IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 June 2023, accepted 19 June 2023, date of publication 23 June 2023, date of current version 29 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289077

RESEARCH ARTICLE

Pursuing Benefits or Avoiding Threats: Realizing Regional Multi-Target Electronic Reconnaissance With Deep Reinforcement Learning

YONGLE XU, MING ZHANG, AND BOYIN JIN
School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China

Corresponding author: Boyin Jin (boyin_jin@just.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62006099.

ABSTRACT Unmanned combat aerial vehicles (UCAVs) are preferred for regional electronic reconnaissance due to their versatility and stealth. This paper proposes a deep reinforcement learning (DRL) method to enable UCAVs to complete regional multi-target electronic reconnaissance (MER) tasks with continuous autonomous maneuvers. Distinguishing from traditional heuristic search algorithms, we first derive the objective function of MER and elucidate sufficient conditions to improve the success rate of reconnaissance recognition. Then, using the original cognitive electronic warfare framework, a three-dimensional MER simulator named Scouer-N is created to satisfy the requirements of dynamic environment training for DRL-based agents. To enable the processing of sequential situation awareness, a generative network is constructed by introducing a partially observable Markov decision process (POMDP) model, which assists the UCAV in filtering the observations from the sensor and predicting the actual states. Finally, we propose a priority-driven state reward shaping method that provides normalized state representation and dense rewards to the agent during training to improve the agent's behavioral knowledge for MER. The experimental results demonstrate a considerable improvement in the task success rate of the trained UCAV relative to the benchmark, proving the efficacy of our approach in helping agents learn the optimal reconnaissance strategy from the potential state space.

INDEX TERMS Multi-target electronic reconnaissance, cognitive electronic warfare, deep reinforcement learning, 3D motion planning, POMDP model.

I. INTRODUCTION

Cognitive electronic warfare (CEW) favors the use of unmanned aerial vehicles (UCAVs) to perform electronic reconnaissance of emitter source information at sensitive places to gain more electromagnetic initiative. The autonomy and real-time performance of UCAV systems are currently under very high demands from multi-target electronic reconnaissance (MER) due to the unknown emission power characteristics of the radar at sensitive places as well as the unknown guidance radius of collateral air-defense threats. In MER, a UCAV relies on passive detection to intercept non-cooperative radar signals and integrate their contents, and then inference and decision-making based on posterior information to reduce its maneuvering risk [1], [2], [3].

Typically, electronic reconnaissance within a target area is characterized as an agent-environment interaction involving limited prior knowledge, with error-introduced target coordinates as the initial input. When performing reconnaissance flights, the UCAV must create efficient flight patterns based on adversary threat information and terrain data to increase effectiveness and safety [4]. Consequently, many studies classify the MER problem as a pure trajectory or motion planning problem with some platform constraints added in to make certain real-world tasks can be completed accordingly [5], [6], [7]. Heuristic evolutionary computing techniques are frequently employed to solve the non-deterministic polynomial hard issues such as trajectory

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

planning that arise in MER [8], [9], [10]. Nonetheless, poor real-time performance is a common fundamental weakness of these heuristic algorithms, as they require discretizing the 3D task space and solving each scene at an extremely high computational cost. Even worse, the above studies have overlooked the dynamic perception of UCAV's reasonable reconnaissance range, i.e., reconnaissance radius, for the target radars, which is seriously inconsistent with reality [11].

Functions of electronic reconnaissance include extracting, classifying, grouping, and recognizing electromagnetic signal features from unknown radiation sources. Most of CEW's studies now focus on real-time analysis of reconnaissance signals, and deep learning networks are considered very useful [12], [13], [14]. However, these studies are heavily biased toward the ability of the UCAV's payload or system to process and sense signals at the back end, which is not closely related to the specific behavior of the UCAV during reconnaissance operations. In other words, in CEW, UCAVs must make good maneuvering judgments to intercept signals from unknown emitter sources to meet the core requirements of autonomous reconnaissance [15].

Recently, Deep reinforcement learning (DRL) technology has attracted the attention of many scholars in the field of CEW due to its excellent performance in complex decision-making tasks [16], [17]. In particular, most DRL algorithms has achieved outstanding achievements in autonomous path planning and end-to-end control of unmanned aerial vehicles or unmanned ships [18], [19]. Through DRL, unmanned agents can optimize control strategies in real-world electronic warfare by shaping cognitive tasks like target search and tracking into Markov processes (MDPs) and tightly integrating the state, control actions, and environmental feedback of UCAVs [20], [21], [22]. Unfortunately, because electronic reconnaissance simulation depends not just on modeling mobile platforms but also on passive receivers as sensors, it's necessary to optimize the partially observable Markov decision process (POMDP) model to shape the electronic reconnaissance task into a multi-layered state transition from perception to motion, which increases the complexity of the MER task [23].

This work aims to determine the most effective strategy for UCAVs to perform multi-target electronic reconnaissance autonomously in a challenging environment. We are geared toward the operational needs of sensitive areas, directing the UCAVs to utilize their agility and reconnaissance capabilities to lock on the target and capture enough electromagnetic data without being destroyed in a constrained exploration space and operational cycle. To handle the intelligent decision-making required for UCAV reconnaissance, we theoretically model the MER problem as a POMDP and develop a DRL network with generative states to solve it. Moreover, an open-source regional MER simulator based on the cognitive electronic warfare simulation framework is developed to address the verisimilitude of the dynamic simulation environment [21]. This simulator can map the processes of physical platform maneuvering and passive signal reception in digital space.

The following are the primary innovations of our work:

- The interaction between a UCAV and its reconnaissance system and multiple radars in continuous space is closed-looped using mathematical modeling, and the dependence between the local optimal reconnaissance strategy and the mission completion probability is derived.
- Determine the state reward shaping equation for pursuing benefits or avoiding threats to make the maneuvering policies of the UCAV for executing MER more directed while maintaining platform and payload constraints.
- A DRL network for MER is proposed to enable UCAVs to comprehend end-to-end maneuvering policies through their own reconnaissance system, such as approaching or escaping when the radar threat signature is unknown.

The rest of this paper is structured as follows. Section II discusses the objective function of local policy optimization for MER tasks and emphasizes the importance of implementing digital electronic reconnaissance models. A detailed description of our DRL framework based on POMDP for handling MER can be found in Section III. Section IV explains the simulation methodology and associated experimental results, followed by an analysis of the behavior recognized by DRL agents based on these data. Finally, we provide a profound summary of this paper in Section V.

II. MATHEMATICAL MODEL AND SIMULATOR OF MER

A. PROBLEM FORMULATION

In this paper, we investigate a unified mathematical model for the MER problem. Typical scenarios for a UCAV executing MER tasks are shown in Figure 1. All radar signals encountered by the UCAV during reconnaissance are utilized to define the electromagnetic environment in which they operate [15]. Assuming there are M ground-based radars in the task space, their pulse signal features satisfy a Gaussian distribution. Thus the electromagnetic features of the nth pulse generated by the mth radar at time step t can be expressed as:

$$\begin{aligned} \bar{x}_{m,n,t} &= \kappa_{m,t\sim t+\Delta T} x_{m,n,t\sim t+\Delta T} \kappa_{m,t\sim t+\Delta T} \in \{0, 1\} \\ \mathbb{E}\left[x_{m,t\sim t+\Delta T}\right] &= \mu_m \end{aligned} \tag{1}$$

where $\kappa_{m,t\sim t+\Delta T}$ is the flag bit for whether the signal can be intercepted by the UCAV, and ΔT is the processing cycle for MER tasks. μ_m is the mean value of the pulse signal features of the mth radar, which can also be regarded as the centroid of the radar signal feature space. $\mathbb{E}$ is the expectation calculation function.

Assuming that the mth radar emits N_m signal pulses and reaches UCAV within the time interval ΔT, the contribution of this radar to electromagnetic space can be defined as:

$$\chi_{m,t} = \sum_{n=0}^{N_m} \bar{x}_{m,n,t} \tag{2}$$

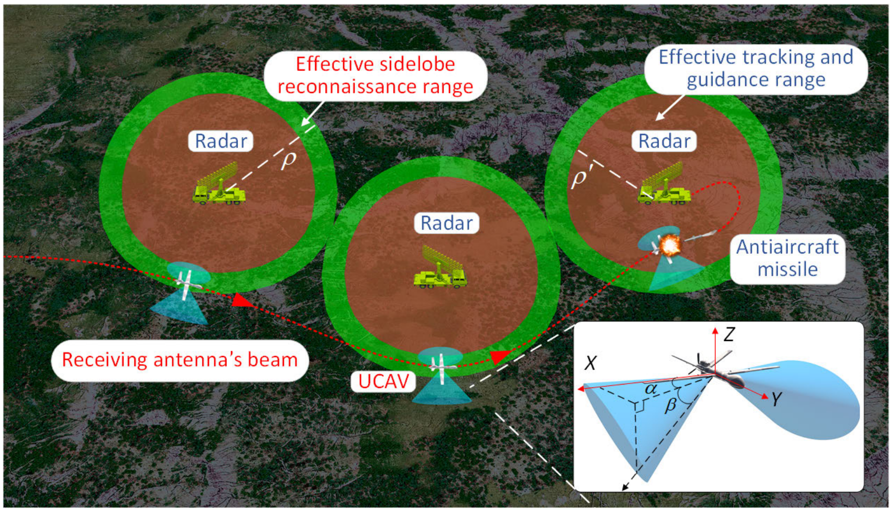


FIGURE 1. Typical scenarios for a UCAV executing MER tasks. The effective reconnaissance range at which the UCAV can intercept radar's sidelobe signals is shown by a green hemisphere, and the effective tracking and guidance range of the radar for the UCAV is shown by a red hemisphere.

where N_m satisfies the Poisson distribution with the pulse flow density F_m of radar.

