Source Term Estimation Using Curiosity-Driven Information-Guided Reinforcement Learning in Turbulent Environments

Junhee Lee[1], Hongro Jang[2], Seunghwan Kim[2], Hyoungho Park[2],
Hyungjin Kim[1], Changseung Kim[2] and Hyondong Oh[1]

***Abstract*— This paper introduces a deep reinforcement learning (DRL) framework for estimating and searching for an invisible gas source using a mobile sensor in turbulent environments. Source term estimation (STE), which aims to estimate key properties of the gas source, is challenging due to environmental uncertainty and sensor noise. Particularly, balancing exploration and exploitation in DRL-based STE is difficult, since the agent makes decisions based on both current and past noisy measurements under turbulence. However, most existing studies have made limited attempts to address these challenging problems and are rarely validated in turbulent or real-world experiments. To address these issues, we propose a curiosity-driven information-guided learning framework that accurately estimates and effectively searches for the gas source in turbulent environments. The proposed method enables efficient exploration by guiding the agent to actively search novel regions where informative source information is likely to exist. Furthermore, the active perception reward function is proposed to ensure the robust source search. Simulations under high turbulence and noise demonstrate that the proposed method outperforms the existing methods in terms of the success rate and the mean travel distance. Moreover, real-world experiments confirm the feasibility and robustness of the proposed framework, highlighting its potential for practical STE problems.**

I. INTRODUCTION

In recent years, numerous incidents of hazardous gas leakage have occurred, causing serious risks to human health. In order to minimize potential damage, it is critical to rapidly identify the exact properties of the gas source. Estimating the key properties (e.g., source location and release strength) is generally referred to as source term estimation (STE) problem. STE problem is inherently challenging, as most hazardous gas leakage is invisible and strongly influenced by turbulent atmospheric conditions. Since it is dangerous for humans to directly identify the gas source, autonomous STE strategies using mobile sensors such as UAVs and UGVs have recently attracted considerable attention [1].

One of the representative strategies for STE is the information-theoretic approach, which employs Bayesian inference to estimate the source term and leverages information theory for planning the search [2]–[5]. In this approach, the mobile sensor determines its next sensing point by selecting an action that reduces the uncertainty of the estimated source term the most. This approach is particularly effective in addressing noisy and sparse sensor measurements in turbulent environments. However, such methods require a substantial computational load to calculate the uncertainty for all action candidates. Therefore, it can impose a significant burden on the real-time processing capabilities of computing boards.

Utilizing deep reinforcement learning (DRL) approaches can provide a promising solution for addressing these issues in STE. DRL can learn an optimal search policy by leveraging neural networks, thereby reducing computational load and ensuring real-time applicability [6]. Furthermore, by empirically determining actions from past experiences and reward feedback, it can guide the agent toward potentially efficient search paths.

For these reasons, many studies have recently applied DRL to the STE problem. However, developing a robust source search policy in turbulent environments with sparse and highly variable sensor measurements still remains a significant challenge. Since STE is a partially observable Markov decision process (POMDP), the decision of the agent depends not only on the current but also on past sensor measurements. Consequently, effective exploration strategies are essential to acquire informative measurements, as gathering unreliable observations due to the turbulence can severely degrade source estimation and search performance. In this context, achieving a proper balance between exploration and exploitation is crucial for DRL-based STE [7].

Nevertheless, most DRL-based STE methods lack effective strategies to tackle these challenges. Many studies have primarily focused on rapidly searching for the source [8]–[11]. But without incorporating a source estimator, these methods are unable to declare when the source search is complete and to identify the source location accurately. Unlike these approaches, PC DQN [12] and MVG-RDDPG [13] utilize DRL to train the search policy while estimating the source by using the particle filter. Despite their effectiveness, these methods rely on random exploration, which can lead the agent into regions that are already explored or uninformative. Thus, it can result in inefficient search

*This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2023R1A2C2003130), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2020-NR049578), and Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea (NRF) and Unmanned Vehicle Advanced Research Center (UVARC) funded by the Ministry of Science and ICT, the Republic of Korea (2020M3C1C1A01082375).

[1]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea ljh0124@kaist.ac.kr, gudwls124z@kaist.ac.kr, h.oh@kaist.ac.kr

[2]Department of Mechanical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea hlomk@unist.ac.kr, kevin6960@unist.ac.kr, gudgh1630@unist.ac.kr, pon02124@unist.ac.kr

in turbulent environments. Furthermore, their methods are trained under ideal conditions with low turbulence, which limits robustness in real-world scenarios. To overcome these issues, Singh et al. [14] introduces a recurrent neural network agent that robustly tracks turbulent gas plumes. Yet, its performance degrades when sensor signals become sparse, as reliance on recovering to sensing regions leads to unstable trajectories. AID-RL [15] introduces information-directed RL that combines reward-driven exploitation with entropy-based exploration, but its ε-greedy action selection still biases the policy toward exploitation, increasing the risk of local minima. Lee et al. [16] improves the source search robustness in turbulent environments by designing the Gaussian mixture model (GMM)-based reward functions. However, it mainly focuses soley on reward designing without addressing effective exploration strategy.

