SACCL:A Novel Reinforcement Learning-Based Approach for Radiation Source Localization

Authors: Ms. Xiaolu Li, Prof. Jian-Wen Huo, Ling, Mr. Mingrun, Guo, Yunlei, Hu, Mr. Xulin, Prof. Jian-Wen Huo

Date: 2025-09-11T16:08:28+00:00

Abstract

With the widespread application of nuclear technology, nuclear safety measures have become increasingly critical. Lost or stolen radioactive sources pose severe threats to public safety, making autonomous radioactive source positioning and search a research focus. However, existing methods exhibit deficiencies in autonomous decision-making and adaptability in unknown complex environments. Therefore, this paper proposes a reinforcement learning strategy (referred to as SACCL) that combines Soft Actor-Critic (SAC) with Curriculum Learning (CL) for autonomous radioactive source positioning. This method achieves positioning through continuous interaction between the agent and the environment. Specifically, it utilizes observed position and radiation intensity data to gradually determine the location of the radiation source. To reduce the computational burden caused by long sequences of historical data, a selective state encoding mechanism is adopted to improve training efficiency. During the training phase, a CL partition strategy is introduced and combined with the Prioritized Experience Replay (PER) mechanism, which aims to enhance sample utilization and the agent's adaptability to unknown environments, enabling post-training search for any radioactive source within the radiation area without retraining. Additionally, a radiation-aware adaptive speed regulation mechanism is designed. It dynamically adjusts the agent' s movement speed based on real-time radiation intensity, enabling rapid coverage in low-radiation areas and meticulous search in high-radiation areas, thereby further improving search efficiency. To verify the effectiveness and reliability of the proposed method, we conducted experiments in both simulated and actual environments. The results showed that the SAC method lacked generalization ability; SAC_{CL} and SAC_{LSTM} had limited search performance in the presence of unknown obstacles; while the SACCL method achieved a success rate of 93% for source point localization in a complex environment with unknown obstacles, with an average localization error of 0.17 meters. This indicates its adaptability and stability.

Full Text

Preamble

SACCL: A Novel Reinforcement Learning-Based Approach for Radiation Source Localization

Xiaolu Li, Jianwen Huo, Mingrun Ling, Yunlei Guo, and Xulin Hu
Southwest University of Science and Technology, Mianyang 621010, China

With the widespread application of nuclear technology, nuclear safety measures have grown increasingly critical. Lost or stolen radioactive sources pose severe threats to public safety, making autonomous radioactive source positioning and search a key research focus. However, existing methods exhibit deficiencies in autonomous decision-making and adaptability in unknown complex environments. This paper proposes a reinforcement learning strategy (referred to as SACCL) that combines Soft Actor-Critic (SAC) with Curriculum Learning (CL) for autonomous radioactive source positioning. This method achieves positioning through continuous interaction between the agent and the environment, utilizing observed position and radiation intensity data to gradually determine the location of the radiation source. To reduce the computational burden caused by long sequences of historical data, a selective state encoding mechanism is adopted to improve training efficiency. During the training phase, a CL partition strategy is introduced and combined with the Prioritized Experience Replay (PER) mechanism, which aims to enhance sample utilization and the agent' s adaptability to unknown environments—allowing post-training search for any radioactive source within the radiation area without re-training. Additionally, a radiation-aware adaptive speed regulation mechanism is designed to dynamically adjust the agent' s movement speed based on real-time radiation intensity, enabling rapid coverage in low-radiation areas and meticulous search in high-radiation areas, thereby further improving search efficiency. To verify the effectiveness and reliability of the proposed method, experiments were conducted in both simulated and actual environments. The results showed that the SAC method lacked generalization ability, while SAC CL and SAC LSTM had limited search performance in the presence of unknown obstacles. In contrast, the SACCL method achieved a 93% success rate for source localization in a complex environment with unknown obstacles, with an average localization error of 0.17 meters, demonstrating its adaptability and stability.

Keywords: Radiation source localization, Soft Actor-Critic, Curriculum Learning, Partition strategy, Generalization ability

Introduction

With the wide application of nuclear technology in fields such as energy, healthcare, and industry, the risks associated with nuclear leaks and the loss of high-intensity radioactive sources have gradually increased, making nuclear safety a major global concern. The potential hazards and complexity of nuclear environ-

ments make research in this area both critically important and highly challenging. Among these efforts, the rapid and accurate localization of lost or missing radioactive sources is a key aspect of nuclear security, attracting significant attention from researchers. Since radioactive sources pose radiation hazards to humans, and given recent advancements in robotics technology, the use of robots equipped with radiation detectors has emerged as a safer and more feasible alternative to traditional manual search methods, which involve extremely high risks. Currently, numerous robotic positioning algorithms have been developed to locate anomalous radioactive sources resulting from accidental events.

