A Source-Seeking Method Combining Angle-of-Arrival Particle Filtering and Reinforcement Learning

Zhao Yaolong, Xiao Yufeng, Zheng Yulai, Yan Dong, Ren Zhenyu, Wang Zhao, Yang Shuang

2026-08-11

ChinaRxiv

Abstract

To address the susceptibility of autonomous radioactive source search in complex occluded environments to radiation-count fluctuations and local optima, a radioactive source search method combining angle-of-arrival particle filtering with reinforcement learning is proposed. The method uses angle-of-arrival (AOA) priors to constrain particle initialization, recursively estimates the posterior distribution of the radioactive source location with a modified radiation observation model, and constructs source-position offset, uncertainty, and confidence as belief features. Meanwhile, a gated recurrent unit (GRU) is introduced to extract historical observation information, and proximal policy optimization (PPO) is adopted to learn motion decision-making policies in complex occluded environments, thereby reducing local stagnation caused by greedy approach behavior. Experimental results show that under single-source low-signal-to-noise-ratio and multi-obstacle conditions, the proposed method maintains a success rate above 97.3%, achieves a localization error below 0.5 m, and demonstrates good obstacle-bypassing and escape capability in L-shaped non-convex obstacle scenarios.

Full Text

A Source-Seeking Method Combining Angle-of-Arrival Particle Filtering and Reinforcement Learning

Zhao Yaolong[1], Xiao Yufeng[1,†], Zheng Yulai[2], Yan Dong[1], Ren Zhenyu[1], Wang Zhao[1], Yang Shuang[1]

(1. School of Information and Control Engineering, Southwest University of Science and Technology, Mianyang, Sichuan 621010;
2. China Institute of Atomic Energy, Beijing 102413)

Abstract: To address the problem that autonomous search for radioactive sources in complex shielding environments is susceptible to fluctuations in radiation counts and local optima, this paper proposes a radioactive-source search method combining angle-of-arrival particle filtering and reinforcement learning. The method uses angle-of-arrival (AOA) prior constraints for particle initialization, combines a corrected radiation observation model to recursively estimate the radioactive-source location posterior, and constructs source-position offset, uncertainty, and confidence as belief features. At the same time, a gated recurrent unit (GRU) is introduced to extract historical observation information, and proximal policy optimization (PPO) is adopted to learn motion decision-making strategies in complex shielding environments, so as to reduce local stagnation caused by greedy approach. Experimental results show that under conditions of a single source with low signal-to-noise ratio and multiple obstacles, the success rate of this method remains above 97.3%, the localization error is less than 0.5

m, and it demonstrates good obstacle-escape ability in L-shaped non-convex obstacle scenarios.

Keywords: radioactive-source search; particle filtering; angle of arrival (AOA); reinforcement learning; PPO algorithm

Chinese Library Classification No.: TL81 **Document identification code:** A **DOI:**

0 Introduction

With the wide application of nuclear technology in medical diagnosis, industrial flaw detection, irradiation processing, environmental monitoring, and nuclear emergency response, the rapid search and localization of out-of-control radioactive sources has become an important issue in nuclear-safety supervision. Out-of-control radioactive sources are highly concealed, highly hazardous, and subject to uncertain on-site environments; if they cannot be discovered and handled in time, they may threaten human health and public safety[1]. Traditional manual search using portable detectors suffers from high exposure risk, low search efficiency, and poor adaptability to complex environments. Using mobile robots equipped with radiation detectors to conduct autonomous source seeking has become an important technical direction for improving the efficiency of nuclear-emergency response[2].

Radioactive-source search usually includes two stages: source-parameter estimation and motion decision-making. Early localization methods mostly relied on fixed sensor networks or multipoint measurement data, calculating the source location through geometric methods, least squares, and maximum-likelihood estimation[3-5], but they are difficult to adapt to continuous source-seeking tasks by mobile robots. For radiation observation noise and source-location uncertainty, Bayesian estimation and particle filtering have been widely used for source-parameter estimation. Wang Mingsheng et al. used adaptive M-H sampling to establish a Bayesian inference model and achieved indoor point-source localization[6]; particle filtering is suitable for sequential estimation under nonlinear and non-Gaussian conditions[7]. Lazna et al. combined active planning to handle localization of multiple sources with unknown quantities[8]; Sheng Shizun et al. introduced angle-of-arrival (AOA) information into particle filtering, improving radioactive-source search efficiency through dynamic contraction of the search area and verifying applicability to multiple sources[9]; Yan Dong et al. combined Voronoi partitioning and used multiple robots to conduct parallel radioactive-source search, improving efficiency in large-area and multi-source scenarios[10].

In terms of motion decision-making, information-driven methods such as Infotaxis and Entrotaxis select the next observation location by maximizing information gain or reducing uncertainty[11-12]. Huo et al. used a convolutional neural network (CNN) to estimate the source location and combined it with the A* algorithm for path planning, obtaining relatively stable source-seeking

performance in concrete-wall shielding scenarios[13]. In recent years, deep reinforcement learning (Deep Reinforcement Learning, DRL) has gradually been applied to autonomous radioactive-source search tasks. Liu et al. proposed a Double Q-Learning radioactive-source search method based on CNN, which significantly shortened the search time compared with uniform search and gradient search[14]; Zhao et al. constructed a multi-objective reward function to achieve balanced optimization of the radioactive-source search path in terms of efficiency, smoothness, and energy consumption[15]; Wu Sunci improved strategic decision-making ability by integrating Double DQN and Dueling DQN, realizing autonomous radioactive-source search in complex environments[16].