As a result, the MER environment in which a UCAV operates can be described by the following formula:

$$\chi_t = \sum_{m=1}^{M}\sum_{n=0}^{N_m} \bar{x}_{m,n,t}, \tag{3}$$

Given the possibility of pulse loss when the UCAV's passive receiver simultaneously receives M radar signals with their respective pulse widths $\tau_m, m = 1, 2, \ldots, M$, and the probability that any current pulse will not be lost is calculated as:

$$\begin{aligned} p_m &= \prod_{m=1}^{M-1} (1-\tau_m F_m) \\ &\approx \exp\left(-\sum_{m=1}^{M} \tau_m F_m\right), \quad \tau_m F_m < 1 \end{aligned} \tag{4}$$

Therefore, the actual electromagnetic space observed by the UCAV is sampled by a joint probability $\prod_{m=1}^{M} p_m$ on χ_t:

$$\hat{\chi}_t = \sum_{m=1}^{M}\sum_{n=0}^{N'_m} \bar{x}_{m,n,t} \sim \left(\sum_{m=1}^{M} \chi_{m,t}, \prod_{m=1}^{M} p_m\right), \quad N'_m \leq N_m \tag{5}$$

where N'_m is the number of signal pulses actually sensed by the UCAV, and its value satisfies a binomial distribution with probability p_m. These sampled signals will form M feature cluster centroids.

The MER's objective function for the UCAV is to improve the similarity between reconnaissance sample points in each cluster centroid generated during signal sorting, which can be characterized by the metric J_t:

$$\begin{aligned} J_t &= \frac{1}{N}\sum_{m=1}^{M}\sum_{n=0}^{N}\left\|\bar{x}_{n,t} - \frac{\chi_{m,t}}{N_m}\right\|^2 \\ &\geq \sum_{m=1}^{M}\left\|\frac{1}{N}\sum_{n=0}^{N}\bar{x}_{n,t} - \frac{\chi_{m,t}}{N_m}\right\|^2 = \sum_{m=1}^{M}\left\|\frac{\chi_t}{N} - \mu_m\right\|^2 \end{aligned} \tag{6}$$

where $\bar{x}_{n,t}$ represents the nth feature sample of $\hat{\chi}_t$, and $N = \sum_{m=1}^{M} N'_m$. Equation (6) indicates that the similarity is measured by the Euclidean distance between the features of the sampled signals and the statistical clustering centroids of all intercepted radar signals. Consequently, minimizing J_t is required to optimize the UCAV's reconnaissance results on target signals. However, (6) also establishes a bound constraint on the value of J_t, demonstrating that J_t cannot be minimized without limitation.

The rule of large numbers states that when N is large enough, $\hat{X}_t/N$ approaches μ_m with a fixed biased error, meaning that the statistical feature distribution of intercepted signals can approximate the actual feature distribution of target signals in the MER environment. Applying (1), to attain a smaller J_t, it is necessary to maximize N under the condition $\kappa_{m,t\sim t+\Delta T} = 1$.

The mathematical expectation of N possesses the following qualities:

$$\begin{aligned} &\mathbb{E}[N] \\ &= \mathbb{E}\left[\sum_{m=1}^{M} N'_m\right] \\ &= \sum_{m=1}^{M}\sum_{N_m=0}^{\mathrm{f_N}(\Delta T/\tau_m)}\sum_{n=N'_m}^{N_m} \frac{\zeta_m^n \exp(-\zeta_m)}{(n-1)!}\left(C_n^{N_m}\right) p_m^{N'_m}(1-p_m)^{n-N'_m} \end{aligned}$$

$$\leq \Delta T \sum_{m=1}^{M} F_m \exp\left(-\sum_{m=1}^{M} \tau_m F_m\right) \tag{7}$$

where $\zeta_m = F_m \Delta T$, and $\mathrm{f_N}(\Delta T/\tau_m)$ means to get the maximum integer smaller than $\Delta T/\tau_m$. From (7), maximizing N for UCAV can only be achieved by increasing ΔT because both F_m and τ_m are the inherent features of the mth radar and independent of the UCAV's reconnaissance abilities.

The essence of regional MER tasks is revealed by (1) to (7) from a physical standpoint: extending the period of reconnaissance (or processing cycle) of the UCAV to each radar will enhance the ability to sort and recognize unknown signals, which is compatible with real operations. Even though the given formula can explain the interaction between UCAV and radar in electromagnetic space, the MER constraints in physical space still need to be rewritten. Based on the spatial parameters of the UCAV, the following will calculate the constraints for radar signal interception.

The coordinates of the mth radar observed by UCAV conform to the following normal distribution:

$$\boldsymbol{p}'_{m,t} \sim \mathcal{N}\left(\boldsymbol{p}_m, \delta_m\right) \tag{8}$$

where $\boldsymbol{p}_m$ is the actual coordinate of the mth radar, and δ_m is the corresponding positioning error.

Define the UCAV's coordinates at time step t as $\boldsymbol{p}_{o,t}$, and the actual relative displacement between the mth radar and the UCAV is $\boldsymbol{p}_{mo,t} = \boldsymbol{p}_m - \boldsymbol{p}_{o,t}$. Then, in UCAV's body coordinate system, the line-of-sight (LOS) vector $\boldsymbol{l}_{m,t} = [l_{m,t,x}, l_{m,t,y}, l_{m,t,y}]$ between the radar and the UCAV is expressed as:

$$\boldsymbol{l}_{m,t} = \boldsymbol{C}_{x,y,z}\boldsymbol{p}_{mo,t} \tag{9}$$

where $\boldsymbol{C}_{x,y,z}$ is the coordinate transformation matrix from the geocentric coordinate system to UCAV's body coordinate system.

As illustrated in Figure 1, the azimuth and pitch mainlobe beamwidths of the UCAV's receiving antenna are represented as α and β, respectively, and the system sensitivity of the UCAV's passive receiver is defined as $P_{\min}$. Thus the fundamental criteria for UCAV to intercept radar signals can be described by:

$$\kappa_{m,t} = \begin{cases} 1, & \text{if} \quad \alpha_{m,t} \leq \alpha, \beta_{m,t} \leq \beta, P_{m,t} \geq P_{\min} \\ 0, & \text{else} \end{cases} \tag{10}$$

where $t \in [t \sim t + \Delta T]$. $\alpha_{m,t}$ and $\beta_{m,t}$ represent the azimuth and pitch angles of the mth radar, respectively, in the UCAV's LOS direction:

$$\begin{cases} \alpha_{m,t} = \arctan \dfrac{los_{m,t,y}}{los_{m,t,x}} \\ \beta_{m,t} = \arcsin \dfrac{los_{m,t,z}}{\left\|\boldsymbol{p}_{mo,t}\right\|} \end{cases} \tag{11}$$

In (10), $P_{m,t}$ represents the radar signal power entering the receiver, which can be calculated by:

$$P_{m,t} = \frac{E_m G_m \lambda_m^2}{\left(4\pi \|\boldsymbol{p}_{mo,t}\|\right)^2} \tag{12}$$

where E_m, G_m, and λ_m represent the effective radiated power (ERP), ratio of the sidelobe to the mainlobe, and signal wavelength, respectively, of the mth radar.

The optimization of the objective function (6) will inevitably be an NP-hard problem if the time for UCAV to execute electronic reconnaissance tasks is divided into multiple uniform time segments and the constraints of (10) are incorporated [1].