Motivated by these issues, we propose a DRL-based STE framework, which enables robust source estimation and efficient search even under sparse and highly fluctuating gas sensor measurements in turbulent environments. Specifically, we leverage curiosity-driven exploration to actively guide the agent toward novel regions where informative source information is likely to exist. By balancing this with the active perception reward, the proposed method enables the agent to rapidly and robustly search for the source, while acquiring reliable gas measurements. To validate the effectiveness and robustness of the proposed framework, we perform various simulations under high turbulence and noisy conditions. We emphasize that, to the best of our knowledge, DRL-based STE studies that conducted real-world experiments in turbulent environments have been rarely reported. To this end, real-world experiments with CO_2 leakage scenarios are conducted to demonstrate the feasibility of the proposed method. In this paper, we term our algorithm the **c**uriosity-driven **i**nformation-**g**uided **s**oft **a**ctor-**c**ritic (**CIG-SAC**). The main contributions of the proposed method is represented as:

1) We introduce a curiosity-driven exploration framework for DRL-based STE problem, which enables more accurate source estimation and efficient source search in turbulent environments;
2) We present the active perception reward function that integrates mutual information with particle filter variance, enhancing the robustness of the source search; and
3) The proposed method demonstrates robust performance compared with existing STE methods through various turbulent simulations, and its feasibility and practical applicability are confirmed in real-world experiments.

II. Problem Statement

In this study, a hazardous gas source is located at $\mathbf{r}_s = [x_s, y_s]^T$ using an agent. The source term, which is the parameter vector to be estimated, is defined as $\theta_s = \left[\mathbf{r}_s^T, q_s\right]^T$. At each time step t, the agent located at $\mathbf{r}_t$ collects gas sensor measurements and estimates the source term using particle filter, based on the predefined gas dispersion model and sensor model. Then, CIG-SAC is applied as the action-selection strategy to optimize the search trajectory.

This section describes the gas dispersion model and the sensor model. Furthermore, it handles the source estimation method utilizing a particle filter.

A. Gas Dispersion Model

For the gas dispersion model, we use the isotropic plume model [2]. In this model, the gas particle propagates with the particle life time τ and the diffusivity D. In addition, the gas particle diffuses at a wind direction χ and an average wind speed V. The gas concentration $C(\mathbf{r}_t|\theta_s)$ acquired at the sensing position $\mathbf{r}_t$ at time step t is defined as:

$$C(\mathbf{r}_t|\theta_s) = \frac{aq_s}{|\mathbf{r}_t - \mathbf{r}_s|} \exp \frac{V(x_t - x_s)\sin\chi}{2D} \cdot \exp \frac{-V(y_t - y_s)\cos\chi}{2D} \cdot \exp \frac{-|\mathbf{r}_t - \mathbf{r}_s|}{\lambda}, \quad (1)$$

where

$$\lambda = \sqrt{\frac{D\tau}{1 + \frac{U^2\tau}{4D}}}. \quad (2)$$

B. Gaussian Sensor Model

Since the sensor measurements obtained by the agent include noise in real-world scenarios, we use the Gaussian noise model as the sensor model. The sensor measurement z_t at the sensing position $\mathbf{r}_t$ at time step t is expressed as:

$$z_t = C(\mathbf{r}_t|\theta_s) + \nu_{sensor} + \nu_{env}, \quad (3)$$

where ν_{sensor} and ν_{env} denote the noise arising from the sensor measuring process and the noise induced by the wind, respectively. Both terms are assumed to follow the white Gaussian noise, as formulated in the following equations:

$$\nu_{sensor} \sim \mathcal{N}(0, \sigma^2_{sensor}), \quad (4)$$

$$\nu_{env} \sim \mathcal{N}(0, \sigma^2_{env}). \quad (5)$$

The standard deviation of the sensor noise, σ_{sensor}, is defined as:

$$\sigma_{sensor} = \beta \cdot C(\mathbf{r}_t|\theta_s), \quad (6)$$

where β and σ_{env} are the level of sensor noise and the instability of the wind conditions, respectively. The probability distribution of the sensor measurement z_t at the sensing position $\mathbf{r}_t$ at time step t is defined as:

$$p(z_t|\theta_s) = \frac{1}{\sigma_T\sqrt{2\pi}} \exp - \frac{(z_t - C(\mathbf{r}_t|\theta_s))}{2\sigma_T^2}. \quad (7)$$

In this context, the overall standard deviation of the noise σ_T is calculated as:

$$\sigma_T = \sqrt{\sigma^2_{sensor} + \sigma^2_{env}}. \quad (8)$$

Fig. 1 illustrates examples of the gas dispersion model and the sensor measurement map with noise.

