

A Reinforcement Learning Approach to Decision Making of Phased Array Radar Jamming

Lihui Huang¹, Changhua Hu^{*2}

¹*Rocket Force University of Engineering*

Xian, China

Wh201410@139.com

²*Rocket Force University of Engineering*

Xian, China

Happyday202411@163.com

Abstract

With the rapid advancement of phased array radar, the guidance and anti-jamming capabilities of phased array radar seekers have been further enhanced. Traditional radar jamming decision-making methods are no longer applicable in electronic warfare. Thus, radar jamming decision-making methods based on reinforcement learning have emerged, which can effectively address the issues of poor performance and low efficiency of traditional jamming approaches. Given the extremely high cost and poor repeatability of physical experiments, the construction of simulation models is frequently adopted for simulation experiments. Unlike most other studies on radar jamming decision-making that employ functional-level simulation modeling, this paper adopts signal-level simulation of radar jamming decision-making, presenting a more realistic and intuitive reflection of the entire process of interference equipment interfering with the missile terminal guidance. Through the signal-level simulation of phased array radar terminal guidance and the introduction of reinforcement learning algorithms, this study investigates how the interference equipment should act to maximize the interference benefit. Through simulation experiments, on the one hand, it is demonstrated that the reinforcement learning algorithm can improve the interference effect. On the other hand, the accuracy of the signal-level simulation model is verified.

Keywords—phased array radar, radar jamming decision-making, signal-level simulation, reinforcement learning

I. INTRODUCTION

In the increasingly complex electronic warfare environment characterized by rapid technological advancements and sophisticated adversarial tactics, traditional radar systems that rely on a single working mechanism and simplistic beam variations are becoming inadequate [1], [2]. These conventional radars struggle to meet the diverse operational requirements of modern warfare scenarios, which demand high levels of adaptability and precision in detection capabilities. Phased array radar represents a significant evolution in radar technology. This type of radar system is designed to dynamically alter its real-time beam direction by adjusting the emitted signals' phase from multiple antennas arranged in an array configuration [3]. The ability to electronically steer the radar beam allows for faster target acquisition and tracking compared to mechanical scanning methods. Consequently, phased array radars are often referred to as electronic scanning array (ESA) radars due to their advanced electronic control mechanisms. Currently, phased array radar has undergone substantial development, evolving into multifunctional systems capable of engaging with various targets across extensive ranges while maintaining high reliability under different operational conditions. These features make them indispensable components within advanced weapon systems utilized by military forces worldwide. Their versatility enables simultaneous engagement with multiple threats—ranging from aircraft and missiles to ground-based assets—thereby enhancing situational awareness on the battlefield. Traditional methods employed for making interference decisions regarding these radars include techniques based on template matching, game theory approaches that analyze strategic interactions between competing entities, and reasoning-based methodologies that utilize logical frameworks [4]. However, all these conventional strategies necessitate a considerable amount of prior data for effective decision-making support. They often fall short when faced with multifunctional radars operating within unknown or unpredictable environments where historical data may be limited or irrelevant. In contrast, interference decision-making methods grounded in reinforcement learning offer distinct advantages over traditional approaches. Reinforcement learning algorithms possess inherent cognitive abilities that allow them not only to learn from past experiences but also adaptively explore optimal interference strategies through continuous trial-and-error processes. By leveraging feedback from previous actions taken during operations against

dynamic adversarial tactics, these algorithms can refine their decision-making capabilities over time without requiring exhaustive datasets upfront.

Q. Xing applied reinforcement learning to intelligent radar confrontation, and the experimental results indicated that reinforcement learning possesses excellent adaptive capabilities in one-on-one confrontations [5]. B. K. Zhang constructed a cognitive interference decision model from the perspective of signal processing, updated the types of interference libraries, analyzed the assessment of interference effects, and played a significant role in the cognitive confrontation of multifunctional radars [6]. S. Y. Zhang combined game theory with reinforcement learning and proposed a multi-agent reinforcement learning algorithm based on equilibrium, which could effectively address the issue of radar observation time scheduling and further demonstrate the effectiveness of reinforcement learning in handling interference decision problems [7]. F. Slimeni proposed an interferer scheme featuring spectrum sensing, offline training, and learning functions. The RL-based interference decision algorithm was simulated on a universal software radio platform. The results indicated that the interferer can effectively learn and interfere without prior information from users [8]. L. Yunjie and X. Qiang introduced prior knowledge based on reinforcement learning and improved the performance of reinforcement learning in radar interference decision-making [9]. X. Qiang, Z. Weigang, and J. Xin proposed an adaptive reinforcement learning algorithm that can solve the interference problem of radars with unknown working modes [10]. W. G. Zhu, confronted with the increasing working modes and states of radars, adopted the deep reinforcement learning approach and effectively addressed the issue of low decision-making efficiency of reinforcement learning [11].

Most of the previous studies on radar interference decision-making methods have employed functional-level simulation approaches, merely simulating the amplitude information of signal transmission, targets, echoes, clutters, and interference signals or presenting a straightforward description of the radar's working status without outputting genuine radio frequency and video signals. It has the typical merits of simplicity, practicality, and strong real-time processing ability. However, it neglects numerous details inherent in waveforms and signal processing, fails to accurately depict the dynamically changing interference confrontation environment created by various electronic interference equipment of both the friendly and hostile sides, and is even less capable of comprehensively and quantitatively elaborating the mechanism of action of various interference factors on the radar system. This paper adopts the coherent video signal analysis method, with the background of the interference equipment on ships countering anti-ship missile attacks, and simulates the entire process of radar interference decision-making from the signal-level perspective, providing a novel idea for subsequent studies on radar interference decision-making.

