Contents lists available at ScienceDirect

Chinese Journal of Aeronautics

Journal homepage: www.elsevier.com/locate/cja

Intelligent Decision-Making Algorithm for Airborne Phased Array Radar Search Tasks Based on a Hierarchical Strategy Framework

Abstract

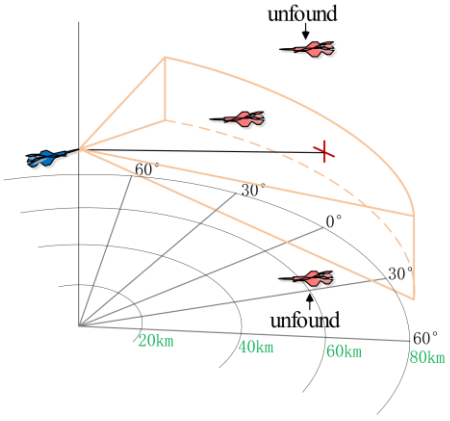
To address the guided search task of airborne phased array radar in the scenarios of large airspace with widespread distribution of cluster targets in beyond visual range (BVR) air combat, a hierarchical strategy framework based on deep reinforcement learning is proposed to guide different stages of search tasks. Firstly, an airspace set-covering model and a radar parameter optimization model for the guided search task of cluster targets are established. Secondly, the hierarchical strategy framework included upper-level and lower-level strategies is constructed based on above models. Finally, the happo-rgs algorithm is proposed for feature extraction from Markov continuous observation sequences, to enhance the training effectiveness and algorithm convergence speed. Simulation results show that the trained agent can make precise autonomous decisions rapidly based on airspace-target covering situation and target guidance information which significantly improves the radar search performance in the forementioned scenarios compared to traditional algorithms.

Keywords: Beyond-visual-range air combat; Phased array radar; Radar search resource optimization; Reinforcement learning; Multi-head attention mechanism;

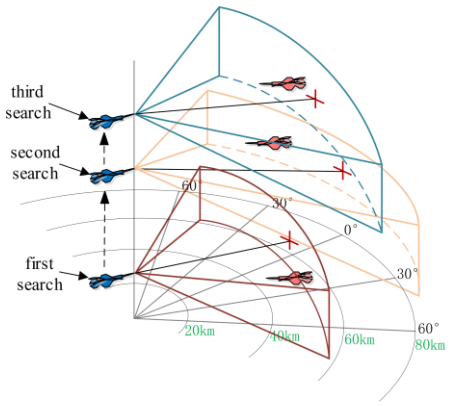
Introduction

In modern beyond visual range (BVR) air combat, the guidance and command of early warning aircraft play a crucial role. Based on the beam agile capability of modern airborne active phased array radar, radar search for small window airspace guided by early warning information has tremendous potential in rapidly completing search tasks [1]. However, in recent years, as the battlefield situation becomes complex and the battlefield domains tend to merge, enemy targets show characteristics such as clustering, stealth, and intelligence [2][3]. The single scan range of airborne phased array radar is limited, making it difficult to quickly complete guided search task of cluster targets, which affects subsequent combat mission processes. By adopting intelligent decision-making algorithms for airspace guided search of cluster targets, radar search parameters are optimized while generating high-precision search airspace coordinates. Intelligent combat with advantages like high agility, high precision, and high cost-effectiveness ratio can achieve an emergent increase in air combat capability [4]. The comparison between single-search and multi-search

areas is shown in Fig. 1. This paper mainly studies the rapid full coverage for cluster targets of radar guided search task widespread distributed in the large airspace scenario, to achieve efficient search of designated airspace in BVR air combat, and quickly generate comprehensive enemy situation information. To our knowledge, this is the first time that intelligent decision-making methods have been discussed for radar guided search task of scenarios with widespread distribution of cluster targets in the open literature.



(a) Single-search area



(b) Multi-search area

Fig. 1 Comparison between single-search and multi-search areas

The guided search task in the scenario with widespread distribution of cluster targets beyond visual range requires complex multi-dimensional decisions including search airspace set-covering decision [5][6] and radar search parameter optimization decision [7][8]. On one hand, it is necessary to determine the sequence and orientation of each search airspace based on the factors such as battlefield situation information and radar search resource which aims to achieve precise coverage of potential target position and maximize radar search performance in the search airspace. On the other hand, to optimize the radar search parameters in the airspace to be searched based on target guidance information, the sub-airspace division strategy and search beam position arrangement strategy in sub-airspaces need to be determined. Then the global radar search parameter optimization model under different search resource loads is constructed to calculate corresponding sub-airspace beam dwell time and beam position search data rate.

The problem of search airspace decision for radar guided search task in the scenario with widespread distribution of cluster targets can be represented as a weighted set-covering problem (SCP) [9]. It is necessary to construct the airspace coverage decision-making model and adopt high-precision real-time solution algorithms to generate online airspace guided search schemes. The set-covering problem, as a NP-complete problem, can be solved by exact solution methods [10][11] and heuristic solution methods [12][13]. Ref. [10] presents and compares computationally three exact algorithms, where two of them are based on Benders decomposition and one uses Benders cuts in the context of a Branch-and-Cut approach. Ref. [11] simplifies the original problem by dual information, and finally uses an exact solution method to solve this simplified problem. Ref. [12] combines the Lagrange strategy and ant colony method to derive better heuristic information. Ref. [13] proposes a hybrid method based on artificial bee colony method and local search solving method, which has achieved good results in the majority of testing problems. However, existing researches mainly focuses on improving the algorithm solution performance for typical problems, lacking the capability to handle the high dynamic constraints of actual operation. There is less studies on constraints related to air combat tasks and radar resources.

Regarding the search parameter optimization problem after airspaces to be searched are determined, traditional radar optimal search models usually need to determine sub-airspace division and beam position arrangement strategies based on target guidance information and radar search performance. Then the relevant search parameters are optimized and each sub-airspace search resources are allocated [14]. Ref. [15] analyzes the impact of false alarm time consumption on search performance and incorporates false alarm probability, search beam dwell time, and search frame period into the optimization model to provide the corresponding optimal search parameters. Ref. [16], based on Swerling III model, focuses on the problem of optimal search resource allocation in multiple airspace scenarios, and optimizes radar search resources for the maximum expected target discovery distance by adjusting beam dwell time of each airspace. Ref. [17] proposes the joint dwell time allocation and detection threshold optimization (JDTADTO) strategy for asynchronous phased array radar networks (PARNs) in multi-target tracking scenarios by considering both false alarm and missed detection probabilities. However, these radar guided search parameter optimization algorithms have limited research on the

cooperative optimization of radar search parameters in the scenarios with widespread distribution of cluster targets, and lack comprehensive global radar search parameter optimization model.

In summary, in the scenario with widespread distribution of cluster targets beyond visual range, the aforementioned radar guided search task needs to construct online decision-making algorithm for search airspace set-covering, and optimize radar search parameters based on its search performance and situation information of both enemy and our sides. For above multi-dimensional decision problem, traditional optimization methods need to determine optimization objectives and constraints in advance, and can only find the optimal solution within the known search space, which is difficult to meet the real-time decision-making needs of future high-dynamic operations [18]. The rise of deep reinforcement learning in recent years has made significant progress in dealing with complex, high-dimensional problems with better generalization and reasoning capabilities in dealing with complex decision-making problems [19]. Among them, multi-agent reinforcement learning has advantages such as collaboration, division of labor, information sharing, and adversarial learning, which can help agents better cope with complex environments and tasks [20][21]. Ref. [22] proposes a waveform design algorithm based on reinforcement learning by modeling the radar target parameter estimation problem as a multi-agent reinforcement learning framework. Ref. [23] uses the Markov Decision Process (MDP) to describe the complex working environment of airborne radar, and proposed a method for designing airborne radar waveforms based on Deep Reinforcement Learning (DRL) under clutter and interference conditions. Ref. [24] proposes a multi-target detection algorithm of MIMO radar based on reinforcement learning for cognitive multi-target detection in the presence of unknown interference distribution. However, most of existing studies use reinforcement learning to optimize radar waveforms or related parameters, with less involvement in radar behavior decision-making for complex airspace guided search tasks of cluster targets. The high dynamics of battlefield situation, the high randomness of cluster target distribution and the time series characteristics of guided search environment observation sequence pose severe challenges to the design of real-time search airspace decision-making algorithms and global radar guided search parameter optimization models.

This paper proposes an intelligent hierarchical decision-making framework that determines search airspace orientation and optimizes radar search parameters in the airspace for the airborne radar guided search task in scenarios with widespread distribution of cluster targets based on deep reinforcement learning respectively. The main innovations of this paper are as follows:

(1) An airspace set-covering model and a global radar search parameter optimization model based on the maximum expected discovery distance and average accumulated discovery probability of cluster targets are established to determine airspace to be searched and optimize sub-airspace beam dwell time and beam search data rate.

(2) Based on above search airspace and parameter optimization models,a hierarchical strategy framework with upper-level and lower-level strategy modules is constructed where upper-level strategy outputs search airspace center coordinates for the lower-level strategy module and lower-level strategy optimizes the search parameters within the airspace.

(3) An improved proximal policy optimization algorithm (happo-rgs) based on proposed hierarchical strategy framework and a real-time beam irradiation environment for above scenario are constructed where a fully connected neural network based on multi-head attention mechanism is designed for the upper-level strategy module decision network and an genetic algorithm expert data set is generated to guide the training process because of the high randomness of target guidance information.

The structure of this paper is as follows: In Sec. 1 we introduce the decision-making problem of radar guided search task and propose models including the airspace set-covering model and the radar search parameter optimization model for cluster targets. In Sec. 2 we propose the hierarchical strategy framework and the happo-rgs training algorithm designed for this framework, with key improvements including the introduction of multi-head attention mechanism and genetic algorithm expert data set. In Sec. 3 we present and discuss simulation results of training and testing experiments. In Sec. 4 we conclude the paper and look forward to some future work.

1. Proposed Model

1.1. Problem Description

The guided search task of airborne phased array radar in scenarios with widespread distribution of cluster targets can be divided into two steps, mainly involving intelligent decision-making of airspace coverage set for radar guided search and optimization of radar search parameters for cluster targets. The specific process is shown in Fig. 2.

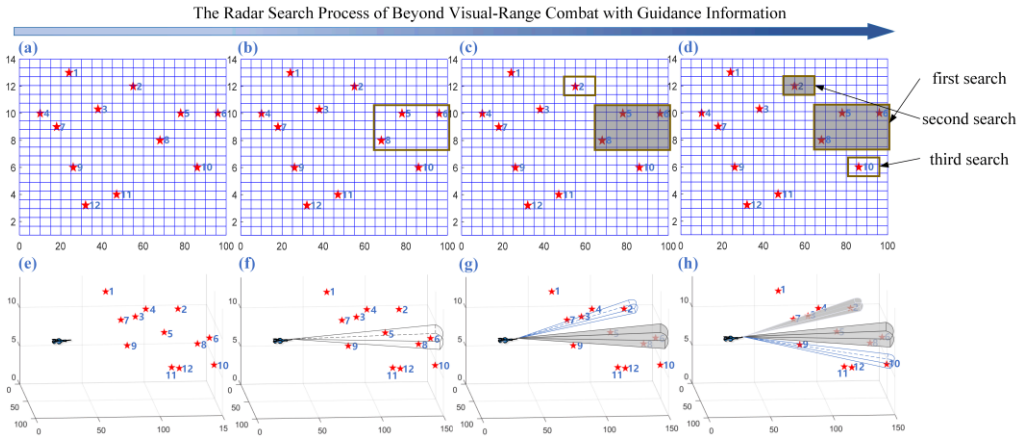


Fig.2 The guided search task process of airborne phased array radar in scenarios with widespread distribution of cluster targets.