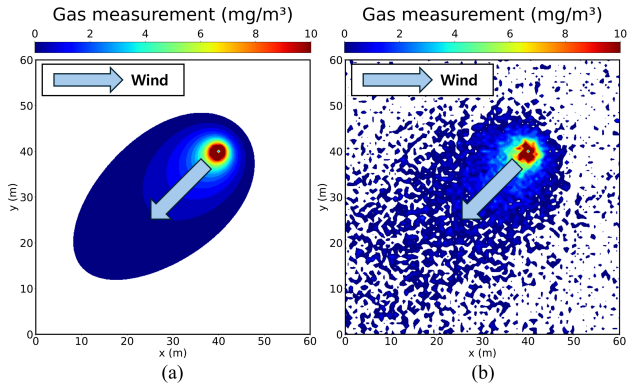


Fig. 1. (a) Gas dispersion model and (b) sensor measurement with noise.

C. Particle Filter

We utilize a particle filter to estimate the source term, as it is well-suited for handling the non-linear characteristics of the source term and remains robust under high levels of sensor measurement noise. The source term probability distribution is expressed by N_p particles as:

$$p(\theta_{t,s}|z_{1:t}) = \sum_{i=1}^{N_p} w_t^i \delta(\theta_{t,s} - \theta_{t,s}^i), \tag{9}$$

where $\delta(\cdot)$ represents the Dirac delta function, $\theta_{t,s}^i$ is each particle representing the source term, and w_t^i denotes the weight of each particle. When the new sensor measurement z_{t+1} is collected at time step $t+1$, the particle weights are calculated with the following equation:

$$\overline{w}_{t+1}^i = p(z_{t+1}|\theta_{t,s}^i) \cdot w_t^i. \tag{10}$$

where $\overline{w}_{t+1}^i$ indicates the unnormalized weight of each particle. The likelihood $p(z_{t+1}|\theta_{t,s}^i)$ is calculated by using the predefined gas dispersion model and sensor model from equations (1) and (7). Then, the normalized weight w_{t+1}^i is given as:

$$w_{t+1}^i = \frac{\overline{w}_{t+1}^i}{\sum_{i=1}^{N_p} \overline{w}_{t+1}^i}. \tag{11}$$

To mitigate the degeneracy problem, in which most particle weights converge toward zero, a resampling is applied. Resampling is performed when the effective number of samples falls below a predefined threshold δ. The effective number of samples N_{eff} is computed as:

$$N_{eff} = \frac{1}{\sum_{i=1}^{N_p} (w_t^i)^2}. \tag{12}$$

Furthermore, the Markov chain Monte Carlo (MCMC) method [17] is utilized after resampling to improve particle impoverishment.

III. Method

In this section, we introduce a curiosity-driven information-guided reinforcement learning framework to achieve robust source search in turbulent environments. The STE problem is formulated as the belief-based Markov decision process (belief-MDP), where the state is defined using particle filter information. State and action utilized in this study is first introduced, and the curiosity-driven exploration for STE is explained. Then, the proposed active perception reward is presented, and the overall learning framework is outlined.

A. State and Action

1) State: The state is defined as the particle filter information and the sensing position. To enhance training stability and the efficiency of source estimation, GMM clustering is applied to extract features from the particle filter [13]. The state in this study is defined as:

$$s_t = [\boldsymbol{M}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\Pi}_t, m_t, \mathbf{r}_t], \tag{13}$$

where $\boldsymbol{M}_t$ denotes the mean of each GMM cluster, $\boldsymbol{\Sigma}_t$ represents the covariance of each GMM cluster, and $\boldsymbol{\Pi}_t$ is the corresponding weight of each GMM cluster. Also, m_t represents the mean of all particles and $\mathbf{r}_t$ indicates the sensing position at the current step.

2) Action: In the STE problem, a continuous action space is more beneficial than a discrete action space, as it increases the possibility of the agent obtaining informative measurements. To enable this, this study adopts a continuous action a_t by applying a fixed-length movement with the heading direction at each step. The next sensing position is determined by the agent's selected action, as given below:

$$\mathbf{r}_{t+1} = \mathbf{r}_t + \begin{bmatrix} \cos(a_t) \\ \sin(a_t) \end{bmatrix} \cdot k, \tag{14}$$

where k is the fixed step size.

B. Curiosity-Driven Exploration for Source Term Estimation

Balancing exploration and exploitation is crucial in DRL-based STE, especially in turbulent environments. To deal with these issues, we adopt curiosity-driven exploration [18], which provides intrinsic motivation that encourages the agent to explore novel but learnable states. As shown in Fig. 2, we design a curiosity network composed of a feature extractor, an inverse network, and a forward network, parameterized by w_{feat}, w_{inv}, and w_{fwd}, repectively.