Alwars et al. [**?**] designed a rapid and accurate single radioactive source detection system using a NaI(Tl) detector carried by a drone, providing a solution to the three-dimensional positioning problem in radioactive source scenarios. Lin et al. [**?**] estimated the position and activity of the radioactive source by recording radiation value changes during the robot's movement and combining the particle filter algorithm with the artificial potential field method. Xu et al. [**?**] employed recursive Bayesian estimation and sequential Monte Carlo methods to estimate the numerical values recorded by the spatial utilization robot at different points, thereby determining the positions of unknown sources in the environment. Fu et al. [**?**] combined Tsallis divergence with the particle filtering algorithm to achieve radioactive source localization. Liu et al. [**?**] utilized unscented particle filter to estimate parameter information of unknown radiation sources. However, these methods are applicable only when the environmental map is known or there are no obstacles in the radiation environment. In actual radiation source search tasks, challenges such as unknown environmental maps and complex terrain are frequently encountered. Obstacles present in the environment can block radiation signals, further increasing the difficulty of positioning. Therefore, robots must possess the ability to autonomously explore the environment and intelligently plan radiation measurement paths. Unlike search methods that rely on predefined paths [**?**], autonomous search strategies require real-time integration of current and historical radiation measurement information and dynamic decision-making on subsequent exploration paths to efficiently locate the target.

In recent years, reinforcement learning algorithms have performed well in various fields. Owing to their capability to optimize decision-making under uncertainty, these algorithms have demonstrated significant potential in the field of autonomous search. Many studies have applied reinforcement learning to positioning and searching tasks with positive results. For instance, Z. Liu et al. [**?**] utilized a CNN-based Double Q-learning algorithm to plan the robot' s path using CNN and historical data for radiation source detection. Gu et al. [**?**] applied double deep Q-network (DDQN) with a dual network structure to radiation source tracking, handling unknown sources with dynamic positions. Sadhu et al. [**?**] proposed the RadDQN architecture, which aims to achieve time efficiency and minimize radiation exposure while identifying optimal paths in radioactive environments. Zhao et al. [**?**] introduced the Particle Clustering-Deep Q-Network (PC-DQN), formulating the search task as a partially observable

Markov decision process (POMDP). Their approach uses DBSCAN clustering to extract confidence state features and a deep Q-network to derive an optimal strategy for source localization. Philippe et al. [**?**] designed a novel neural network architecture based on the A2C framework to search for radiation sources in non-convex environments, significantly reducing search time and improving success rate. Minkyu Park et al. [**?**] proposed a source term estimation strategy based on deep reinforcement learning, using GMM to approximate the particle filter to estimate the source term, with DDPG used as the decision-maker to find the source. Shi Yiwei [**?**] combined multi-step deep reinforcement learning with Bayesian inference to estimate source information, where the multi-step update mechanism improved positioning efficiency (increased success rate and reduced number of time steps). Jeremy Marquardt et al. [**?**] introduced reinforcement learning (RL) into the Bayesian optimization (BO) search framework, significantly reducing estimation error. The above research has fully demonstrated the effectiveness of reinforcement learning in solving the problem of locating radioactive sources, making significant progress in improving search efficiency, success rate, reducing radiation exposure, and adapting to the complexity of specific environments. However, most studies (as shown in Table 1) were conducted in relatively simple and obstacle-free radiation environments for verification, and their adaptability to complex real-world scenarios is limited. The trained intelligent agents have weak migration capabilities and poor environmental adaptability; once the radiation scenario (such as obstacle layout or source characteristics) changes, it often requires a significant amount of time to retrain the model. Moreover, training a mature and effective intelligent agent usually demands vast amounts of samples and computing resources, resulting in extremely high time costs. Therefore, the time efficiency of robot search for radioactive sources, including the time efficiency of training mature intelligent agents, is the core bottleneck for practical application of this technology. Some studies (such as those by Atefeh Fathi [**?**], Mohammed Shurrab [**?**], Hu [**?**], and others) have begun to focus on the transfer capability or environmental adaptability of intelligent agents, aiming to enable trained agents to localize and search for radiation sources in altered environments or settings with unknown obstacles. However, these attempts still exhibit certain limitations—for instance, reliance on prior information (Atefeh Fathi' s method requires training with radiation data, which limits its practicality in real scenarios where information about the radiation source, such as intensity and type, is unknown) or lengthy training time required to develop a mature agent (Mohammed Shurrab' s approach still demands training on the order of 10^6 steps, which is too time-consuming to meet rapid-response requirements).

In response to key issues such as weak environmental adaptability, poor transfer capability, and low training efficiency in existing reinforcement learning methods for radioactive source search—as well as the resulting high time costs that hinder practical application—this paper investigates strategies for efficient localization and retrieval of radioactive sources. The core objectives are to significantly improve the temporal efficiency of the search process (including reducing training

costs and accelerating source detection) while enhancing the intelligent agent' s adaptability and generalization ability in complex and dynamic environments.

In reinforcement learning algorithms, the main problem addressed is the Markov decision framework issue. The search for radioactive sources is essentially a partially observable Markov decision process (POMDP) [**?**]. The robot can only obtain partial environmental states, making the decision-making process very challenging. Current approaches to solving POMDP can be mainly categorized into two types [**?**, **?**]: one is to approximate the belief state as a fully observable MDP using particle filtering, but it is difficult to avoid the problem of particle degeneration; the other is to use network coding such as RNN/LSTM to encode the historical observation sequence, but this faces the problem of high computational complexity caused by long sequences, which leads to low training efficiency. In response to these issues, inspired by the three-point positioning principle of radioactive sources [**?**], this paper proposes a selective historical encoding strategy. This method also utilizes neural networks to extract features but discards lengthy historical sequences and selectively retains and encodes key information from the agent's recent three steps, including the radiation values of positions passed in the last three steps and the actions performed. This design aims to assist the intelligent agent in learning underlying states more effectively, thereby enabling better subsequent decisions while avoiding the computational burden and training instability caused by handling long sequences.

