\beta$ 
4: Unpack trajectories  $Tr$  into states  $s_{i,t}$ , actions  $a_{i,t} = (a_{i,t}^{\text{move}}, a_{i,t}^{\text{scan}})$ , old log probabilities
    $\log \pi_{\theta_i}^{\text{old}} = (\log \pi_{\theta_i, \text{move}}^{\text{old}}, \log \pi_{\theta_i, \text{scan}}^{\text{old}})$ , rewards  $r_{i,t}$ , next states  $s_{i,t+1}$ 
5: for  $j = 1, \dots, M$  do
6:   Compute  $\pi_{\theta_i, \text{move}}^{\text{new}}(a_{i,t}^{\text{move}} | s_{i,t})$ ,  $\pi_{\theta_i, \text{scan}}^{\text{new}}(a_{i,t}^{\text{scan}} | s_{i,t})$  using  $\pi_{\theta_i}$ 
7:   Compute  $V(s_{i,t})$  and  $V(s_{i,t+1})$  using  $V_{\omega_i}$ 
8:   Compute advantages  $A_{t,i}^{\text{norm}}$  using Equation (22)
9:   Compute surrogate losses  $L_{\text{move}}^{\text{CLIP}}(\theta_i)$ ,  $L_{\text{scan}}^{\text{CLIP}}(\theta_i)$  using Equations (23) and (24)
10:  Compute policy loss  $L_{\text{total}}^{\text{CLIP}}(\theta_i)$  using Equation (25)
11:  Compute entropy bonus  $S[\pi_{\theta_i}](s_t)$  using Equation (29)
12:  Compute total policy loss  $L^{\text{PF}}(\theta_i) = L_{\text{total}}^{\text{CLIP}}(\theta_i) + 0.01 \cdot S[\pi_{\theta_i}](s_t)$ 
13:  Update policy network parameters:  $\theta_i \leftarrow \theta_i + \alpha \nabla_{\theta} L^{\text{PF}}(\theta_i)$ 
14:  Compute value loss  $L^{\text{VF}}(\omega_i)$  using Equation (17)
15:  Update value network parameters:  $\omega_i \leftarrow \omega_i - \beta \nabla_{\omega} L^{\text{VF}}(\omega_i)$ 
16: end for

```

Algorithm 2 MURPPO Algorithm Execution Process

```

1: Input: system parameters
2: Output: localization time duration  $I(t)$ 
3: Initialize:  $I(t) \leftarrow 0$ ,  $Tr \leftarrow \emptyset$ ,  $t \leftarrow T$ 
4: while  $t > 0$  do
5:   for each UAV  $i$  do
6:     Observe current state  $s_{i,t}$ 
7:     Select actions  $a_{i,t} = (a_{i,t}^{\text{move}}, a_{i,t}^{\text{scan}})$  using  $\epsilon$ -greedy policy (26)
8:     Compute log probabilities  $\log \pi_{\theta_i}^{\text{old}} = (\log \pi_{\theta_i, \text{move}}^{\text{old}}, \log \pi_{\theta_i, \text{scan}}^{\text{old}})$ 
9:     Measure RSSI and calculate distance  $d_i(t)$  using Equation (11)
10:    Calculate  $r_{i,t}$  as the sum of immediate rewards and penalties
11:   end for
12:   if  $|VM(t)| \geq 3$  and  $d_i(t) < d_{\text{max}}$ , where  $VM(t)$  is defined in Equation (13) then
13:     Execute three-point localization process
14:     Update  $r_{i,t}$  by distribute localization rewards
15:      $I(t) \leftarrow I(t) + 1$ 
16:   end if
17:   for each UAV  $i$  do
18:     Observe next state  $s_{i,t+1}$ 
19:     Store  $(s_{i,t}, a_{i,t}, \log \pi_{\theta_i}^{\text{old}}, r_{i,t}, s_{i,t+1})$  in trajectories  $Tr$ 
20:   end for
21:   if  $|Tr| = M$  then
22:      $\theta_i, \omega_i \leftarrow \text{UpdateParameters}(Tr, M, \pi_{\theta_i}, V_{\omega_i})$  {See Algorithm 1}
23:     Clear  $Tr$ 
24:   end if
25:    $t \leftarrow t - 1$ 
26: end while

```

6. Simulation Results

This section presents the simulation experiments for the proposed algorithm. We begin by simulating the environment described in this paper, including configuration and parameter settings. We then conduct simulation experiments to evaluate the proposed algorithm's performance in terms of reward, localization rate, and KL divergence. The localization rate of our algorithm is compared and analyzed against the baseline methods.