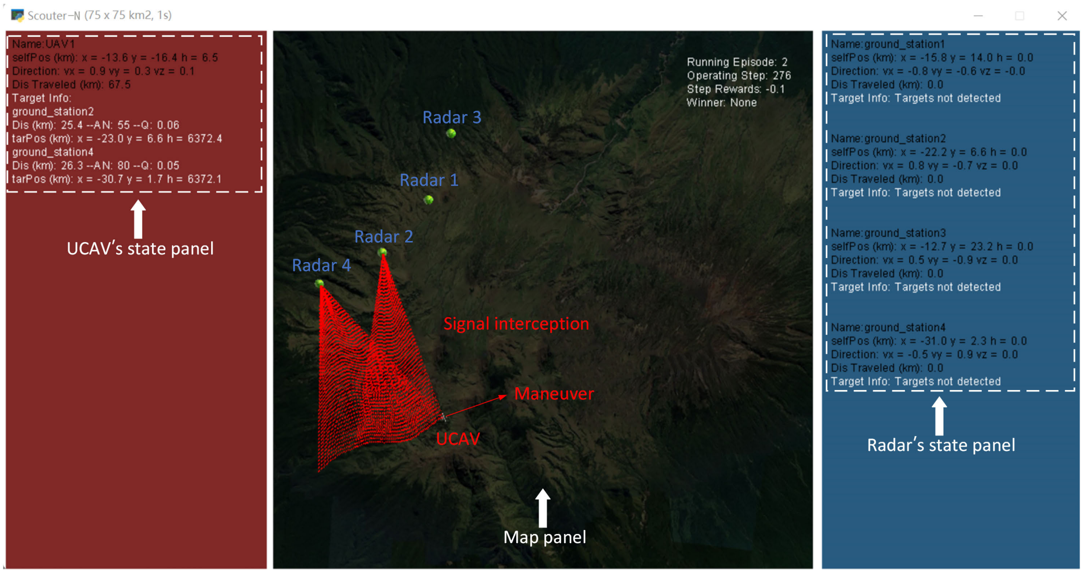
B. SIMULATOR "SCOUTER-N"

The previous subsection reveals that the evaluation of UCAV in MER processes can be represented by a continuous-time mathematical model, necessitating a dynamic simulation environment. Moreover, to handle the tasks of long-term, time-continuous, and mixed scenarios, DRL-based control decisions also have strict requirements for simulation environments with complete state transitions [16], [24]. Thus, using the classic CEW framework [21], [22], we develop a simulator for MER called Scouter-N and make sure it meets the functional mapping equations from (8) to (12).

As Figure 1 illustrated, the characteristics of a MER task are described as follows:

- A 3D space with the dimensions of length L, width W, and height H represents the target area. Two or more radars exist in this area used for ground-to-air surveillance, and air-defense firepower, such as anti-aircraft missiles, is deployed near the radars to protect them. The imprecise locations of these radars can be obtained through satellites and other intelligence sources, while their detection capabilities, the guidance radius of air-defense firepower, and the strike conditions remain unknown.
- Autonomous mobile platforms such as UCAV gradually move toward the target with an electronic reconnaissance system, intercept the radar signals by changing its position and orientation, and then do long-term reconnaissance and surveillance as needed. Note that, because of its inherent energy constraints, the UCAV has a maximum amount of time, $Ttextmax$, to finish an operation.
- There are three possible results for the conclusion of a MER task: 1) when the UCAV intercepts signals from all target radars, the task is considered complete and the UCAV side wins; 2) if the UCAV is destroyed by the radars' air-defense firepower during reconnaissance, the radar side wins; 3) it will be declared a draw if the UCAV does not finish the task in the required time but returns safely.

The activation of radar air-defense firepower is mainly connected to the radar's tracking threshold in a real-world



```
Name:UAV1
selfPos (km): x = -13.6 y = -16.4 h = 6.5
Direction: vx = 0.9 vy = 0.3 vz = 0.1
Dis Traveled (km): 67.5
Target Info:
ground_station2
Dis (km): 25.4 --AN: 55 --Q: 0.06
tarPos (km): x = -23.0 y = 6.6 h = 6372.4
ground_station4
Dis (km): 26.3 --AN: 80 --Q: 0.05
tarPos (km): x = -30.7 y = 1.7 h = 6372.1
```

```
Running Episode: 2
Operating Step: 276
Step Rewards: -0.1
Winner: None
```

```
Name:ground_station1
selfPos (km): x = -15.8 y = 14.0 h = 0.0
Direction: vx = -0.8 vy = -0.6 vz = -0.0
Dis Traveled (km): 0.0
Target Info: Targets not detected

Name:ground_station2
selfPos (km): x = -22.2 y = 6.6 h = 0.0
Direction: vx = 0.8 vy = -0.7 vz = 0.0
Dis Traveled (km): 0.0
Target Info: Targets not detected

Name:ground_station3
selfPos (km): x = -12.7 y = 23.2 h = 0.0
Direction: vx = 0.5 vy = -0.9 vz = 0.0
Dis Traveled (km): 0.0
Target Info: Targets not detected

Name:ground_station4
selfPos (km): x = -31.0 y = 2.3 h = 0.0
Direction: vx = -0.5 vy = 0.9 vz = 0.0
Dis Traveled (km): 0.0
Target Info: Targets not detected
```

FIGURE 2. Scouter-N's interface displays the UCAV status panel on the left, the map dynamic display panel in the center, and the radar status panel on the right.

scenario. In this case, the CEW framework's parametric data processing system (PDPS) [20] is used to determine the conditions for activation as:

$$\begin{cases} ||\boldsymbol{p}_{mo,t}||q_t \leq \Delta d_m \\ \dfrac{E_m \lambda_m^2 \sigma}{(4\pi)^3 \, ||\boldsymbol{p}_{mo,t}||^4} \geq P_{m,\min} \end{cases} \tag{13}$$

where q_t is a finite-time convergent factor for describing radar positioning inaccuracy and Δd_m is the required tracking precision for air-defense firepower guidance. σ is the UCAV's radar cross section (RCS), and $P_{m,\min}$ is the system sensitivity of the mth radar.

Equation (13) indicates that in order for the radar side to win, two conditions must be satisfied: the UCAV must access the radar's detection range, and the radar must be able to track the UCAV with less error than the guidance threshold. Therefore, the UCAV in MER have the option of maintaining a safe distance from radar stations or employing effective reconnaissance strategies to accomplish their tasks.

Consider a circular deployment layout for multiple radars to cover the core area. Generate the initial layout in ''Scouter-N'' via the following equation:

$$\boldsymbol{p}_m = [x_c + d\cos\varphi_m, y_c + d\sin\varphi_m, z] \tag{14}$$

where $z = \sqrt{d_e - (x_c + d\cos\varphi_m)^2 - (y_c + d\sin\varphi_m)^2}$, $\varphi_m = \frac{\pi m}{2M}(-1)^m + 2\pi\text{rand}(0, 1)$. $[x_c, y_c]$ is the core area's ground projection coordinate, d is the designed radar shielding radius, d_e is the Earth radius (6371 km), and function rand$(0, 1)$ generates a number randomly between 0 and 1.

High-speed motion models in Scouter-N are independently equipped with a thrust-vectoring controller that allows the UCAV to maneuver continuously within its dynamic acceleration boundaries by using two-dimensional variables $[\vartheta_t, \phi_t]$ as control inputs. The calculation of dynamic acceleration boundaries is rely on the UCAV's attitude as well as its physical overload. Specific method for updating the UCAV's motion equations can refer to [21].

TABLE 1. Fundamental configuration parameters of Scouter-N.

Description	Value	Affiliation
3D MER space	[75 km,75 km,12 km]	Scouter-N
Maximum flight speed	250 m/s	UCAV
2D overload coefficient	[6 g, 3 g], g=9.8 m/s^2	UCAV
System sensitivity of passive receivers	-75 dBm	UCAV
RCS	0 dBsm	UCAV
2D receiving antenna beamwidths	$[90°, 45°]$	UCAV
Signal wavelength	50 cm	Radars
Ratio of sidelobe to mainlobe	-38 dBc	Radars
ERP	110 dBm	Radars
System sensitivity of active receivers	-125 dBm	Radars

Figure 2 shows the interaction between the UCAV and its target radars in physical and electromagnetic space. Table 1 summarizes some fundamental configuration parameters for Scouter-N. Despite the fact that the simulator's setup is quite cumbersome, to promote collaboration and knowledge sharing amongst scholars, we plan to share the source code of Scouter-N through our team's email.

III. DEEP REINFORCEMENT LEARNING NETWORK BASED ON POMDP

Since the UCAV in MER detects the electromagnetic energy of radar signals via a passive receiver, only a portion of the environment is observable at anytime. Given the significant difficulties in directly optimizing the objective function of MER, we consider describing the MER tasks in Scouter-N via a POMDP model and employing a novel DRL-based approach to solve it.