In conclusion, this paper makes the following contributions:

1. A radiation source localization framework (referred to as SACCL) is proposed, which integrates the Soft Actor-Critic (SAC) algorithm with prioritized experience replay and a curriculum learning strategy. By leveraging reinforcement learning for robotic motion decision-making, the framework enhances environmental adaptability and autonomy in source search.

2. Instead of long historical sequences, the work selectively encodes key state information of the robot. This approach allows the agent to effectively learn underlying environmental states while reducing computational overhead and improving training stability.

3. The robot' s movement speed and direction are decoupled, and a radiation-aware adaptive speed mechanism is designed to realize intelligent adjustment of the robot' s search speed. Combined with movement direction output by the SAC network, this mechanism enhances overall search efficiency.

The structure of this paper is organized as follows: Section 1 introduces the research background and motivation; Section 2 provides relevant theoretical foundations; Section 3 elaborates on the proposed methodological framework; Section 4 verifies the framework's effectiveness through comparative experiments with different benchmark methods and corresponding metric analyses; Section 5 summarizes the work presented in this paper.

II. Theoretical Descriptions

Typically, lost or stolen radioactive sources are very small compared to the search area [**?**]. Therefore, in this study, the radioactive sources are treated as point sources.

In a radioactive environment, various types of particles are produced by processes such as alpha decay, beta decay, gamma decay, neutron emission, and nuclear fission. Among them, the gamma rays produced by gamma decay have strong penetrating power, with longer propagation distances compared to other types and higher energy. They are widely used in irradiation sterilization, industrial flaw detection, and so on. Therefore, most lost or stolen radioactive sources are mainly of the gamma type. Hence, this paper focuses on gamma-ray radioactive sources as the experimental subjects for detection and conducts detection research by equipping robots with nuclear radiation detectors. When gamma rays undergo radioactive decay, the exposure rate at a certain measurement point within the radiation field can be expressed by Eq. (1):

$$\dot{X} = \frac{\Gamma A}{R^2}$$

Here, X represents exposure dose, $\dot{X}$ ($\mathrm{C \cdot kg^{-1} \cdot s^{-1}}$) denotes the gamma ray exposure rate; A (Bq) is the activity of the point radiation source; Γ ($\mathrm{C \cdot m^2 \cdot kg^{-1} \cdot Bq^{-1} \cdot s^{-1}}$) refers to the exposure rate constant of the point radiation source; R (m) represents the distance from the point radiation source to the measurement point.

When gamma rays pass through different media, they experience varying degrees of attenuation. This attenuation follows an exponential decay pattern, as described by Eq. (2):

$$\dot{X} = \frac{\Gamma A}{R^2} e^{\sum u_i R_i}$$

Here, R_i represents the distance that the gamma ray travels through the ith medium, and u_i denotes the linear attenuation coefficient of the ith medium for the gamma ray. From the formula above, it can be observed that the radiation intensity at any observation point is inversely proportional to the square of the distance from that point to the radiation source.

The exposure dose X is converted into the internationally standardized dose equivalent H (μSv) for further characterization as follows:

$$H = w \cdot D = w \cdot f \cdot X$$

Here, w represents the radiation weighting factor, which varies for different types of radiation; D ($\mathrm{J \cdot kg^{-1}}$) denotes the absorbed dose, which describes the amount

of radiation energy absorbed by the substance and the resulting radiation effects. Exposure is a measure of the outcome of the interaction between radiation and matter. Both describe the energy effect produced by a unit mass of the substance from different perspectives. They are linked through the conversion factor f ($\mathrm{J} \cdot \mathrm{C}^{-1}$), that is, $D = f \cdot X$.

When detecting decay particles emitted by a radioactive source using radiation detectors, not all particles enter the detector. Therefore, the detection process is stochastic, and the recorded count in radiation measurement is a random variable. Suppose within a measurement time t, all N particles emitted by the source decay and are incident on the detector. If the detection efficiency for these particles is ε, then the detection of each individual particle follows a Bernoulli trial. With N independent particles incident on the detector, the process is equivalent to N independent Bernoulli trials. Therefore, the detected count n follows a binomial distribution:

$$P_N(n) = \frac{N!}{(N-n)!n!}\varepsilon^n(1-\varepsilon)^{N-n}$$

However, the number of particles N that enter the detector is not a constant but a random variable following a Poisson distribution. Denoting its expected value as M, we obtain:

$$P_N(n) = \sum_{N=n}^{\infty} \frac{N!}{(N-n)!n!}\varepsilon^n(1-\varepsilon)^{N-n} \cdot \frac{M^N e^{-M}}{N!}$$

Then, the probability function for the detector's output count n can be obtained as:

$$P(n) = \frac{(M\varepsilon)^n e^{-M\varepsilon}}{n!}$$

It can be seen that when the number of incident particles N follows a Poisson distribution with an average value of M, the count n of the detector also follows a Poisson distribution, and its average value m and variance σ^2 are:

$$m = M\varepsilon, \quad \sigma^2 = M\varepsilon$$

Suppose the measured count rate of the radiation detector at a certain position is z, and the theoretical count rate at that position is μ. The expected number of detections within a time interval of τ seconds is $\lambda = \mu\tau$. According to Poisson statistics, the probability detection rate of the radiation detector is:

$$P(z;\lambda) = \frac{\lambda^z e^{-\lambda}}{z!}$$

The counting rate is measured in CPM (counts per minute). It can be converted to and correlated with the dose equivalent via Eq. (3), using the energy-dependent conversion factor known as the Energy Response:

$$\text{CPM} = \dot{X} \cdot \text{Energy Response}$$

III. Method

The goal of reinforcement learning is to learn strategies that can maximize cumulative rewards by interacting with the environment. A strategy refers to the decision-making process of the intelligent agent based on the observed state—that is, how to select an action from the action space to optimally achieve the goal. In recent years, reinforcement learning algorithms have developed rapidly, with numerous algorithms such as DQN, Q-Learning, Actor-Critic, DDPG, and TRPO emerging one after another. Their outstanding performance has made them highly favored in fields such as control and game theory. In the autonomous source-seeking task of robots, the agent is confronted with a complex and partially observable environment. A key challenge lies in designing algorithms that enable efficient exploration of unknown regions to locate the target source while avoiding local optima or disorientation due to sensor noise and environmental randomness. Therefore, there is an urgent need for robust algorithms that can effectively utilize past experience, maintain sufficient exploration ability, and handle environmental randomness and sensor noise. The Soft Actor-Critic (SAC) algorithm, with its unique advantages, has become one of the ideal solutions for such problems.

SAC [**?**, **?**] is an off-policy deep reinforcement learning algorithm that innovatively incorporates the principle of maximum entropy within the Actor-Critic framework. While the standard objective of reinforcement learning is to maximize cumulative rewards, the training process in its early stages often suffers from high variance in reward signals and an unconverged policy network. As a result, the agent may prematurely favor actions with the highest estimated value—i.e., over-exploit known "good" actions—leading the policy to converge to deterministic or low-entropy behaviors. This suppresses exploration of potentially superior actions and states, often resulting in convergence to local optima. To mitigate this issue, SAC introduces an entropy term of the policy as a regularization term in addition to the conventional cumulative reward objective (as shown in Eq. (10)). By incorporating maximum entropy into the optimization target, the agent is encouraged to balance reward maximization with behavioral randomness. This approach helps maintain exploratory behavior while pursuing high returns, prevents premature convergence to suboptimal policies, and enhances the algorithm' s capacity to explore the state-action space more thoroughly. Furthermore, it improves the agent' s generalization capability and robustness across varying initial conditions, thereby strengthening its overall adaptability to environmental changes.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t \gamma^t (r(s_t, a_t) + \alpha H(\pi(\cdot|s_t)))\right] \quad (10)$$

Here, $r(s_t, a_t)$ represents the reward obtained by the agent when taking action a_t in state s_t, γ represents the discount factor for future rewards, the term $H(\pi(\cdot|s_t))$ corresponds to the entropy of the policy π at state s_t, and α is the temperature parameter that balances the relative importance between the entropy regularization term and the cumulative reward. The design of this objective function encourages the strategy to pursue high cumulative rewards while maintaining a certain degree of randomness (high entropy). This mechanism offers two key advantages: First, it significantly enhances the agent' s exploration capability, prompting it to experiment with new actions that may yield higher long-term returns, thereby effectively preventing the policy from converging prematurely to suboptimal solutions. Second, entropy regularization can automatically adjust the exploration intensity, thus naturally striking a balance between exploring unknown information and exploiting existing knowledge during the learning process. However, the temperature coefficient α controls the stochasticity of the output action. A larger α increases the uncertainty of the action distribution, encouraging the agent to explore more extensively, whereas a smaller α leads the agent to focus more on maximizing immediate rewards. A fixed value of α is often insufficient to accommodate the varying requirements at different stages of the learning process. Ideally, the agent should explore more actively in the early phases and progressively converge toward the optimal policy in later stages, reducing exploration and adopting a more deterministic strategy.

To address this issue, an adaptive adjustment mechanism for the temperature parameter α is introduced (with its loss function given in Eq. (11)):

$$L(\alpha) = \mathbb{E}_{a\sim\pi_\phi}[-\alpha \log \pi_\phi(a|s) - \alpha\bar{H}] \quad (11)$$

Here, $\bar{H}$ represents the target entropy value. Through this equation, α can be dynamically adjusted: in the early training stage, a larger α value promotes a more uniform distribution of actions, thereby enhancing exploration; as the agent gains a deeper understanding of the environment and the strategy gradually improves, the value of α decreases, reducing the randomness of actions, and the strategy eventually converges to the high-reward region. When α approaches 0, the maximum entropy objective reduces to the standard reinforcement learning objective.

In autonomous source-seeking tasks of robots based on reinforcement learning algorithms, the target source signals are often spatially sparse. The uniform experience replay mechanism in traditional algorithms samples all experience samples equally, making it difficult to effectively utilize samples containing key information. To improve the learning efficiency of the agent in high-value states, this paper incorporates a prioritized experience replay mechanism into the SAC

framework. This mechanism dynamically allocates sample priorities based on temporal difference errors (as shown in Eq. (12)) and focuses on experiences with high potential for efficient learning through non-uniform sampling:

$$p_i = |\delta_i| + \epsilon \quad (12)$$

Here, δ_i represents the TD error, and ϵ is a positive constant that prevents the situation where samples will not be accessed when the TD error is 0.