6.1. Experimental Configuration and Parameters

In this section, we present the configuration of the experimental environment and platform. The experiments were conducted using an Intel Core i9-12900K processor and an NVIDIA RTX 3080 GPU with approximately 24 h of training time. The duration of each round is 10 s, discretized into 100 time steps (0.1 s per step). At each time step, each UAV executes a sequence of operations, including perception, localization, and decision-making. The experimental platform used UAVs with a conservative setting of maximum speed at 10 m/s (1 m/step) and maximum angular change of π rad/step, which are lower than the actual UAV capabilities to ensure algorithm reliability. These parameters were configured based on our discretized experimental design (1 m $\times$ 1 m grid) to ensure both algorithm reliability and efficient performance evaluation. Each UAV was equipped with a directional radiation detector operating at 2.4 GHz band, with a conservative maximum effective radius of 200 m to ensure reliable detection in complex environments. The detector could either rotate in eight directions at $\pi/4$ intervals or be turned off, providing 360° horizontal plane coverage while maintaining computational efficiency. The base station was positioned outside the 1000 $\times$ 1000 m² operation area, utilizing the UAVs' transmission range to ensure stable communication throughout the missions. For different rewards and penalties in the reward function, their weights are determined through systematic training process and adjusted based on specific mission scenarios to achieve a balance between task speed and accuracy. The simulation parameters are summarized in Table 3.

Table 3. Hyperparameter setting.

System Parameters	Numerical Settings	Description
$X \times Y$	1000 $\times$ 1000 (m ²)	map size
$\Delta\phi_{\max}$	π (rad/step)	the maximum angle change per step
$\Delta v_{\max}$	1 (m/step)	the maximum speed change per step
m	3	number of red UAVs
n	1	number of blue UAVs
<i>episodes</i>	1000	number of training episodes
T	100	time steps per round
ϵ	1.0	greedy factor
$\epsilon_{\min}$	0.1	minimum greedy factor
ϵ_{decay}	0.99	greedy factor decay rate
α	0.0001	actor learning rate
β	0.0005	critic learning rate
ϵ_{clip}	0.2	clip ratio
$d_{\max}$	200 (m)	maximum detection distance
γ	0.9	discount factor
r_p	0.5	perception reward
r_{fp}	0.5	first perception reward
r_{pos}	0.5	three-point localization reward
p_p	−0.2	penalty for non-detection
p_{ff}	−0.2	penalty for inconsistent action after detection
p_c	−0.2	collision penalty
D	1 (m)	safe distance

This study develops a platform for multi-agent collaboration, deploying three red team UAVs and one blue team UAV. During each mission evolution, the system logs critical data. Figure 5a displays the global localization map, illustrating the distribution of red and blue UAVs within the defense area; blue rectangles indicate blue team device locations, while red rectangles represent red team UAV positions. Figure 5b shows the

red team's perception and detection map, where UAVs achieve flexible detection through movement and orientation, with fan-shaped areas representing their sensing range. The red areas indicate where blue team UAVs are detected within the sensing range, while the orange areas show where blue team UAVs are not detected within the sensing range. Figure 5c provides an illustrative example of possible UAV trajectories in a multi-UAV scenario. The paths of four UAVs are shown, three red team UAVs and one blue team UAV. The trajectories shown here represent connected line segments with position points marked along their paths in a 2D coordinate system measured in meters, demonstrating potential spatial relationships between UAVs in the operational area. The successful three-point localization of the anomalous radiation source by UAVs is demonstrated as shown in Figure 5d. When three UAVs simultaneously detect the anomalous radiation source, the three-point localization function is triggered to precisely determine its position. In the diagram, the three green UAVs represent the red team UAVs, serving as the centers of dashed circles. The blue point indicates the location of the anomalous radiation source. Red lines connect the green UAVs to the blue point, representing the distance relationships calculated by the UAVs using Equation (11) based on the perceived RSSI, which also serve as the radius of the dashed circles. The intersection point of the dashed circles corresponds to the position of the blue point, demonstrating the localization result of the anomalous radiation source.