Fig 2 (a)-(d) illustrate the airspace cross-section at various stages from obtaining guidance information of cluster targets to the third search, while Fig. 2 (e)-(h) illustrate the corresponding three-dimensional airspace shown at the top. Red pentagrams represent targets, with a total of 12 targets in this example.

As shown in Fig. 2, in the BVR air combat scenario, the pilot will request approximate target position information from the early warning aircraft before conducting the search task. By adjusting radar working mode, azimuth and pitch search centers, and scan angle range in coordination, the radar search airspace can accurately cover the possible position of cluster targets [25]. The radar search range in the vertical airspace is directly related to corresponding radar performance parameters. (As shown in Fig. 2 (a)-(d), each radar search airspace range, combined with target guidance information, aims to cover targets as much as possible). It is necessary to comprehensively consider the target distribution characteristics of search airspace and current search task progress to reasonably select search airspace based on guidance information. Moreover, the radar search orientation needs to be adjusted multiple times to achieve full coverage of cluster targets. (As shown in Fig. 2 (e)-(h), search airspaces of three times cover all targets, which completes the selection of each search airspace and search sequence, and determines the precise position of targets by optimizing the search parameters in each search airspace).

Once the search airspace is determined, the airspace is divided into several sub-airspaces for beam position arrangement. Common beam position arrangement methods [26] mainly include columnar beams, interleaved beams, and low-loss beams, as shown in Fig. 3. Then, based on target guidance information, the probability of targets falling into each beam position are calculated. Finally, sub-airspace beam dwell time and search data rate of each beam position are optimized for maximum target expected discovery distance and accumulated discovery probability respectively. Taking columnar beam arrangement as an example, the radar parameter optimization process for the search airspace is shown in Fig. 3.

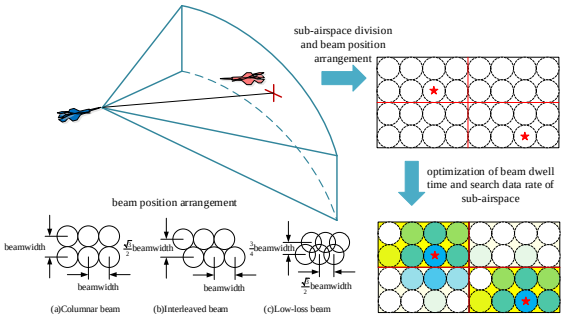


Fig. 3 Optimization process of radar parameters in search airspace

As shown in Fig. 3, the search airspace contains two targets and is divided into four sub-airspaces. The color depth of the beam position indicates the level of search data rate, with darker color and higher search data rate for beam positions with higher target distribution probability. The sub-airspace color depth indicates the length of beam dwell time, with darker color and longer dwell time for sub-airspaces containing more targets.

This paper establishes an airspace set-covering model and a radar search parameter optimization model for different stages of above radar guided search task. For the airspace set-covering model constructed based on the minimum weight set covering problem [27] (MWSCP), a method for generating discrete search airspace sets is proposed, and an airspace cost matrix is calculated based on the Gaussian mixture model of cluster target distribution probability. For the search parameter optimization model, the sub-airspace beam dwell time and beam position search data rate are optimized based on expected discovery distance and average accumulated discovery probability of cluster targets respectively.

1.2. Airspace Set-covering Model Based on Guided Search of Cluster Targets

Assuming that there are n targets distributed in the entire airspace which is divided into m smaller

airspaces that can be covered by radar single-search area, airspace-target covering situation is represented by matrix $A=(a_{ij}), i\in N, j\in M$, where $N=\{1,2,...,n\}$ and $M=\{1,2,...,m\}$ represent the sets of targets and airspaces in matrix A respectively. $a_{ij}=1$ means i-th target is covered by j-th airspace; conversely, $a_{ij}=0$ means i-th target is not covered by j-th airspace. The cost matrix $C=(c_j), j\in M$, where c_j represents the cost of airspace j, with $c_j>0$, $\forall j\in M$. The goal of airspace ser-covering model is to find such a subset of airspaces $X(X\subseteq M)$ so that each target in N is covered by at least one airspace in X, and the cost sum of all airspaces in X is minimized. The optimization objective function can be written as follows [28]:

$$\min \quad \sum_{j=1}^{m} c_j x_j \tag{1-1}$$

$$s.t \quad \sum_{j=1}^{m} a_{ij} x_j \geq 1, i=1,2,...,n, \tag{1-2}$$

$$x_j \in \{0,1\}, j=1,2,...,m. \tag{1-3}$$

Objective function (1-1) aims to minimize the total airspace cost, where c_jx_j is the cost of each airspace which is included or not; constraint (1-2) ensures that each target in matrix A is covered by at least one airspace in set X; in constraint (1-3), $x_j=1$ indicates that j-th airspace is included in airspace set X, and $x_j=0$ means airspace j is not included in X.

Based on the above airspace set-covering model, this paper proposes the method for generating discrete search airspace set M and the method for calculating airspace cost matrix C based on the probability distribution model of target guidance information.

a. Generation of search airspace set

The principle of search airspace set generation is as shown in Fig. 4.

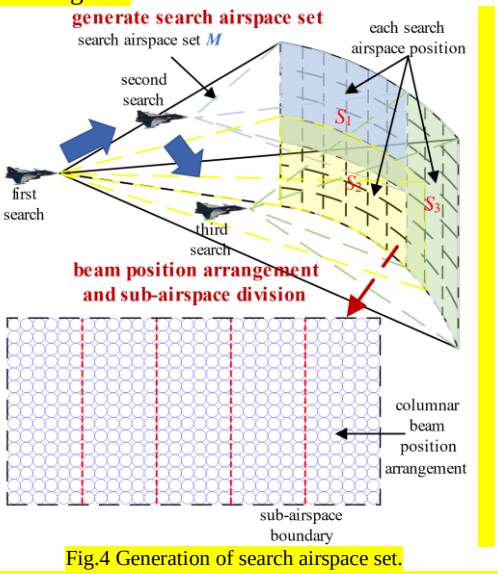


Fig.4 Generation of search airspace set.

(1) Numerous search airspace subsets are generated within entire airspace by different division accuracies, as shown in Fig. 4 as $M=\{S_1,...,S_m\}$;

(2) Beam positions of divided airspaces are arranged and parameters like beam position centers and sub-airspace boundaries are calculated.

b. Calculation of Airspace Cost Matrix

The principle of airspace cost matrix calculation is as shown in Fig. 5.

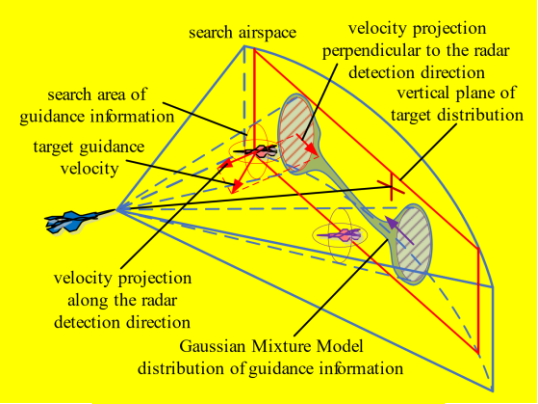


Fig. 5 Calculation of airspace cost matrix

The application scenario of airborne active phased array radar in this paper is the small window airspace search guided by warning information. Assuming that target position and velocity are included in the warning information, a 3D Gaussian sphere model of warning observation error is constructed as follows:

$$\begin{cases} f(\boldsymbol{r}_T)=\dfrac{1}{(\sqrt{2\pi}\sigma_r)^3}\exp[-\dfrac{\|\boldsymbol{r}_T-\boldsymbol{r}_p\|^2}{2\sigma_r^2}] \\ f(\dot{\boldsymbol{r}}_T)=\dfrac{1}{(\sqrt{2\pi}\sigma_v)^3}\exp[-\dfrac{\|\dot{\boldsymbol{r}}_T-\dot{\boldsymbol{r}}_p\|^2}{2\sigma_v^2}] \end{cases} \tag{1-4}$$

Where the true target position and velocity are $\boldsymbol{r}_T,\dot{\boldsymbol{r}}_T$; the corresponding predicted position and velocity are $\boldsymbol{r}_p,\dot{\boldsymbol{r}}_p$; σ_r,σ_v represent standard deviation of position and velocity errors respectively.

Assuming that the target motion is always within the radar detection range, the depth error along the radar detection direction can be ignored. Formula (1-4) can be converted into a 2D normal distribution error circle in a spherical coordinate system centered on radar, as follows:

$$\begin{cases} f_0(\boldsymbol{X}_0;t_0)=\dfrac{1}{2\pi\sigma^2}\exp[-\dfrac{(\boldsymbol{X}_0-\bar{\boldsymbol{X}}_0)(\boldsymbol{X}_0-\bar{\boldsymbol{X}}_0)^T}{2\sigma^2}] \\ \omega_0(\boldsymbol{V})=\dfrac{1}{2\pi\mu^2}\exp[-\dfrac{(\boldsymbol{V}-\bar{\boldsymbol{V}})(\boldsymbol{V}-\bar{\boldsymbol{V}})^T}{2\mu^2}] \end{cases} \tag{1-5}$$

Where the target motion and search space is set as $\boldsymbol{X}=[x,y]\in\boldsymbol{R}^2$, time domain $t\geq 0$. The single point target velocity space is set as $\boldsymbol{V}=[v_x,v_y]\in\boldsymbol{R}^2$, initial moment $t_0\geq 0$. Where σ and μ represent the standard deviations of position and velocity guidance errors respectively; $\boldsymbol{X}_0$ is the true value of the target position at t_0, $\bar{\boldsymbol{X}}_0$ is the guidance position at t_0; $\boldsymbol{V}$ is the true value of target velocity, $\bar{\boldsymbol{V}}$ is the guidance velocity.

Based on Ref.[29], the real-time position probability distribution model for the constant speed target based on two-dimensional normal distribution is established as follows:

$$f(\boldsymbol{X};t)=\frac{1}{2\pi[\mu^2(t-t_0)^2+\sigma^2]}\exp\{-\frac{[\boldsymbol{X}-\bar{\boldsymbol{X}}_0-\bar{\boldsymbol{V}}(t-t_0)][\boldsymbol{X}-\bar{\boldsymbol{X}}_0-\bar{\boldsymbol{V}}(t-t_0)]^T}{2[\mu^2(t-t_0)+\sigma^2]}\} \quad (1\text{-}6)$$

Secondly, we define the Gaussian Mixture Model (GMM) [30] of two-dimensional constant speed target position distribution probability as follows:

$$p(\boldsymbol{X};t)=\sum_{i=1}^{n}\delta_i f_i(\boldsymbol{X};t) =\frac{1}{2\pi n[\mu^2(t-t_0)^2+\sigma^2]}\sum_{i=1}^{n}\exp\{-\frac{[\boldsymbol{X}-\bar{\boldsymbol{X}}_{i0}-\bar{\boldsymbol{V}}_i(t-t_0)][\boldsymbol{X}-\bar{\boldsymbol{X}}_{i0}-\bar{\boldsymbol{V}}_i(t-t_0)]^T}{2[\mu^2(t-t_0)+\sigma^2]}\} \quad (1\text{-}7)$$

Then, the comprehensive interception probability P_{dj} for all targets in airspace j can be defined as follows:

$$P_{dj}=\iint_{\boldsymbol{S}_j} p(\boldsymbol{X};t)d\boldsymbol{X} \quad (1\text{-}8)$$

Where $\boldsymbol{S}_j$ is the covering area of airspace j.