The feature extractor encodes the current and next states, s_t and s_{t+1}, into feature representations $\varphi(s_t)$ and $\varphi(s_{t+1})$. The inverse network predicts the action $\hat{a}_t$ that caused the transition between consecutive states given $\varphi(s_t)$ and $\varphi(s_{t+1})$, and is trained with the following loss function:

$$\mathcal{L}_{inv} = \text{MSE}(a_t, \hat{a}_t). \tag{15}$$

The forward network is designed to predict the next feature representation $\hat{\varphi}(s_{t+1})$ from $\varphi(s_t)$ and a_t. The prediction error between $\varphi(s_{t+1})$ and $\hat{\varphi}(s_{t+1})$ is used as the loss of the forward network:

$$\mathcal{L}_{fwd} = \|\varphi(s_{t+1}) - \hat{\varphi}(s_{t+1})\|_2^2. \tag{16}$$

For the STE problem, note that the state is defined as the belief information extracted by GMM clustering and the sensing position of the agent. As illustrated in Fig. 3,

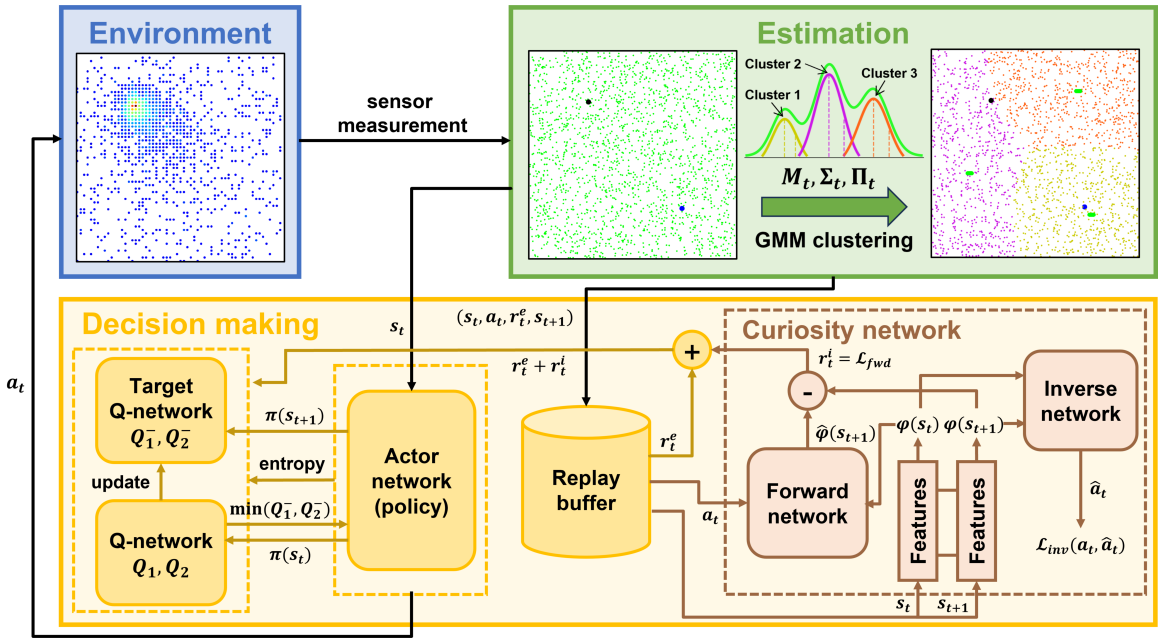


Fig. 2. System architecture of CIG-SAC for source term estimation.

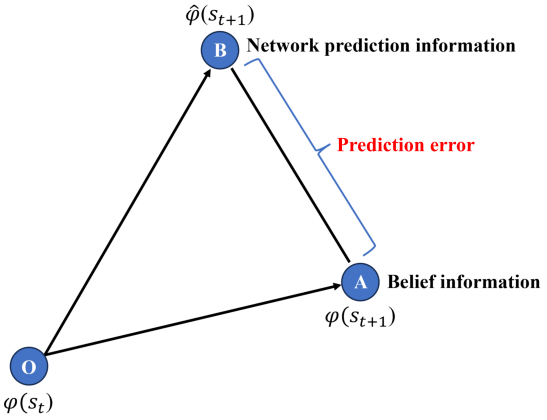


Fig. 3. Prediction error between the belief state and the predicted state in curiosity-driven exploration.

point A represents the belief information (i.e., source information), while point B denotes the predicted information by the forward network. In the early stages of training, a noticeable gap exists between point A and B, representing the prediction error (i.e., uncertainty) measured as $|\overline{AB}| = \|\varphi(s_{t+1}) - \hat{\varphi}(s_{t+1})\|_2^2$. Rewarding $|\overline{AB}|$ can induce the agent to visit novel regions, where previously unknown source information is expected to be obtained. Importantly, since $|\overline{AB}|$ corresponds to the loss of the forward network, it tends to decrease as the forward network is trained. Consequently, it can guide the agent to learn in a way that gradually reduces this uncertainty. Based on this insight, we adopt the forward network loss, representing the prediction error, as the intrinsic reward:

$$r_t^i = \|\varphi(s_{t+1}) - \hat{\varphi}(s_{t+1})\|_2^2 . \tag{17}$$

The overall optimization objective for the curiosity network is denoted as:

$$\min_{w_{feat}, w_{inv}, w_{fwd}} \mathcal{L}_c = (1-\varepsilon)\,\mathcal{L}_{inv} + \varepsilon\,\mathcal{L}_{fwd}, \tag{18}$$

where ε balances the two terms.