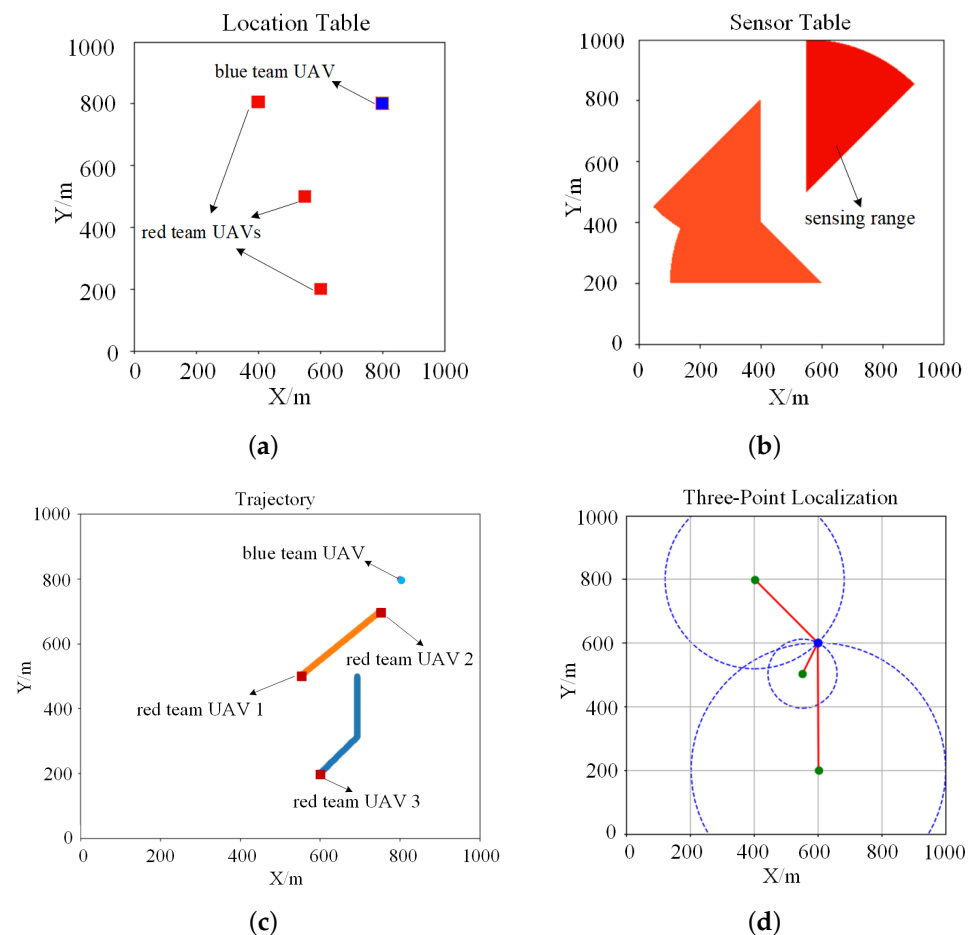


Figure 5. Environment and information for intelligent spectrum warfare platform: (a) global location information; (b) global sensing information; (c) UAVs' movement trajectory; (d) schematic diagram of three-point localization.