Existing studies have provided a variety of estimation and decision-making methods for autonomous radioactive-source search, but challenges remain in complex shielding environments. On the one hand, some deep reinforcement learning methods emphasize end-to-end motion decision modeling, often directly inputting raw observation data into the network and ignoring the explicit guidance of the physical laws of the radiation field for search decision-making, which leads to a lack of clear logical association between physical estimation results and policy inputs. On the other hand, existing hierarchical strategies or end-to-end decisions in

---

Date received: 2026-06-04; **Date revised:** 2026-08-04
Funding project: National Natural Science Foundation of China funded project (12175187)
Author profile: Zhao Yaolong (2000-), male, from Dingxi, Gansu, master' s student, engaged in research on special robot technology; E-mail: 1046769574@qq.com
† **Corresponding author:** Xiao Yufeng, E-mail: xiaoyf_swit1@163.com

When facing trapped regions containing non-convex obstacles, the system often falls into stagnation because it lacks temporal memory capability. To this end, this paper proposes a radioactive-source search method based on an AOA particle filter and a PPO-GRU decision network. This method uses AOA geometric prior constraints for particle initialization, improving the utilization rate of the initial particles; it extracts source-position offset, uncertainty, and confidence as belief features, thereby realizing an explicit mapping from particle posterior to policy input. Meanwhile, a PPO-GRU network with temporal memory capability is introduced, using belief features and historical observation information to assist the policy network in avoiding local traps and making obstacle-avoidance decisions, thus achieving closed-loop coordination between perception estimation and motion decision-making.

Fig. 1 (online color) Framework of the AOA-PF-PPO autonomous radioactive-source search system

Figure 1: Fig. 1 (online color) Framework of the AOA-PF-PPO autonomous radioactive-source search system

1 Overall System Architecture

The AOA-PF-PPO radioactive-source search system proposed in this paper adopts a closed-loop architecture with decoupled perception and decision-making, as shown in Fig. 1.

Fig. 1 (online color) Framework of the AOA-PF-PPO autonomous radioactive-source search system

The system inputs include the robot pose, local laser-radar ranging, radiation counts, and the AOA direction prior obtained from the initial scan. First, the AOA-PF perception module uses the angle-of-arrival prior and radiation count observations to recursively estimate the radioactive-source position, and outputs belief features b_t composed of source-position offset, uncertainty, and confidence.

Subsequently, the belief features output by AOA-PF are combined with the robot pose, current radiation observation, AOA prior features, and local obstacle information into an 18-dimensional state vector S_t, which is input to the PPO-GRU decision network. The GRU is used to extract temporal features from historical observations; its output hidden state h_t encodes historical observation information and, together with the current belief state, serves as the input to the policy network, alleviating state aliasing caused by insufficient single-step observations in occluded environments. PPO outputs the action probabilities of eight discrete motion directions according to the current belief state, and the robot samples an action and executes movement.

During system operation, the robot first completes one horizontal scan using the collimated detector, obtains the direction with the maximum count as the AOA direction prior, and initializes the particle distribution accordingly. After entering the motion-search stage, the detector obtains radiation observations in an omnidirectional counting manner; the AOA-PF module recursively updates the posterior source position based on continuous radiation counts; and the PPO-GRU decision network outputs movement actions according to the current belief state and local obstacle information. After the robot executes the action, new radiation observations and environmental feedback are obtained, and the estimation and decision process is entered again, thereby forming a closed-loop process of "observation–estimation–decision–re-observation."

2 Method Design

2.1 Radioactive-Source Directional Detection Model

This paper uses ^{137}Cs as the experimental radioactive source, with an activity of 3.7×10^7 Bq. In actual radioactive-source detection, the radiation count is not only related to the source activity, but is also affected by factors such as detector efficiency, effective area, collimator response, and electronics threshold [17]. The measured radiation count rate cps at a specific distance R from a radioactive source with activity A follows the classical point-source geometric efficiency conversion equation:

$$\mathrm{cps} = A \cdot P_\gamma \cdot \varepsilon_{\mathrm{int}} \cdot \frac{S_{\mathrm{det}}}{4\pi R^2} + n_b \qquad (1)$$

where P_γ is the branching ratio of the main γ ray of ^{137}Cs, S_{det} is the cylindrical light-receiving cross-sectional area of the probe, $\varepsilon_{\mathrm{int}}$ is the intrinsic detection efficiency of the detector in this energy region, and n_b is the background count rate. Considering that parameters such as the effective detection area and detection efficiency of the detector are difficult to calculate, the calibrated count rate at 1 m from the radioactive source is often used.

as the effective source-strength parameter, denoted I_s, with the unit cps@1m. In this paper, a NaI(Tl) scintillation detector was used for actual measurement calibration under unobstructed conditions; its calibrated count rate at 1 m was 15000 cps.

In the radiation-transport model, the radioactive source is regarded as an isotropic point source. Air attenuation and scattering contributions are ignored, and only the shielding attenuation of obstacles on γ rays is considered [17 – 18]. Let the radioactive-source parameter be $X = [x_s, y_s, I_s]^T$. The expected count of the detector at the i-th observation point with a single measurement time τ is:

$$\lambda_i(X) = \left(\frac{I_s e^{-\sum_{m=1}^{M} \mu_m \Delta_{im}}}{R_i^2} + n_b \right) \tau \qquad (2)$$

where R_i is the Euclidean distance between the i-th observation point and the radioactive source, in m; μ_m is the linear attenuation coefficient of the m-th obstacle, in m^{-1}; Δ_{im} is the equivalent thickness through which the i-th radiation-propagation path passes in the m-th obstacle, in m; n_b is the background count rate, in cps; and τ is the detector single-measurement time, in s.

Because the emission and detection processes of γ rays exhibit statistical fluctuation characteristics, the actual count z_i obtained within the measurement time τ can be approximated as following a Poisson distribution [17 – 18]. Its observation likelihood function is:

Fig. 2 schematic of the collimated detector structure and detection modes

Figure 2: Fig. 2 schematic of the collimated detector structure and detection modes

$$p(z_i \mid X) = P(z_i; \lambda_i(X)) = \frac{\lambda_i(X)^{z_i}}{z_i!} e^{-\lambda_i(X)} \qquad (3)$$

where z_i is the actual count obtained at the i-th observation point within the measurement time τ.