C. Active Perception Reward Function

In turbulent environments, where gas measurements are highly variable and sparse, the agent requires an effective search strategy, as obtaining unreliable measurements can easily mislead the search process. In this context, we introduce the active perception reward function, which enables robust source search.

The mutual information is employed as a reward to drive the policy toward reducing the uncertainty of the estimated source term distribution, motivated by a study in [16]. The mutual information is defined as:

$$\begin{aligned} I(\mathbf{r}_{t+1}) = &-\sum_{i=1}^{N_p} w_t^i \log w_t^i \\ &+ \sum_{\hat{z}_{t+1}=0}^{z_{\max}} p(\hat{z}_{t+1}|\theta_{t,s}) \left(\sum_{i=1}^{N_p} \hat{w}_{t+1}^i \log \hat{w}_{t+1}^i \right), \end{aligned} \tag{19}$$

where $\mathbf{r}_{t+1}$ is the sensing position at the next step, $\hat{z}_{t+1}$ denotes all possible future measurements at the next sensing position, and $\hat{w}_{t+1}^i$ represents the potential particle weight. Furthermore, to accelerate the source search, we introduce a distance term between the agent's position and the estimated source location, defined as:

$$d_t = -\left|\hat{\mathbf{r}}_{t,s} - \mathbf{r}_t\right|, \tag{20}$$

where $\hat{\mathbf{r}}_{t,s}$ denotes the estimated source location.

To enable an efficient balance between the mutual information term and the distance term, we introduce an automatic adjustment method based on the particle filter variance. The variance of all particles at time step t is calculated as:

$$\mathrm{Cov}(\theta_{t,s}) = \sum_{i=1}^{N_p} w_t^i \left(\theta_{t,s}^i - m_t\right)\left(\theta_{t,s}^i - m_t\right)^T. \tag{21}$$

Ultimately, we define the active perception reward function as:

$$r_{t,step} = I(\mathbf{r}_{t+1}) - tr(\mathrm{Cov}(\theta_{t,s}))^{-1} d_t. \tag{22}$$

A high particle filter variance indicates high uncertainty in the estimated source term, so the mutual information term dominates and encourages exploration to reduce this uncertainty. As the variance decreases, the distance term becomes more influential, guiding the agent to exploit its knowledge and move rapidly to the estimated source.

Finally, the total extrinsic reward function is defined as:

$$r_t^e = \begin{cases} +10, & \text{find the source,} \\ I(\mathbf{r}_{t+1}) - tr(\mathrm{Cov}(\theta_{t,s}))^{-1} d_t, & \text{for each step,} \\ -5, & \text{outside boundary.} \end{cases} \tag{23}$$

D. Learning Framework for CIG-SAC

To train a robust source search policy while promoting active exploration under uncertainty, we propose the learning-based framework that integrates the soft actor-critic (SAC) method [19] with curiosity-driven exploration for STE.

We design the replay buffer D that stores tuples of agent-environment interaction in the form $(s_t, a_t, r_t^e, s_{t+1})$. It facilitates diverse experiences and efficient sample reuse by updating the policy with randomly sampled mini-batches. As shown in Fig. 2, the extrinsic active perception reward r_t^e is sampled from the replay buffer, whereas the intrinsic reward r_t^i is updated by the curiosity network. This method prevents intrinsic reward values that were high in previously uncertain regions from being repeatedly reused, thereby avoiding bias toward excessive exploration. The total reward is then updated as:

$$r_t^{total} = r_t^e + \eta r_t^i, \tag{24}$$

where η is the scaling factor between the extrinsic and intrinsic rewards.

The target Q-network is utilized to stabilize training by offering a temporally smoothed estimate of future rewards. The target Q-value is computed as:

$$y_t = r_t^{total} + \gamma\, \mathbb{E}_{a_{t+1}\sim\pi_\psi}\left[Q_{\bar{\phi}}(s_{t+1}, a_{t+1}) - \alpha \log \pi_\psi(a_{t+1}|s_{t+1})\right], \tag{25}$$

where $\bar{\phi}$ denotes the parameter of the target Q-network, ψ represents the parameter of the actor network, and α is the temperature parameter which balances the entropy term against the reward. The critic network is trained by minimizing the following loss function:

$$\mathcal{L}_{critic}(\phi) = \mathbb{E}_{(s_t,a_t)\sim D}\left[\left(Q_\phi(s_t, a_t) - y_t\right)^2\right]. \tag{26}$$

To optimize the policy, the actor network is trained by minimizing the following equation:

$$\mathcal{L}_{actor}(\psi) = \mathbb{E}_{s_t\sim D,\, a_t\sim\pi_\psi}\left[\alpha \log \pi_\psi(a_t|s_t) - Q_\phi(s_t, a_t)\right]. \tag{27}$$

Finally, we adopt automatic entropy tuning to dynamically adjust the temperature parameter α, which balances the trade-off between exploration and exploitation for the expected reward maximization. The objective function for tuning α can be written as:

$$\mathcal{L}_\alpha = \mathbb{E}_{a_t\sim\pi_\psi}\left[-\alpha \log \pi_\psi(a_t|s_t) + \widetilde{\mathcal{H}}\right], \tag{28}$$

where $\widetilde{\mathcal{H}}$ is the target entropy which determines the desired level of policy stochasticity. The overall learning process of the proposed framework is shown in Fig. 2.