A. POMDP MODEL

Define the POMDP by a 6-element tuple, i.e., $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z}, \gamma \rangle$, where $\mathcal{S}$ denotes the actual state space, and $\mathcal{Z}$ is the observation set of the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the probability distribution of state transition, $\mathcal{R}$ is the reward function, and $\gamma \in (0, 1]$ is the discount factor [25], [26]. At time step t, the actual state of the UCAV in Scouter-N environment is $S_t = \boldsymbol{s} \in \left\{\boldsymbol{s}^1, \boldsymbol{s}^2, \ldots, \boldsymbol{s}^{|\mathcal{S}|}\right\} \subseteq \mathcal{S}$, and $Z_t = \boldsymbol{z} \in \mathcal{Z}$ is the observation processed by the UCAV's reconnaissance system. According to (10) to (12), there exists a certain analytical relationship between the observation and the radar's ERP, which can be represent by:

$$Z_t = \left[\boldsymbol{p}_{o,t}, \kappa_{1,t}, P_{1,t}, \kappa_{2,t}, P_{2,t}, \ldots, \kappa_{M,t}, P_{M,t}\right] \tag{15}$$

As shown in (12), the dimension of Z_t is $2M + 3$, and $P_{m,t}$ is a function of $\|\boldsymbol{p}_{mo,t}\|$, thus the following equation is satisfied:

$$Z_t \sim \left[\boldsymbol{p}_{o,t}, \kappa_{1,t}, \boldsymbol{p}_{1o,t}, \kappa_{2,t}, \boldsymbol{p}_{2o,t}, \ldots, \kappa_{M,t}, \boldsymbol{p}_{Mo,t}\right] \tag{16}$$

The UCAV chooses the optimal action $A_t = \boldsymbol{a}$ in the action space $\mathcal{A}$ based on Z_t and its controller's policy function $\pi(\cdot | Z_t = \boldsymbol{z})$, and then transfers the current state to the next stage $S_{t+1} = \boldsymbol{s}'$, the probability of describing the state transition is $p^{\boldsymbol{a}}_{\boldsymbol{ss}'} = P[S_{t+1} = \boldsymbol{s}' \mid S_t = \boldsymbol{s}, A_t = \boldsymbol{a}] \in \mathcal{P}$. Meanwhile, a timely reward will be provided by the environment to evaluate the action's quality, i.e., $R_{t+1} = r = R(S_{t+1} = \boldsymbol{s}', S_t = \boldsymbol{s}, A_t = \boldsymbol{a})$. When environmental observation deviates significantly from reality, the UCAV-generated action decisions will contain a large number of unstable factors. Consequently, for a continuous task, we expect to predict the current state S_t by a sequence of observations and actions, i.e., $\boldsymbol{h} = \{A_0, Z_1, R_1, \ldots, A_{t+1}, Z_t, R_t\}$, while evolving the POMDP problem into a statistical MDP problem.

To reduce the quantity of historical data required for state prediction, belief state $B_t = \boldsymbol{b} \in \mathcal{B}$ is defined to characterize the effect of $\boldsymbol{h}$ on $\boldsymbol{s}$:

$$B_t\left(\boldsymbol{s}\right) = \{p^{\boldsymbol{s}^1}_{\boldsymbol{h}}, p^{\boldsymbol{s}^2}_{\boldsymbol{h}}, \ldots, p^{\boldsymbol{s}^{|\mathcal{S}|}}_{\boldsymbol{h}}\}, \sum_{\boldsymbol{s} \in \mathcal{S}} p^{\boldsymbol{s}}_{\boldsymbol{h}} = 1 \tag{17}$$

where $p^{\boldsymbol{s}}_{\boldsymbol{h}}$ is the probability of observing the sequence $\boldsymbol{h}$ and predicting the actual state as $\boldsymbol{s}$.

Even if $p^{\boldsymbol{s}}_{\boldsymbol{h}}$ and $p^{\boldsymbol{a}}_{\boldsymbol{ss}'}$ are unknown, a DRL-based agent can still learn the prior distribution of the optimal policy from the execution experience of MER tasks through the accumulation of a large number of historical state transition pairs and the application of Monte Carlo sampling.

The objective function of a POMDP can be defined as [27]:

$$Q_\pi\left(B_t, A_t\right) = \mathbb{E}_{\pi,\boldsymbol{b}}\left[\sum_{k=0}^{\infty} \gamma R_{t+k+1} \mid B_t = \boldsymbol{b}, A_t = \boldsymbol{a}\right] \tag{18}$$

where Q is known as the action-value function in reinforcement learning, and a higher value of the Q indicates that the agent's decision has a better influence on the future. The optimal policy required to maximize Q-function for agent is:

$$\pi^*\left(\boldsymbol{a}|B_t = \boldsymbol{b}\right) = \underset{\boldsymbol{a} \in \mathcal{A}}{\operatorname{argmax}} Q \tag{19}$$

Figure 3 shows the logical framework for modeling a MER task in Scouter-N as POMDP.

Apparently, the observation feedback Z_t and B_t in Scouter-N is based on real-world operational situations, and we expect to use a DRL network to extract enough favorable policies to guide the UCAV in completing MER tasks.

B. STATE REWARD SHAPING

According to (7), a UCAV executing MER must maintain sufficient processing cycles ΔT for each radar to intercept a sufficient number of pulses to complete clustering and recognition, and intermittent reconnaissance methods will inevitably result in pulse loss. Due to the characteristics of radar deployment, the UCAV can only intercepts intercept radar signals rather than all within a given region, i.e., $\prod_{m=1}^{M} \kappa_{m,t} = 0$, necessitating a deliberate search for the current most important target.

The prerequisite for state reward shaping is how to train the UCAV to learn a trade-off between pursuing benefits and avoiding threats. Drawing on the priority-driven approach proposed in [28], we can provide a new idea to define the state and reward. Assuming that the most important reconnaissance target at time step t is the m'-th radar, i.e., the mth radar possesses the highest priority.

To avoid conceptual confusion, this work treats trained UCAVs as intelligent agents capable of autonomous thinking like humans. Define the input state vector of the DRL-based agent as:

$$S_t = \left[\boldsymbol{p}_{o,t}, \bar{\kappa}_{1,t}, \boldsymbol{p}_{1o,t}, \bar{\kappa}_{2,t}, \boldsymbol{p}_{2o,t}, \ldots, \bar{\kappa}_{M,t}, \boldsymbol{p}_{Mo,t}\right] \tag{20}$$

where $\bar{\kappa}_{m,t}$ is a flag bit obtained using the formula below:

$$\bar{\kappa}_{m,l} = \begin{cases} 1, & if\ m = m' \\ -1, & \text{else if } m = m' \text{ and } \kappa_{m,t} = 1 \\ 0, & \text{else} \end{cases} \tag{21}$$

There are various methods for determining m', and here we present a simple but reasonable design for radar's priority: the closer the target to UCAV the higher the priority. The original intent of this design is to perform reconnaissance on the nearest target first, allowing the MER task to be accomplished more quickly and confidently. As a result, the flag bit must be constantly updated by $m'_t = \underset{m}{\operatorname{argmin}} \|\boldsymbol{p}'_{mo,t}\|$, $m = 1, 2, \ldots, M$.

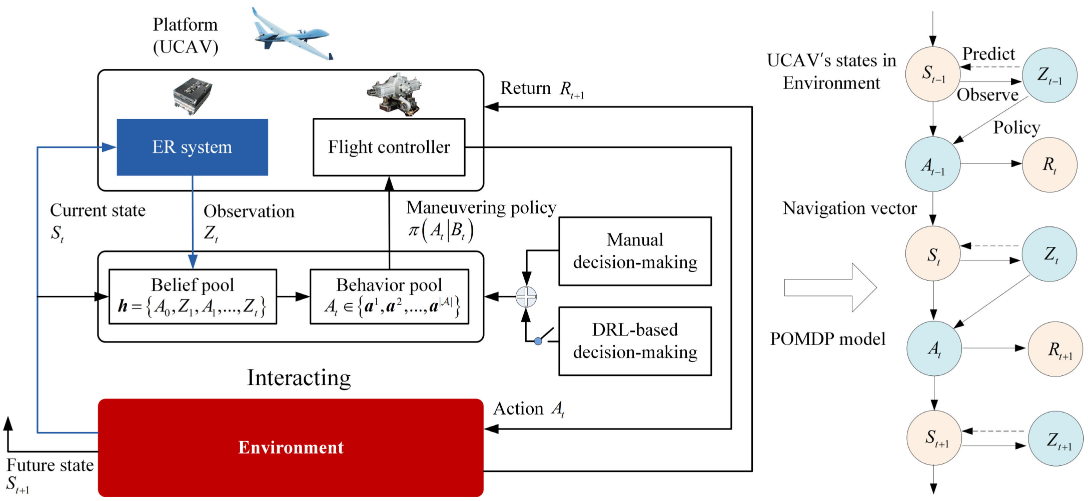


FIGURE 3. The logical framework for modeling a MER task in Scouter-N as POMDP. The left describes the UCAV's dynamic interaction in the Scouter-N environment, and the right displays the corresponding state transition of POMDP.

With the completion of state shaping, rewards must be designed to interpret the behaviors of pursuing benefits or avoiding threats. we present the reward shaping of avoiding threats as follows:

$$r_d=\begin{cases}-\sum_{m=1}^{M}\left(\left(\frac{\|\boldsymbol{p}_{mo,t}\|-\rho}{L_{map}-\rho}\right)_{\text{clip}}+\left(\frac{\rho'-\|\boldsymbol{p}_{mo,t}\|}{\rho'}\right)_{\text{clip}}\right), & \text{if } \bar{\kappa}_{m,t}=1\\ -\sum_{m=1}^{M}\left(\frac{\rho-\|\boldsymbol{p}_{mo,t}\|}{\rho}\right)_{\text{clip}}, & \text{else if } \bar{\kappa}_{m,t}=-1\\ 0, & \text{else}\end{cases} \tag{22}$$

where $(\cdot)_{\text{clip}}$is a operator for clipping variable values to the range [0, 1], and $L_{\text{map}} = \sqrt{L^2+W^2+H^2}$ is the maximum distance between any two points in the target area. $\rho = \frac{\lambda}{4\pi}\sqrt{\frac{Erp_iG_m}{P_{m,t}}} \sim \boldsymbol{p}_{mo,t}$ and ρ' are the agent-predicted reconnaissance radius of the UCAV and guidance radius of radar's air-defense firepower, respectively. Note that in this paper, we assume that the UCAV possesses the same reconnaissance radius for each radar and is larger than the threat radius of the radar's air-defense firepower because passive detection typically has much smaller distance attenuation than active detection, i.e., $\rho > \rho'$. Then, ρ can be further defined as $\rho' = w\rho, w = (0, 1)$.