The cost c_j of airspace j can be expressed as follows:

$$c_j=f_{c_j}(n_j,S_j,P_{dj})=e^{[-(\alpha n_j+\beta\frac{P_{dj}}{S_j})]} \quad (1\text{-}9)$$

Where α,β are both positive constants which determine between target number and interception probability, which one has a greater influence on the airspace cost; n_j represents the number of targets calculated to fall into airspace j based on guidance information at the start of guided search task; P_{dj}/S_j is the comprehensive interception probability per unit airspace of airspace j as the search progresses. Formula (1-9) indicates that the cost c_j of airspace j is negatively correlated with the number of targets in airspace j and comprehensive interception probability per unit airspace, i.e., the more covered targets and the higher interception probability, the lower airspace cost.

1.3. Radar Parameter Optimization Model Based on Guided Search of Cluster Targets

The principle of radar search parameter optimization model for cluster targets proposed is shown in Fig. 6.

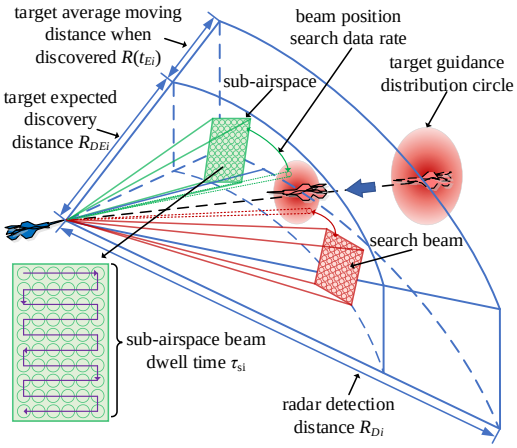


Fig. 6 The principle of radar search parameter optimization model

Phased array radar search performance optimization indicator can be expressed in various forms, such as the maximum starting distance for tracking target or the minimum time for discovering target. Focused on the statistical aspect of search optimization problem, this paper represents the search performance as follows [16][31]: the maximum expected distance for radar to detect approaching targets during the search process, defined as the expected target discovery distance. This represents the average distance from the radar when a breakthrough target enters the detection range until it is finally detected. Under predetermined search beam position arrangement strategy, the search data rate of each beam position is optimized to maximize the accumulated discovery probability of the target within the radar search airspace update cycle, defined as the target accumulated discovery probability.

1.3.1 Sub-airspace Beam Dwell Time Optimization Model Based on Maximum Expected Discovery Distance for Cluster Targets

Assuming the radar surveillance airspace is divided into N discrete nonoverlapping sub-airspaces, R_{Di} ($i\in1,...,N$)(corresponding detection probability is p_d) represents the radar detection distance for sub-airspace i. Assuming the uniform beam scanning strategy is used, the expected target discovery distance can be expressed as follows:

$$\bar{R}_{DE}=\sum_{i=1}^{N}\alpha_i R_{DEi}=\sum_{i=1}^{N}\alpha_i[R_{Di}-R(t_{Ei})] =\sum_{i=1}^{N}\alpha_i[R_{Di}-R(0.5t_f)] \quad (1\text{-}10)$$

Where α_i represents normalized weight coefficient of each sub-airspace assigned based on target guidance information, satisfying $\sum_{i=1}^{N}\alpha_i=1$, which means the search resource sum of all sub-airspaces is 1. t_f is the radar search frame period. t_E=0.5t_f is the expected target discovery time.

According to the radar equation [32], R_{Di} satisfies:

$$R_{Di}=\sqrt[4]{\frac{\Omega_i\tau_{si}}{SNR_D}},\Omega_i=\frac{P_{av}G_{ti}G_{ri}\lambda^2\sigma}{(4\pi)^3kT_0F_nL} \quad (1\text{-}11)$$

Where the radar system constant Ω_i of each sub-airspace is related to the radar system [33]; P_{av} is the average transmission power; G_{ti} and G_{ri} are the transmission and reception antenna gains, respectively; λ is the radar wavelength; σ is the target RCS; k is Boltzmann constant; T_0 is receiver noise temperature (at room temperature, 290K); F_n is the receiver noise figure; L is the radar system loss; SNR_D is the echo signal-to-noise ratio at the radar detection distance satisfying $p_d=p_{fa}^{\frac{2}{SNR}[1-\frac{2\ln(p_{fa})}{SNR}]}$ [34]. It is a constant when the false alarm probability p_{fa} and detection probability p_d are given; τ_{si} is the beam dwell time of each sub-airspace;

Based on target guidance information, weighting the expected target discovery distance of all sub-airspaces to obtain:

$$\bar{R}_{DE}=\sum_{i=1}^{N}\alpha_i R_{DEi}=\sum_{i=1}^{N}\alpha_i\sqrt[4]{\frac{\Omega_i\tau_{si}}{SNR_D}}-\frac{t_f}{2}\sum_{k=1}^{n}w_k v_k \quad (1\text{-}12)$$

Where v_k represents the target speed; n is the number of targets; w_k is the normalized target threat coefficient, satisfying $\sum_{k=1}^{n} w_k = 1$. The threat weight coefficient of each sub-airspace (α_i) can be calculated based on the internal target threat information as follows:

$$\begin{cases} \alpha_i' = \mathrm{clip}\left(\dfrac{\sum_{q\in Q_i}^{n} w_q}{\sum_{p=1}^{n} w_p}, \alpha_{\min}, \alpha_{\max}\right) \\ \alpha_i = \dfrac{\alpha_i'}{\sum_{j=1}^{N} \alpha_j'} \end{cases} \tag{1-13}$$

Where Q_i represents the target collection covered by sub-airspace i. Formula (1-13) simultaneously restricts the upper and lower limits of the sub-airspace threat coefficient to prevent individual sub-airspaces from occupying too much search resources and affecting the search effect of other sub-airspaces, as well as to avoid allocating too few search resources to individual sub-airspaces and thus preventing missed targets.

Assuming the number of search beam positions N_s in the sub-airspace are equal, the total beam position number of search airspace is $N*N_s$ which is usually set to three times the guidance standard deviation, i.e., the 3σ rule. The relationship between beam dwell time τ_{si} of each sub-airspace and entire radar search frame period t_f is as follows:

$$N_s \sum_{i=1}^{N} \tau_{si} = t_f \tag{1-14}$$

Substituting equation (1-14) into equation (1-12), we get:

$$\bar{R}_{DE} = \sum_{i=1}^{N} \alpha_i \sqrt[4]{\frac{\Omega_i \tau_{si}}{SNR_D}} - \frac{1}{2} N_s \sum_{i=1}^{N} \tau_{si} \sum_{k=1}^{n} w_k v_k \tag{1-15}$$

Deriving equation (1-15) with respect to τ_{si} yields:

$$\frac{\partial \bar{R}_{DE}}{\partial \tau_{si}} = \alpha_i \frac{1}{4} \left(\frac{\Omega_i}{SNR_D}\right)^{\frac{1}{4}} \tau_{si}^{-\frac{3}{4}} - \frac{1}{2} N_s \sum_{k=1}^{n} w_k v_k = 0 \tag{1-16}$$

Solving for τ_{si} gives:

$$\tau_{si} = \frac{\alpha_i^{\frac{4}{3}} \left(\frac{\Omega_i}{SNR_D}\right)^{\frac{1}{3}}}{\left(2N_s \sum_{k=1}^{n} w_k v_k\right)^{\frac{4}{3}}} \tag{1-17}$$

From equation (1-17), the optimal beam dwell time τ_{si} of each sub-airspace can be determined. Section 1.3.2 will solve for the search data rate of each beam position within each radar search frame period based on τ_{si}.

1.3.2 Sub-airspace Beam Position Search Data Rate Optimization Model Based on Maximum Average Accumulated Discovery Probability for Cluster Targets

In general, the guided search window is small, and the search time is relatively short. Additionally, targets indicated by the guidance information are typically far from the radar, so it is assumed that targets do not exit the coverage range of single beam position during the search frame period [35]. For target i, the optimal beam position search data rate model in sub-airspace j based on the maximum accumulated discovery probability of target i is as follows:

Assuming the target detection probability of radar is p_{d0}, and the radar irradiates k-th beam position of sub-airspace j for n^i_{jk} times within a radar search frame period t_f, the accumulated discovery probability of target i at this beam position is as follows:

$$P^i_{jk} = p^i_{jk}\left[1-(1-p_{d0})^{n^i_{jk}}\right] \tag{1-18}$$

Where p^i_{jk} is the probability of target i appearing at this beam position. Assuming that the coverage area of this beam position is $\boldsymbol{S}_{jk}$, the probability p^i_{jk} can be obtained based on the constant speed target position distribution probability model from section 1.2:

$$p^i_{jk} = \iint_{\boldsymbol{S}_{jk}} f(\boldsymbol{X};t)d\boldsymbol{X} \tag{1-19}$$

Then the optimization model can be established as follows:

$$\begin{aligned} &\max_{n^i_{jk}} P^i = \sum_{k=1}^{N_s} P^i_{jk} = \sum_{k=1}^{N_s} p^i_{jk}\left[1-(1-p_{d0})^{n^i_{jk}}\right] \\ &s.t. \sum_{k=1}^{N_s} \tau_{sj} n^i_{jk} = t_f \cdot \alpha_j = N_s \alpha_j \sum_{j=1}^{N} \tau_{sj} \end{aligned} \tag{1-20}$$

The search data rate for each beam position is then calculated as:

$$n^i_{jk} = \frac{\alpha_j}{\tau_{sj}} \sum_{m=1}^{N} \tau_{sm} + \frac{\frac{1}{N_s}\ln\left(\prod_{l=1}^{N_s} p^i_{jl}\right) - \ln\left(p^i_{jk}\right)}{\ln(1-p_{d0})} \tag{1-21}$$

Based on equation (1-21), the optimal search data rate n^i_{jk} for beam position k corresponding to target i in the sub-airspace j is determined. The search data rates of beam positions corresponding to each target are then summed and normalized to obtain cooperative beam position search data rate n_{jk} based on the maximum average accumulated discovery probability for cluster targets as follows:

$$n_{jk} = \frac{\sum_{i=1}^{n} n^i_{jk}}{\sum_{k=1}^{N_s}\sum_{i=1}^{n} n^i_{jk}} N_s \tag{1-22}$$

Based on target guidance information and search airspace position, radar search parameter optimization models are constructed for the cluster target scenario, which optimize both sub-airspace beam dwell time and beam position search data rate to improve the expected discovery distance and average accumulated discovery probability for cluster targets simultaneously.

2. Happo-rgs Algorithm

2.1. Algorithm Structure

In airborne radar guided search task, reinforcement learning agents determine the radar search airspace and optimize relevant radar search parameters based on guidance information and target discovery status. The decision-making process of radar guided search task in scenarios with widespread distribution of cluster targets satisfies the Markov decision process, which can be decomposed into two sub-problems. First sub-problem is to determine the current search airspace orientation for the optimal search efficiency ratio, i.e., the search airspace decision problem. Second sub-problem is to optimize sub-airspace beam dwell time after dividing the sub-airspaces according to a fixed strategy based on the target guidance information within the search airspace. Then, a sub-airspace beam position search model is established based on the radar capability to solve for the search data rate of each beam position, i.e., the search parameter optimization problem.

Considering different sub-problem characteristics, the radar guided search task decision-making problem in the above scenario is divided into two sub-strategies, defined as upper-level and lower-level strategies. And a new hierarchical reinforcement learning framework is proposed to guide the agent training process.