6.2. Experimental Results and Analysis

In this section, we present the performance of the previously proposed MURPPO algorithm based on branch training. All results were recorded via Tensorboard during the training process. The reward curves for three red team UAVs are illustrated in Figure 6, with the x-axis representing the number of training episodes and the y-axis showing the corresponding reward values. Each episode consists of 100 time steps, and the reward values represent the cumulative rewards obtained by UAVs through multiple mechanisms including perception rewards, first detection rewards and penalties, as well as additional localization rewards when three-point localization is achieved. Figure 6a–c individually illustrate the reward progression for UAVs 1, 2, and 3, reflecting each UAV’s independent learning performance. The learning process shows two distinct phases, a rapid improvement phase in the first 200 episodes where the agents quickly learn effective strategies, followed by a stabilization phase where the rewards fluctuate around converged values. UAVs 1 and 2 converge to approximately 70 and 80, respectively, while UAV 3 exhibits superior performance with more rapid convergence. Figure 6d displays the aggregate rewards for all three UAVs, which ultimately stabilize around 250. The overall trend indicates that, notwithstanding intermittent fluctuations, the system effectively enhances task execution capabilities through multi-UAV collaboration, resulting in successful policy convergence. Overall, the red team’s cooperative mechanism is effective, enabling them to gradually gain a competitive edge in the complex electromagnetic environment. This validates the effectiveness of multi-UAV cooperation and the algorithmic improvements. The system demonstrates its ability to learn and adapt in a challenging, adversarial setting.

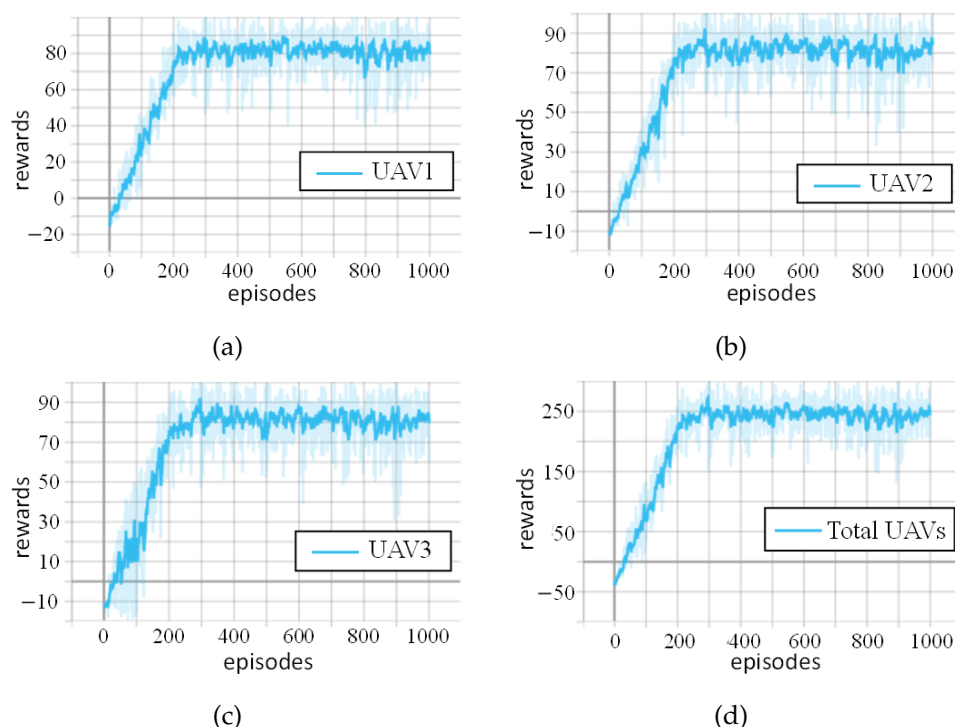


Figure 6. Rewards obtained by red team UAVs over 1000 training episodes: (a) rewards obtained by UAV 1; (b) rewards obtained by UAV 2; (c) rewards obtained by UAV 3; (d) total reward.