II. MODELING OF PHASED ARRAY RADAR SEEKER

A. Signal transmission and signal reception model

1) Antenna model

The general antenna radiation pattern function is shown below:

$$F(\alpha, \beta) = \left| \frac{\sin\left(\pi \frac{\alpha}{\theta_\alpha}\right)}{\pi \frac{\alpha}{\theta_\alpha}} \right| \cdot \left| \frac{\sin\left(\pi \frac{\beta}{\theta_\beta}\right)}{\pi \frac{\beta}{\theta_\beta}} \right| \quad (1)$$

where α and β represent the azimuth and elevation offsets between the target and the antenna center, expressed in radians; θ_α and θ_β represent the azimuth and elevation beamwidths of the antenna, also expressed in radians. The beamwidth parameter can be estimated directly from the antenna size as follows:

$$\theta_{\alpha, \beta} = \frac{\lambda}{L_{\alpha, \beta}} \quad (2)$$

where L_α and L_β are the lengths of the antenna in the azimuth and elevation dimensions, and λ is the wavelength of the transmitted signal. In phased array antenna mode, the beamwidth will widen if the antenna spacing cannot meet the requirement of $\Delta d \leq \lambda/2$. Meanwhile, the antenna gain can also be estimated simply by the antenna size as follows:

$$G = \frac{4\pi L_\alpha L_\beta}{\lambda^2} \quad (3)$$

In the three-dimensional orientation, the amplification power varies in different directions. The angular information of the antenna pointing to the target can be solved by the sum-and-difference beam method. The directional patterns of the four sub-beams are respectively recorded as:

$$O_1: F_1(\alpha, \beta) = G(-\Delta_1, \Delta_2, \alpha, \beta, \alpha_0, \beta_0) \quad (4)$$

$$O_2: F_2(\alpha, \beta) = G(\Delta_1, \Delta_2, \alpha, \beta, \alpha_0, \beta_0) \quad (5)$$

$$O_3: F_3(\alpha, \beta) = G(-\Delta_1, -\Delta_2, \alpha, \beta, \alpha_0, \beta_0) \quad (6)$$

$$O_4: F_4(\alpha, \beta) = G(\Delta_1, -\Delta_2, \alpha, \beta, \alpha_0, \beta_0) \quad (7)$$

where α_0 and β_0 are respectively the center pointing of the phased array antenna (coordinates in the coordinate system of the array surface azimuth), α and β are the coordinates in the coordinate system of the pointing azimuth and Δ_1 and Δ_2 are also the coordinates in the coordinate system of the pointing azimuth. Then, the sum beam pattern is as follows:

$$F_{\Delta 1}(\alpha, \beta) = F_1(\alpha, \beta) + F_2(\alpha, \beta) + F_3(\alpha, \beta) + F_4(\alpha, \beta) \quad (8)$$

The azimuthal difference beam direction pattern is:

$$F_{\Delta 1}(\alpha, \beta) = F_1(\alpha, \beta) + F_2(\alpha, \beta) - F_3(\alpha, \beta) - F_4(\alpha, \beta) \quad (9)$$

The elevation-azimuth direction beam pattern is:

$$F_{\Delta 2}(\alpha, \beta) = F_1(\alpha, \beta) - F_2(\alpha, \beta) + F_3(\alpha, \beta) - F_4(\alpha, \beta) \quad (10)$$

Furthermore, the proportional coefficient needed for angle error signal resolution in the angle measuring module can be obtained as follows:

$$\mu_1 = \frac{F_{\Sigma}(0,0)}{\left. \frac{dF_{\Delta 1}}{d\alpha} \right|_{\alpha=0, \beta=0}}, \mu_2 = \frac{F_{\Sigma}(0,0)}{\left. \frac{dF_{\Delta 2}}{d\beta} \right|_{\alpha=0, \beta=0}} \quad (11)$$

2) Signal transmission modeling

To accurately simulate the strike process of the terminal guidance of a radar seeker in complex electromagnetic environments, signal-level simulation of the radar seeker is necessary. Coherent video signal simulation technology is a type of signal-level simulation. This technology simulates the entire process of signal reception and processing through video signals[12]. The time-domain and frequency-domain characteristics of radar signals can typically be expressed using analytical mathematical expressions, as shown in equation 1, thereby generating the signal's I and Q channel signals.

$$\begin{aligned} S(t) &= A(t) \cdot e^{j(\omega_0 t + \phi(t))} = [A_c(t) + jA_s(t)] \cdot e^{j\omega_0 t} \\ &= u(t)e^{j2\pi f_0 t} \end{aligned} \quad (12)$$

For the envelope $a(t)$ of the real signal $x(t)$, the envelope of its exponential signal $S(t)$ can be depicted as:

$$u(t) = a(t) \cdot e^{j\theta(t)} \quad (13)$$

Therefore, the signal $x(t)$ can be written as:

$$\begin{aligned} x(t) &= a(t)\cos[2\pi f_0 t + \theta(t)] = \frac{1}{2}[S_e(t) + S_e^*(t)] \\ &= \text{Re}[S_e(t)] \end{aligned} \quad (14)$$

Meanwhile, the signal $x(t)$ can also be written as:

$$x(t) = a_i(t)\cos(2\pi f_0 t) - a_q(t)\sin(2\pi f_0 t) \quad (15)$$

$$a_I(t) = a(t)\cos\theta(t) \quad (16)$$

$$a_Q(t) = a(t)\sin\theta(t) \quad (17)$$

In the above formula, a_I and a_Q are the in-phase and quadrature components of the baseband signal, respectively. This is the signal sampling sequence that we need to generate. Correspondingly, when the carrier frequency is eliminated through downconversion (quadrature demodulation) processing to obtain the zero intermediate frequency (coherent video) signal as:

$$S_0(t) = A(t) \cdot e^{j\phi(t)} = A_c(t) + jA_s(t) \quad (18)$$

When using full digital simulation, the coherent video signal is sampled at a certain sampling rate, thus forming a sequence of coherent video signals (complex signals) that can be represented as follows:

$$s[n] = s_r[n] + js_i[n] \quad (19)$$

Typical phased array radar transmits signals using an LFM (linear frequency modulation) signal pattern. The radar transmitted signal is represented by a complex signal as follows:

$$S_t(t) = \sqrt{\frac{2P_t}{4\pi L_t}} g_{vt}(\theta) v(t) \cdot e^{j\omega_{ck}t} \omega_{ck} \quad (20)$$

where ω_{ck} denotes the current pulse carrier frequency, P_t indicates the peak power of the transmitter, L_t stands for the combined loss of the transmitter, $g_{vt}(\theta)$ represents the transmitting antenna pattern (voltage gain), and $v(t)$ represents the complex modulation function as follows:

$$v(t) = \text{Rect}\left(\frac{t}{T_p}\right) \cdot \exp\left(j\pi F_m t^2\right) \quad (21)$$

the rectangular function is defined as:

$$\text{Rect}(t) = \begin{cases} 1, & t \in (0,1) \\ 0, & \text{else} \end{cases} \quad (22)$$

where T_p represents the pulse width, and F_m represents the frequency modulation slope. Based on the above formulas, the coherent video signal pattern adopted in the system is as follows:

$$s_t(t) = \sqrt{\frac{P_t}{4\pi L_t}} g_{vt}(\theta) \cdot \text{Rect}\left(\frac{t}{T_p}\right) \cdot \exp\left(j\pi \frac{BW_{rg}}{T_p} t^2\right) \quad (23)$$