To obtain prior information on the direction of the radioactive source, this paper uses a NaI(Tl) scintillation detector combined with a lead collimator for directional detection. The NaI(Tl) detector is used to obtain ray counts, while the collimator enhances directional selectivity by suppressing rays incident from outside the aperture. To clearly characterize the acquisition principle of the AOA directional prior during the algorithm-verification stage, the angular response function is simplified in this paper as an idealized binary model:

$$G(\phi) = \begin{cases} 1, |\phi| \le \alpha \\ \eta, |\phi| > \alpha \end{cases} \qquad (4)$$

where ϕ is the included angle between the ray incidence direction and the central axis of the collimator; α is the half-angle of the collimator aperture; and $\eta \in [0, 1)$ is the equivalent transmittance of the lead shielding body. For an actual lead collimator, the transmittance of γ rays at different incidence angles changes continuously rather than abruptly in a stepwise manner; at the same time, finite-thickness penetration and Compton-scattering effects exist. Equation (4) is an idealized approximation for the algorithm-verification stage, and the above factors are ignored.

The wall thickness of the lead collimator is 0.7 cm. For the 662 keV γ rays of ^{137}Cs, the transmittance is approximately 50% (corresponding to the half-value-layer thickness). At this thickness, there is a clear count difference between the inside and outside of the collimator aperture, which is sufficient to obtain the AOA directional prior through 360° scanning. The collimated detector is shown in Fig. 2.

(a) Directional detection (b) omnidirectional detection

Fig. 2 (online color) Schematic diagram of the collimated detector structure and detection modes

In the initial scanning stage of source-search detection, as shown in Fig. 3.

Figure 3 Initial scanning for source-seeking by (online color figure)

Figure 3. Initial scanning for source-seeking by (online color figure)

Figure 3: Figure 3. Initial scanning for source-seeking by (online color figure)

The collimated detector is controlled to rotate once in the horizontal plane, while synchronously recording the azimuth angle and radiation count rate, forming the scanning dataset $\{(\theta_k, D_k)\}_{k=1}^{K}$. The direction corresponding to the maximum count rate is taken as the current arrival-angle estimate:

$$k^* = \arg\max_{1\le k\le K} D_k, \quad \theta^* = \theta_{k^*}, \quad D^{\max} = D_{k^*} \tag{5}$$

where θ^*, as the AOA-direction prior, is input into the subsequent particle-filtering module to constrain particle initialization and improve the initial convergence efficiency [9]; $D^{\max}$ serves as the radiation-intensity feature of the current directional scan. This scanning process is used to obtain the initial direction prior for searching. After entering the motion-search stage, the detection model is switched to the omnidirectional counting mode; the collimated directional response is no longer introduced, and only radiation-count observations are continuously acquired.

2.2 Design of the Arrival-Angle Particle-Filtering Algorithm

To address the problems of low particle utilization and slow initial convergence caused by globally uniform initialization in traditional particle filtering, this paper introduces the arrival-angle (AOA) prior obtained by collimated scanning into the particle-filtering process. The particle state is defined as:

$$P_j(t) = \left[x_j(t), y_j(t), I_j(t)\right]^{\mathrm{T}}, \quad j = 1, 2, \cdots, N \tag{6}$$

where $\left(x_j(t), y_j(t)\right)$ is the radiation-source position corresponding to the j-th particle, $I_j(t)$ is the effective source-strength parameter corresponding to that particle, and N is the number of particles.

1) **AOA-constrained initialization**

The particle source strength $I_j(0)$ is uniformly sampled within the preset prior range $[I_{\min}, I_{\max}]$. In this paper, $I_{\min} = 500$ cps@1m and $I_{\max} = 20000$ cps@1m.

At the initial moment, with the detector position $P_{\mathrm{det}}(0) = (x_{\mathrm{det}}, y_{\mathrm{det}})$ as the vertex and the θ_0^* obtained by collimated scanning as the central direction, a sector-shaped prior region with opening angle β is constructed. In this paper, $\beta = 30^\circ$. Particles are then initialized within this region. During sampling, invalid particles that cross the boundary or lie inside obstacles are removed. The initial weight is set to $w_j(0) = 1/N$. This approach can restrict the initial

particle distribution to the vicinity of the possible direction of the radiation source and improve particle utilization in the initial stage.

2) **Particle prediction and weight updating**

During filtering recursion, the prediction step adopts an additive random-walk model, injecting bounded process noise into the particle set to maintain diversity:

$$x_k^{(j)} = x_{k-1}^{(j)} + \varepsilon_x, \quad y_k^{(j)} = y_{k-1}^{(j)} + \varepsilon_y, \quad I_k^{(j)} = I_{k-1}^{(j)} + \varepsilon_I \tag{7}$$

where $\varepsilon_x, \varepsilon_y \sim \mathcal{N}(0, \sigma_{\mathrm{pos},k}^2)$, and $\varepsilon_I \sim \mathcal{N}(0, \sigma_{I,k}^2)$. The process noise adopts an iterative attenuation strategy:

$$\sigma_{\mathrm{pos},k} = \max\left(\sigma_{\mathrm{pos,min}}, \frac{\sigma_{\mathrm{pos},0}}{\sqrt{k+1}}\right) \tag{8}$$

$$\sigma_{I,k} = \max\left(\sigma_{I,\mathrm{min}}, \frac{\sigma_{I,0}}{\sqrt{k+1}}\right) \tag{9}$$

The initial value of the position noise is $\sigma_{\mathrm{pos},0} = 80$ cm, referring to the single-step movement distance of the detector (1 m), ensuring that the exploration radius does not exceed the distance scale between adjacent observation points. The lower bound $\sigma_{\mathrm{pos,min}} = 5$ cm is far smaller than the localization-accuracy target (0.5 m), thereby ensuring steady-state convergence accuracy. The initial value of the source-strength noise is $\sigma_{I,0} = 0.05 \cdot (I_{\max} - I_{\min})$, so that particles can sufficiently cover the uncertainty in source strength during the initial exploration range. The lower bound $\sigma_{I,\mathrm{min}} = 200$ cps@1m ensures that excessive jitter does not occur after convergence. This group of parameters achieves a good balance between convergence speed and steady-state accuracy.

This attenuation mechanism is equivalent to introducing a forgetting factor into Bayesian recursive estimation: current observations gradually dominate during weight updating, while the influence of old observations is diluted as the number of iterations increases, thereby adapting to position inference for a static source.