The purpose of (22) is to enable the UCAV to agilely avoid threatening radars and enter an area where it can perform effective electronic reconnaissance on the m'-th radar while maintaining its own security. The calculation for pursuing benefits rewards is related to the UCAV's maneuvering policy. After intercepting the most important radar signals, unmanned aerial vehicles must keep their attitude stable so that the current target can be locked, thus We can design a reward function to reinforce this type of conduct:

$$r_p=\begin{cases}1-\sum_{m=1}^{M}\left(1-\frac{\langle \boldsymbol{l}_{m,t},\boldsymbol{v}_t\rangle}{90}\right)_{\text{clip}}, & \text{if } \|\boldsymbol{p}_{mo,t}\|\in[w\rho,\rho]\\ 0, & \text{else}\end{cases} \tag{23}$$

where $\langle \boldsymbol{l}_{m,t},\boldsymbol{v}_t\rangle \in (0, 90°)$ is the angle formed by the UCAV's velocity $\boldsymbol{v}_t$ and the LOS $\boldsymbol{l}_{m,t}$. $\|\langle \boldsymbol{l}_{m,t},\boldsymbol{v}_t\rangle - 90°\|$ can be referred to as the reconnaissance angle (RSA).

As shown in Figure 1, for the UCAV, the normal direction of the receiving antenna's beam is perpendicular to the its head direction, while the velocity vector is parallel to that direction. Therefore, in order to align the beam mainlobe of the receiving antenna with target radar, the maneuvering direction of the UCAV should be as perpendicular as possible to the LOS, as encouraged by (23).

Give the agent a one-time task completion reward as well:

$$r_a=\begin{cases}100, & \text{if } N_m = \mathrm{f_N}(\Delta T/\tau_m)\\ -100, & \text{else if Eq. (13) is satisfied}\\ 0, & \text{else}\end{cases} \tag{24}$$

Equation (24) indicates that 100 points will be awarded if the UCAV triumphs and 100 points will be deducted if it is destroyed. If a stalemate occurs, no points are awarded.

The state reward of the UCAV can be calculated at any time by summing the results of the three aforementioned reward functions, i.e., $R_t = r_d + r_p + r_a$.

C. DECISION-MAKING NETWORK BASED ON DRL

Using the network architecture of traditional DRL algorithms along with the state, action, and reward functions designed in the previous sections [25], we propose a new DRL network to solve POMDP problems such as MER.

TABLE 2. **Software and hardware in the testing environment.**

Name	Version/Model	Description
CPU	Intel Core i5-10400 (2.9 GHz)	Central processing unit
GPU	NVIDIA GeForce GTX 1660	Graphics processing unit
Memory	DDR4 RDIMM (16 GB)	Storage
Python	v3.9	Programming Language
Pytorch	v1.13.1	An open-source machine learning library
Pyglet	v1.5.26	A multimedia framework under Python

At time step t, the ideal input state is S_t, a $(4M+3)$-dimensional vector based on the UCAV's observation, and the output action is $A_t = [\vartheta_t, \phi_t]$, a two-dimensional vector based on $\pi(\cdot|B_t)$. Observation-based estimation of the UCAV's relative displacement between radars is $\boldsymbol{p}'_{mo,t} = \boldsymbol{p}'_{m,t} - \boldsymbol{p}_{o,t}$. Then, it can be inferred that $\boldsymbol{p}'_{mo,t} \sim \mathcal{N}\left(\boldsymbol{p}_{mo}, \delta_m\right)$, and $Z_t \sim S_t$ holds using (8), (16), and (20). Furthermore, based on the belief state defined in (17), we will use a generative model to estimate the probability distribution of the real states and use a particular length of $\boldsymbol{h}$ as input sample to predict $\boldsymbol{s}$. Simultaneously, consider employing a Gaussian distribution to simplify the analysis of the model, i.e., $S_t \sim \mathcal{N}\left(f_{\theta_S^\mu}(B_t), f_{\theta_S^\Sigma}(B_t)\right)$, where $\theta_S = \{\theta_S^\mu, \theta_S^\Sigma\}$ represents the parameters of the generative network, which corresponds to the mean and variance of the states. Similarly, the reward function associated with the reconnaissance radius ρ must be estimated by the sequence $\boldsymbol{h}$, i,e., $\rho \sim \mathcal{N}\left(f_{\theta_\rho^\mu}(B_t), f_{\theta_\rho^\Sigma}(B_t)\right)$.

Notably, ρ' in the reward function is highly dependent on the $P_{m,min}$ of the mth radar, which cannot be measured and has no prior knowledge and can only be learned through the consequences of being shot down by air-defense firepower. In practice, however, this is untenable because the cost of obtaining posterior information is too high and the environmental feedback is too sparse, neither of which is conducive to the agent's learning. To achieve a flexible estimation of ρ' and assist the agent in completing reconnaissance tasks to the greatest extent possible, we regard $\rho' - \rho = (1-w)\rho$ as a fixed tolerance that restricts the UCAV maneuverability. A greater w will make the reconnaissance behaviors of the agent more cautious and safe, but it will also make it difficult to search for optimal maneuvering policies.

Based on the configuration parameters of Scouter-N listed in Table 2, we can calculate that the maximum detection range of each radar to the UCAV is 25 km (ρ'=25 km), while the maximum reconnaissance range of the UCAV to each radar is 30 km (ρ=30 km), which means that the space tolerance necessary for the UCAV to conduct continuous electronic reconnaissance of the target radar is a circle of 5km in width. In light of the margin design, the value of w in this paper is set to 0.9.

We adopt the rapidly convergent soft actor-critic (SAC) algorithm as the primary body of the DRL network in Scouter-N in order to equip the agent with superior exploration capabilities and robustness [29]. The SAC algorithm was proposed in 2018 and has demonstrated exceptional performance in end-to-end continuous action control [30]. The network structure of the SAC consists of two critic networks, two target critic networks, and one actor network, with network parameters denoted by $\theta_{Q,i}$, $\theta'_{Q,i}$, $i \in \{1, 2\}$, and θ_π, respectively.

Each critic network can produce a precise estimation of Q-function by off-policy training using the following loss function:

$$\frac{1}{|\mathcal{D}|} \sum_{S_t,A_t,S_{t+1},r\sim\mathcal{D}} \left(Q(S_t, A_t \mid \theta_{Q,i}) - \Delta Q_{\min}\right)^2 \tag{25}$$

where $\mathcal{D}$ is a batch of transitions sampled from replay buffer with a size of $|\mathcal{D}|$, and $\Delta Q_{\min}$ is the targets of the Q functions constructed by the critic networks:

$$\begin{aligned}\Delta Q_{\min} = r + \gamma(&\min Q_{\theta_{Q,i}}(S_{t+1}, A_{t+1}) \\ &- \eta \log(\pi_{\theta_\pi}(A_{t+1}|S_{t+1})))\end{aligned} \tag{26}$$

where $Q_{\theta_{Q,i}}$ represents the action-value function's estimation for the ith critic network, and η represents the target entropy.

As replicas of the critic networks, the target networks themselves do not participate in training and instead update the network weights $\theta'_{Q,i}$ via a soft update [31].

The actor network is primarily utilized to improve policies, which can be updated by one step of gradient ascent using $\nabla_{\theta_\pi} \frac{1}{|\mathcal{D}|} \sum_{S_t\in\mathcal{D}} \Delta Q'_{\min}$, and $\Delta Q'_{\min}$ can be calculated as follows:

$$\Delta Q'_{\min} = \min Q_{\theta_{Q,i}}(S_t, \tilde{A}(S_t)) - \eta \log \pi_{\theta_\pi}(\tilde{A}(S_t)|S_t) \tag{27}$$

where $\tilde{A}(S_t)$ is a sample from the actor network $\pi_{\theta_\pi}(\cdot|S_t)$ via the reparameterization trick.

Note that both inputs of the critic network and the actor network involve the generative network's state prediction S_t, thus the original input of these networks will be B_t, consistent with the POMDP model described in (18) and (19). We train a network as our MER agent, the basic DRL network framework is shown in Figure 4, and the detailed hyperparameters and configurations of all networks in Figure 4 are given below.