The sampling probability of prioritized experience replay is determined according to the priority distribution (as in Eq. (13)) to ensure that samples with smaller TD errors can still be sampled, guaranteeing the diversity of samples during algorithm training. At the same time, to eliminate the distribution deviation caused by priority sampling, importance sampling weights (as in Eq. (14)) are introduced for bias correction:

$$P(i) = \frac{p_i^{\alpha_{\text{per}}}}{\sum_k p_k^{\alpha_{\text{per}}}}, \quad w_i = \left(\frac{1}{N \cdot P(i)}\right)^{\beta}, \quad \hat{\nabla} L(\theta) = \sum_i \min(w_j, 1)\delta_i \nabla_\theta Q_\theta(s_t, a_t) \quad (13, 14)$$

Here, α_{per} represents the intensity of regulatory priority, and N represents the capacity of the experience pool.

The Prioritized Experience Replay (PER) mechanism was combined with the SAC framework, achieving coordinated optimization of exploration efficiency in the action space and the state space. Specifically, the maximum entropy policy promotes thorough exploration of the action space to avoid premature convergence; the PER mechanism, on the other hand, guides the agent to focus on key regions within the state space that possess higher informational value (such as areas near the signal source or states with significant signal changes) for efficient experience mining. The combination of the two significantly enhances the search efficiency of the source-finding path and the convergence robustness of the final strategy.

As illustrated in Fig. 1 [Figure 1: see original paper], the radiation field is divided into four distinct regions, denoted as U_k ($k = 1, 2, 3, 4$). These regions are sorted according to the distance from the entrance of the radiation field. The agent first trains in region U_1 until it either completes a preset number of episodes or reaches a predefined success threshold. Subsequently, it progressively advances to U_k ($k = 2, 3, 4$) for learning. After training is completed across all four regions, a global training phase is conducted in the complete environment. This final phase integrates the knowledge acquired from each local region and fine-tunes the policy to approximate the global optimum.

The entire algorithm framework is shown in Fig. 2 [Figure 2: see original paper]. The proposed framework consists of three core modules, with their specific

functions and collaboration logic as follows: The curriculum learning environment (CL ENV) adopts a progressive regional strategy based on the progress of training rounds, randomly generating the positions of radiation sources within a selected range and initializing the robot' s initial position. The robot then obtains the current radiation intensity measurement value through detectors and inputs the position information, historical sequence information, and state vector together into the Soft Actor-Critic (SAC) intelligent agent. The policy network (Actor) of the SAC intelligent agent receives the state vector and outputs continuous two-dimensional actions for forward direction and speed, dynamically adjusting the maximum speed based on the current radiation intensity. Meanwhile, the system evaluates decisions through the reward function and determines the rewards and punishments for the agent and whether to terminate the round. After the round ends, the system stores the experience tuple $(s_t, a, s_{t+1}, r, \text{Done})$, and when the experience pool capacity is met, the prioritized experience replay (PER) buffer will sample based on sample importance priority, providing the agent with training data to improve sample utilization and learn high-reward decisions (the update steps of SAC per step are shown in Table 2). The three elements work together to form a complete closed loop of "perception-decision-learning," helping the framework improve its capabilities through progressive learning.

IV. Experiment and Discussion

To verify the correctness and effectiveness of the algorithm, this paper conducts simulation tests on the radioactive source search model based on the SAC algorithm. The framework design of the SAC algorithm is shown in Table 3 .

To make the designed reinforcement learning framework applicable to radiation field source localization, the agent's decisions are mainly based on environmental radiation information. Therefore, the state space design includes the following key elements: the current radiation measurement value detected by the agent' s current position, motion direction information representing the robot' s movement state, its positional coordinates in the environment, and the position of obstacles within its sensing range. The design also retains key information from the agent' s most recent three steps, including the radiation values of positions passed in the last three steps and the actions performed. This approach aims to help the agent learn implicit environmental states more effectively, thereby making better subsequent decisions and avoiding the computational burden and training instability caused by processing long sequences.

As shown in Table 3, the action space is continuous with independent dimensions. The agent outputs a unit vector a_t representing the movement direction through the policy network to control the continuous motion of the robot. To improve source-seeking efficiency, an adaptive mechanism for movement speed based on radiation perception was introduced: the maximum speed $v_{\max}$ for the next action is dynamically adjusted according to the current radiation intensity I_t based on predefined rules. Thus, the robot' s displacement is jointly determined

by the direction vector a_t and the adaptive speed magnitude. The update rule for the agent' s position is given by Eq. (15):

$$P_t = P_{t-1} + a_t V_{\max} \Delta t \quad (15)$$

Here, $P_t = [x_t, y_t]$ represents the position coordinates at time t, while $P_{t-1} = [x_{t-1}, y_{t-1}]$ represents the position coordinates at time $t-1$. The vector a_t is the unit direction vector output by the policy network at time t. The scalar $v_{\max}$ (m/s) denotes the magnitude of velocity determined based on the observed radiation intensity I_t according to a predefined rule with a fixed time step $\Delta t = 1$s. The velocity rule can be expressed as: $v_{\max} = \{1.0 \text{ for low-radiation area}, 0.8 \text{ for medium-radiation area}, 0.6 \text{ for high-radiation area}\}$. That is, when $I_t < 1000$ CPM (in the low radiation area), a higher speed ($v_{\max} = 1.0$ m/s) is adopted to quickly pass through this area, accelerating the source search process; when $1000 \leq I_t \leq 5000$ CPM (in the medium radiation area), a medium speed ($v_{\max} = 0.8$ m/s) is used; when $I_t > 5000$ CPM (in the high radiation area), a lower speed ($v_{\max} = 0.6$ m/s) is adopted to perform more precise positioning operations.