After prediction, particles that cross the boundary or fall inside obstacles are first reflected and then truncated, so that the particle set always remains within the feasible region. Weight updating is then performed:

$$\tilde{w}_j(t) = w_j(t-1)\frac{\left[\lambda_t^{(j)}\right]^{z_t}}{z_t!}\exp\left[-\lambda_t^{(j)}\right] \tag{10}$$

where $\tilde{w}_j(t)$ is the unnormalized weight, $w_j(t-1)$ is the particle weight at the previous time, z_t is the actual observed count, and $\lambda_t^{(j)}$ is the theoretical expected

count. After weight updating, the particle weights are normalized, and the effective particle number is used to evaluate the degree of particle degeneracy:

$$N_{\text{eff}} = \frac{1}{\sum_{j=1}^{N} \left(w_j\right)^2} \qquad (11)$$

where w_j is the weight of the j-th particle after normalization. When $N_{\text{eff}} < 2N/3$, the particle set is considered to have undergone degeneracy, and resampling is performed to alleviate particle degeneracy and maintain the diversity of the particle population.

3) Source-position estimation and belief-feature output

After weight updating is completed, the weighted particle set is used to compute the radiation-source position estimate:

$$\mu_t = \sum_{j=1}^{N} w_j(t) p_j(t), \qquad p_j(t) = \left[x_j(t), y_j(t)\right]^{\mathrm{T}} \qquad (12)$$

where μ_t is the radiation-source position estimate, and $p_j(t)$ is the position component of the j-th particle.

To map the high-dimensional particle distribution $P(x_t \mid z_{1:t})$ into an input that can be processed by the decision network, this paper extracts a four-dimensional posterior belief feature vector b_t from the particle set $\{P_j(t), w_j(t)\}_{j=1}^{N}$, defined as follows:

$$b_t = [\Delta x_t, \Delta y_t, u_t, c_t]^{\mathrm{T}} \qquad (13)$$

where the position-offset feature $[\Delta x_t, \Delta y_t]$ is defined as the normalized deviation between the posterior expectation (weighted center) μ_t of the radiation-source position and the robot' s current pose $P_{\text{det}}(t)$:

$$\Delta x_t = \frac{\mu_{t,x} - x_{\text{det}}(t)}{L_{\text{diag}}}$$

$$\Delta y_t = \frac{\mu_{t,y} - y_{\text{det}}(t)}{L_{\text{diag}}} \qquad (14)$$

where L_{diag} is the diagonal length of the search region, used for magnitude normalization.

The particle-population uncertainty u_t is defined as the maximum dispersion of the particle distribution relative to the weighted mean center μ_t:

$$u_t = \frac{\max_j \left\| p_j(t) - \mu_t \right\|_2}{L_{\text{diag}}} \qquad (15)$$

The estimated confidence c_t measures the reliability of the localization result:

$$c_t = \frac{1}{1 + u_t} \qquad (16)$$

This value lies in the interval $[0, 1]$; when the particles are highly clustered, c_t tends to 1.

Therefore, AOA-PF recursively infers the historical radiation observations into the posterior of the source position, and realizes the mapping from particle posterior to policy input through the belief feature b_t.

2.3 PPO-GRU Decision-Network Design

The position of a radiation source is difficult to determine directly from a single radiation observation. In complex occlusion environments, problems such as count-rate fluctuation, local occlusion, and state aliasing still exist.

Therefore, this paper uses the source-position belief output by AOA-PF as part of the input to the policy network, and introduces a GRU temporal memory unit to extract historical observation information, enabling the robot to make obstacle-avoidance and source-seeking decisions under partially observable conditions.

To reduce the influence of absolute coordinate scale on policy learning, normalized pose and local perception features are used to construct the state space. The state vector consists of the current radiation observation, robot pose, AOA prior features, particle-filter belief features, and local-obstacle information:

$$S_t = \left[\tilde{r}_t, S_{\text{pos},t}^{\mathrm{T}}, S_{\text{scan},t}^{\mathrm{T}}, b_t^{\mathrm{T}}, S_{\text{obs},t}^{\mathrm{T}} \right]^{\mathrm{T}} \qquad (17)$$

where $\tilde{r}_t$ is the normalized current radiation count rate; $S_{\text{pos},t} \in \mathbb{R}^2$ is the normalized robot position; $S_{\text{scan},t} \in \mathbb{R}^3$ is the AOA prior feature; $b_t \in \mathbb{R}^4$ is the belief feature output by the particle filter, including the normalized offset of the estimated source position relative to the robot, the particle-cloud uncertainty, and the confidence; and $S_{\text{obs},t} \in \mathbb{R}^8$ is the local obstacle distance measured by lidar in eight discrete directions. Thus, the total dimension of the state is 18. This state representation does not directly input a complete environment map, but instead uses local obstacle distances, AOA prior features, and AOA-PF belief features for online decision-making.

The AOA prior feature is defined as:

$$S_{\text{scan},t} = \left[\log\left(1 + D^{\max}\right), \sin\left(\theta^*\right), \cos\left(\theta^*\right)\right]^{\mathrm{T}} \qquad (18)$$

where $D^{\max}$ is the maximum radiation count rate obtained during the initialization scanning stage, and θ^* is the corresponding estimated angle of arrival. Using $\sin(\theta^*)$ and $\cos(\theta^*)$ to represent the angle can avoid the discontinuity problem near angles 0 and 2π.

The action space adopts an 8-neighborhood discrete-action modeling:

$$A = \left\{ a_i \mid a_i = i \times \frac{\pi}{4},\ i = 0, 1, \ldots, 7 \right\} \qquad (19)$$

The robot selects an action $a_t \in A$ according to sampling, and controls the mobile chassis to move forward with a fixed step length of 1 m.