IV. Numerical Simulation

A. Simulation Environment

The numerical simulation is conducted in two different environments by adjusting the parameters of gas, wind, and noise. Both environment 1 and 2 are designed with high turbulence, while environment 2 includes higher environmental and sensor noise as shown in Table I. These settings are designed to reflect challenging real-world scenarios and to train a robust policy under such conditions. The agent moves to the next sensing position with a fixed step size of 2 m within a 60 m × 60 m search area. The simulation terminates if the agent exceeds 300 steps or the standard deviation of the particle filter is below 0.1. At the end of each simulation, the source search is considered successful if the distance between the estimated source and the true source is within 1 m. The agent is trained with 60,000 number of training episodes, and the hyperparameters for training are summarized in Table II. Moreover, the locations of both the source and the mobile agent are randomly initialized in each training episode, and the parameters in Table I are randomly selected at the beginning of each episode.

TABLE I
PARAMETER VALUES OF ENVIRONMENTS 1 AND 2

Symbol	Environment 1	Environment 2	Unit
q_s	$U(500, 3000)$	$U(500, 3000)$	mg/s
τ	$U(200, 1500)$	$U(200, 1500)$	s
D	$U(2, 15)$	$U(2, 15)$	m^2/s
V	$U(0, 5)$	$U(0, 5)$	m/s
χ	$U(0, 360)$	$U(0, 360)$	deg
σ_{env}	0.4	0.5	mg/s
β	0.25	0.4	-

B. Ablation Study

The ablation study is conducted using two key metrics: success rate (SR) and mean travel distance (MTD). SR denotes the percentage of episodes where the agent successfully estimates the source location, whereas MTD represents the average distance the agent travels to achieve a successful source estimation. For each test, the parameters listed in

TABLE II

HYPERPARAMETER LISTS AND CORRESPONDING VALUES

Hyperparameter	Value
The size of first fully connected layer	256
The size of second fully connected layer	64
Learning rate (actor, critic network)	0.0003
Learning rate (curiosity network)	0.0001
Replay buffer size	100,000
Minibatch size	256
Curiosity loss weighting factor (ε)	0.2
Discount factor (γ)	0.99
Soft update coefficient (τ)	0.005
Intrinsic reward weight (η)	2.5
Optimizer	Adam

TABLE III

PERFORMANCE EVALUATION OF SAC-BASED METHODS

Method	Environment 1		Environment 2	
	SR (%)	MTD (m)	SR (%)	MTD (m)
SAC	86.0	94.1	84.4	98.7
IG-SAC	93.0	84.4	90.2	92.6
C-SAC	91.8	87.8	90.3	94.8
CIG-SAC	**98.3**	**80.2**	**95.0**	**86.2**

TABLE IV

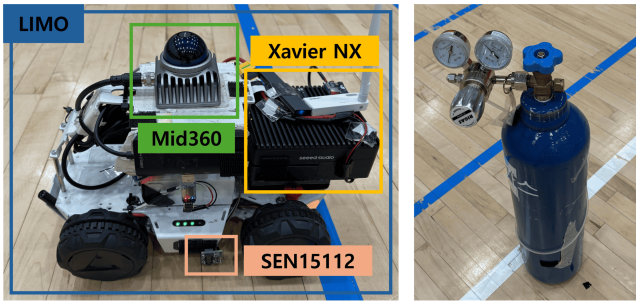
PERFORMANCE COMPARISON WITH EXISTING METHODS

Method	Environment 1		Environment 2	
	SR (%)	MTD (m)	SR (%)	MTD (m)
Infotaxis [2]	87.6	155.6	83.0	169.0
Entrotaxis [4]	85.4	142.3	82.7	149.2
PC-DQN [12]	82.3	100.5	78.5	112.4
MVG-RDDPG [13]	90.6	91.8	84.0	98.4
AID-RL [15]	73.9	120.4	63.4	138.2
Lee et al. [16]	91.6	95.8	86.0	100.4
CIG-SAC	**98.3**	**80.2**	**95.0**	**86.2**

Table 6 are randomly initialized, and 1,000 random scenarios are executed to ensure statistical reliability.

First, ablation analysis is conducted in both environments by selectively removing different components in CIG-SAC. Particularly, to verify the contributions of the curiosity-driven exploration and the active perception reward, SAC is fixed as the baseline and evaluated with each component removed. As shown in Table III, in SAC without the active perception reward and curiosity-driven exploration, the agent receives a positive reward only when it successfully finds the source. This corresponds to a typical sparse reward problem, which provides insufficient guidance for effective source term estimation. Consequently, it exhibits the worst performance. IG-SAC with the active perception reward, and C-SAC, which employs curiosity-driven exploration, both achieve improved performance compared to SAC. In IG-SAC, the proposed active perception reward facilitates robust policy learning by guiding the agent to reduce the uncertainty of the estimated source term distribution. However, relying solely on this reward can cause the policy to overlook informative states. In contrast, CIG-SAC achieves the most robust performance by encouraging the agent to explore novel regions where informative source term is likely to exist, while ensuring robust policy learning in turbulent environments.