The movement direction output by the strategy network and the speed amplitude set by the rules jointly determine the robot' s movement. This design of decoupling the movement direction decision from speed amplitude control effectively reduces the learning complexity of the strategy network, allowing it to focus on direction decision-making, which helps improve the stability of the training process. The test results of this mechanism are shown in Fig. 3 [Figure 3: see original paper].

The reward function consists of two parts: a sparse reward R_s and an exploration reward R_e. The sparse reward is granted only upon successful localization of the radiation source. The exploration rewards are determined based on radiation readings detected by the robot, obstacle encounters, and time consumption. A reward R_d is given when the current radiation reading is higher than that of the previous time step. Additionally, if the current radiation value exceeds the maximum historical reading, an incremental reward R_i is provided. The next position of the robot is computed based on its current state. If obstacles are detected within 1.2 m along the intended path, a penalty R_o is applied, scaled inversely with the distance to the obstacle. It is worth noting that to avoid over-penalizing potentially valid paths near obstacles, this penalty is only activated within the 1.2-meter threshold. A time penalty term R_t is also introduced to incentivize the robot to find the target efficiently and discourage lingering merely to accumulate rewards.

The training parameters for the SAC algorithm are designed as shown in Table 4 .

This study assumes that an approximate map of the accident site has been obtained through radiation-resistant robots, with known boundaries and major

obstacles. Subsequently, in the simulation environment, the search area is simplified to a two-dimensional planar region, ignoring the influence of height. The radiation field was set as a 20 m × 20 m area. The radiation source was parameterized as $\theta_s = [x_s, y_s, I_s]^T$, with an intensity of $I_s = 5520$ CPM at 1 meter from the source, and the environmental background radiation was $\mu_b = 20$ CPM. Experiments were conducted in both obstacle-free and obstacle-present environments. The simulation robot started from the initial position, autonomously deciding its movement positions and step sizes based on sensor measurements until the radiation source was located. To enhance the environmental adaptability of the robot, the positions of the radiation sources are randomly generated throughout the entire radiation area according to the zone training strategy. The starting position of the robot is also expanded (the x-direction is within the range of 0 to 20 meters, and the y-direction is limited to 0 to 5 meters).

1. Obstacle Environment Test

In the early stage of this paper, the proposed algorithm was preliminarily verified in an environment without obstacles, and the test results are presented in Appendix 1. Subsequently, two basic square obstacles were added to the obstacle-free environment, with their positions fixed. The attenuation coefficient of the obstacles was set to 0.1 cm^{-1}, and the intelligent agent was trained. Due to the slightly more complex radiation environment after adding obstacles, the number of training episodes was also increased accordingly. According to Fig. 1, the source was randomly generated in the regions U_k $(k = 1, 2, 3, 4)$ for 500 episodes of training each, and then randomly generated throughout the entire area for training up to 8000 episodes. The training data was visualized to obtain the reward function curve, the final source-seeking error, and the success rate curve, as shown in Fig. 4 [Figure 4: see original paper].

As shown in Fig. 4, during the transformation process, all parameters fluctuated. In the obstacle environment, the first transformation caused fluctuations but quickly adjusted. The second and third regional transformations also caused fluctuations, especially in region U_4, because of the increased Euclidean distance from the entrance and larger data volume, and the training in the third region was not fully completed, the error fluctuations were more obvious. However, as the global regional training gradually repaired, the agent gradually found the global optimal strategy and converged. Around 3000 episodes, parameters such as the agent' s reward function, error, and success rate all tended to stabilize.

The trained model was tested to verify its effectiveness. As shown in Fig. 5 [Figure 5: see original paper], in an obstacle environment, the robot can locate the source at any position without being restricted by the robot' s starting position. Thus, it can be seen that this algorithm framework remains effective in complex environments.

To verify the effectiveness of this algorithm, a comparative experiment was conducted by comparing the proposed algorithm with the SAC CL algorithm and

the SAC LSTM algorithm. The SAC CL algorithm refers to the adoption of a regional curriculum learning training strategy on the basis of the pure SAC algorithm. The SAC LSTM algorithm refers to the introduction of an LSTM network on the basis of the SAC CL algorithm to extract state sequences and patterns from historical data. Since the SAC algorithm was found to lack transferability and environmental adaptability in the initial obstacle-free environment (as shown in Appendix VI), it was no longer used as the comparison baseline in the obstacle environment comparison test. Under the condition of ensuring fair comparison (with consistent reward functions, region sizes, training times, and neural network parameters), training was carried out, and the three algorithms were trained and tested. Each algorithm model was subjected to 100 random tests (randomly specifying the location points of the radiation source and the initial position of the robot), and the robot' s source localization performance was statistically analyzed. The results are shown in Table 5 . The distribution of source locations generated during the test is shown in Fig. 8 Figure 8: see original paper. As can be seen from the figure, the source generation locations cover most of the area throughout the region.

As can be seen from Table 5, after training with the SAC CL algorithm, the robot' s source-finding success rate was 72%, with an average positioning error of 2.73 m. This result indicates that in a highly random and partially observed obstacle environment, adding the traditional curriculum learning strategy still cannot enable the robot to adopt a highly robust strategy. The SAC LSTM algorithm performed better than SAC CL, with a success rate of 83% and an average positioning error of 2.18 meters. This suggests that processing historical sequence information is helpful for dealing with complex environments. However, the training model of the algorithm proposed in this paper had a significantly higher success rate than the comparison benchmark algorithms, at 98%, and the average positioning error was 0.38 meters. This proves that the proposed SACCL algorithm still has better performance in complex environments.