The reward function consists of a success reward, collision penalty, source-seeking progress reward, and step penalty:

$$R_{\text{total}} = \omega_1 R_{\text{succ}} + \omega_2 R_{\text{obs}} + \omega_3 R_{\text{prog}} + \omega_4 R_{\text{step}} \qquad (20)$$

where R_{succ} is the success reward: when the robot enters the effective decision radius of the radiation source, a positive reward is given and the episode is terminated; R_{obs} is the collision penalty: when the action executed by the robot causes a collision with an obstacle, a negative reward is given; R_{step} is the step penalty, used to reduce unnecessary movement steps; and $\omega_1 \sim \omega_4$ are the reward weights, determined based on the principle of magnitude alignment. Specifically, $\omega_3 = 0.2$, so the cumulative progress reward for finding the source within the maximum 120 steps in a single episode is $120 \times 0.2 = 24$, and after subtracting the step penalty the net gain is zero, preventing the policy from wandering inefficiently near the source to obtain redundant rewards. $\omega_1 = 20$ is equivalent to the reward of continuously approaching the source for 100 steps, forming the dominant advantage. $\omega_2 = -5.0$ is equivalent to the reward of continuously approaching the source for 25 steps ($25 \times 0.2 = 5.0$); a single collision cancels out a large amount of positive accumulation, enforcing safe obstacle avoidance. $\omega_4 = -0.2$ is symmetric in magnitude with ω_3, encouraging policy efficiency while avoiding an excessively strong penalty that would cause the policy to terminate the episode early when facing non-convex obstacles.

To alleviate the sparsity of rewards in the initial source-search stage, a source-seeking progress reward based on distance change is introduced:

$$R_{\text{prog}} = \kappa \left(d_{t-1} - d_t\right) \qquad (21)$$

where d_t is the path distance between the robot position and the radiation-source position at time t, and κ is the guidance weight. When the robot approaches

the radiation source along a traversable region, $R_{\text{prog}} > 0$; otherwise, negative feedback is given. This progress reward is used only for policy optimization in the simulation-training stage. During test execution, the policy input does not include the true radiation-source position.

In terms of network structure, after the current state S_t is embedded by features, it is input into the GRU, where the hidden state h_t integrates historical observation information; subsequently, the policy network Actor outputs the action distribution $\pi_\theta(a_t \mid s_t)$, and the value network Critic outputs the state value $V_\theta(s_t)$, which is used for subsequent policy updates. PPO improves training stability by restricting the update magnitude between the old and new policies. The probability ratio between the old and new policies is defined as:

$$r_t(\theta) = \frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \qquad (22)$$

The clipped surrogate objective of PPO is:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right] \qquad (23)$$

where $\hat{A}_t$ is the advantage function calculated based on GAE, and ε is the clipping coefficient. During training, an entropy reward is added to enhance exploration ability:

$$L(\theta) = L^{\text{CLIP}}(\theta) + \alpha H\left(\pi_\theta\right) \qquad (24)$$

where $H(\pi_\theta)$ is the entropy of the policy distribution, and α is the entropy reward coefficient.

3 Experimental Results and Analysis

3.1 Simulation Results and Performance Comparison

The training and simulation parameter settings are shown in Table 1.

Table 1 Training and simulation parameters

Parameter	Value	Parameter	Value
GRU hidden dimension	32	Scene size	20 m × 20 m
PPO clipping coefficient	0.10	Number of obstacles	1–7

Parameter	Value	Parameter	Value
Entropy reward coefficient	0.01	Obstacle size	$1.5\,\mathrm{m} \times 5\,\mathrm{m}$
Discount factor	0.99	Radiation source intensity	15000–20000 cps@1m
GAE parameter	0.90	Background count rate	50–500 cps
Actor learning rate	2.0×10^{-4}	Obstacle attenuation coefficient	$0.3\,\mathrm{m}^{-1}$
Critic learning rate	1.0×10^{-3}	Single-step displacement	1 m
Movement speed	1 m/s	Success determination radius	1 m
Reward weight ω_1	20	Reward weight ω_2	−5.0
Reward weight ω_3	0.2	Reward weight ω_4	−0.2
Number of particles	2000	Fan-shaped opening angle	30°
Number of training episodes	3000 episodes	Maximum steps per episode	120 steps

In Table 1, the obstacle attenuation coefficient is set to $\mu = 0.3\,\mathrm{m}^{-1}$, which is lower than typical values for real materials such as concrete for 662 keV γ-rays (approximately $10 \sim 14\,\mathrm{m}^{-1}$). This value is not a physically measured parameter, but an engineering attenuation setting used in simulation training. If a true high attenuation value were adopted, the low-activity source signal would rapidly be submerged in the background after penetrating obstacles, and the robot would lose radiation-gradient guidance in the occluded region, causing reinforcement learning to fail to converge within a limited number of steps. If the obstacle were made thinner to maintain the signal, the geometric realism required for LiDAR obstacle avoidance would be compromised. Therefore, while retaining the actual thickness of the obstacle, this paper appropriately reduces the attenuation coefficient so that the radiation field retains a weak but perceptible penetration gradient, thereby balancing policy learning and obstacle realism. This coefficient is a simulation setting and must be calibrated separately in actual deployment.

The measured calibration value is $I_s = 15000$ cps@1m. During training, I_s is randomly sampled within the range of $15000 \sim 20000$ cps@1m, so as to simulate

Figure 4 training return curve

Figure 4: Figure 4 training return curve

the fluctuation in radiation detection and avoid overfitting the policy to a single calibration value.

In Table 1, the Critic learning rate (1.0×10^{-3}) is set to five times the Actor learning rate (2.0×10^{-4}), mainly based on the characteristic that the value function in a radiation environment has a large dynamic range: radiation observations span several orders of magnitude from the background region to the near-source region, so the value network must converge faster to provide stable advantage estimates. This setting was determined through experimental tuning and can balance the convergence speed and stability of the Critic.

In addition, the GRU hidden dimension is set to 32, consistent with the hidden dimensions of the policy network and the value network, so as to maintain structural symmetry and control the number of parameters. This dimension provides an encoding space of approximately 3:1 for the 10-dimensional core temporal information (radiation count, AOA prior, belief features, and pose), sufficient to capture the task' s ···

radiation-change trends and key temporal patterns such as obstacle-avoidance decisions.

In constructing the simulation environment, the environmental boundaries and obstacle contours are used to build the simulation map, perform collision detection, remove invalid particles, and compute attenuation due to shielding. The experimental hardware platform is an Intel Core i7-10700 CPU (2.90 GHz). Model inference and particle-filter updating are both run on the CPU. Under 10-core parallel training, 3000 rounds take about 74.2 h in total, completing 1.44×10^7 environment interactions. In the evaluation mode, after removing gradient computation, single-step inference takes about 5–10 ms (including recursive updating of 2000 particles and 32-dimensional GRU inference), and the total inference time for one complete search is about 0.1–0.3 s, far less than the robot' s physical movement time (step length 1 m, speed 1 m/s). Thus, on the current CPU hardware, it does not constitute a real-time bottleneck. The return curve after training is shown in Figure 4.