3) Receiving Signal Model

The received signal mainly consists of target echoes, interference signals, various types of clutter, and receiver noise. This paper mainly considers the aspects of target echoes and interference signals. With respect to a specific transmitted pulse, the RF signal received by the radar can be represented as follows:

$$r_{RF}(t) = S_{RF}(t) + J_{RF}(t) + n_{RF}(t) \quad (24)$$

where $S_{RF}(t)$ represents the echo signal received after the transmitted pulse is reflected by the target, $J_{RF}(t)$ represents the received interference signal, which is the combined interference signal formed by various active interference and passive interference, and $n_{RF}(t)$ represents the receiver noise. The receiver noise is Gaussian-limited white noise, which means that in the radar receiver's passband, the power spectral density of the noise is uniform, and its amplitude follows a Rayleigh distribution. The variance of the noise can be calculated from the receiver noise coefficient and the receiver bandwidth. In the simulation, a sample function of a Gaussian process can be used to represent it specifically, and the band-limited noise signal is represented as:

$$n(t) = \text{Re}[\tilde{n}(t) \cdot e^{j\omega_c t}] \quad (25)$$

Therefore, in coherent video simulation, the noise at the receiver can be represented as:

$$\tilde{n}(t) = n_d(t) - jn_q(t) \quad (26)$$

In this model, $n_d(t)$ and $n_q(t)$ are independent Gaussian random processes with zero mean and variance σ_N^2 . The variance of the noise σ_N^2 can be calculated from the receiver noise coefficient N_F and receiver bandwidth Δf as follows:

$$\sigma_N^2 = kT_0 N_F \Delta f \quad (27)$$

where K is the Boltzmann constant, and T_0 is the reference temperature of the receiver, which $T_0 = 290$ K. Combining the target echo signal and the receiver thermal noise signal, we finally obtain the radar received signal as:

$$r(t) = \text{Rect}\left(\frac{t - \frac{2R}{c}}{T_p}\right) \cdot \left[\sqrt{\frac{2P_t}{(4\pi)^3 L}} \frac{g_{vt}(\theta)g_{vr}(\theta)}{R^2} \lambda_k \sqrt{\sigma} \right] \cdot \exp\left[j\pi \frac{BW_{rg}}{T_p} \left(t - \frac{2R}{c}\right)^2\right] \cdot \exp\left[2\pi f_d t - 2\pi \frac{2R}{\lambda_k}\right] + \tilde{n}(t) \quad (28)$$

B. Radar Signal Processing Modeling

1) Pulse compression

The pulse width of the transmitted signal influences the range resolution of radar. The narrower the pulse width of the signal, the higher the range resolution of the radar. However, when the pulse width of the signal becomes smaller, the maximum operating range of the radar will also be reduced, which leads to the fact that the pulse width of the signal cannot be too small. Therefore, the commonly adopted method for current radar is to employ a high-power transmitter to emit a wide pulse signal and then perform compression on the original wide pulse signal after receiving the echo signal, converting the signal into a narrow pulse signal, thereby achieving the functions of enhancing both the maximum operating range and the range resolution. This process is known as pulse compression[13]. In the commonly used radars at the present stage, a matched filter is typically employed to carry out pulse compression processing of signals. Suppose the transfer function of the matched filter is $H(f)$, the impulse response is $h(f)$, and the input signal is $s(t)$. The process of the input signal passing through the matched filter is depicted in Fig. 1

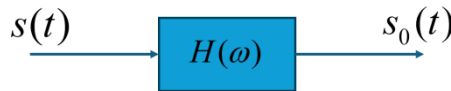


Fig. 1. Flowchart of Matched Filter

$s_0(t)$ is the output result of the target signal after it has passed through a matched filter, which can be expressed as follows:

$$s_0(t) = \int_{-\infty}^{\infty} H(f) S(f) e^{j2\pi f t} df \quad (29)$$

Taking the inverse Fourier transform of the transfer function $H(f)$ results in its impulse response function as

$$h(t) = k s^*(t_0 - t) \quad (30)$$

For the sampled digital signal, denoting the input signal after sampling and quantization as $s(n)$, the unit impulse response of the matched filter can be represented as $h(n) = s^*(n)$.

2) demodulation model

Consider the squaring law and envelope detection methods. The squaring law demodulation process is as follows:

$$x(n) = [A_{re}^2(n) + A_{im}^2(n)], n = 0, \dots, N - 1 \quad (31)$$

The process of envelope detection is as follows:

$$x(n) = \sqrt{[A_{re}^2(n) + A_{im}^2(n)]}, n = 0, \dots, N - 1 \quad (32)$$

3) Constant False Alarm Handling and Detection Model

Constant False Alarm Handling and Detection Model Since the signals received by the radar comprise both target echo signals and clutter signals, not all processing steps in the signal processor can completely filter out the clutter signals. Thus, a threshold value is often set in the radar. The portions of the signal that are higher than the threshold value are retained, while those lower than the threshold value are filtered out. This threshold value is the false alarm probability. To achieve this aim, the false alarm threshold must be calculated in real time based on the received signal to adjust the radar detection threshold accordingly to obtain the desired false alarm probability[14]. The detection processor that can maintain a constant false alarm probability is referred to as the Constant False Alarm Rate (CFAR) processor.

C. Output Model

1) Distance Output

The target echo signal will have a time delay t_r due to the distance between the target and the radar, which can be expressed as $t(r) = 2R/c$. Among them, R represents the relative distance between the target and the radar, and c represents the speed of light. Thus, in the case of a known time delay, the distance between the target and the radar can be inversely deduced based on the time delay t_r of the echo signal.

2) Angle output

It has been mentioned above that since the radar adopts the sum and difference beam angle measurement method when measuring the angle, the pitch angle θ and the yaw angle ϕ of the target relative to the radar can be obtained based on the amplitude of the corresponding position of the target in the processed sum beam signal and the amplitudes of the corresponding positions of the target in the pitch difference beam signal and the yaw difference beam signal:

$$\theta = \frac{\Delta F(\delta)}{\Sigma_{\theta} \frac{dF(\theta)}{d\theta} \Big|_{\theta=\delta}}, \quad \phi = \frac{\Delta F(\delta)}{\Sigma_{\phi} \frac{dF(\phi)}{d\theta} \Big|_{\phi=\delta}} \quad (33)$$

3) Velocity output

If relative motion exists between the target and the radar, the frequency of the target's echo will undergo changes. The Doppler frequency shift f_d caused by the relative velocity v_r can be expressed as $f_d = 2v_r/\lambda$. In the case where the Doppler frequency shift of the echo signal is known, the relative velocity between the target and the radar can be inversely deduced.