2. Test for Environments with Unknown Obstacles

In actual situations, radiation-resistant robots may fail to detect obstacles, or new obstacles may appear after an accident. Therefore, it is assumed that there are undetected obstacles. The experiment sets up an obstacle that is randomly generated during the training process to train the agent to be able to deal with sudden obstacles. The same model as previously set is adopted. Because there are unknown obstacles, the obstacle attenuation coefficient is set in the range of 0.1-0.2 cm^{-1}. During the training process, the positions, shapes, and attenuation coefficients of the random obstacles are all unknown. This environment is trained for 8,000 episodes, and the training model' s reward function curve, error curve, and success rate curve are shown in Fig. 6 [Figure 6: see original paper].

From the data results of the training process, it can be seen that compared with fixed and known obstacles, the reward function fluctuates more significantly,

especially when training in regions U_3 and U_4, where the single-episode reward value of the robot reaches a new low and is extremely unstable. This is related to the interference of dynamic obstacles, but as the number of training episodes increases, the robot gradually acquires more regional information and also gradually learns how to avoid dynamic obstacles to obtain greater benefits.

After training, the intelligent agent was tested, and the results are shown in Fig. 7 [Figure 7: see original paper]. From the test results of several groups in Fig. 7, it can be seen that the robot has learned how to avoid fixed obstacles and unknown randomly generated obstacles and will search for the radiation source along the shortest path. When the SAC CL and SAC LSTM algorithms are applied in an environment with unknown obstacles, the robot is unable to find the source. During training, it even gets stuck in a vicious cycle and wanders in an area far from the source.

Therefore, only the SACCL algorithm was tested. Then, 100 radiation source positions were specified, among which the robot' s starting position and the positions of dynamic obstacles were also randomly generated, including the shape of the dynamic obstacles. The attenuation coefficient was also randomly generated, and the test results are shown in Table 6 . The distribution of the source locations generated during the test is shown in Fig. 8 Figure 8: see original paper.

Out of these 100 tests, there were 30 cases of dynamic obstacles with low attenuation (the attenuation coefficient was set to 0.1 cm^{-1}), and the success rate of source localization was 96.67%; 50 cases were high-attenuation obstacles (the attenuation coefficient was set to 0.2 cm^{-1}), and the success rate of source localization was 90%. This indicates that the attenuation degree of the obstacles also has an impact on source localization detection. However, the average source localization error of SACCL remained at a relatively low level, with an error of 0.69 m. The above test results show that within the range of attenuation coefficients set (0.1–0.2 cm^{-1}), regardless of the position and shape of the obstacles, the robot can still accurately locate the radiation source. This proves that the algorithm still has effectiveness and robustness in an environment with unknown obstacles. Compared to the fixed obstacle environment, due to the existence of unknown obstacles for the intelligent agent, the relative source localization error is slightly larger, which is in line with our expectations.

3. Verification of Realistic Complex Scenarios

After the algorithm' s rationality was verified in the simulation environment, it was applied to physical experiments for further verification. The detector used was the NaI radiation detector. The experimental scene was a 6 m × 6 m area. To facilitate a more intuitive view of the robot' s position during the experimental test, the boundary and scale were drawn in the experimental area in advance. The radioactive source Co-60 was selected as the experimental object, with an activity of 6.6×10^7 Bq. The source position was (1.5 m, 5.0

m). The robot entered the radiation area from any point in the XA direction for source localization. Two wooden boards and one iron plate were selected as obstacles, and their placement positions are shown in Fig. 9 [Figure 9: see original paper].

Before conducting the source search in the actual scenario, due to the high error caused by the detector being installed on the robot, the selected source and the actual scenario data were verified through detection before the experiment. The measured data was visualized, and the results are shown in Fig. 10 [Figure 10: see original paper]. Since the radioactive source used in the experiment was enclosed in a casing, the data deviation was relatively large near the source (within 1 m of the source), but the fitting was good at a distance of 1 m from the source and beyond. At the same time, the one-dimensional scenario was expanded to a two-dimensional scenario to obtain the radiation distribution map. Compared with the simulation experiment (as shown in Fig. 5 and Fig. 7), it was found that in the real scenario, the radiation source would attenuate more significantly as the distance increases.

At the same time, tests were conducted on the obstacle scenarios, and the radiation detection data were fitted. To present the attenuation effect of obstacles more intuitively, the radiation values in the obstacle-free environment and in the two obstacle environments were compared. The results are shown in Fig. 11 [Figure 11: see original paper]. By comparing the data measured in Fig. 10 and Fig. 11, it is found that the attenuation and obstruction effects of the iron plate and the wooden plate are extremely obvious within a distance of 1 meter from the source. Beyond 1 meter, the data are relatively close. Therefore, the data after 1 meter are enlarged for visualization, and the difference in attenuation effects between the two obstacle materials and the environment without obstacles at a distance of 1.4 meters is close. Moreover, the gap gradually decreases after 1.4 meters, and after 2 meters, the robot' s detection data in both obstacle environments hardly changes, which can be considered as close to the attenuation value without obstacles.

The measured data from the actual scene are imported into the simulation environment of this paper. A trained agent is used to locate and search for the radiation source in the experimental scene in Fig. 9, and the robot' s detection point data (shown in Table 7) and path information are obtained. The visualization result is shown in Fig. 12 [Figure 12: see original paper].