- The generative network has four layers, from top to bottom, a input layer of $16M+40$ units, a long short term memory (LSTM) layer of 128 units, a fully connected hidden layer of 300 units and an output layer with a combination of four elements (two output sublayers of $4M + 3$ units and two output sublayers of 1 unit). The learning rate of the generative network is 0.002, the batch size for training is 128, the output activation function is ReLu, and the optimizer type is AdaDelta.
- Each critic network has four layers, with an input layer of $4M + 5$ units, two fully connected hidden layer of 300 units, and an output layer of 1 unit from bottom

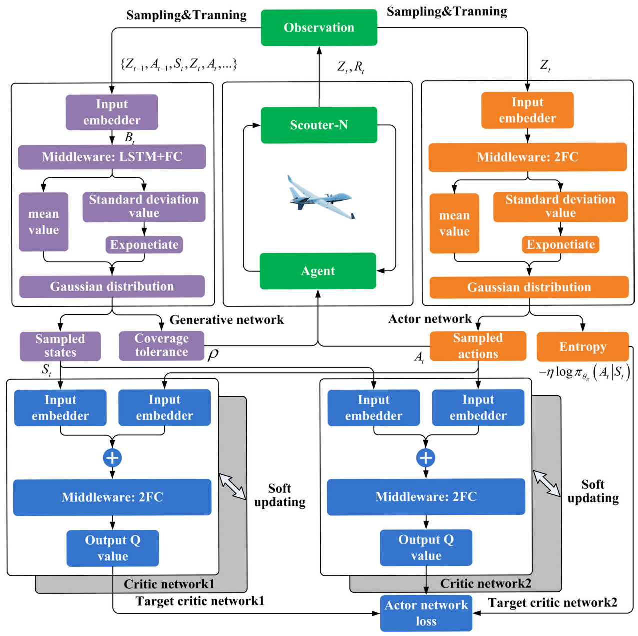


FIGURE 4. The DRL network framework.

to top. The learning rate of the critic network is 0.001, the learning rate of the target entropy is 0.01, the batch size for training is 128, the output activation function is ReLu, and the optimizer type is Adam.

- The actor network has four layers, with an input layer of $4M + 3$ units, two fully connected hidden layer of 300 units, and an output layer of 2 units from bottom to top. The learning rate of the actor network is 0.0004, the batch size for training is 128, the output activation function is Tahn, and the optimizer type is Adam.

IV. SIMULATION VERIFICATION

This section first introduces the simulation environment and the hardware and software configuration used in the experiment, then it designs multiple simulation scenarios of different difficulty according to the required MER metrics, and finally presents the results and analysis of these scenarios to verify the effectiveness of our proposed DRL network in Scouter-N.

Since the primary concern for DRL is the algorithm's computational burden, the software and hardware versions in the testing environment must be guaranteed to be consistent when comparing the performance of different algorithms under the same conditions. In the experiment, we use the software and hardware of the version/model shown in Table 2.

A. SIMULATION SETTINGS AND EVALUATION METRICS

Due to the concurrent amplification of state dimensions and learning samples, an increase in the number of targets in finite-time MER tasks not only raises the decision-making complexity of the agent, but also dramatically enhances the convergence difficulty of the DRL algorithm. Thus, three MER scenarios with radar numbers of 2, 3, and 4 are set up in Scouter-N, respectively, to investigate the variance in the performance of the agent and the characteristics of the reconnaissance behavior it mastered. Apparently, the difficulty of the task increases as the radar number increases.

TABLE 3. After training, the performance of the proposed DRL network at different tasks, metrics includes the SR (%), MTD (km), CT (h), and MDT (s).

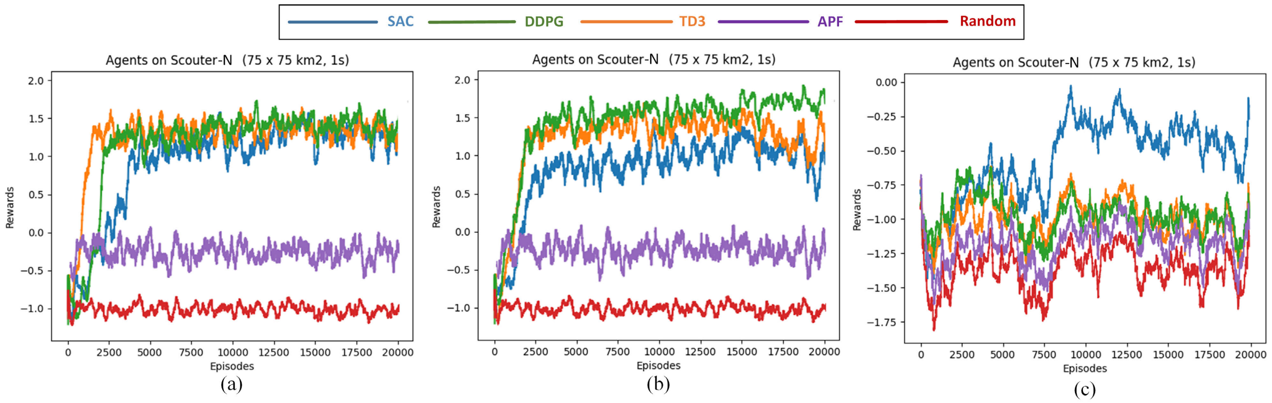
Number of radars	Metrics	Baseline	Algorithms			
		Random	APF	DDPG	TD3	SAC
2	SR (%)	0.2	10.7	83.8	**85.8**	84.5
	MTD (km)	170.6	125.1	83.3	**75.2**	78.5
	CT (h)	-	-	**24.7**	33.5	53.5
	MDT (s)	0.04	0.11	**0.32**	0.45	0.69
3	SR (%)	0	2.9	**81.1**	79.7	78.6
	MTD (km)	-	165.0	**113.5**	115.9	117.5
	CT (h)	-	-	**30.4**	39.8	58.0
	MDT (s)	0.04	0.12	**0.35**	0.49	0.77
4	SR (%)	0	0.7	15.4	12.6	**44.5**
	MTD (km)	-	196.8	164.6	149.5	**139.4**
	CT (h)	-	-	53.2	**52.6**	68.2
	MDT (s)	0.05	0.12	**0.51**	0.56	0.87

Based on a specific random seed, 21000 randomly initialized episodes are carried out for each task, and each episode is run independently for no more than 1000 operating steps. The first 20000 episodes are used to train the DRL-based agent while the remaining 1000 episodes are used to test it.

A real-time CEW requires signal processing to be completed in 1 s after the UCAV intercepts the target radar signal. The operating cycle of all tasks in Scouter-N is therefore set at 1 s, and the maximum time $T_{\max}$ allowed for each task is 1000 s. According to the conclusion in Section II-A, the UCAV must choose a greater ΔT as much as possible while meeting the condition $\Delta T < T_{\max}$ to complete MER, thus We compromise between following the practical situation and the convergence ease of DRL algorithms by setting ΔT to 100 s, and N_m=100 s/1 s=100. Additionally, if the UCAV reconnaissance of each target radar lasts for 100 cycles or the UCAV is destroyed by anti-aircraft missiles, the episode will be regarded as ending ahead of time.

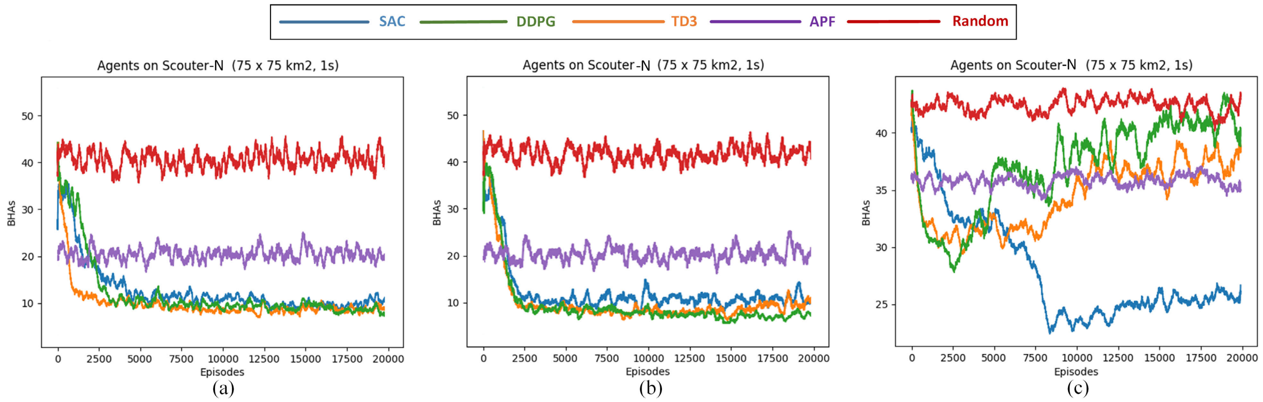
To improve the generalizability of our proposed DRL network, it is emphasized that the state information of all objects in Scouter-N will be initialized randomly at the beginning of each episode under the following reset conditions:

$$\begin{cases} \kappa_{m,t} = 0 \\ ||\boldsymbol{p}_{mo,t}||q_t > \Delta d_m \\ \dfrac{E_m \lambda_m^2 \sigma}{(4\pi)^3 \, ||\boldsymbol{p}_{mo,t}||^4} < P_{m,\min} \end{cases}, m = 1, 2, \ldots, M \qquad (28)$$



(a) (b) (c)

FIGURE 5. Evaluating the DRL network's convergence performance in Scouter-N. The three subfigures of MAR represent the results of various tasks.