D. Guidance method

The proportional guidance method refers to the fact that the angular velocity of the missile's speed-changing direction in space is proportional to the angular velocity of the target's relative radar position angle rotation[15]:

$$\Delta\theta_M = k' \cdot \Delta\theta_T \quad (34)$$

where $\Delta\theta_M$ denotes the angular velocity of the change in the direction of the missile's velocity in space, k' is the proportional coefficient, and $\Delta\theta_T$ represents the angular velocity of the rotation of the target's relative position angle to the radar, also known as the line-of-sight angular velocity. Since the proportional guidance method alters the direction of the missile's velocity change based on the variation of the target's angle, it basically avoids situations where the missile makes large turns during the guidance process and is convenient for implementation in practical applications. As depicted in Fig. 2, assume that at time k , the target is positioned at point T , the missile is at point M , the velocity of the target is v_T , the velocity of the missile is v_M , and the distance between the target and the radar is r .

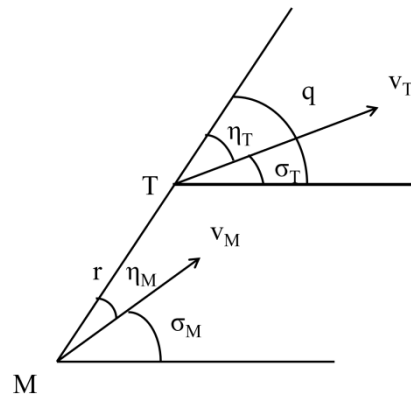


Fig. 2. Example of a figure caption

A diagram of proportional guidance method.

Denote the angle between the target's velocity and the line connecting the target and the missile as η_T , the angle between the missile's velocity and the line connecting the target and the missile as η_M , and q as the line-of-sight angle at the time k . At time $k + 1$, suppose the angle between the target velocity and the baseline is η'_M , the angle between the missile velocity and the baseline is η'_M , and q' is the line-of-sight angle at the time $k + 1$. Then the line-of-sight angular velocity at time $k + 1$ is:

$$\Delta\theta_T = q' - q \quad (35)$$

It is possible to further obtain the angular velocity of the rotation of the missile's velocity vector at time $k + 1$:

$$\Delta\theta_M = k \cdot (q' - q) \quad (36)$$

It can be known from the above formula that the angle between the velocity of the missile's motion and the reference line at time $k + 1$ is

$$\eta'_M = \eta_M + \Delta\theta_M \quad (37)$$

E. Jamming pattern modeling

1) Amplitude Modulation Interference

Noise amplitude modulation interference refers to the signal produced by modulating the amplitude of the carrier signal with a noise signal[16]. Generally, Gaussian white noise is often used to modulate the signal's amplitude. The definition of the noise amplitude modulation interference signal is as follows:

$$J(t) = (U_0 + U_n(t)) \cos(\omega_j t + \phi) \quad (38)$$

where $U_n(t)$ is a generalized stationary random process that follows $(0, \sigma^2)$ distribution, U_0 is the amplitude of the signal basis, ω_j is the carrier signal frequency, and ϕ is the signal phase.

2) Frequency-modulated noise interference

Noise frequency modulation interference is currently the most widely used type of suppressive interference signal, which has a broader interference bandwidth and makes it easy to achieve a more considerable noise power. The mathematical model for noise frequency modulation interference is represented as follows:

$$J(t) = U_j \cos\left(\omega_j t + 2\pi K_{FM} \int_0^t u(t') dt' + \phi\right) \quad (39)$$

where U_j is the amplitude coefficient of the signal, ω_j is the carrier frequency, K_{FM} is the frequency modulation slope, and modulation noise $u(t)$ is a wide-sense stationary random process with zero mean. ϕ is the signal phase.

3) Agile noise jamming

Agile noise interference is a compromise technique that combines deception and noise interference. Its essence is to combine forwarding-style interference with random pulse interference. The agile noise interference expression based on frequency modulation noise is as follows:

$$J(t) = U_0 \cos \left[2\pi f_j t + 2\pi K_{FM} \int u(t') dt' + \varphi \right] \cdot e^{-j\pi K^2} \text{rect}\left(\frac{t}{T}\right) \quad (40)$$

By changing the frequency modulation slope K_{FM} , the bandwidth of the FM noise can be changed, and therefore, the entire bandwidth of the agile noise is also changed.

III. REINFORCEMENT LEARNING

Inspired by the laws of biological learning, reinforcement learning interacts with the environment through a trial-and-error mechanism and learns and optimizes by maximizing cumulative rewards, ultimately achieving the optimal strategy. In reinforcement learning, the decision-maker or learner is defined as the "learning agent," and everything outside the learning agent is defined as the "environment," with the system integrating with the environment [17]. The interaction process between the learning agent and the environment can be described by three elements: state s , action a , and reward r . The learning agent acts a_0 based on the initial state s_0 and interacts with the environment, obtaining reward r_1 and updated state s_1 . At time t , based on the current state s_t and reward r_t , the learning agent provides the current action a_t , and then the system state transitions from s_t to s_{t+1} . r_{t+1} is the feedback reward from interacting with the environment. The basic principle of reinforcement learning is shown in Fig. 3.

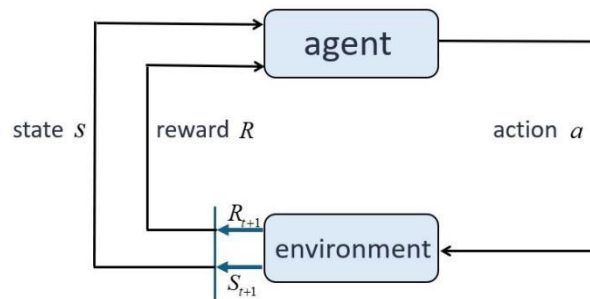


Fig. 3. Reinforcement learning system

Generally speaking, reinforcement learning emphasizes the interaction between the agent and the environment, expressed as a series of sequences of states, actions, and rewards: $s_0, a_0, r_1, s_1, a_1, \dots, s_{n-1}, a_{n-1}, r_{n-1}$. Although n can tend to infinity, a terminal state $s_n = s_T$ is usually defined in practice for limitation. This sequence of states, actions, and rewards, starting from the initial state and ending at the terminal state, is called an Episode or training cycle. The policy is usually denoted as π , which is a mapping from state s to action a . At the current time step, the learning agent interacts with the environment, learns through trial and error, and iteratively optimizes the current policy π to make the new policy π_1 superior to the current policy π . This process is called "policy update" and is repeatedly executed during reinforcement learning until the learning agent cannot find a better policy. In the interaction with the environment, the learning agent receives a feedback reward r each time t until the terminal state s_T . However, the reward at each step does not represent the long-term reward gain. To express the long-term gain of the learning agent, the return at time step t is introduced as follows:

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^{T-t} r_T \quad (41)$$

Among them, γ is the discount factor that satisfies $0 < \gamma < 1$. When γ is close to 1, the learning machine tends to place more emphasis on long-term returns, while when it is close to 0, it tends to put more emphasis on short-term returns. A value function usually represents the merit of a strategy. The state value function used to evaluate the merit of a strategy in state s is expressed as:

$$V_\pi(s) = E[G_t | s_t = s, \pi] \quad (42)$$

According to this formula, the optimal strategy can be obtained:

$$\pi^* = \operatorname{argmax}_\pi V_\pi(s) \quad (43)$$

Another type of value function is used to evaluate the degree of excellence of taking an action a in state s , called the state value function, also known as the Q function:

$$Q_{\pi}(s, a) = E[G_t \mid s_t = s, a_t = a, \pi] \quad (44)$$

The optimal strategy at this point is represented as:

$$\pi^* = \underset{a}{\operatorname{argmax}} Q_{\pi^*}(s, a) \quad (45)$$

A. Q-learning

Q-learning, also called Action-dependent Heuristic Dynamic Programming (ADHDP) [18], does not require a model. Q-learning does not wait until the end of an episode for an update but instead updates using the Temporal Difference (TD) approach at each step, achieving a faster convergence effect. The TD learning utilizes the previous estimate to update the current state value function. The TD learning method aims to obtain the value function. When confronted with control decision problems, the state-action value function is more instructive for selecting actions. Q-learning employs the Bellman optimality principle to make the current value function directly approach the value function of the optimal policy. The update method is as follows:

$$Q(s_t, a_t) \leftarrow \alpha Q(s_t, a_t) + (1 - \alpha) \left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a') \right) \quad (46)$$

where α represents the update rate, which satisfies the condition $0 < \alpha < 1$, where a' is the action that maximizes the Q function in state S_{t+1} .

B. Policy gradient methods

Q-learning is a value-function-based approach. In practical applications, a table is frequently employed to store the state or state-action value function, thus being relatively inefficient for complex problems with ample action space. Policy gradient methods do not rely on the value function. They directly parameterize the policy π as $\pi(s \mid \theta)$ and then calculate the gradient concerning the policy performance metric. The policy parameters are adjusted based on the gradient direction to obtain the optimal policy [19]. Parameterized policies can be classified into stochastic policy $\pi(s \mid \theta) = P[a \mid s, \theta]$ and deterministic policy $a = \mu(s \mid \theta)$. A policy objective function $J(\theta)$ is set to evaluate the parameterized policies. For stochastic policies, the action a in the current state s obeys a particular probability distribution with parameter θ . For deterministic policies, the action corresponding to each state is definite. According to the policy gradient theorem, the gradient of stochastic policies is expressed as [20]:

$$\nabla_{\theta} J(\theta) = E_{s,a \sim \pi} [\nabla_{\theta} \ln \pi(s \mid \theta) Q_{\pi}(s, a)] \quad (47)$$

The deterministic policy gradient is represented as [1], [21]:

$$\nabla_{\theta} J(\theta) = E_{s,a \sim \mu} \left[\nabla_{\theta} \mu(s \mid \theta) \nabla_a Q_{\mu}(s, a) \Big|_{a=\mu(s \mid \theta)} \right] \quad (48)$$

When computing the gradient, a genuine state-action value function $Q_{\pi}(s, a)$ or $Q_{\mu}(s, a)$ is required. Nevertheless, this function is still being determined in practice. One method is to employ the return values over certain steps to estimate the state-action value function. Another approach is to utilize the actor-critic architecture, where the critic approximates the state-action value function, and the actor represents the policy. The critic is represented as a function $Q(s, a \mid \omega)$ of parameter ω and is updated using the temporal difference method. The temporal difference error δ_t is expressed as:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1} \mid \omega) - Q(s_t, a_t \mid \omega) \quad (49)$$

The update formula for parameter ω in the evaluator is:

$$\omega \leftarrow \omega + \alpha \delta_t \nabla_{\omega} Q(s_t, a_t \mid \omega) \quad (50)$$

Replace the learned evaluation function $Q(s_t, a_t \mid \omega)$ with the actual value function $Q_{\pi}(s, a)$ or $Q_{\mu}(s, a)$ in the policy gradient formula to update the policy.

IV. CONSTRUCTION OF ADVERSARIAL MODELS

A. Environmental description

In the model, the environment is precisely the target radar seeker. The seeker's state is constituted by two variables: velocity and position. The simulation system generates corresponding emission signals based on the configuration of simulation parameters. The antenna gain amplifies and transmits the emission signals until they reach the target position. Based on the distance between the target and the radar and the scattering cross-sectional area of the target, the emission signals are processed correspondingly to generate the echo signal. Subsequently, the echo signal is sent to the radar receiver, while the receiver also receives clutter signals from the environment and external interference signals. The receiver transmits all the received signals to the signal processor, which processes the received signals and extracts the contained target information. The proportional guidance law then adjusts the velocity, and the corresponding position and velocity will also change. This paper considers that when the altitude of the seeker is zero, the missile detonates, and this condition serves as the termination condition for one simulation. The workflow diagram of the seeker is presented as Fig. 4:

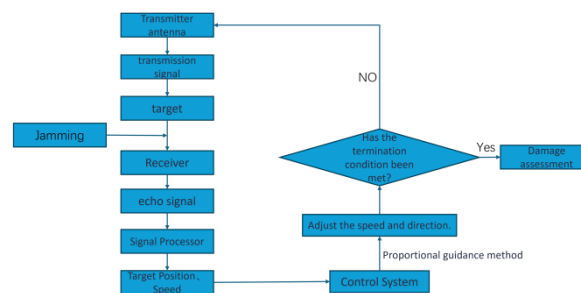


Fig. 4. The workflow diagram of the seeker

B. Action description

We selected the three types of interference mentioned above to disrupt the seeker. The action was represented by the implementation status of the three interferences at each simulation moment, which a 1x3 vector could indicate. For instance, (1, 0, 1) signified that the radio frequency interference was activated, the noise frequency modulation interference was deactivated, and the intelligent noise interference was activated. To better conform to the actual situation, it was stipulated that the duration from activation to deactivation of each interference type was no less than 5 seconds, which, to a certain extent, restricted the occurrence of some interference combinations and enhanced the simulation efficiency. Assuming that the duration of one simulation was 5 seconds, the random interference strategy could be presented in Tab. 1.