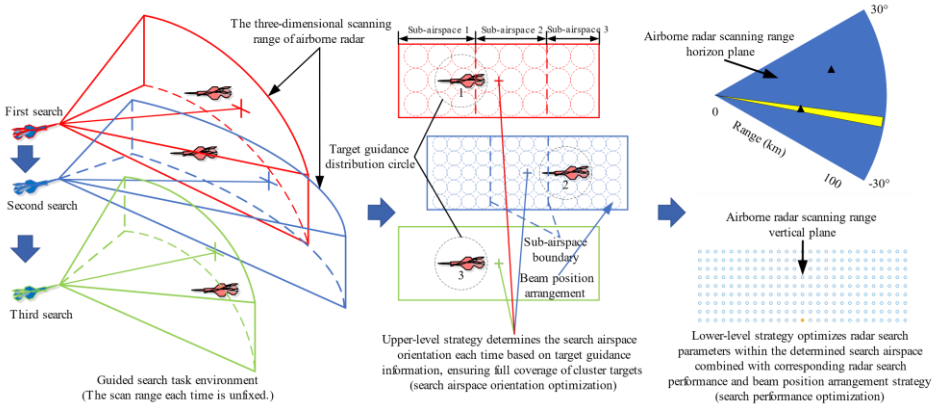


Fig. 7 Schematic diagram of upper-level strategy and lower-level strategy

As shown in Fig. 7, the upper-level strategy selects the airspace to be searched based on target guidance information and the airspace set set-covering model from section 1.2. The lower-level strategy optimizes sub-airspace beam dwell time and beam search data rate based on the target distribution probability within the search airspace and search parameter optimization model from section 1.3. These strategies correspond to different sub-tasks in the radar guided search task.

2.2. Airspace Search Environment for Cluster Targets Based on Happo-rgs

2.2.1 Observation Space and Action Space

The entire airspace search strategy consists of two parts: an upper-level reinforcement learning radar search orientation optimization strategy and a lower-level fixed radar search parameter optimization strategy. The target guidance information combined with target discovery status serves as the observation space for upper-level reinforcement learning strategy module. The optimization indicator for the upper-level strategy module is defined as search airspace number required to complete the cluster target search task. By adjusting the search airspace coordinates, the goal is to quickly discover targets and reduce the redundancy of the search airspaces. To avoid redundant observation information and make full use of guidance information, we design a dual observation space s_t^{up} to combine the current search process and guidance information for the upper-level strategy module. s_t^{up} can be expressed as follows:

$$s_t^{up} = [\{\varepsilon_i, \hat{x}_i, \hat{y}_i\}_{i=1}^{n}] \tag{2-1}$$

Where ε_i indicates the number of target covered times; $\hat{x}_i, \hat{y}_i$ represent the target azimuth and pitch guidance coordinates respectively. The dual observation space is visualized as shown in Fig. 8.

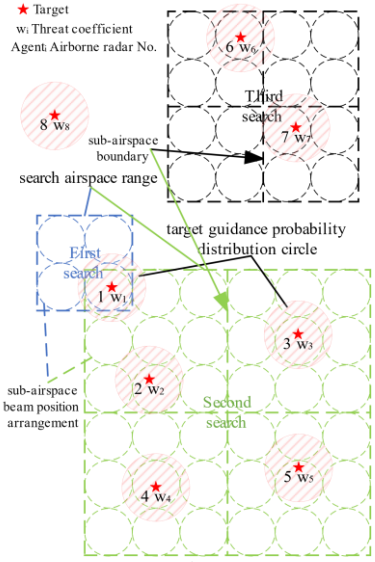


Fig.8 Visualization of observation space.

As illustrated, three airspaces have been searched: The first search includes targets 1; The second search partially overlaps with the first, rediscovering target 1 and newly discovering target 2,3,4,5 which means the first search can be considered as invalid search; The third search newly discovers target 6,7 with target 8 unfound. The parameters change as follows:

$$\xi=[\varepsilon_1,\varepsilon_2,...,\varepsilon_n]=[2,1,1,1,1,1,1,0]$$
$$\hat{X}=[\hat{x}_1,...,\hat{x}_n],\hat{Y}=[\hat{y}_1,...,\hat{y}_n],s_t^{up}=[\xi,\hat{X},\hat{Y}]$$

To simplify task format and enable policy network to further explore optimal decision actions, we design a continuous action space to solve the airspace set-covering problem. The azimuth and pitch coordinates of the next search airspace are taken as the executable action a_t^{up} for the upper-level strategy module. With the search airspace size fixed, a_t^{up} can be expressed as:

$$a_t^{up}=[\text{azimuth\_center},\text{pitch\_center}] \tag{2-2}$$

The lower-level strategy module needs to divide the airspace to be searched into multiple sub-airspaces using a fixed strategy at first. Then, based on the target expected discovery distance and average accumulated discovery probability respectively, the sub-airspace beam dwell time and the search data rate of each beam position are optimized. The observation and action spaces of upper-level strategy module are combined as the observation input s_t^{low} for the lower-level strategy module. The optimized radar search parameters are set as the executable action a_t^{low}. The s_t^{low} and a_t^{low} can be respectively expressed as:

$$s_t^{low}=[s_t^{up},a_t^{up}],a_t^{low}=[\{\tau_{si}\}_{i=1}^{N},\{n_{ij}\}_{i=1,j=1}^{N,N_s}] \tag{2-3}$$

The observation space and action space of entire airspace search environment can be expressed as follows:

$$s_t=s_t^{up},a_t=[a_t^{up},a_t^{low}] \tag{2-4}$$

2.2.2 Reward Function

The optimization goal of the upper-level strategy module is to minimize the number of search airspaces while ensuring all targets discovered. Therefore, the search times are the primary factor to be considered. However, different initial distributions of targets will affect the final search times. Denser distributed targets need fewer final search times, otherwise the opposite. Considering training efficiency, this paper combines a greedy algorithm (selecting the airspace containing the most undiscovered targets each time) as the baseline search times to define the following reward function:

$$reward_1=\alpha(n_{greedy}-n_{rl}) \tag{2-5}$$

Where α is a positive constant, n_{greedy} is the number of search times required by the greedy algorithm, and n_{rl} is the search times of current RL episode.

In the early training stage, policy network is in the exploration phase, while formula (2-5) only provides reward feedback at the end of each episode. To prevent slow learning due to sparse rewards, a process reward is defined as follows:

$$reward_2=\beta(n_{start}-n_{end}) \tag{2-6}$$

Where β is a positive constant, n_{start} is the number of undiscovered targets before the current search, while n_{end} is the number of undiscovered targets after search.

In the later training stage when the policy network can output more effective episode strategies, reducing search airspace redundancy becomes a goal to prevent overfitting and encourage exploration. The redundancy reward function for the search airspace per episode is defined as follows:

$$reward_3=-\gamma S\tanh(\frac{t}{t_0})=-\gamma S\frac{e^{\frac{t}{t_0}}-e^{-\frac{t}{t_0}}}{e^{\frac{t}{t_0}}+e^{-\frac{t}{t_0}}} \tag{2-7}$$

Where γ and t_0 are positive constants; S represents the overlapping area of search airspaces in the episode; t is the number of training steps.

Combining the target guidance information to select appropriate search airspace can significantly improve the efficiency of the lower-level strategy in optimizing radar search parameters. The policy network outputs search airspace coordinates as the action of upper-level strategy module and partial observation of lower-level strategy module. Therefore, the optimization effect of lower-level strategy module can significantly reflect the quality of upper-level strategy. The optimization effect reward of lower-level is defined as:

$$reward_4=\frac{\bar{R}_{DE}}{\bar{R}_{DE}'}\times\frac{\sum_{i=1}^{n}P_i}{\sum_{i=1}^{n}P_i'} \tag{2-8}$$

Where $\bar{R}_{DE},\bar{R}_{DE}'$ are the weighted target expected discovery distance of entire search airspace achieved by algorithm proposed and a uniform search algorithm, respectively. $\frac{1}{n}\sum_{i=1}^{n}P_i,\frac{1}{n}\sum_{i=1}^{n}P_i'$ are the comprehensive accumulated discovery probability of cluster targets based on the two algorithms.

To prevent coupling of various rewards and improve task completion rate, the episode task completion rate reward function is defined as follows:

$$reward_5=\begin{cases}r_0 & \text{if done \& } t\le \text{step}_{\max}\\ -r_0 & \text{if } t>\text{step}_{\max}\end{cases} \tag{2-9}$$

Where $\text{step}_{\max}$ is the maximum steps of an episode; r_0 is a positive constant.

To sum up, the comprehensive reward function for the episode is as follows:

$$\begin{cases}R=[reward_1+\sum_{i=1}^{T}reward_2^i]\cdot reward_4+reward_3+reward_5\\ reward_2^i=\beta(n_{start}^i-n_{end}^i)\end{cases} \tag{2-10}$$

Where T is the actual episode steps; n^i_{start} is the number of undiscovered targets before each search step; n^i_{end} is the number of undiscovered targets after search.

2.3. Hierarchical Strategy Structure for Airspace Guided Search of Cluster Targets Based on Happo-rgs

2.3.1 Upper-level Strategy Module

a. Upper-level Strategy Module Structure

Upper-level strategy module analyzes the impact of radar parameters and air combat situation on search airspace and studies multi-level airspace set generation methods. It selects search airspace based on airspace set-covering model proposed in section 1.2, and the model solution serves as a partial input for lower-level strategy module. After generating the search airspace set and calculating the cost matrix, upper-level strategy module solves the airspace set-covering problem based on the probability distribution of target position within divided airspace and radar search performance. It seeks the optimal airspace set to ensure that search task is completed with the least number of search airspaces. The structure of the upper-level strategy module is as shown in Fig. 9.

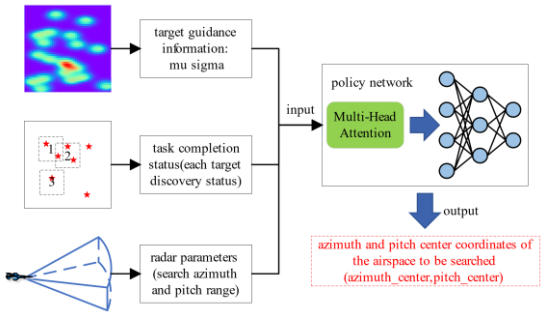


Fig.9 Upper-level strategy module structure.

As shown in Fig. 9, we design the policy network based on multi-head attention mechanism to directly learn the solving process of the airspace set-covering model in section 1.2. The search airspace solution does not rely on branch-and-bound or heuristic methods for complex calculation but is directly transformed into iterative learning of network parameters. The model solution accuracy only relies on the division accuracy of discrete search airspace set and the strategy learning effectiveness. After interacting with the environment for training, inputting target guidance information and radar search parameters allows the policy network to output the azimuth and pitch center coordinates of the airspace to be searched based on current task progress which greatly simplifies the solving process for the set-covering problem.

b. Policy Network Structure

The radar search task decision process satisfies the Markov decision process because of the continuity of state space sequence within a single episode. The task progress states are strongly correlated with the states at previous moment. Traditional reinforcement learning agents cannot identify correlated positions, resulting in poorer learning effectiveness. By introducing learnable weights, the attention mechanism helps agents more accurately select and learn useful information to improve model performance. For example, within a single episode, the target guidance coordinates do not change, and the attention mechanism can help agents focus on the guidance information of uncovered targets to better complete the task; after the environment resets, the agent can pay attention to updated target guidance coordinates to generate better feasible airspace set. The policy network structure incorporating the attention mechanism is as shown in Fig. 10:

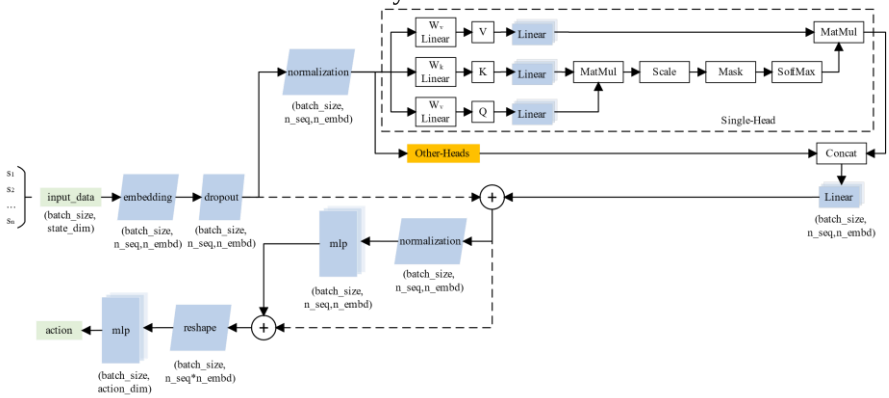


Fig.10 Policy network structure incorporating attention mechanism

As shown in Fig. 10, two-dimensional input data consisting of multiple continuous state sequences is mapped to three-dimensional vector by an embedding layer to extract features from guidance information of undiscovered target. After the embedding layer is a dropout layer to enhance model generalization ability and prevent overfitting. Then input data is normalized and input into the multi-head attention module where data dimension doesn't change. Finally, corresponding actions are output after passing by a residual structure to reduce data dimension. The multi-head attention mechanism is a variant of attention mechanism that can simultaneously focus on different aspects of input data and capture interaction and dependency between these aspects, which helps the model better handle complex inputs [36].