From the results in Fig. 12, it can be seen that the robot located the radioactive source at (1.4 m, 4.8 m), with a positioning error of 0.22 m. It successfully found the radioactive source, verifying that the algorithm proposed in this paper has autonomy and environmental adaptability for the search and location of radioactive sources.

V. Summary

This paper focuses on the problem of single-robot radioactive source localization and proposes a SAC reinforcement learning framework (SACCL) that integrates curriculum learning (CL) to enable robots to autonomously locate radioactive sources in complex radiation environments. The experiments were conducted in a radiation field with known boundaries, and during the training phase, the radioactive source positions were randomly generated. After training, the radioactive source positions were randomly specified to evaluate the performance of radioactive source localization. The results show that the robot trained by SACCL can efficiently locate any position of the radiation source within the radiation area, even in an environment with unknown obstacles, maintaining good performance. This proves that the algorithm has environmental adaptability. The SACCL framework only requires the known boundary information of the radiation field, and after training, the robot can complete the inference of the source position and path planning. By introducing selective historical sequence encoding and radiation perception speed adaptive mechanism, the training efficiency and convergence speed have been enhanced, and the training time has been reduced. However, the current research only focuses on single-source scenarios, and there is a limit in the assumption of the number of sources, which restricts its practicality in more widely applied scenarios. Future work aims to expand the SACCL framework to multi-radiation source localization tasks and further study its application in more complex and realistic environments.

VI. Appendix 1: Accessibility Environment Test

In an accessible environment, following the curriculum training strategy, 500 episodes of training were conducted in region U_1, followed by the same duration of training in regions U_k $(k = 2, 3, 4)$. After 2000 episodes, the agent began to adjust the strategy throughout the entire region, and the training reward function value curve, sourcing error, and success rate curve were obtained, respectively, as shown in Fig. 13 [Figure 13: see original paper].

As shown in the figure, during the first change of the training area (at the 500th episode), the reward value, error, and success rate parameters all fluctuated. This is because the agent had fallen into a local optimum in area U_1. Changing the area disrupts the existing strategies it has learned, so the agent needs to adjust its strategies to adapt to the new area. After the area change learning, the agent has gradually learned how to approach the target. In subsequent area changes, the fluctuations have decreased significantly and even stabilized. At the same time, the error has decreased to around 0. This indicates that the agent has learned how to find the radiation source and possesses generalization ability and environmental adaptability.

The trained intelligent agent was tested. By changing the position of the radiation source and resetting the robot' s position, it was verified whether the robot could intelligently and autonomously locate the source. The results are shown

in Fig. 14 [Figure 14: see original paper]. As can be seen from the figure, the trained intelligent agent can quickly lock onto the target and move toward the target source along the optimal path.

To verify the effectiveness of this algorithm, a comparative experiment was conducted by comparing the proposed algorithm with the SAC algorithm, the SAC CL algorithm, and the SAC LSTM algorithm. The results obtained by training the four algorithms under the same conditions (consistent reward functions, consistent region sizes, consistent training times, and consistent neural network parameters) are shown in Table 8 .

From the table, it can be seen that when only the SAC algorithm is added and random sources are generated within the area under the same training times, the robot is almost unable to find these sources, with a success rate of only 38%, and is unable to learn useful strategies. When the training strategy of curriculum learning is added to the SAC algorithm, the robot improves, but the success rate of finding the sources is 63%. After adding the LSTM strategy, the robot can almost find the radiation sources at any position, with a success rate of 98%, but compared with the algorithm used in this paper, the number of steps to find the sources is more, the average error is also larger, and the path is longer. The above comparative test shows that this method enables intelligent robots to successfully locate the target position in an obstacle-free environment, regardless of how the target position changes, and is not limited by the robot' s starting position.

VII. Appendix 2: Physical Test Design

Because the radiation from the source spreads out radially from the source as a whole, to reduce the workload, this paper only conducts test radiation data in a one-dimensional scenario, thereby reconstructing the entire radiation field. As shown in Fig. 15 [Figure 15: see original paper], the source is placed at the origin of the coordinate system, and radiation values are detected at intervals of 0.2 meters. The obtained data is then subjected to visual analysis.

During the autonomous exploration process carried out by mobile robots equipped with detectors, the radiation model can be expressed by Eq. (17):

$$\dot{X} = \frac{\Gamma A}{4\pi R^2} e^{\sum u_i m_i} + \dot{H}_b \quad (17)$$

In the formula, u_i represents the attenuation factor of the medium shielding, m_i is the thickness of the shielding medium, and $\dot{H}_b$ is the local environmental radiation dose.

When performing the fitting of the detection data, to visually display the integrated processing of some parameters in Eq. (1), the expression is shown as Eq. (18):

$$\dot{X} = \frac{K}{4\pi R^2} + \dot{H}_b \quad (18)$$

where $K = \Gamma A e^{\sum u_i m_i}$.

The measured data were processed visually according to Eq. (18), and the relevant conditions of the source in the actual situation were analyzed. At the same time, the experiment also tested the obstacle scenarios. In this study, two types of obstacles were selected, namely wooden boards and iron plates. The experimental placement scene is shown in Fig. 16 [Figure 16: see original paper]. The visual results of the obtained data are shown in Fig. 17 [Figure 17: see original paper] and Fig. 18 [Figure 18: see original paper].

Note: Figure translations are in progress. See original paper for figures.