Figure 7 demonstrates the superior performance of our proposed MURPPO algorithm compared to baseline methods (MAA2C and MAPPO) over 1000 training episodes. The horizontal axis represents the number of training episodes, while the vertical axis indicates the localization rate. From an overall performance perspective, MURPPO demonstrated the highest localization rate throughout most of the training phases, followed by

MAPPO, while MAA2C's performance was comparatively subpar. During the initial training period (approximately episodes 0–200), all three algorithms exhibited an upward trend, with MURPPO showing the fastest convergence rate, surpassing both MAPPO and MAA2C. As training progressed, the performance disparities among the algorithms became increasingly apparent. In the later stages of training (after approximately 700 episodes), MURPPO's advantages became more pronounced, not only maintaining a high success rate but also exhibiting exceptional stability. Although MAPPO's performance experienced some decline, it generally remained superior to MAA2C. Notably, MURPPO's learning curve was the most consistent, with minimal fluctuations. In contrast, MAPPO's performance fluctuated considerably, especially during the middle and later stages. While MAA2C exhibited relatively small fluctuations, its overall performance consistently remained lower. In the final 300 episodes of training, MURPPO achieved an average localization rate of 56.5%. This was approximately 38.5% higher than MAPPO's and 41.6% higher than MAA2C's. In summary, the MURPPO algorithm surpasses the benchmark algorithms in key metrics such as localization rate, stability, and convergence speed.

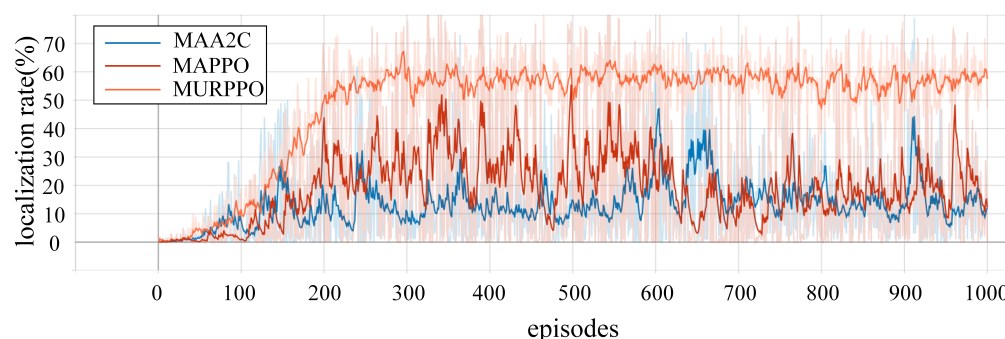


Figure 7. The comparison of localization time ratio per episode among the proposed algorithm and two baseline methods.

Figure 8 presents a novel analysis of policy convergence stability through KL divergence tracking for each UAV independently, offering unique insights into the training dynamics of multi-agent systems in electromagnetic environments. KL divergence measures the difference between the current policy and the previous policy, where MURPPO ensures training stability by limiting these changes. In the early stages of training, all UAVs exhibit significant fluctuations in KL divergence, indicating substantial policy updates consistent with the exploration process. As training progresses, KL divergence gradually decreases and stabilizes, suggesting policy convergence for all three UAVs. This validates the effectiveness of the MURPPO algorithm in balancing exploration and exploitation in a complex, multi-UAV reinforcement learning environment.