TABLE I. RANDOM JAMMING STRATEGY

time	jamming1	jamming2	jamming3
1	0	0	0
2	0	0	0
3	0	1	0
4	1	1	0
5	1	1	0
6	1	1	1

C. Reward Function Building

The reward value function is composed of two components. The first part is the reward function attributed to the measurement error, namely the difference between the measurement information output by the radar seeker based on the echo signal under the application of interference and the actual information of the target. The more significant the difference, the more influential the interference is. The measurement information encompasses the target's distance, speed, and angle. The second part is the detection probability of the interference device by the seeker. To avoid being discovered and detected by the

missile, the activation time of each interference device should not be too early. An earlier activation would result in more obvious exposure of the target. Therefore, the activation time is also considered within the reward function. The reward function can be formulated as:

$$r = k_1\Delta L + k_2\Delta v + k_3\Delta\theta + k_4\Delta\varphi + b_1T_{on1} + b_2T_{on2} + b_3T_{on3} \quad (51)$$

where Δv represents the velocity error value, ΔL represents the distance error value, $\Delta\theta$ $\Delta\varphi$ represents the pitch and azimuth angle error values, and $T_{on1}, T_{on2}, T_{on3}$ represents the first startup time of the three interference patterns. $k_1, k_2, k_3, k_4, b_1, b_2, b_3$ are proportional coefficients. Through numerous experiments, a set of relatively reasonable values for them can be obtained as : $k_1, k_2, k_3, k_4, b_1, b_2, b_3 = 0.02, 0.1, 20, 20, 0.2, 0.1, 0.1$.

V. EXPERIMENT

A. Parameter settings

The parameters for the phased array radar seeker are listed in Tab. 2.

TABLE II. PARAMETERS FOR THE PHASED ARRAY RADAR SEEKER

Project	Major Parameters
Transmitter	Center frequency: 18 GHz
	Bandwidth: 250 MHz
	Maximum power: 30 kW
	Pulse width: 100 ns
Entry 2	Beam width: 0.2 rad
	Angular measurement range: $-\pi/2 \sim \pi/2$

The LFM signal parameters are as shown in Tab. 3.

TABLE III. PARAMETERS OF LFM SIGNAL

Parameters	Pulse width	Bandwidth	Pulse Repetition Period	Transmission frequency	Number of Pulse Emissions
Value	10us	10MHz	100us	15GHz	16

TABLE IV. JAMMING PARAMETERS

Types of jamming	Amplitude Modulation jamming	Frequency-modulated noise jamming	Agile noise jamming
Pulse width	50us	200us	10us
Bandwidth	40MHz	1MHz	30MHz

The parameters of jamming are as shown in Tab. 5:

TABLE V. JAMMING PARAMETERS

Types of jamming	Amplitude Modulation jamming	Frequency-modulated noise jamming	Agile noise jamming
Pulse width	50us	200us	10us
Bandwidth	40MHz	1MHz	30MHz

The initial state of the missile and the ship is as shown in Tab. 6:

TABLE VI. JAMMING PARAMETERS

Parameters	Missile	Ships
Magnitude of velocity	2500 km/h	50 km/h
The direction vector of velocity	(0, 1, 0)	$(\sqrt{2}, \sqrt{2}, 0)$
Location coordinates	(0, 0, 10 km)	(8 km, 8 km, 0)
Radar Cross-Section	—	10 m ²

The parameters in reinforcement learning are: $\alpha = 0.02, \gamma = 0.8, \varepsilon = 0.5$.

B. Experimental results

This paper employs four jamming strategies, namely, jamming decision without jamming means, jamming decision with random jamming patterns, jamming decision based on Q-learning and jamming decision based on policy gradient. The Fig. 5 shows the relationship between the distance between the missile and the ship when it explodes and the number of simulation experiments conducted in 400 trials.

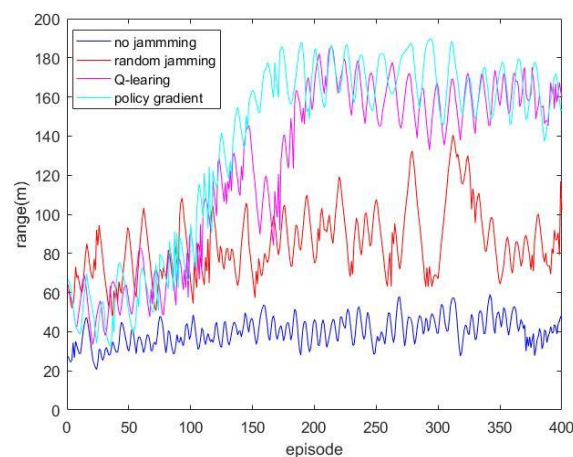


Fig. 5. Diagram of Variation in Missile Landing Point Distance

It can be observed from the figure that when no jamming measures are adopted, the missile is capable of causing damage to the target. Due to the presence of noise in the received signal, the landing distance of the missile fluctuates, which, to a certain extent, validates the accuracy of the radar seeker signal-level simulation. When the random jamming strategy is employed, it can exert a particular influence on the guidance of the radar seeker, but the randomness is considerable. When the Q-learning method is utilized, as the number of simulations increases, the distance the missile lands gradually converges, and it can stably and effectively interfere with the missile. When the policy gradient method is adopted, compared with the Q-learning method, it has a faster convergence speed, and the interference effect is similar to that of the Q-learning method. Finally, the optimal jamming strategies derived based on Q-learning and policy gradient learning are presented in Tab. 7.

TABLE VII. THE BEST JAMMING STRATEGY

Time	Q-learning	Policy gradient
1	000	000
2	000	000
3	100	100
4	100	100
5	101	100
6	101	101

Time	Q-learning	Policy gradient
7	111	111
8	111	111
9	111	111
10	111	111
11	111	111
12	111	111
13	111	111
14	111	111
15	111	111

The optimal interference strategy obtained above was subjected to 100 simulation experiments, and the resulting diagram of the missile landing point and the position of the ship is shown in Fig. 6.

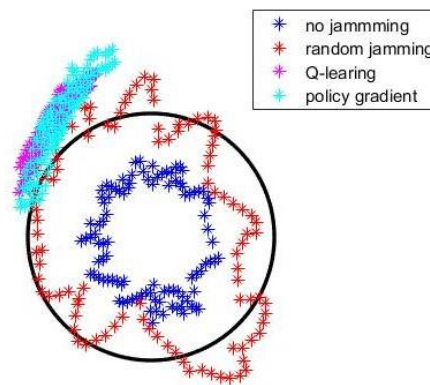


Fig. 6. Distribution Map of Missile Landing Points

The center of the circle represents the target location where the missile detonates, and the data obtained from the simulation can be used to establish the following evaluation indicators for missiles[22]. To simplify the analysis, the ship target is regarded as a circle with a radius of 100m, and any landing point within a 100m range can inflict effective damage on the target.