Figure 4 (color online) Training return curve

As can be seen from Figure 4, in the initial stage of training, because the reward is sparse and obstacle penalties affect learning, the average episodic return is relatively low; after about 500 rounds, the return rises rapidly and gradually stabilizes, indicating that the constructed state representation and reward function can effectively guide the policy network to learn obstacle-avoidance source-seeking behavior.

To further evaluate the performance of the algorithm under different shielding

Figure 5 comparison of source-seeking success rates of three algorithms under different numbers of obstacles

Figure 5: Figure 5 comparison of source-seeking success rates of three algorithms under different numbers of obstacles

complexities, scenes with 1, 3, 5, and 7 obstacles are selected for testing. The initial position of the robot is at least 10 m away from the radiation source, and the remaining environmental parameters are kept consistent with the training stage. In this paper, the ratio of the radiation count rate at the robot' s initial position to the background count rate is set to 1–1.5 as a low signal-to-noise condition. For each number of obstacles, 100 test scenes are randomly generated, and 10 Monte Carlo repeated experiments are conducted in each scene; that is, each group of statistical results contains 1000 test samples.

The comparison algorithms are set as gradient search (GS) and AOA-PF-Greedy. GS selects the moving direction according to the local count-rate gradient and is a conventional baseline method commonly used in autonomous radiation-source search. AOA-PF-Greedy uses the same AOA-PF sensing module as the method in this paper, but does not use the PPO-GRU decision network; instead, at each step it selects the feasible action that brings the robot closer to the current estimated source position.

This paper does not introduce other DRL methods as comparison baselines, mainly because of the essential differences between their methodological frameworks and task assumptions: such methods mostly adopt end-to-end mapping and lack an explicit state-estimation mechanism; in contrast, this paper uses AOA-PF to filter high-noise observations into belief features, so the information-processing pipelines of the two are not equivalent. In addition, some methods rely on global maps for path planning, which is inconsistent with the information assumption of an unknown environment (only local LiDAR) in this paper, making it difficult to achieve a fair quantitative comparison.

The success rates of the three algorithms under different numbers of obstacles are shown in Figure 5.

Figure 5 (color online) Comparison of the source-seeking success rates of three algorithms under different numbers of obstacles

As can be seen from Figure 5, as the number of obstacles increases, the source-seeking success rate of GS decreases significantly, indicating that the stability of the local count-rate-gradient strategy in shielded environments is

insufficient; although AOA-PF-Greedy can use source-position estimation to improve search directionality, in complex occlusion regions it is still prone to falling into search stagnation, and its success rate likewise decreases markedly as the number of obstacles increases. In contrast, AOA-PF-PPO maintains a relatively high success rate in multi-obstacle scenarios, indicating that the strat-

Figure 6. Comparison of the average source-seeking time of the three algorithms under different numbers of obstacles (online color figure)

Figure 6: Figure 6. Comparison of the average source-seeking time of the three algorithms under different numbers of obstacles (online color figure)

egy decision-making that integrates posterior beliefs, local obstacle information, and historical observations can improve the task completion rate in complex occlusion environments.

On the premise of successful source seeking, the average source-seeking times of the three algorithms under different numbers of obstacles are shown in Figure 6.

Figure 6 (online color figure) Comparison of the average source-seeking time of the three algorithms under different numbers of obstacles

As can be seen from Figure 6, the average source-seeking time of GS increases significantly as the number of obstacles increases, and is always higher than that of the other two methods; the average source-seeking times of AOA-PF-Greedy and AOA-PF-PPO are generally close, indicating that the AOA-PF posterior information can reduce blind search. Combined with Figure 5, it can be seen that the advantage of AOA-PF-PPO is mainly reflected in improving the task success rate while maintaining relatively high source-seeking efficiency.

To evaluate the source localization capability of the algorithms, this paper defines the Euclidean distance between the final estimated radioactive source position after successful source seeking and the true radioactive source position as the localization error:

$$e_{\mathrm{loc}} = \|\hat{s} - s\|_2 = \sqrt{(\hat{x}_s - x_s)^2 + (\hat{y}_s - y_s)^2} \qquad (25)$$

where $s = (x_s, y_s)$ is the true source position, and $\hat{s} = (\hat{x}_s, \hat{y}_s)$ is the final position estimated by the algorithm. Since GS makes motion decisions only according to the local count gradient, does not include a source-position estimation module, and cannot output a final localization result, error statistics are therefore conducted only for AOA-PF-Greedy and AOA-PF-PPO, which have source-estimation localization capability, as shown in Table 2.

Table 2 Localization estimation errors and statistical tests under different numbers of obstacles

Number of obstacles / pcs	AOA-PF-Greedy error mean ± standard deviation / m	AOA-PF-PPO error mean ± standard deviation / m	t value	p value	d value	Effect size
1	0.39 ± 0.21	0.30 ± 0.19	9.96	<0.001	0.45	small
3	0.37 ± 0.19	0.30 ± 0.18	8.43	<0.001	0.38	small
5	0.43 ± 0.26	0.34 ± 0.23	8.09	<0.001	0.37	small
7	0.39 ± 0.21	0.32 ± 0.20	7.41	<0.001	0.34	small

Table 2 gives the final localization errors and statistical test results for successful source-seeking trials. As the number of obstacles increases, the localization errors of the two methods do not show an obvious monotonically increasing trend, indicating that the number of obstacles mainly affects the source-seeking success rate and the average source-seeking time, while its influence on the final localization accuracy of successful trials is relatively limited. Welch' s t-test shows that, under all four obstacle-number conditions, the mean localization error of AOA-PF-PPO is statistically significantly lower than that of AOA-PF-Greedy ($p < 0.001$), and Cohen' s d effect sizes range from 0.34 to 0.45. Because the two methods use the same particle-filter estimation module, the error difference mainly comes from the observation sequences formed by the motion strategies; the effect sizes are small, which is as expected. AOA-PF-PPO can reduce ineffective wandering near occlusion boundaries and enable the particle filter to obtain more continuous effective observations, so the final localization error remains at a relatively low level.