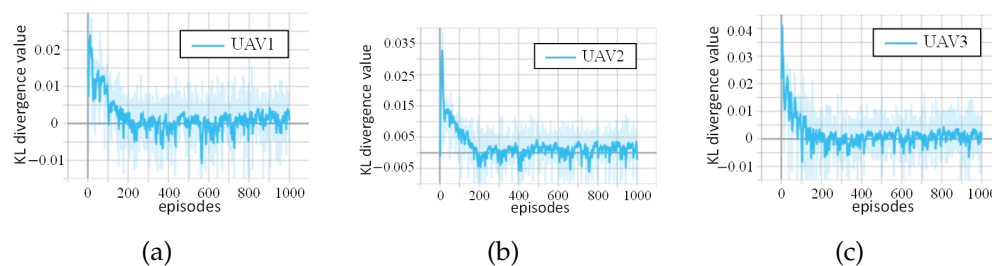


Figure 8. KL divergence trend of red UAVs over 1000 training episodes: (a) KL divergence of UAV 1; (b) KL divergence of UAV 2; (c) KL divergence of UAV 3.

7. Conclusions

This paper addressed the challenge of intelligent decision-making for UAV defense in electromagnetic spectrum warfare within complex urban environments. It introduced a MURPPO algorithm based on branch training. By applying Markov modeling to the task process, this paper elucidated how UAVs performed closed-loop perception, localization, and tracking of anomalous radiation sources under designed reward conditions. Simulation results demonstrated that the MURPPO algorithm effectively perceived anomalous radiation sources, achieved detection and localization through multi-UAV collaboration, and attained stable convergence. Moreover, it outperformed baseline algorithms in key metrics such as localization rate, stability, and convergence speed. The proposed algorithm exhibited a 38.5% higher localization rate compared to MAPPO in the later stages of training. The MURPPO algorithm's ability to handle sparse rewards and coordinate multiple UAVs highlights its potential for enhancing autonomous decision-making capabilities in challenging electromagnetic environments. Future research directions could explore hierarchical structures and formation control to further improve performance in large scale applications, as well as enhanced sensing capabilities with finer angular resolution or multiple frequency bands for better source discrimination.

Author Contributions: Conceptualization, J.C.; methodology, J.C.; software, J.C., Z.Z. and X.Z.; validation, J.C., Z.Z. and D.F.; formal analysis, Z.Z.; investigation, J.C. and Z.Z.; resources, Z.Z.; data curation, Z.Z. and D.F.; writing—original draft preparation, D.F. and C.H.; writing—review and editing, J.C., Y.Z. and X.Z.; visualization, T.H.; supervision, D.F., T.H. and J.Z.; project administration, J.C.; funding acquisition, Y.Z., D.F., C.H. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under grant 2022YFC3301404.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Dan Fan was employed by the company China Satellite Network E-Link Co.Ltd., and author Jun Zhao was employed by the company The 22th Research Institute of China Electronics Technology Group Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
PPO	Proximal Policy Optimization
MURPPO	Multi-UAV Reconnaissance Proximal Policy Optimization
FFA	Firefly Algorithm
PSO	Particle Swarm Optimization
ACO	Ant Colony Optimization
ABC	Artificial Bee Colony
IMUCS-BSAE	Improved Multi-UAV Collaborative Search Algorithm based on Binary Search Algorithm
MPSO	Motion-encoded Particle Swarm Optimization
DRL	Deep Reinforcement Learning
2D	Two-Dimensional
MDP	Markov Decision Process

RSSI	Received Signal Strength Indication
FC	Fully Connected Layer
NL	Normalization Layer
BL	Branch Layer
MAA2C	Multi-Agent Advantage Actor-Critic
MAPPO	Multi-Agent Proximal Policy Optimization
KL divergence	Kullback–Leibler divergence

References

- Meng, Y.; Qi, P.; Zhou, X.; Li, Z. Capability Analysis Method for Electromagnetic Spectrum. In Proceedings of the 2021 8th International Conference on Dependable Systems and Their Applications (DSA), Yinchuan, China, 11–12 September 2021; pp. 739–740.