The query matrix Q, key matrix K, and value matrix V input to the masked multi-head attention layer are obtained from the input sequence X by different linear layers. The sizes of the three matrices are respectively: $Q \in R^{ns \times hs}, K \in R^{ns \times hs}, V \in R^{ns \times hs}$, where ns is sequence length

of the input state vector, and *hs* is the size of the matrix. Assuming the head number of multi-head model is *I*, *I* sets of identical *Q,K,V* are generated and sent to each head. The self-attention mechanism is applied to *i*-th head to calculate the weighted attention result A_i. The outputs from *I* heads are then concatenated and passed through a fully connected layer $W^O \in R^{I*hs\times n\_embd}$ to fit and match different dimensional features produced by the previous heads for unified output result.

$$A_i = Attention(Q_i, K_i, V_i) = \text{softmax}(\text{mask}(\frac{Q_i K_i^T}{\sqrt{d_k}}))V_i \quad (2\text{-}11)$$
$$MultiHead(Q,K,V) = \text{Concat}(A_1,...,A_I)W^O$$

Since the state space sequence within single episode is a causal sequence that satisfies the Markov process and is only affected by states from previous times, we introduce the mask layer in the processing to ensure the model causality. A frame mask with size of *ns*×*ns* is added on the temporal dimension of input sequences, so that the current and previous part of the timeline is set to 1 and the future part to -∞. Only current and previous time points are involved in attention weight calculation to reflect the model causality [37].

2.3.2 Lower-level Strategy Module

The upper-level strategy module selects airspace to be searched, while the lower-level strategy module, based on target guidance information and radar search performance parameters, divides the search airspace into secondary sub-airspaces by a fixed strategy. Then, based on the maximum expected discovery distance and the average accumulated discovery probability of cluster targets, the sub-airspace beam dwell time and search data rate are optimized respectively. The lower-level strategy module is constructed as shown in Fig.11.

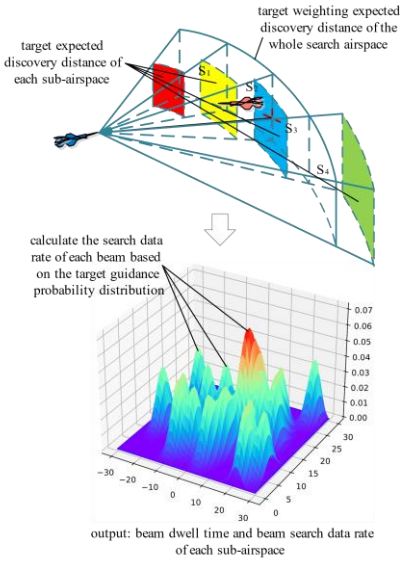


Fig.11 Lower-level strategy module structure.

As shown in Fig. 11, lower-level strategy module maximizes the weighted expected discovery distance of cluster targets in selected airspace by adjusting the beam dwell time of each sub-airspace under fixed search resource load. After determining optimal beam dwell time for each sub-airspace, lower-level strategy module optimizes the search data rate of each beam position based on the target falling into probability to achieve better guided search performance. Combined with target guidance information and search airspace selected by upper-level strategy, lower-level strategy module is constructed based on the radar guided search parameter cooperative optimization model for cluster targets.

2.4. Happo-rgs Algorithm Process

2.4.1 Expert Data Set Generation for the Upper-level Strategy Module Based on Genetic Algorithm

Considering the high randomness of cluster target guidance distributions and the significant influence of initial search airspace selected by upper-level strategy module on the result reward, the happo-rgs algorithm incorporates a genetic algorithm to solve the airspace set-covering problem [38]. Firstly, numerous solutions are generated by genetic algorithm as an expert data set based on different initial target distributions. During the initial training phase, upper-level strategy module learns the fixed expert search strategies from the data set. Once the policy network can stably complete the search task, the network parameters are iteratively updated based on objective function of the PPO-clip algorithm.

The genetic algorithm adopts binary encoding and heuristic methods for population initialization, where half of the population is generated randomly, and the other half directly uses feasible solutions. It employs a dual fitness function based on linear transformation, f_1 and f_2, to evaluate individual fitness, where f_1 can be represented as follows:

$$f_1 = \begin{cases} b/(\sum_{j=1}^{m} c_j x_j + d) & \sum_{j=1}^{m} a_{ij} x_j \geq 1, i = 1,2,...,n, x_j \in \{0,1\}, j = 1,2,...,m. \\ 0 & \text{else} \end{cases} \quad (2\text{-}12)$$

Where *b,d* are positive constants. Individual fitness is negatively correlated with its cost.

To prevent premature convergence, fitness function f_1 is subjected to linear scaling transformation. In the early stage of evolution, the individual fitness below population average is enhanced, otherwise the opposite. In the later stage, to ensure population diversity, the fitness differences are increased, and the evolutionary stage is distinguished based on the standard deviation of the sample [39]:

$$\sigma = \sqrt{\frac{1}{P_s - 1}\sum_{i=1}^{P_s} | f_1(X_i) - \bar{f}_1 |} \quad (2\text{-}13)$$

Given a critical value δ, the linear variation of fitness f_1 is defined as follows:

$$f_2 = \alpha \times f_1 + \beta \quad (2\text{-}14)$$

Where α, β are determined as follows:

(1) If $\sigma > \delta : \alpha = r, \beta = (1-r) \times \bar{f}_1; r = \text{random}[0,1]$

(2) If $\sigma \le \delta$, then

$$\begin{cases} \alpha = 0.5 f_{1avg} / (f_{1avg} - f_{1min}) \\ \beta = f_{1avg} \times (0.5 f_{1avg} - f_{1min}) / (f_{1avg} - f_{1min}) \end{cases}$$

Where f_{1avg}, f_{1max}, f_{1min} are the average, maximum, and minimum of fitness f_1, respectively. It is evident that when $f_1=f_{1min}$, $f_2=0.5f_{1avg}$; when $f_1=f_{1avg}$, $f_2=f_{1avg}$; when $f_1=f_{1max}$, $f_2=f_{1avg}+ 0.5f_{1avg}(f_{1max} - f_{1avg})/(f_{1avg} - f_{1min})$, at this point $f_2>f_{1avg}$.

The selection operation adopts the roulette wheel selection method. Given that improving the solution precision of the set-covering problem implies reducing the division step of airspace to increase the airspace discretization, the feasible solutions contain a much smaller number of selected airspaces compared to the total divided airspaces. Therefore, to avoid damaging good gene strings, single-point crossover is adopted in the crossover operation. On the contrary, the mutation operation employs adaptive multi-bit mutation, where the number of mutation bits are different for feasible and infeasible solutions. The specific operation is randomly selecting a corresponding number of gene positions for inversion. The number of mutation positions in adaptive multi-bit mutation is given by:

$$\begin{cases} round(mu \times \exp(me \times \dfrac{f_{1max} - f_{1avg}}{f_{1max} - f_{1min}}) / (1+\sigma/\delta)) & f_1 > 0 \\ round(mu / (1+\sigma/\delta)) & f_1 = 0 \end{cases} \quad (2\text{-}15)$$

Where *mu*, *me* are positive constants; $f_1>0$ indicates the individual to be mutated is feasible; $f_1=0$ indicates the opposite; and *round* represents rounding to the nearest integer.

2.4.2 Training Process of the Hierarchical Strategy Framework Based on Happo-rgs

The happo-rgs (Hierarchical Attention PPO-Radar Guided Search) algorithm framework proposed in this paper is as shown in Fig. 12.

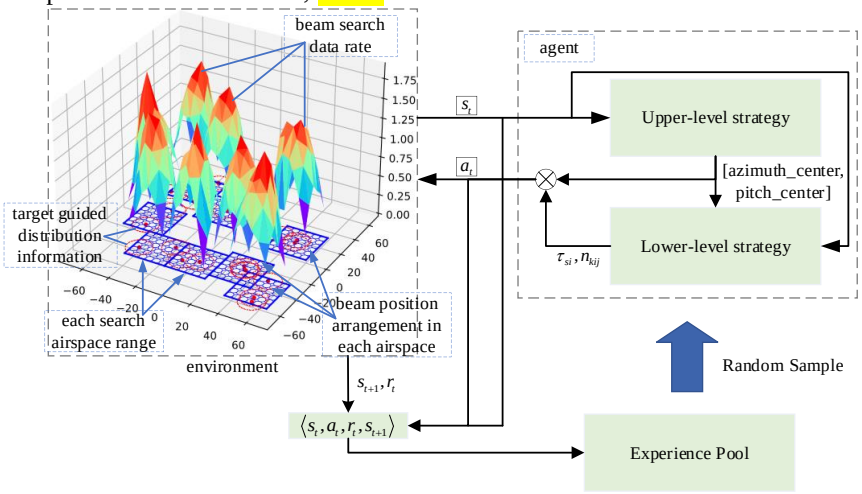


Fig.12 Happo-rgs algorithm architecture.

The happo-rgs training algorithm adopts improved PPO-clip algorithm based on the Actor-Critic structure [40]. The Actor network, i.e., the policy network, is denoted as $\pi_\theta(s_t)$, where s_t represents the state at time *t*, and θ represents the policy network parameters. The policy network outputs action $a_t \sim \pi_\theta(s_t)$. The Critic network, i.e., the value network, is denoted as $V_\varphi(s_t)$, where φ represents value network parameters. The value network is used to estimate the current policy return R_t, which can be expressed as follows:

$$R_t = E_{a \sim \pi_\theta(\cdot|s)}(\sum_{t'=t}^{\infty} \gamma^{t'} r(s_{t'}, a_{t'})) \quad (2\text{-}16)$$

Where $E(.)$ represents the mathematical expectation; γ is the discount factor to ensure convergence of Markov decision process; r is the reward function. The goal of reinforcement learning algorithm is to maximize the episode return.

The PPO-clip algorithm uses the advantage function A^θ to evaluate policy quality and improve algorithm stability. A^θ is defined as follows:

$$A^\theta(s_t, a_t) = E_\theta(R_t \mid s_t, a_t) - V^\theta(s_t) \quad (2\text{-}17)$$

In practice, the Generalized Advantage Estimation (GAE) method is used to estimate A^θ where defines the estimated value $\hat{A}_t$ as follows:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + ... + (\gamma\lambda)^{T-t+1}\delta_{T-1} \quad (2\text{-}18)$$

Where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$, and the parameter λ is used to balance variance and bias.