- probability of hitting; The hit probability of the missile is a comprehensive index for evaluating the quality of the missile's landing point. Its meaning is the double integral of the probability density function of the missile's landing point over the idealized circular area of the target ship. The calculation formula is as follows:

$$P_{\text{hit}} = \iint_{x^2+y^2 \leq r_s^2} f(x, y) dx dy \quad (52)$$

where r_s represents the idealized circular radius of the ship, and $f(x, y)$ represents the probability density function of the landing coordinates of the missile.

- Accuracy of landing point; The accuracy of missile landing point is an index reflecting the degree to which the landing point of an anti-ship missile is close to the center of the target ship. The accuracy of missile landing point can be derived by averaging the distances of all the landing points relative to the center of the target ship. The specific calculation method is as follows:

$$m_R = \frac{\sum_{i=1}^{N_{\text{tal}}} R_i}{N_{\text{total}}} \quad (53)$$

where R_i is the distance of the missile landing point relative to the center of the ship in the 6th simulation, and N_{total} is the total number of simulations.

- Circularity deviation probability; the circular error probable (CEP) can be defined as the radius R_{cep} of the integral circle at which the hit probability reaches 0.5. This indicator is a comprehensive reflection of the accuracy and density of the landing point of the missile. The integral equation for calculating the circular probability deviation is as follows:

$$\iint_{x^2+y^2 \leq R_{\text{cep}}^2} f(x, y) dx dy = 0.5 \quad (54)$$

- Damage degree; The damage degree is for assessing the destruction degree inflicted on a ship by an anti-ship missile. Supposing that when the center of the ship is hit, the damage degree of the anti-ship missile is 1. Then, with the distance distribution of the anti-ship missile's landing point relative to the center of the target ship, the damage degree of the anti-ship missile follows a certain probability distribution, and its distribution can be approximated as the following distribution:

$$D(x, y) = \exp\left(-\frac{x^2 + y^2}{2r_s^2}\right) \quad (x^2 + y^2 \leq r_s^2) \quad (55)$$

The damage degree S of anti-ship missiles is calculated as follows:

$$S = \iint_{x^2+y^2 \leq r_s^2} f(x, y) D(x, y) dx dy \quad (56)$$

By conducting statistical analysis on the data in Fig. 5, the damage effects of the seeker can be obtained as presented in Tab. 8.

TABLE VIII. ASSESSMENT RESULTS OF DAMAGE EFFECTS

Damage indicators	No jamming	Random jamming	Q-learning	Policy Gradient
Probability of hitting	100%	75.3%	5.6%	6.8%
Accuracy of landing point	38.5 m	72.4 m	5.6 m	6.8 m
Circularity deviation probability	40.3 m	84.5 m	156.2 m	154.6 m
Damage degree	70.8%	58.2%	2.6%	3.3%

It can be seen that the interference strategy based on the reinforcement learning algorithm can have a significant impact on the guidance effect of the radar seeker, and the interference effect is stronger than the random interference pattern. Since the optimal policy based on the policy gradient method is similar to that of Q-learning, we select the interference policy obtained from Q-learning as the optimal interference policy and conduct a simulation experiment. Compared with the situation where no interference measures are taken and the situation where random interference measures are adopted, the difference between the output signal of the seeker and the actual value at each simulation moment is recorded. The distance error values are presented in Fig. 7, the Velocity error values in Fig. 8, the pitch angle error values in Fig. 9, and the yaw angle error values in Fig. 10.

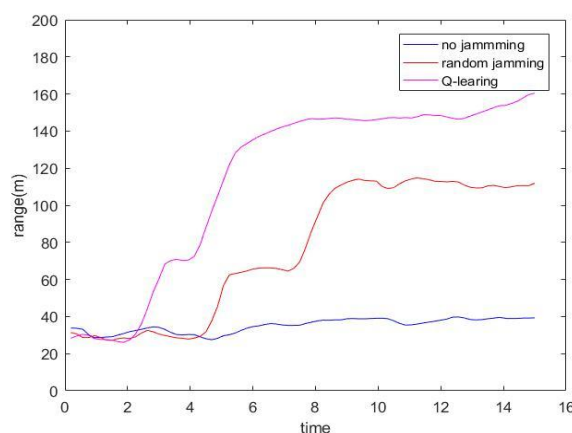


Fig. 7. Diagram of Variation in distance error

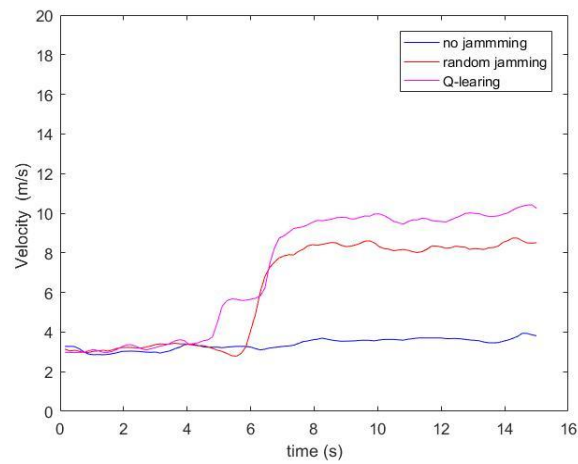


Fig. 8. Diagram of Variation in Velocity error

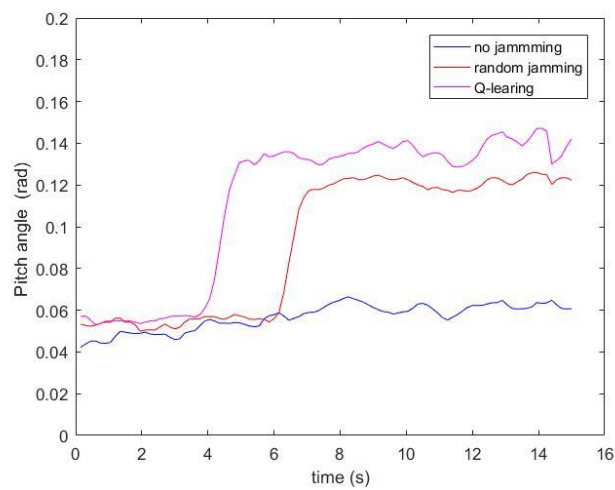


Fig. 9. Diagram of Variation in Pitch angle error

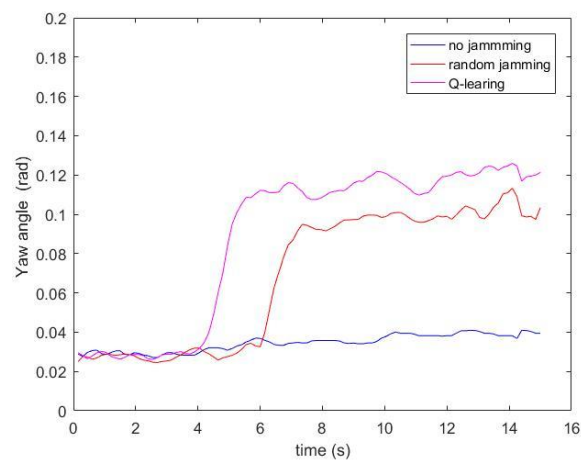


Fig. 10. Diagram of Variation in Yaw angle error

From the above 4 figures, it can also be seen that the best interference strategy obtained by reinforcement learning has a better interference effect.