(a) (b) (c)

FIGURE 6. Evaluating the reconnaissance behaviors learned by agents in Scouter-N. The three subfigures of RSA represent the results of various tasks.

The following five metrics are employed to evaluate the experiments for the three different scenarios mentioned above:

1) **Success rate (SR)**: SR refers to the proportion of reconnaissance tasks that are completed successfully, and its value can only be obtained after all test episodes are exhausted.
2) **Mean traveled distance (MTD)**: MTD refers to the mean distance traveled by the agent to complete a task, and its value can only be obtained after all test episodes are exhausted.
3) **Computation time (CT)**. CT refers to the time necessary for the DRL network to complete all training episodes, and its value can only be obtained after all training episodes are exhausted.
4) **Mean decision-making time (MDT)**. MDT is the time it takes the agent to make a reconnaissance decision. MDT is a measure of the processing time of the algorithm at a single time step for each test episode and can be used to characterize the algorithm's complexity.
5) **Mean accumulated reward (MAR)**. MAR is used to evaluate the DRL algorithms' convergence in different scenarios, and its calculation method is given in [22].

A simulation test of three scenarios is run in Scouter-N with the traditional algorithms for artificial potential field (APF) [32], the regular deep deterministic policy gradient (DDPG) [33], and the twin delayed DDPG (TD3) as the comparison calculation examples [34]. Additionally, we use a completely random strategy as a baseline in the comparative experiment to verify the learning and cognitive abilities of the agent.

B. CONVERGENCE ANALYSIS

During the training episode, we focus on the change in convergence of a continuous metric like MAR. Effective agents tend to show a convergence tendency before exhausting all episodes. In other words, if the agent's MAR value does not converge within an appropriate number of iterations, the agent is not competent for the current task.

Figure 5 shows the convergence results of the algorithms on MAR. The detailed performance results of our DRL network in three tasks, such as SR and MTD values, are reported in Table 3. Although the number of radar stations and the difficulty of the MER task are directly connected, DRL-based agents are able to execute electronic reconnaissance tasks with a high SR, which is at least 43.8% higher than traditional algorithms and up to 75.1% higher. Unfortunately, the MAR curve does not converge in the given episodes when four radars emerge in the Scouter-N environment. A plausible explanation for this is that as the number of radars gradually increases, the percentage of the radar air-defense area compared to the overall task space increases, i.e., the maneuverable space is compressed relative to the UCAV, and the priority-driven reward shaping tends to trap the UCAV's maneuvering policy in a local optimum. Overall, our DRL network is ultimately able to train the most potent UCAV despite the severe performance loss in the most challenging task.

Note that, although the traditional algorithms almost ineffective in Scouter-N, the deterministic-policy-based DDPG and TD3 perform exceptionally well in the less difficult two types of tasks [33], [34], and the SAC has slightly inferior performance due to the enormous amount of action exploration required [29]. However, in complex MER scenarios with four radars, due to insufficient action exploration, DDPG and TD3 will overemphasize low-reward policies, resulting in far inferior performance compared to SAC.

Although we propose a network framework that is compatible with multiple DRL algorithms, its actor network becomes more complicated than DDPG and TD3 due to the computational requirements of the SAC for probability distribution generation and action sampling. As verified in Table 3, all DRL algorithms have significant time consumption, and DDPG consumes less due to its simpler network structure and faster convergence. Moreover, judging from the changes in MTD and MDT values, the SAC algorithm is better adapted to MER.

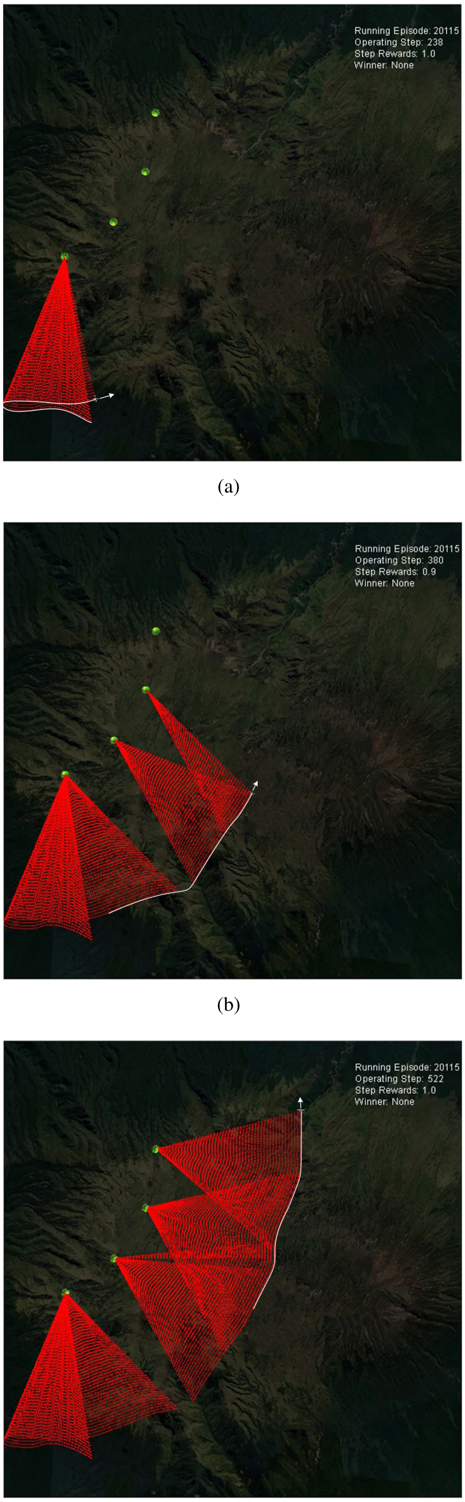
C. BEHAVIOR ANALYSIS

Table 2 reveals that the DRL-based agent has the most advanced MER capabilities in Scouter-N, but we still wish to investigate what kind of behavioral convergence drives it to win and possess a high MAR value.

In this section, we employ the RSA mentioned in (23) as the key metric to perform a behavior analysis for the UCAV. Using data smoothing, Figure 6 illustrates that the trained UCAV's reconnaissance behavior changes significantly.

The mean RSA value for the benchmark strategy is 45° since it is completely random. The APF optimizes the local policy based on vector navigation under limited capability to reduce its RSA value compared to the baseline. From the three subfigures in Figure 6, we can determine that the DRL-based approaches enable the agent to understand RSA minimization.

To guarantee that the target radar signal can be intercepted continuously throughout the MER task, the converged RSA value oscillates around 10°, indicating that the UCAV attempts to stabilize the orientation of its receiving antenna mainlobe by attitude control (little oscillations in RSA



(a)

(b)

(c)

FIGURE 7. Reconnaissance trajectories of the agent in a MER task with four radars. The white lines represent the flying path of the agent, while the red dashed line indicates that the agent is intercepting radar signals and has not been detected by any radar. The starting and ending operating steps of the corresponding trajectory segments in three subgraphs involving: (a) 1st to 238th; (b) 239th to 380th; (c) 381st to 522nd.

indicate this same trend, even in the most challenging task that ended in failure). Although the ideal RSA value from the perspective of god should be 0, the existence of observation errors makes this wish impossible to achieve. Furthermore, a smaller RSA can enable the UCAV to perform circling maneuvers, thereby allowing it to keep a safe reconnaissance distance from the target.

We can extract and analyze the maneuvering patterns of electronic reconnaissance preferred by the agent by segmenting high-quality trajectories, a MER task of four radars illustrated in Figure 7 is a good example:

(a) At the beginning of the task, the trained agent first aims at the closest and highest-priority target for a quick approach, then quickly adjusts its posture, conducting reconnaissance in the form of a lateral circling flight and turning back within a certain range, and smoothly follows and locks in the leftmost radar.

(b) The agent detects a new radar in the forward direction, but due to the incomplete reconnaissance of the current target, it chooses a threat avoidance maneuver after evaluating the priority of the new target while maintaining an effective reconnaissance distance from the current target until the 100 times of electronic reconnaissance have been completed. The agent then moves with the same agility into the reconnaissance area of the next target in a lateral surround.

(c) Maintaining a safe distance from each radar, the agent continuously traverses the reconnaissance area of the 2nd, 3rd, and 4th radars using horizontal maneuvers. Note that if the agent intercepts signals from other radars before the locking of the current radar is over, the agent will maneuver a short distance away from the target area after weighing the pros and cons of continuous reconnaissance and threat avoidance.