Additionally, we compare our method with both information-theoretic and DRL-based approaches. The information-theoretic methods include infotaxis [2] and entrotaxis [4], whereas the DRL-based methods include PC-DQN [12], MVG-RDDPG [13], AID-RL [15], and Lee et al. [16]. Infotaxis and entrotaxis tend to exhibit higher MTD than DRL-based methods in both environments, as their limited action candidates can cause the agent to miss opportunities to reach optimal sampling positions. PC-DQN achieves more efficient MTD compared to infotaxis and entrotaxis, but still exhibits low SR. During training, it relies on random exploration, which causes the agent to miss informative states. Moreover, without utilizing rewards directly related to source information in this approach, it can result in suboptimal policy learning. MVG-RDDPG incorporates a gated recurrent unit (GRU) memory network to better exploit past measurements, achieving higher SR and MTD than PC-DQN. However, it still relies on random exploration and a concentration-based reward that is vulnerable to turbulence, resulting in lower performance than CIG-SAC. On the other hand, AID-RL selects actions with probability $1-\varepsilon$ by maximizing the Q-function, and with probability ε by reducing belief uncertainty the most. Thus, since this strategy is biased toward exploitation, its performance degrades in turbulent environments, as shown in Table IV. Finally, Lee et al. introduces GMM-based reward functions that improve robustness compared with other algorithms; however, the lack of effective exploration still limits its overall performance.



(a) Mobile sensor (b) Gas tank

Fig. 4. Setup for the experiments.

V. REAL-WORLD EXPERIMENTS

In the real-world experiments, a CO_2 gas tank is used to generate gas dispersion. The AgileX Limo UGV, equipped with a SEN15112 gas sensor, is utilized as the mobile sensing platform. The NVIDIA Jetson Xavier NX is employed for onboard computation, while FAST-LIO2 [20] and a Livox Mid-360 LiDAR are used for mobile sensor localization and environmental perception, respectively. Each algorithm is tested 10 times in an $8\,\mathrm{m} \times 8\,\mathrm{m}$ indoor gym, and the mobile sensor is set to move $0.5\,\mathrm{m}$ at each step. To ensure that CO_2 is sufficiently dispersed and reaches a steady state, the gas is released for 4 minutes before each experiment begins.

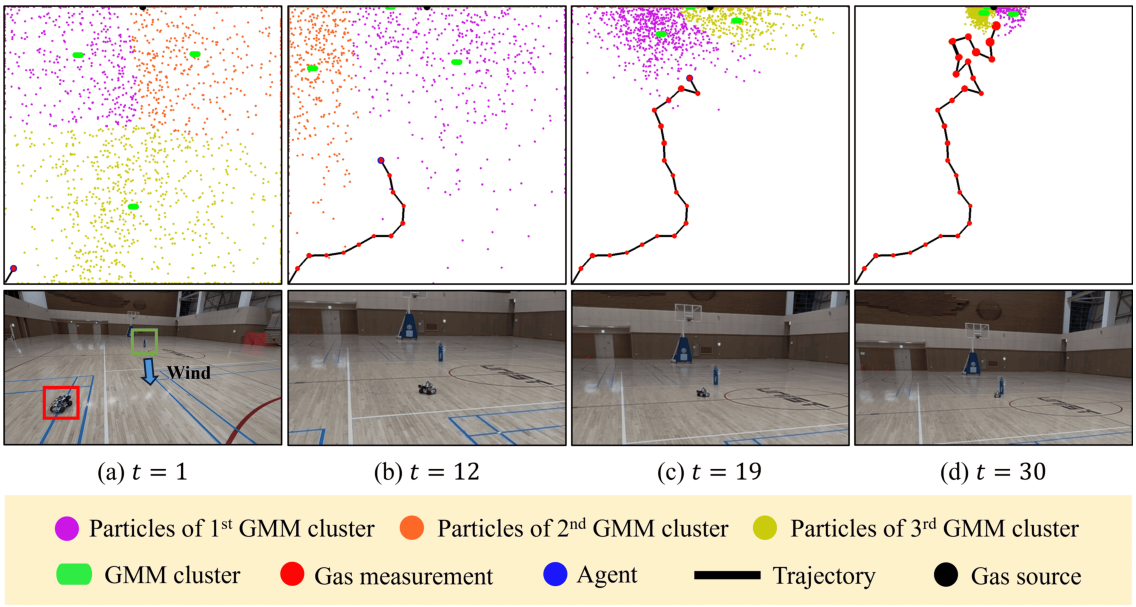


Fig. 5. Sample result of CIG-SAC for real experiments.