To evaluate the limitations of the algorithm, failure cases in the 7-obstacle scenario were analyzed. In 1000 experiments, there were 27 failures in total, of which 21 were caused by the action being rejected by obstacles for 10 consecutive times and thus terminating early (stuck type), and 6 were due to timeout after 120 steps were exhausted (timeout type). Figure 7(a) presents a stuck failure: after the robot is blocked by obstacles, the particle filter has a bias in direction estimation under continuous radiation observations; the strategy repeatedly attempts the blocked direction and fails to, within the stuck limit,

turn into a passable bypass channel. Figure 7(b) gives a timeout-type failure: the robot has entered the near-source region, but an obstacle in front of the

Figure 7

Figure 7: Figure 7

Figure 8

Figure 8: Figure 8

source blocks the final approach path, and the bypass is not completed within 120 steps.

(a) stuck-type failure (b) timeout-type failure

Figure 7 (online color) Trajectories of two types of source-search failure cases

3.2 Trajectory Analysis in Obstructed Scenarios

To verify the local escape capability of the algorithm in a non-convex obstructed environment, an L-shaped non-convex obstacle scenario was constructed, and the source-search trajectories of GS, AOA-PF-Greedy, and AOA-PF-PPO were compared. In this test scenario, the effective source strength of the radioactive source was $I_s = 1.75 \times 10^4$ cps@1m, the background count rate was set to 300 cps, and the remaining parameters were kept consistent with Table 1. The results are shown in Figures 8, 9, and 10.

Figure 8 Source-search trajectories of the GS algorithm

Figure 9 (online color) Source-search trajectories of the AOA-PF-Greedy algorithm

Figure 10 (online color) Source-search trajectories of the AOA-PF-PPO algorithm

As can be seen from Figure 8, GS moves only according to the local radiation-count gradient. When the radioactive source is blocked by obstacles, the count variation is prone to attenuation by shielding and Poisson fluctuations.

is affected, causing the local-gradient direction to be inconsistent with the actual detour direction, and the robot hovers at the edge of the obstacle. As can be seen from Fig. 9, AOA-PF-Greedy can use the source-position estimation result to make the robot as a whole move toward the radiation source, but its action selection still takes the currently estimated source position as the direct target. In an L-shaped nonconvex obstacle scenario, the robot is easily guided by the estimated source direction into the concave region; when the estimated

Figure 9

Figure 9: Figure 9

Figure 10

Figure 10: Figure 10

Fig. 11 (online color) Search process in a dual-source scenario

Figure 11: Fig. 11 (online color) Search process in a dual-source scenario

source is located on the other side of the obstacle, the greedy strategy continuously drives the robot toward the impassable region, causing it to make repeated adjustments near the L-shaped concave region. By contrast, in Fig. 10, AOA-PF-PPO does not take the estimated position of AOA-PF as a fixed target point for greedy approach, but instead uses the source-position offset, uncertainty, and confidence as dynamic posterior features of the source position, and combines them with local obstacle information and historical observations for decision-making. When the estimated source direction is blocked, the robot can temporarily deviate from the source direction, detour around the local depression along the passable region, and then gradually approach the radiation source.

This result indicates that a strategic decision-making method combining posterior source-position estimation with historical observations can alleviate the greedy-stagnation problem in complex obstructed environments.

3.3 Extended validation of dual-source search

To verify the adaptability of the AOA-PF-PPO framework in multi-source scenarios, this paper further constructs a dual-radiation-source simulation scenario, retrains the policy network for the dual-source scenario, and adopts a sequential search strategy: the robot first searches for the dominant radiation source according to initialized AOA scanning and the AOA-PF posterior belief; after the first radiation source is localized, it is removed from the set to be searched and from the radiation-field contribution (this operation is a simulation assumption and does not mean that the actual system has the ability to distinguish particles from different radiation sources), and AOA scanning and particle-filter initialization are performed again to continue searching for the second radiation source. Figure 11 gives a test scenario after training is completed. In this scenario, the source intensities are $I_{s1} = 1.95 \times 10^4$ cps@1m, $I_{s2} = 1.73 \times 10^4$ cps@1m, and the background count rate is set to 80 cps.

Fig. 11 (online color) Search process in a dual-source scenario

In Fig. 11, the particle cloud after radiation source 1 is localized is used only to display the posterior estimation result and does not participate in the filtering update for radiation source 2; when searching for radiation source 2, AOA scanning and particle initialization are performed again. The results show that, while keeping the network structure unchanged and retraining it for the dual-

source task, the proposed framework can complete sequential dual-source search. Table 3 gives the overall statistical results for all 4000 test samples in dual-source scenarios with 1, 3, 5, and 7 random obstacles.

Table 3 Sequential search results in the dual-source scenario

Scenario type	Number of samples / cases	Success rate / %	Average total source-search time / s	Average localization error / m
Dual-source search	4000	89.5	32.2 ± 10.5	0.65 ± 0.23

In Table 3, the average localization error in the dual-source scenario is 0.65 ± 0.23 m, which is larger than that in the single-source scenario. The main reason is that the superposition of the radiation-field signals from the two sources causes the particle-filter likelihood constructed based on the single-source observation model to remain biased. There is an intrinsic difference between the superposed count rate of the two true sources and the theoretical expectation of the single-source model, which shifts the weighted center of the particle cloud toward the weighted radiation centroid of the two sources, making it difficult to converge as accurately to a single true source position as in the single-source scenario. This

The bias arises because the single-source observation model cannot fully characterize the multi-source superposed radiation field, which is an inherent limitation when applying a single-source likelihood function to a superposed field. Within a search area of 20 m $\times$ 20 m, this error is within a reasonable range relative to the physical size of the artificial radiation source itself and the spatial resolution of the detector, and the dual-source localization success rate reaches 89.5%, verifying that the proposed method still has good applicability potential in multi-source scenarios.