Additionally, the important sampling method is adopted to directly clip the probability amplitude of old and new policies. The loss functions for the policy and value networks of PPO-clip are as follows:

$$\begin{aligned} L^{ppo}_{policy}(\theta) &= E_t[\min(c_t(\theta)\hat{A}_t, \text{clip}(c_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)] \\ L^{ppo}_{value}(\varphi) &= E_t[\frac{1}{2} \| \hat{R}_t - V_\varphi(s_t) \|^2] \end{aligned} \quad (2\text{-}19)$$

Where $c_t(\theta) = \pi_\theta(a_t \mid s_t) / \pi_{\theta,old}(a_t \mid s_t)$

The hierarchical strategy training process based on the happo-rgs algorithm is shown in Fig. 13.

Training process of the hierarchical strategy framework based on happo-rgs

1.Initialize policy network parameters θ_0 and value network parameters ϕ_0;

2.Initialize hyperparameters: expert data set E generated based on genetic algorithm, experience replay pool D, policy network learning rate l_{actor}, value network learning rate l_{critic}, maximum training steps T_{max}, maximum episode steps $step_{max}$, pre-training steps $step_0$, current training step t;

3.**while** $t < T_{max}$ **do**

4. Reset the environment and initialize observation state $s_t \leftarrow s_0$;

5. **while** not done **do**

6. The policy network selects the upper-level policy action $a_t^{up}=\pi_{\theta k}(s_t^{up})$ based on the upper-level observation state s_t^{up};

7. Execute a_t^{up} to obtain the lower-level observation state $s_t^{low}=[s_t^{up},a_t^{up}]$;

8. Calculate the lower-level policy action a_t^{low} based on formulas (1-17), (1-22), and s_t^{low} to obtain the composite action $a_t=[a_t^{up},a_t^{low}]$;

9. Execute a_t to obtain the composite reward r_t and transition state $s_{t+1}=[s_{t+1}^{up},s_{t+1}^{low}]$, and combine them into a tuple (s_t,a_t,r_t,s_{t+1}) to store in the experience replay pool D;

10. $t=t+1$;

11. **if** update net **do**

12. if $t < step_0$: goto 13, else: goto 14

13. $t < step_0$, update policy network parameters θ_k by minimizing following loss function;

$$L_t^{\pi}(\theta_k)=\sum_{s_t,a_t}(\pi_{\theta_k}(s_t)-E(s_t))^2$$

14. $t \geq step_0$, update policy network parameters θ_k by maximizing following loss function;

$$L_t^{\pi}(\theta_k)=\sum_{s_t,a_t}\min\left(\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_k}(a_t\mid s_t)}A^{\pi_{\theta_k}}(s_t,a_t),\mathrm{clip}\left(\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_k}(a_t\mid s_t)},1-\varepsilon,1+\varepsilon\right)A^{\pi_{\theta_k}}(s_t,a_t)\right)$$

15. end if

16. Update value network parameters ϕ_k by minimizing following loss function;

$$L_t^{V}(\phi_k)=\sum_{s_t,a_t}(V_{\phi_k}(s_t)-\hat{R}_t)^2=\sum_{s_t,a_t}(V_{\phi_k}(s_t)-\sum_{l=0}^{\infty}\gamma^l r_{t+l})^2$$

17. **end if**

18. **end while**

19.**end while**

Fig.13 Training process of the hierarchical strategy framework based on happo-rgs.

3. Algorithm Performance Verification of Intelligent Hierarchical Strategy Decision-Making Algorithm for Radar Guided Search Tasks Based on Happo-rgs

3.1. Parameter Settings

3.1.1 Upper-level Strategy Module

a. Fixed Target Distribution Scenarios

Genetic algorithm parameter settings: population size P_s=100; crossover probability *pcross*=0.6; mutation probability *pmutation*=0.1; fitness function f_1 parameters: b=20, d=-10; adaptive multi-bit mutation parameters: *mu*=10.0, *me*=1.0, δ=1.0; maximum evolutionary epochs: $epoch_{max}$=100

Airspace search environment parameter settings: target number n=10;distribution ranges $x,y\in[-30^\circ,30^\circ]$; maximum episode steps $step_{max}$=20; radar azimuth and pitch search ranges rx=ry=30°; γ=1.0/(rx*ry), α=1.0, β=1.0, r_0=10. The azimuth and pitch coordinates of targets are listed in Table 4-1.

Table 4-1 Target azimuth and pitch coordinate parameters.

Target	1	2	3	4	5	6	7	8	9	10
Azimuth	21.01	28.84	-3.25	-22.30	22.17	29.00	-7.09	28.99	23.33	-27.29
Pitch	6.58	-12.18	8.38	-2.66	29.65	21.91	-22.25	28.44	-2.43	-0.30

For the fixed and random target distribution search scenarios, the following RL algorithms are used for comparison with happo-rgs algorithm: PPO-clip[40], SAC[41], TD3[42], DDPG[43], and compare whether adding multi-head attention module with themselves respectively. Main hyperparameters of each reinforcement learning algorithm are set as Table 4-2:

Table 4-2 Reinforcement learning algorithm hyperparameters

Parameter Name	PPO-clip	SAC	happo-rgs	TD3	DDPG
n_steps	1024		256		
batch_size	64		64		
n_epochs	10		10		
gamma	0.98	0.98	0.98	0.98	0.98
gae_lambda	0.95		0.95		
clip_range	0.2		0.2		
learning_rate	0.001	0.001	0.001	0.0001	0.0001
learning_starts		100		500	500
tau		0.005		0.005	0.005
buffer_size		10^6		10^6	10^6
action_noise_mean				0.0	0.0
action_noise_std				0.1	0.1
target_policy_noise				0.2	0.2
target_noise_clip				0.5	0.5
train_steps	3×10^5	3×10^5	3×10^5	3×10^5	3×10^5

Multi-head attention module parameter settings: The number of embedding dimensions *n_embd*=16; the maximum length of state space sequences *ns*=4; the number of self-attention heads *I*=4; output dimension of each head hs=*n_embd*/*ns*=4. The No. of compared algorithms are shown in Table 4-3.

Table 4-3 The No. of compared algorithms

Algorithm Name	PPO-clip	PPO-clip-ATT	SAC	SAC-ATT	TD3	TD3-ATT	DDPG	DDPG-ATT	happo-rgs
No.	1-1	1-2	2-1	2-2	3-1	3-2	4-1	4-2	5

Simulation environment: CPU Intel i7-10700, RAM 16G; GPU Nvidia RTX3060, memory size 6G. The algorithms PPO, SAC, TD3 and DDPG adopt the deep reinforcement learning library of Stable-Baselines3.0 based on the Pytorch deep learning framework.

b. Random Target Distribution Scenarios

RL and GA algorithm parameter settings in random scenarios are consistent with those in fixed scenarios. The number of cluster targets is increased to 20 and initial positions are randomly generated; distribution ranges $x,y\in[-60^\circ,60^\circ]$; maximum episode steps $step_{max}$ =20; pre-training steps $step_0=10^5$; radar azimuth and pitch search ranges rx=ry=30°; α=1.0, β=1.0, r_0=20, γ=1.0/(rx*ry). For simplicity, radar beamwidth=5°, and each search airspace is divided into 4 sub-airspaces. The beam position arrangement of single radar search airspace and sub-airspace division is shown in Fig. 14.

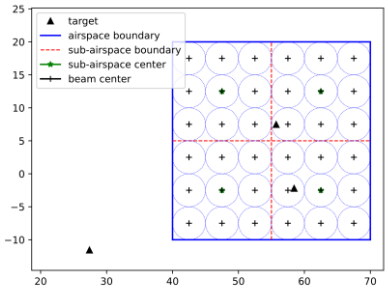


Fig. 14 Search airspace beam position arrangement and sub-airspace division.

3.1.2 Lower-level Strategy Module

Radar parameters: maximum transmission power P_t=600kW, duty cycle pwm=5%, transmission antenna gain G_t=40.0dB, reception antenna gain G_r=40.0dB, radar wavelength λ=0.09m, target RCS=1m^2, Boltzmann constant k=1.380649×10^{-23} J/K, receiver noise temperature T_0=290K, receiver noise factor F_n=3dB, radar system loss L=5dB, detection probability p_d=0.9, false alarm rate p_{fa}=10^{-6}. The radar detection distance echo signal-to-noise ratio can be computed as SNR_D=130.13 by p_d and p_{fa}. The radar system constant Ω_0=4.85×10^{26} is calculated from above parameters. Radar beamwidth=2°.

Number of targets n=20, standard deviation of position and speed guidance information σ=1°, μ=0.3°. The threat coefficients of each target are calculated and normalized based on their respective speed and relative distance to the radar. Specific parameters and airspace settings for the targets are provided in Table 4-4.

Table 4-4 Target parameters

Target No.	1	2	3	…	20
Velocity(m/s)	300	315	330	…	585
Relative Distance (km)	200	210	220	…	385
Threat Coefficient	0.0	0.01286	0.02466	…	0.01571
Belonging Sub-airspace	1	3	1	…	5

Search airspace parameters: Search airspace range azimuth [-30°,30°], pitch [0°,30°]. Assuming the entire search airspace is divided into 5 sub-airspaces, with α_{max}=0.3, α_{min}=0.1, the parameters of each sub-airspace are listed in Table 4-5.

Table 4-5 Search sub-airspace parameters

Airspace No. W_i	Number of Search Beam Positions N_s	Radar System Constant Ω_i	Airspace Threat Coefficient α_i	Azimuth Range	Pitch Range
Sub-Airspace 1	90	$0.8\Omega_0$	0.1046	[−30°,−18°]	[0°,30°]
Sub-Airspace 2	90	$0.9\Omega_0$	0.1888	[−18°,−6°]	[0°,30°]
Sub-Airspace 3	90	$1.0\Omega_0$	0.2226	[−6°,6°]	[0°,30°]
Sub-Airspace 4	90	$1.1\Omega_0$	0.2258	[6°,18°]	[0°,30°]
Sub-Airspace 5	90	$1.2\Omega_0$	0.2579	[18°,30°]	[0°,30°]

The cluster target guidance position distribution and the relative position of the airspace beam position are shown in Fig. 15. The blue dashed line represents the beam position division within the airspace adopting the columnar beam policy; the red dashed line represents the boundary of each sub-airspace which divides entire search airspace into 5 equal sub-airspaces; the black dashed and solid lines respectively represent the error circle of target position distribution before and after search.