VI. CONCLUSION

This paper initially establishes a signal-level simulation model for the entire end guidance process of the radar seeker from several aspects, such as signal transmission, signal reception, signal processing, and measurement information output. Subsequently, simulation experiments are conducted without interference, and it is shown that the missile can hit the target with sure accuracy, verifying the accuracy of the established model. Then, based on this signal-level simulation model, in the interference decision-making of the interference equipment against the radar seeker, Q-learning, and policy gradient learning are introduced to obtain the optimal interference strategy. Simulation experiments demonstrate that the interference effect using this optimal strategy on the radar seeker is significantly better than that of the random interference strategy, thereby proving the superiority of reinforcement learning algorithms in the aspect of radar interference decision-making.

REFERENCES

- [1] G. M. Reich, M. Antoniou, and C. Baker, "Memory-enhanced cognitive radar for autonomous navigation," *IET Radar Sonar Navig.*, vol. 14, pp. 1287–1296, 2020.
- [2] S. Z. Gurbuz, H. D. Griffiths, and A. Charlish, "An overview of cognitive radar: Past, present, and future," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 34, pp. 6–18, 2019.
- [3] Z. Li, S. Perera, and Y. Zhang, "Phased-array radar system simulator (PASIM): Development and simulation result assessment," *IREMOTE SENSING*, vol. 11, no. 4, 2019.
- [4] B. K. Zhu, W. G. Zhu, and W. Li, "A review of reinforcement learning-based radar jamming decision technology," *Electro-Opt. Control.*, vol. 29, pp. 52–58, 2022.
- [5] Q. Xing, W. G. Zhu, and X. Jia, "Research on method of intelligent radar confrontation based on reinforcement learning," in *2017 2nd IEEE international conference on computational intelligence and applications*, CHINA, 2017, pp. 8–11.
- [6] B. K. Zhang and W. Z. Zhu, "Construction and key technologies of cognitive jamming decision-making system against MFR," *Syst. Eng. Electron.*, vol. 42, pp. 1969–1975, 2020.
- [7] S. Y. Zhang, H. Tian, X. Q. Chen, and Z. Y. Du, "Design and implementation of reinforcement learning-based intelligent jamming system," *IET COMMUNICATIONS*, vol. 14, pp. 3231–3238, 2020.
- [8] S. Feten, S. Bart, and C. Zied, "A modified q-learning algorithm to solve cognitive radio jamming attack," *International Journal of Embedded Systems*, vol. 10, pp. 41–51, 2018.
- [9] Y. J. Li, Y. P. Zhu, and M. G. Gao, "Design of cognitive radar countermeasures based on q-learning algorithm," *J. Beijing Inst. Tech.*, vol. 35, pp. 1194–1199, 2015.
- [10] Q. Xing, W. G. Zhu, and X. Jia, "Intelligent countermeasure design of radar working-modes unknown," in *2017 IEEE international conference on signal processing, communications and computing*, CHINA, 2017, pp. 22–25.
- [11] B. K. Zhang and W. G. Zhu, "A DQN cognitive interference decision method for multifunctional radars," *Syst. Eng. Electron. Technol.*, vol. 42, pp. 819–825, 2020.
- [12] T. Shan, S. Liu, and Y. D. Zhang, "Efficient architecture and hardware implementation of coherent integration processor for digital video broadcast-based passive bistatic radar," *IET RADAR SONAR AND NAVIGATION*, vol. 10, pp. 97–106, 2016.
- [13] A. S. Mudukutore, V. Chandrasekar, and R. J. Keeler, "Pulse compression for weather radars," *IEEE TRANSACTION ON GEOSCIENCE AND REMOTE SENSING*, vol. 36, pp. 125–142, 1998.
- [14] G. V. Weinberg, "Geometric mean switching constant false alarm rate detector," *DIGITAL SIGNAL PROCESSING*, vol. 69, pp. 1–10, 2017.
- [15] K. Y. Li and G. J. Zhou, "State estimation with a destination constraint imposed by proportional navigation guidance law," *DIGITAL SIGNAL PROCESSING*, vol. 58, pp. 58–73, 2022.
- [16] K. Abratkiewicz, P. J. Samczynski, and R. Rytel-Andrianik, "Multipath interference removal in receivers of linear frequency modulated radar pulses," *IEEE SENSORS*, vol. 21, pp. 19000–19012, 2021.
- [17] C. X. Mu, Z. Ni, C. Y. Sun, and H. B. He, "Air-breathing hypersonic vehicle tracking control based on adaptive dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 584–598, 2017.
- [18] C. Mu, Q. Zhao, C. Sun, and Z. Gao, "A novel q-learning algorithm for optimal tracking control of linear discrete-time systems with unknown dynamics," *Applied Soft Computing*, vol. 82, pp. 1–13, 2019.
- [19] Y. D. Wang, J. Sun, and H. B. He, "Deterministic policy gradient with integral compensator for robust quadrotor control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 50, pp. 3713–3725, 2017.
- [20] J. Q. Cao, Q. Liu, and L. Wu, "Generalized gradient emphasis learning for off-policy evaluation and control with function approximation," *NEURAL COMPUTING and APPLICATIONS*, vol. 35, pp. 23599–23616, 2023.

- [21] A. H. Hosseinloo, Q. Liu, and M. A. Dahleh, “Deterministic policy gradient algorithms for semi-markov decision processes,” *INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS*, vol. 37, pp. 4008–4019, 2021.
- [22] T. L. Yao, Q. Liu, W. L. Wang, and R. Miao, “Damage effectiveness assessment method for anti-ship missiles based on double hierarchy linguistic term sets and evidence theory,” *JOURNAL OF SYSTEMS ENGINEERING AND ELECTRONICS*, vol. 33, pp. 393–405, 2022.