- Ye, N.; Miao, S.; Pan, J.; Xiang, Y.; Mumtaz, S. Dancing with Chains: Spaceborne Distributed Multi-User Detection under Inter-Satellite Link Constraints. *IEEE J. Sel. Top. Signal Process.* **2025**. [\[CrossRef\]](#)
- Ostrometzky, J.; Messer, H. Opportunistic Weather Sensing by Smart City Wireless Communication Networks. *Sensors* **2024**, *24*, 7901. [\[CrossRef\]](#) [\[PubMed\]](#)
- Alsarhan, A.; Al-Dubai, A.Y.; Min, G.; Zomaya, A.Y.; Bsoul, M. A New Spectrum Management Scheme for Road Safety in Smart Cities. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3496–3506. [\[CrossRef\]](#)
- Jasim, M.A.; Shakhathreh, H.; Siasi, N.; Sawalmeh, A.H.; Aldalbahi, A.; Al-Fuqaha, A. A Survey on Spectrum Management for Unmanned Aerial Vehicles (UAVs). *IEEE Access* **2022**, *10*, 11443–11499. [\[CrossRef\]](#)
- Zhou, L.; Leng, S.; Wang, Q.; Quek, T.Q.S.; Guizani, M. Cooperative Digital Twins for UAV-Based Scenarios. *IEEE Commun. Mag.* **2024**, *62*, 112–118. [\[CrossRef\]](#)
- Mohsan, S.A.H.; Othman, N.Q.H.; Li, Y.; Alsharif, M.H.; Khan, M.A. Unmanned Aerial Vehicles (UAVs): Practical Aspects, Applications, Open Challenges, Security Issues, and Future Trends. *Intell. Serv. Robot.* **2023**, *16*, 109–137. [\[CrossRef\]](#)
- Xu, C.; Liao, X.; Tan, J.; Ye, H.; Lu, H. Recent Research Progress of Unmanned Aerial Vehicle Regulation Policies and Technologies in Urban Low Altitude. *IEEE Access* **2020**, *8*, 74175–74194. [\[CrossRef\]](#)
- Kang, H.; Joung, J.; Kim, J.; Kang, J.; Cho, Y.S. Protect Your Sky: A Survey of Counter Unmanned Aerial Vehicle Systems. *IEEE Access* **2020**, *8*, 168671–168710. [\[CrossRef\]](#)
- Kang, B.; Ye, N.; An, J. Achieving Positive Rate of Covert Communications Covered by Randomly Activated Overt Users. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 2480–2495. [\[CrossRef\]](#)
- Li, J.; Yang, L.; Wu, Q.; Lei, X.; Zhou, F.; Shu, F.; Mu, X.; Liu, Y.; Fan, P. Active RIS-Aided NOMA-Enabled Space-Air-Ground Integrated Networks with Cognitive Radio. *IEEE J. Sel. Areas Commun.* **2025**, *43*, 314–333. [\[CrossRef\]](#)
- Zhang, W.; Zhang, W. An Efficient UAV Localization Technique Based on Particle Swarm Optimization. *IEEE Trans. Veh. Technol.* **2022**, *71*, 9544–9557. [\[CrossRef\]](#)
- Abdelhakim, A. Heuristic techniques for maximum likelihood localization of radioactive sources via a sensor network. *Nucl. Sci. Tech.* **2023**, *34*, 127. [\[CrossRef\]](#)
- Yu, Y.; Lee, S. Efficient Multi-UAV Path Planning for Collaborative Area Search Operations. *Appl. Sci.* **2023**, *13*, 8728. [\[CrossRef\]](#)
- Phung, M.D.; Ha, Q.P. Motion-encoded Particle Swarm Optimization for Moving Target Search Using UAVs. *Appl. Soft Comput.* **2020**, *97*, 106705. [\[CrossRef\]](#)
- Chen, J.X. The Evolution of Computing: AlphaGo. *Comput. Sci. Eng.* **2016**, *18*, 4–7. [\[CrossRef\]](#)
- Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; Tang, Y. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1122–1136. [\[CrossRef\]](#)
- Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face Recognition Systems: A Survey. *Sensors* **2020**, *20*, 342. [\[CrossRef\]](#)