By comparing the trajectory segments in Figure 7 (a), (b), and (c), we conclude that circling and reciprocating turnback are the preferred behaviors for the agent in Scouter-N, as they allow for a longer and more stable MER with simpler maneuvers while ensuring that the mainlobe of the receiving antenna is always aligned with the target radar.

V. CONCLUSION

In this paper, we innovatively design a DRL network to solve the problem of regional MER. Based on the test results at three difficulty levels, the superior adaptability of the SAC algorithm's network for MER is verified. Meanwhile, we are surprised to find that the trained agent's electronic reconnaissance behaviors match those of artificially manipulated aircraft after behavior analysis on high-quality trajectories. Unfortunately, because the behaviors of avoiding threats are frequently coupled with the behaviors of pursuing benefits, even though priority-driven rewards enable the agent to work out MER quickly, this insufficiently greedy strategy makes it difficult for the agent to excel in MER tasks with dense radars. Consequently, the two most significant challenges to be addressed for future MER implementations utilizing artificial intelligence are designing more instructive reward functions and creating DRL algorithms with better action exploration.

ACKNOWLEDGMENT

The authors would like to thank the No. 8511 Research Institute for supporting them during the development of the Scouter-N simulator.

REFERENCES

[1] L. Swartzentruber, J. L. Foo, and E. Winer, "Multi-objective UAV path planning with refined reconnaissance and threat formulations," in *Proc. 51st AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn., Mater. Conf., 18th AIAA/ASME/AHS Adapt. Struct. Conf. 12th*, 2010, p. 2758.

[2] Z. Shi, X. Huang, Y. Hua, and D. Xu, "Statistical physics method for multi-base multi-UAV cooperative reconnaissance mission planning," in *Proc. IEEE Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Dec. 2015, pp. 64–68.

[3] P. Sun and A. Boukerche, "Performance modeling and analysis of a UAV path planning and target detection in a UAV-based wireless sensor network," *Comput. Netw.*, vol. 146, pp. 217–231, Dec. 2018.

[4] I. Mahmud and Y.-Z. Cho, "Detection avoidance and priority-aware target tracking for UAV group reconnaissance operations," *J. Intell. Robotic Syst.*, vol. 92, no. 2, pp. 381–392, Oct. 2018.

[5] J. Happe and J. Berger, "CoUAV: A multi-UAV cooperative search path planning simulation environment," in *Proc. SummerSim*, 2010, pp. 86–93.

[6] W. H. van Willigen, M. C. Schut, A. Eiben, and L. J. Kester, "Online adaptation of path formation in UAV search-and-identify missions," in *Proc. Int. Conf. Adapt. Natural Comput. Algorithms*. Ljubljana, Slovenia: Springer, 2011, pp. 186–195.

[7] G. Varela, P. Caamaño, F. Orjales, Á. Deibe, F. López-Peña, and R. J. Duro, "Autonomous UAV based search operations using constrained sampling evolutionary algorithms," *Neurocomputing*, vol. 132, pp. 54–67, May 2014.

[8] K. Obermeyer, "Path planning for a UAV performing reconnaissance of static ground targets in terrain," in *Proc. AIAA Guid., Navigat., Control Conf.*, Aug. 2009, p. 5888.

[9] L. Lin and M. A. Goodrich, "Hierarchical heuristic search using a Gaussian mixture model for UAV coverage planning," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2532–2544, Dec. 2014.

[10] A. Majeed and S. Lee, "A new coverage flight path planning algorithm based on footprint sweep fitting for unmanned aerial vehicle navigation in urban environments," *Appl. Sci.*, vol. 9, no. 7, p. 1470, Apr. 2019.

[11] Y. Cao, "UAV circumnavigating an unknown target under a GPS-denied environment with range-only measurements," *Automatica*, vol. 55, pp. 150–158, May 2015.

[12] L. Gao, X. Zhang, J. Gao, and S. You, "Fusion image based radar signal feature extraction and modulation recognition," *IEEE Access*, vol. 7, pp. 13135–13148, 2019.

[13] Y. Ren, W. Jiang, and Y. Liu, "Complex-valued parallel convolutional recurrent neural networks for automatic modulation classification," in *Proc. IEEE 25th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2022, pp. 804–809.

[14] K. Chen, J. Zhang, S. Chen, and S. Zhang, "Deep metric learning for robust radar signal recognition," *Digit. Signal Process.*, vol. 137, Jun. 2023, Art. no. 104017.

[15] E. Dimperio, G. Gunzelmann, and J. Harris, "An initial evaluation of a cognitive model of UAV reconnaissance," in *Proc. 17th Conf. Behav. Represent. Model. Simul.*, 2008, pp. 165–173.

[16] D. Ebrahimi, S. Sharafeddine, P. Ho, and C. Assi, "Autonomous UAV trajectory for localizing ground objects: A reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1312–1324, Apr. 2021.

[17] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826–3839, Sep. 2020.

[18] T. M. Ho, K. Nguyen, and M. Cheriet, "UAV control for wireless service provisioning in critical demand areas: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 7138–7152, Jul. 2021.

[19] D. Wu, Y. Lei, M. He, C. Zhang, and L. Ji, "Deep reinforcement learning-based path control and optimization for unmanned ships," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–8, May 2022.

[20] S. You, M. Diao, and L. Gao, "Completing explorer games with a deep reinforcement learning framework based on behavior angle navigation," *Electronics*, vol. 8, no. 5, p. 576, May 2019.

[21] S. You, M. Diao, and L. Gao, "Deep reinforcement learning for target searching in cognitive electronic warfare," *IEEE Access*, vol. 7, pp. 37432–37447, 2019.

[22] S. You, M. Diao, L. Gao, F. Zhang, and H. Wang, "Target tracking strategy using deep deterministic policy gradient," *Appl. Soft Comput.*, vol. 95, Oct. 2020, Art. no. 106490.

[23] Z. Mou, Y. Zhang, F. Gao, H. Wang, T. Zhang, and Z. Han, "Deep reinforcement learning based three-dimensional area coverage with UAV swarm," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3160–3176, Oct. 2021.

[24] N. Imanberdiyev, C. Fu, E. Kayacan, and I. Chen, "Autonomous navigation of UAV by using real-time model-based reinforcement learning," in *Proc. 14th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2016, pp. 1–6.

[25] P. Zhu, X. Li, P. Poupart, and G. Miao, "On improving deep reinforcement learning for POMDPs," 2017, *arXiv:1704.07978*.

[26] S. Bhattacharya, S. Badyal, T. Wheeler, S. Gil, and D. Bertsekas, "Reinforcement learning for POMDP: Partitioned rollout and policy iteration with application to autonomous sequential repair problems," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 3967–3974, Jul. 2020.

[27] G. Singh, S. Peri, J. Kim, H. Kim, and S. Ahn, "Structured world belief for reinforcement learning in POMDP," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 9744–9755.

[28] Z. Liu, C. Liu, W. Zhao, and A. Li, "A user-priority-driven multi-UAV cooperative reconnaissance strategy," *Int. J. Aerosp. Eng.*, vol. 2021, pp. 1–14, Oct. 2021.

[29] 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.*, 2018, pp. 1861–1870.

[30] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.

[31] E. Marchesini, D. Corsi, and A. Farinelli, "Genetic soft updates for policy evolution in deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–15.

[32] T. Paul, T. R. Krogstad, and J. T. Gravdahl, "Modelling of UAV formation flight using 3D potential field," *Simul. Model. Pract. Theory*, vol. 16, no. 9, pp. 1453–1462, Oct. 2008.

[33] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[34] S. Bai, S. Song, S. Liang, J. Wang, B. Li, and E. Neretin, "UAV maneuvering decision-making algorithm based on twin delayed deep deterministic policy gradient algorithm," *J. Artif. Intell. Technol.*, vol. 2, no. 1, pp. 16–22, Dec. 2022.

YONGLE XU is currently pursuing the bachelor's degree in the Internet of Things engineering with the Jiangsu University of Science and Technology, China. He is also participating in a number of national research programs under the supervision of college teachers. His research interests include machine learning, evolutionary computation, and the cross-application of deep reinforcement learning and electronic warfare.

MING ZHANG received the B.S. and M.S. degrees in computer science from the Jiangsu University of Science and Technology, in 2002 and 2005, respectively, and the Ph.D. degree in pattern recognition and machine intelligence from the Nanjing University of Science and Technology, in 2013. He is currently an Associate Professor with the School of Computer, Jiangsu University of Science and Technology. His current research interests include granular computing, pattern recognition, reinforcement learning, and bioinformatics.

BOYIN JIN received the master's degree from the Graduate School of Engineering, Hiroshima University, Japan, and the Ph.D. degree in engineering from Hiroshima University, specializing in reinforcement learning and collective intelligence. He is currently a Faculty Member with the School of Computer Science, Jiangsu University of Science and Technology. In his research, he is dedicated to conceptualizing, designing, and implementing deep reinforcement learning algorithms to train artificial intelligence to exhibit intelligent behavior. Since 2016, he has been conducting research in various fields of machine learning and artificial intelligence. He collaborates with multiple international institutions, including Hiroshima University and Toyama University, Japan.