4 Conclusion

This paper proposes a radiation-source search framework that integrates AOA-prior particle filtering with PPO-GRU policy decision-making, explicitly incorporating source-position posterior beliefs into the reinforcement-learning state representation, thereby improving the stability of autonomous source search in complex occluded environments. Experimental results show that, under low signal-to-noise ratios and scenarios with 1–7 obstacles, the proposed method achieves a success rate of more than 97.3% in single-source search, with a single-source localization error below 0.5 m. Meanwhile, the robot can select a detour path by combining historical observations and local obstacle information, effectively alleviating the problem of local stagnation in complex occluded environments. Sequential dual-source search experiments further demonstrate that the

framework has certain potential for multi-source extension. This study provides a feasible technical solution for autonomous radiation-source search by mobile robots in nuclear emergency environments, and lays a foundation for subsequent validation on physical platforms and in irregular obstacle environments.

References

[1] IAEA. Nuclear Security Review 2023[R/OL]. Vienna: International Atomic Energy Agency, 2023. https://www.iaea.org/sites/default/files/gc/gc67-inf3.pdf.

[2] Hu H, Wang J, Chen A, et al. Nuclear Engineering and Technology, 2023, 55(1): 285-294. doi: 10.1016/j.net.2022.09.010

[3] Rao N S V, Sen S, Prins N J, et al. Nuclear Instruments and Methods in Physics Research Section A, 2015, 784: 326-331. doi: 10.1016/j.nima.2015.01.037

[4] Howse J W, Ticknor L O, Muske K R. Automatica, 2001, 37(11): 1727-1737. doi: 10.1016/S0005-1098(01)00134-0

[5] Bai E W, Yosief K, Dasgupta S, et al. The Maximum Likelihood Estimate for Radiation Source Localization: Initializing an Iterative Search[C]//Proceedings of the IEEE Conference on Decision and Control. Los Angeles: IEEE, 2014: 277-282.

[6] Wang Mingsheng, Xiao Yufeng, Liu Ran, et al. Measurement and Control Technology, 2019, 38(6): 44-48.

(Wang Mingsheng, Xiao Yufeng, Liu Ran, et al. Measurement and Control Technology, 2019, 38(6): 44-48. doi: 10.19708/j.ckjs.2019.00.005)

[7] Ristic B, Arulampalam S, Gordon N. Beyond the Kalman Filter: Particle Filters for Tracking Applications[M]. Boston: Artech House, 2004.

[8] Lazna T, Zalud L. Nuclear Engineering and Technology, 2025, 57(2): 103171. doi: 10.1016/j.net.2024.08.040

[9] Sheng Shizun, Xiao Yufeng, Yan Dong, et al. Nuclear Physics Review, 2025, 42(1): 111-122.

(Sheng Shizun, Xiao Yufeng, Yan Dong, et al. Nuclear Physics Review, 2025, 42(1): 111-122. doi: 10.11804/NuclPhysRev.42.2024028)

[10] Yan D, Xiao Y F, Sheng S Z, et al. Applied Radiation and Isotopes, 2024, 212: 111475. doi: 10.1016/j.apradiso.2024.111475

[11] Vergassola M, Villermaux E, Shraiman B I. Nature, 2007, 445(7126): 406-409. doi: 10.1038/nature05464

[12] Hutchinson M, Oh H, Chen W H. Information Fusion, 2018, 42: 179-189. doi: 10.1016/j.inffus.2017.10.009

[13] Huo J, Hu X, Wang J, et al. Nuclear Engineering and Technology, 2023, 55(8): 3030-3038. doi: 10.1016/j.net.2023.05.017

[14] Liu Z, Abbaszadeh S. Sensors, 2019, 19(4): 960. doi: 10.3390/s19040960

[15] Zhao Y. Deep Reinforcement Learning-Based Path Planning for Radioactive Source Search[C]//2025 IEEE 3rd International Conference on Sensors, Electronics and Computer Engineering (ICSECE). Chongqing: IEEE, 2025: 570-574.

[16] Wu Sunci. Research on Mobile Robot Autonomous Source Searching Algorithm Based on Deep Reinforcement Learning[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2022.

(Wu Sunci. Research on a Mobile Robot Source-Searching Algorithm Based on Deep Reinforcement Learning[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2022.)

[17] Knoll G F. Radiation Detection and Measurement[M]. 4th ed. New York: John Wiley & Sons, 2010.

[18] Morelande M R, Ristic B. IEEE Transactions on Signal Processing, 2009, 57(11): 4220-4231. doi: 10.1109/TSP.2009.2026618

Radioactive Source Search Method Based on Angle-of-Arrival Particle Filtering and Reinforcement Learning

ZHAO Yaolong^1, XIAO Yufeng^1,†, ZHENG Yulai^2, YAN Dong^1, REN Zhenyu^1, WANG Zhao^1, YANG Shuang^1

(1. School of Information and Control Engineering, Southwest University of Science and Technology, Mianyang 621010, Sichuan, China;
2. China Institute of Atomic Energy, Beijing 102413, China)

Abstract: To address the problems that autonomous radioactive source search in complex occluded environments is susceptible to radiation-count fluctuations and local optima, a radioactive source search method combining angle-of-arrival particle filtering and reinforcement learning is proposed. The proposed method uses an angle-of-arrival (AOA) prior to constrain particle initialization, recursively estimates the posterior distribution of the radioactive source position based on a modified radiation observation model, and takes source-position offset, uncertainty, and confidence as the belief features. Meanwhile, a gated recurrent unit (GRU) is introduced to extract historical observation information, and proximal policy optimization (PPO) is adopted to learn motion decision-making

strategies in complex occluded environments, thereby reducing local stagnation caused by greedy source approaching. Experimental results show that, under single-source, low-signal-to-noise-ratio, and multi-obstacle conditions, the proposed method achieves a success rate of more than 97.3% and a localization error of less than 0.5 m. In addition, the method demonstrates good obstacle-detouring and escape capability in an L-shaped non-convex obstacle scenario.

Key words: radioactive source search; particle filter; angle of arrival (AOA); reinforcement learning; PPO algorithm

Received date: 04 Jun. 2026; **Revised date:** 04 Aug. 2026
Foundation item: National Natural Science Foundation of China (12175187)
† **Corresponding author:** XIAO Yufeng, E-mail: xiaoyf_swit1@163.com

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