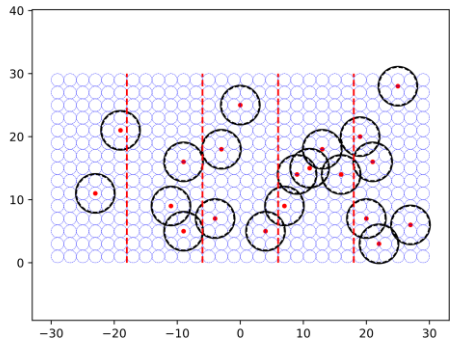


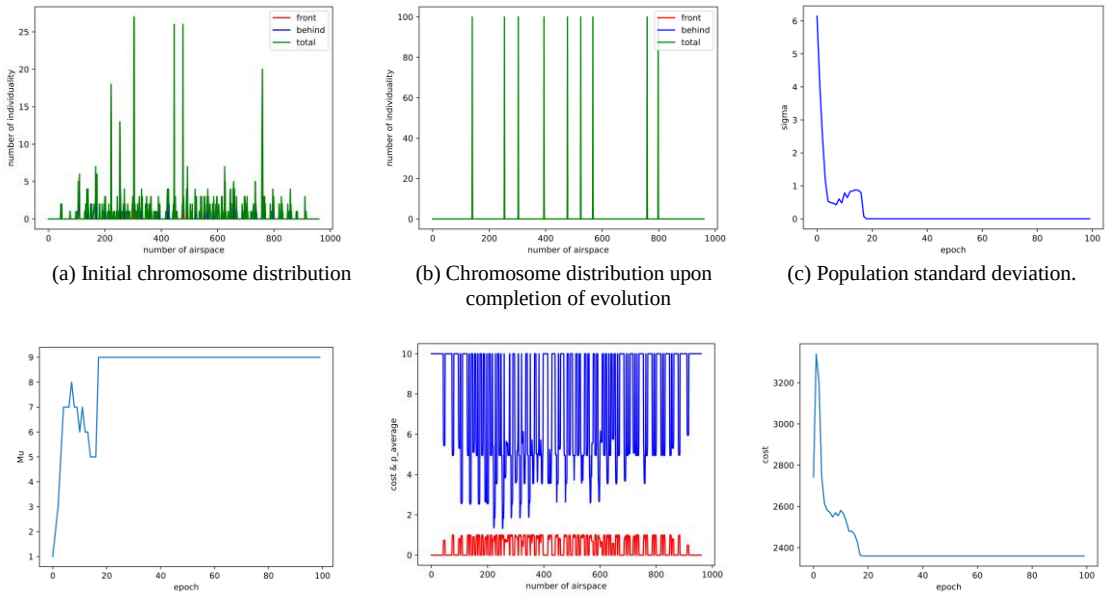
Fig. 15 Relative position of target distribution and airspace beam position division

3.2. Simulation Results

3.2.1 Algorithm Performance Verification in Fixed Target Distribution Scenarios

a. Expert Data Set Generation for the Upper-level Strategy Module Based on Genetic Algorithm

The parameter changes of genetic algorithm population chromosomes and cost matrix are shown in Fig. 16.



(a) Initial chromosome distribution (b) Chromosome distribution upon completion of evolution (c) Population standard deviation.

(d) Number of adaptive mutation bits (e) Cost matrix and average target interception probability (f) Airspace cost sum of best chromosome

Fig. 16 The parameter changes of genetic algorithm population chromosomes and cost matrix

Fig.16(a),(b) show that during evolutionary process, the chromosomes evolve from an initially disordered state to more optimal solutions. Initially, the top 50 individuals are generated by heuristic algorithm with more concentrated initial airspace distribution; the latter 50 are generated randomly with more chaotic initial distribution. After evolution, airspace sets of all individuals concentrate around the best individual set, which validates the effect of evolutionary mechanisms designed in this paper.

Fig.16(c),(d) show that the population standard deviation decreases and converges with evolutionary epochs. Individuals differ significantly in early stage, requiring fewer mutation bits to ensure the crossover operator effect. As the population standard deviation decreases with smaller differences between individuals, the number of adaptive mutation bits are increased to ensure population diversity and prevent local optima. Also, a cap is imposed on the number of mutation bits in a chromosome to prevent excessive mutation from damaging superior genes.

Fig.16(e),(f) show that the costs of airspaces with lower average target interception probability are higher. Conversely, airspaces with higher average interception probability and more targets have lower costs. Total cost of the population decreases and converges with evolutionary epochs. These results show the effect of the proposed evolutionary mechanism and the ability for converging to the optimal solution of the airspace set-covering problem in radar guided search task.

b. Upper-level Strategy Module

The changes in various training parameters of upper-level strategy module under fixed scenarios are shown in Fig. 17:

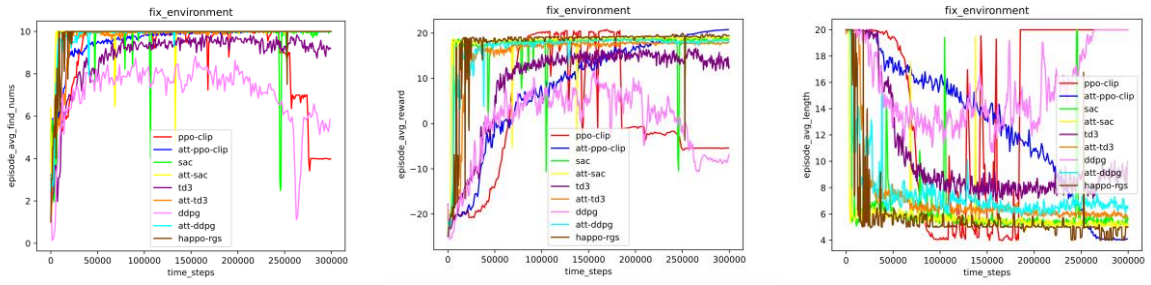


Fig. 17 (a) The number of targets found in an episode (b) Episode reward (c)Episode length

The comparison of various algorithm training parameters is shown in Table 4-6:

Table 4-6. Various algorithm training parameters

Algorithm No.	Episode found num of targets	Episode reward	Episode length	Convergence steps
1-1	9.980±0.230	19.477±0.695	4.505±0.722	100k
1-2	9.999±0.001	20.297±0.283	4.034±0.061	260k
2-1	9.988±0.241	18.486±0.135	5.395±0.169	50k
2-2	9.997±0.053	19.007±0.083	5.193±0.115	30k
3-1	9.524±1.676	15.078±0.790	8.062±0.405	150k
3-2	9.998±0.046	17.778±0.130	5.930±0.126	100k
4-1	7.876±2.817	5.711±1.578	12.866±0.714	100k
4-2	9.995±0.114	18.467±0.155	6.377±0.272	70k
5	**9.996±0.056**	**19.410±0.240**	**4.544±0.439**	**40k**

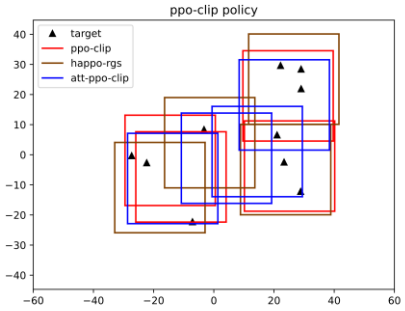
From Fig. 17 and Table 4-6, it can be seen that:

(1) The introduction of multi-head attention mechanism and expert data set of genetic algorithm significantly improves the performance of PPO-clip algorithm, with noticeable improvements in algorithm convergence speed and training stability. For example, the convergence episode reward of algorithm 5 is not much different from algorithms 1-1 and 1-2, but the convergence speed is significantly improved. Besides, the training stability of algorithm 5 is higher than that of algorithm 1-1, and the training reward fluctuates less. Compared with the SAC series algorithms, which are also distributed strategies, the PPO-clip series algorithms have slower convergence speed and higher convergence episode reward, indicating that PPO-clip algorithm is easier to jump out of the local optimum than the SAC algorithm. Since the PPO-clip algorithm is on-policy, the training trajectory needs to adopt the data generated by current policy, resulting that the training efficiency is lower than the SAC algorithm, which is off-policy. These results suggest that although the multi-head attention mechanism is introduced to extract features from continuous state sequences, the fixed target distribution results in limited extraction of guidance information features and less improvement of algorithm performance.

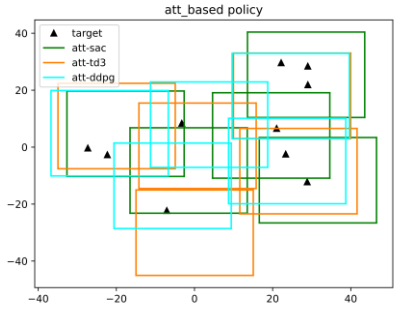
(2) For the TD3 and DDPG algorithms which are deterministic strategies, convergence episode reward and training stability of algorithm 3-2 are higher than those of other three algorithms, which shows the best overall performance. Algorithm 4-2 has the fastest convergence speed, but the training variance is large, with a downward trend in episode reward during the later training stage, which results in increasing episode length. Both DDPG algorithms are inferior to the corresponding TD3 algorithm in terms of convergence

episode reward and length, since TD3 algorithm can train the deep deterministic strategy more stably by using double-Q network and delayed update target policy network strategy to reduce the overestimation of target policy network which improves the algorithm adversariality and makes it more suitable for dealing with problems of the continuous action space.

After completing training, the effect of the above algorithm strategies is verified in Fig. 18. As shown in Fig. 18, the three PPO-clip based algorithms require 4 search times to complete the guided search task for cluster targets. The SAC, TD3, and DDPG algorithms with multi-head attention mechanism all require 5 times to complete the task. The above four types of algorithms do not search zero-target airspaces or repeat searches in fixed target distribution scenarios, showing better strategy effectiveness. The PPO-clip algorithm is easier to jump out of the local optimum, with more average number of targets covered in a single search airspace, showing higher search efficiency. As shown in Fig.17,18 and Table 4-6, all the deep reinforcement learning algorithms compared can converge to higher episode reward and complete episode task in fixed scenarios, which indicates that changes in rewards can encourage agent to learn and act towards completing guided search task. Moreover, the changes of various parameters during the training process are relatively stable. There are less instances of excessive motivation or punishment, which indicates that the process and completion rewards are relatively balanced. Based on the above analysis, the parameter changes during the training process and trained strategies can both reflect the rationality of reward mechanism set in this paper. The reward changes can accurately reflect the behavior of agent and prevent falling into local optimal solution.



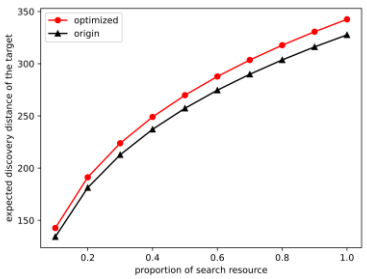
(a) PPO-clip policy



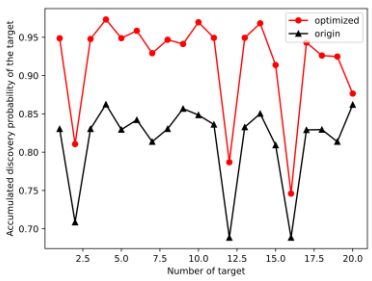
(b) ATT_based policy

Fig.18 Comparison between upper-level strategies trained by above algorithms

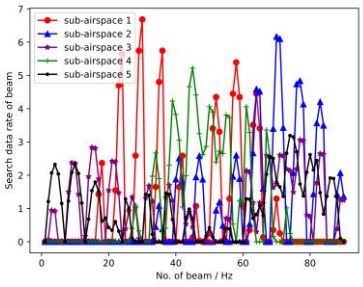
c. Lower-level Strategy Module



(a) Target expected discovery distance. with changing search resource



(b) Target accumulated discovery probability when the search resource is 1



(c) Search data rate for each beam position when the search resource is 1

Fig. 19 The changes in various parameters of lower-level strategy module

The weighted target expected discovery distance of the entire search airspace as a function of the search resource ratio is shown in Fig. 19(a). The accumulated average discovery probability of targets in each sub-airspace when the search resource ratio is 1 is shown in Fig.19(b). Fig.19(a) shows that weighted expected discovery distances calculated by the method proposed and the method of evenly distributing the search beam dwell time based on the number of sub-airspaces both increases as search resource increases. The proposed method performs significantly better, which achieves a greater expected discovery distance for targets at fixed search resource ratio, with a higher advantage as the search resource increases. When the search resource ratio is 1, the comparison of various parameters before and after optimization is listed in Table 4-7.