- Guo, H.; Wang, Y.; Liu, J.; Liu, C. Multi-UAV Cooperative Task Offloading and Resource Allocation in 5G Advanced and Beyond. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 347–359. [\[CrossRef\]](#)
- Guo, H.; Zhou, X.; Wang, J.; Liu, J.; Benslimane, A. Intelligent Task Offloading and Resource Allocation in Digital Twin Based Aerial Computing Networks. *IEEE J. Sel. Areas Commun.* **2023**, *41*, 3095–3110. [\[CrossRef\]](#)
- Guo, H.; Chen, X.; Zhou, X.; Liu, J. Trusted and Efficient Task Offloading in Vehicular Edge Computing Networks. *IEEE Trans. Cogn. Commun. Netw.* **2024**, *10*, 2370–2382. [\[CrossRef\]](#)
- Abdelhakim, A. Machine learning for localization of radioactive sources via a distributed sensor network. *Soft Comput.* **2023**, *27*, 10493–10508. [\[CrossRef\]](#)
- Zhang, W.; Zhao, T.; Zhao, Z.; Wang, Y.; Liu, F. An Intelligent Strategy Decision Method for Collaborative Jamming Based on Hierarchical Multi-Agent Reinforcement Learning. *IEEE Trans. Cogn. Commun. Netw.* **2024**, *10*, 1467–1480. [\[CrossRef\]](#)

24. Wu, J.; Sun, Y.; Li, D.; Shi, J.; Li, X.; Gao, L.; Yu, L.; Han, G.; Wu, J. An Adaptive Conversion Speed Q-Learning Algorithm for Search and Rescue UAV Path Planning in Unknown Environments. *IEEE Trans. Veh. Technol.* **2023**, *72*, 15391–15404. [[CrossRef](#)]
25. Moon, J.; Papaioannou, S.; Laoudias, C.; Kolios, P.; Kim, S. Deep Reinforcement Learning Multi-UAV Trajectory Control for Target Tracking. *IEEE Internet Things J.* **2021**, *8*, 15441–15455. [[CrossRef](#)]
26. Liu, B.; Zhang, Y.; Fu, S.; Liu, X. Reduce UAV Coverage Energy Consumption through Actor-Critic Algorithm. In Proceedings of the 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), Shenzhen, China, 11–13 December 2019; pp. 332–337.
27. Hou, Y.; Zhao, J.; Zhang, R.; Cheng, X.; Yang, L. UAV Swarm Cooperative Target Search: A Multi-Agent Reinforcement Learning Approach. *IEEE Trans. Intell. Veh.* **2024**, *9*, 568–578. [[CrossRef](#)]
28. Alagha, R.; Mizouni, R.; Singh, S.; Bentahar, J.; Otrók, H. Adaptive target localization under uncertainty using Multi-Agent Deep Reinforcement Learning with knowledge transfer. *Internet Things* **2025**, *29*, 101447. [[CrossRef](#)]
29. Liu, S.; Lin, Z.; Wang, Y.; Huang, W.; Yan, B.; Li, Y. Three-body cooperative active defense guidance law with overload constraints: A small speed ratio perspective. *Chin. J. Aeronaut.* **2025**, *38*, 103171. [[CrossRef](#)]
30. Katircioglu, O.; Isel, H.; Ceylan, O.; Taraktas, F.; Yagci, H.B. Comparing Ray Tracing, Free Space Path Loss and Logarithmic Distance Path Loss Models in Success of Indoor Localization with RSSI. In Proceedings of the 19th Telecommunications Forum (TELFOR), Belgrade, Serbia, 22–24 November 2011; pp. 313–316.
31. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
32. Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1999; Volume 12.
33. Sutton, R.S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **1988**, *3*, 9–44. [[CrossRef](#)]
34. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
35. Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *arXiv* **2015**, arXiv:1506.02438.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.