TABLE V
RESULTS OF REAL-WORLD EXPERIMENTS

Method	SR (%)	Step Number	CT (ms)
Infotaxis	70	49.3	420.3
MVG-RDDPG	80	34.8	**80.4**
CIG-SAC	**100**	**26.4**	82.6

Due to the recovery time of the gas sensor, measurements are collected 5 seconds after the mobile sensor arrives at each sensing position. Furthermore, to introduce external airflow, the indoor gym windows are kept open during the experiments. The experiment setup is shown in Fig. 4.

In the real-world experiments, we compare infotaxis, MVG-RDDPG, and CIG-SAC. Unlike in the simulations, we additionally evaluate the averaged computation time for one step decision making (CT) as a performance metric. Since infotaxis must compute the uncertainty for all action candidates by considering all possible future measurements, it shows the highest CT, which is inefficient for real-time applications. In contrast, both MVG-RDDPG and CIG-SAC achieve low CT by learning an optimal policy through DRL. However, MVG-RDDPG, which relies on random exploration during training, exhibits inefficient step number and low success rate in turbulent environments. In contrast, since CIG-SAC effectively balances curiosity-driven exploration with the active perception reward, it achieves robust performance even in real-world experiments. As shown in Fig. 5, despite sparse measurements up to $t = 12$, the mobile sensor guided by CIG-SAC follows an efficient trajectory toward the source.

VI. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a curiosity-driven information-guided reinforcement learning framework for robust source search in turbulent environments. The proposed method effectively guides the agent to explore novel regions where informative source information is likely to be obtained, while the active perception reward enables robust and efficient source search. Through extensive turbulent simulations and real-world experiments, CIG-SAC demonstrated superior performance compared to existing approaches, highlighting its potential applicability to practical STE problems. For future work, we plan to extend the framework to multi-agent systems to further enhance source search performance.

REFERENCES

[1] M. Hutchinson, H. Oh, and W.-H. Chen, "A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors," *Inf. Fusion*, vol. 91, pp. 83–100, 2017.

[2] M. Vergassola, E. Villermaux, and B. I. Shraiman, "'Infotaxis' as a strategy for searching without gradients," *Nature*, vol. 445, no. 7126, pp. 406–409, 2007.

[3] B. Ristic, A. Skvortsov, and A. Gunatilaka, "A study of cognitive strategies for an autonomous search," *Inf. Fusion*, vol. 28, pp. 1–9, 2016.

[4] M. Hutchinson, H. Oh, and W.-H. Chen, "Entrotaxis as a strategy for autonomous search and source reconstruction in turbulent conditions," *Inf. Fusion*, vol. 42, pp. 179–189, 2018.

[5] M. Park, S. An, J. Seo, and H. Oh, "Autonomous source search for UAVs using Gaussian mixture model-based Infotaxis: algorithm and flight experiments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 4238–4254, 2021.

[6] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: a survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.

[7] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: a survey," *Inf. Fusion*, vol. 85, pp. 1–22, 2022.

[8] H. Hu, S. Song, and C. L. P. Chen, "Plume tracing via model-free reinforcement learning method," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 8, pp. 2515–2527, 2019.

[9] X. Chen, C. Fu, and J. Huang, "A deep Q-network for robotic odor/gas source localization: modeling, measurement and comparative study," *Measurement*, vol. 183, p. 109725, 2021.

[10] H. Li, J. Yuan, and H. Yuan, "An active olfaction approach using deep reinforcement learning for indoor attenuation odor source localization," *IEEE Sens. J.*, vol. 24, no. 9, pp. 14 561–14 572, 2024.

[11] Y. He, L. Cheng, Y. Pan, D. Wang, Y. Li, and H. Zheng, "Gas source localization using dueling deep Q-network with an olfactory quadruped robot," *Int. J. Adv. Robot. Syst.*, vol. 21, no. 3, p. 17298806241255797, 2024.

[12] Y. Zhao, B. Chen, X. Wang, Z. Zhu, Y. Wang, G. Cheng, R. Wang, R. Wang, M. He, and Y. Liu, "A deep reinforcement learning based searching method for source localization," *Inf. Sci.*, vol. 588, pp. 67–81, 2022.

[13] M. Park, P. Ladosz, and H. Oh, "Source term estimation using deep reinforcement learning with Gaussian mixture model feature extraction for mobile sensors," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8323–8330, 2022.

[14] S. H. Singh, F. van Breugel, R. P. Rao, and B. W. Brunton, "Emergent behaviour and neural dynamics in artificial agents tracking odour plumes," *Nat. Mach. Intell.*, vol. 5, no. 1, pp. 58–70, 2023.

[15] Z. Li, W.-H. Chen, J. Yang, and Y. Yan, "AID-RL: Active information-directed reinforcement learning for autonomous source seeking and estimation," *Neurocomputing*, vol. 544, p. 126281, 2023.

[16] J. Lee, H. Jang, M. Park, and H. Oh, "Enhanced reward function design for source term estimation based on deep reinforcement learning," *IEEE Access*, 2025.

[17] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: particle filters for tracking applications*. Artech House, 2003.

[18] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2017, pp. 2778–2787.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2018, pp. 1861–1870.

[20] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "FAST-LIO2: Fast direct lidar-inertial odometry," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2053–2073, 2022.