Table 4-7 Optimization parameters

Sub-Air-space No.	Beam Dwell Time (s) (This Paper/ Unoptimized)	Expected Target Dis-covery Distance (km) (This Paper/ Unoptimized)	Weighted Expected Target Discovery Distance (km) (This Paper/Unoptimized)
1	0.3860/1.1088	85.741/173.381	
2	0.9856/1.1088	309.100/314.466	
3	1.3213/1.1088	386.371/356.911	342.499/327.581
4	1.2894/1.1088	369.267/343.675	
5	1.5614/1.1088	410.180/360.696	

Fig. 19(b) indicates that the proposed sub-airspace beam position search data rate is significantly more effective than the traditional uniform strategy. When

the number of sub-airspace search beam positions is fixed, proposed method achieves higher accumulated discovery probability for targets. Due to the upper and lower limits for threat coefficient of each sub-airspace, instances of over-scanning or miss-scanning key beam positions are markedly reduced. The distribution of the sub-airspace beam position search data rate is shown in Fig.19(c), where beam positions with higher target appearance probability have higher search data rate. This suggests that multiple searches of beam positions with higher target appearance probability within radar search frame period can effectively increase the target accumulated discovery probability. Above simulation results validate the effectiveness of proposed methods in optimizing sub-airspace beam dwell time and beam search data rate in cluster target scenarios. When given the specified radar search airspace ranges and search resource ratio, the method proposed can achieve better radar guided search performance.

3.2.2 Algorithm Performance Verification in Random Target Distribution Scenarios

The changes in various training parameters of upper-level strategy module in random scenarios are shown in Fig. 20:

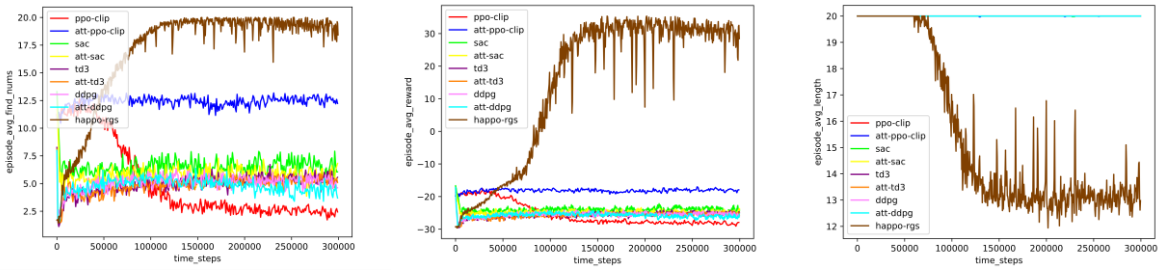


Fig. 20 (a) The number of targets found in an episode (b) Episode reward (c)Episode length

The comparison of various algorithm training parameters is shown in Table 4-8:

Table 4-8. Various algorithm training parameters

Algorithm No.	episode found num of targets	episode reward	episode length	convergence steps
1-1	2.371±2.038	-17.434±0.280	20.0±0.0	200k
1-2	12.587±2.347	-7.314±0.214	19.795±0.203	50k
2-1	6.436±4.324	-13.524±0.546	20.0±0.0	30k
2-2	5.798±3.746	-14.186±0.318	20.0±0.0	30k
3-1	5.440±2.878	-14.560±0.280	20.0±0.0	250k
3-2	5.279±2.729	-14.729±0.267	20.0±0.0	200k
4-1	5.402±3.219	-14.596±0.370	20.0±0.0	100k
4-2	5.384±2.976	-14.524±0.234	20.0±0.0	50k
5	**19.391±0.905**	**26.348±9.253**	**13.985±1.616**	**200k**

From Fig. 20 and Table 4-8, it can be seen that:

(1) The episode reward and task completion rate of happo-rgs algorithm are significantly higher than other baseline algorithms which is followed by the PPO-clip algorithm with multi-head attention mechanism. The other baseline algorithms all explore twists and turns, due to the high randomness of initial target guidance information which make it difficult to increase episode reward. The introduction of expert data set and multi-head attention enables happo-rgs algorithm to learn fixed search strategy while extracting target guidance feature information for rapid increases in episode task completion rate and reward. The happo-rgs algorithm episode reward converges at around 200k steps, with significant fluctuation during the first 100k steps due to the limited initial target distribution in the expert data set; after 100k steps, the network iteratively learns based on the hierarchical strategy framework proposed in this paper, to stabilize and converge episode reward. Finally, the episode reward and number of discovered targets of happo-rgs algorithm converge near 30 and 20, respectively, indicating that trained agent can complete the radar guided search task in random scenarios with widespread distribution of cluster targets, which shows superior decision-making capability compared to other baselines algorithms.

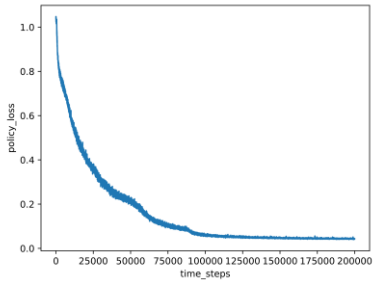
(2) The multi-head attention module introduced to the happo-rgs algorithm enables the policy network to focus on the observation information of undiscovered targets and output corresponding action to improve the task completion rate. The stability of the algorithm is evident in the later training stages with little fluctuation in episode reward and completion rate. The network loss of happo-rgs algorithm during the training process are shown in Fig. 21.

The changes in network loss indicate that the multi-head attention module designed significantly reduces policy network loss, which enables the agent to learn GA expert strategy and accelerates the training process. The reward mechanism is effectively designed, and the value network loss decreases over the training process, with smaller fluctuation in the later stage. In the early stage of training, the policy network incurs larger loss and cannot complete the episode task. At this stage, $reward_5$ from the reward function (2-10) has a higher proportion, resulting in lower episode reward. And the update in the policy network predominantly influences $reward_2$, with minimal variation in episode reward and

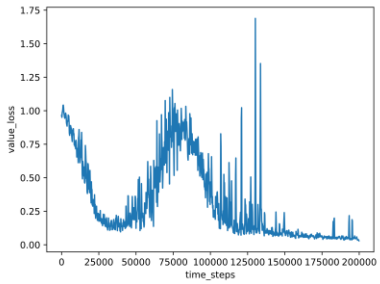
consequent gradual decrease in the value network loss. In the later stage of training, the policy network incurs lesser loss, and certain episodes successfully fulfill the search task, which leads to significant fluctuation in $reward_5$ and the value network loss. However, as the search policy continues to be optimized, the fluctuation in episode reward diminishes, and the value network loss stabilizes again. As the policy network updates, episode reward gradually stabilizes and converge.

The upper-level strategy module and the lower-level strategy module trained by the happo-rgs algorithm generate the search airspace selection strategy and search parameter optimization strategy based on target guidance information and radar search performance respectively. The effect of above hierarchical strategy is verified in scenarios with widespread distribution of cluster targets, where target positions are randomly generated at the beginning of each search task.

Assuming that the radar search resource ratio is 1, take the following scenario as an example. Selected search airspace positions and target weighted expected discovery distance and average accumulated discovery probability of each search airspace are shown in Fig 22.

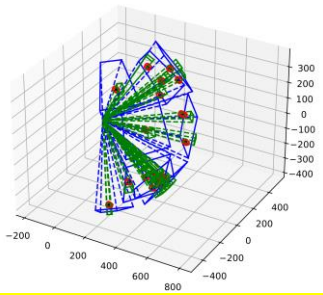


(a) Policy network loss

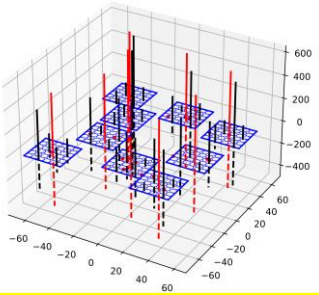


(b) Value network loss

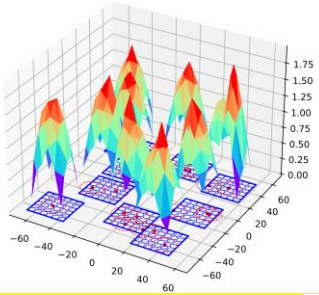
Fig. 21 Network loss of happo-rgs algorithm



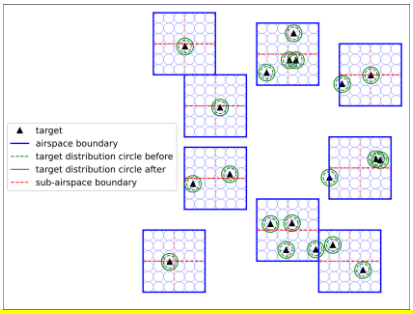
(a) Three-dimensional search airspaces



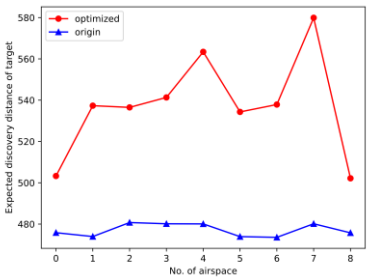
(b) Expected target discovery distance of each sub-airspace



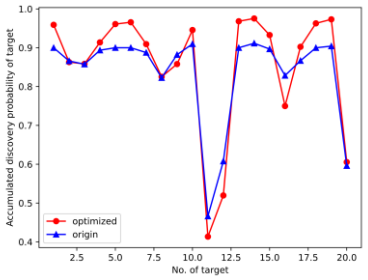
(c) Search data rate for each beam position



(d) Two-dimensional search airspaces



(e) Weighted expected target discovery distance



(f) Accumulated discovery probability of targets

Fig. 22 Example scenario of radar guided search task

Table 4-9 Target discovery status

Search Airspace No.	Target No.	Beam Position No.
1	7,1,17,12	14,16,27,30
2	2,4,11	17,17,19
3	3,13,18,16	4,22,22,25
4	20,5	19,22
5	6	22
6	15,8	17,19
7	9,10	8,23
8	14	22
9	19	21

Fig. 22(a),(d) show the 3D and 2D search airspaces and beam positions respectively.The detailed airspaces and beam positions where each target is discovered are shown in Table 4-9.

In Fig. 22(b), the red lines represent the weighted expected target discovery distance of each search airspace while the black lines represent the expected target discovery distance of each sub-airspace. Solid lines represent the radar search parameter optimization method proposed in this paper while the dashed lines represent the method of directly distributing the search beam dwell time based on the number of sub-airspaces. Fig.22(b),(e) show that sub-airspaces of higher target

distribution probability are allocated more resources for farther target expected discovery distance, which improves weighted target expected discovery distance of entire search airspace. The search airspaces selected by upper-level strategy module with higher probability distribution of targets help lower-level strategy module achieve farther target expected discovery distance to improve radar search performance under fixed search load. Fig. 22(e) shows the precise values of weighted expected target discovery distance by the two methods.

In Fig. 22(c), the warmer the color tone, the higher the search data rate at that beam position, indicating that beam positions with denser target distribution, i.e., higher distribution probability, are searched frequently to achieve higher accumulated discovery probability of cluster targets. Fig. 22(f) shows the precise values of the accumulated discovery probability for all targets in all search airspaces. It can be seen that the happo-rgs algorithm significantly improves average accumulated discovery probability of most targets by slightly sacrificing the search performance of targets with lower distribution probability and reallocating search resources to beam positions where other targets are more likely to appear.

4. Conclusions

In this study, a hierarchical strategy framework based on deep reinforcement learning is designed for the decision of radar guided search task in scenarios with widespread distribution of cluster targets. The proposed mothed has advantages as follows:

(1) Upper airspace selection and lower performance optimization decisions achieve organic integration by exploring multi-level strategies for radar guided search.

(2) The training-based deep reinforcement learning method can meet the high real-time needs of future high dynamic air combat.

(3) The algorithm framework has strong scalability which is easy to extend to situations such as multi-aircraft cooperation and heterogeneous aircrafts.

Shortcomings and future works are as follows:

(1) At present work target maneuver is simplified as constant velocity motion. A real-time accurate position distribution model for random maneuver targets based on guidance information needs to be constructed.

(2) At present the impact of radar detection depth is ignored. A realistic and complete real-time radar beam search 3D environment needs to be constructed where 3D search airspace geometric redundancy needs to be described accurately.

(3) The maneuver characteristics of carrier aircraft and enemy cluster targets in the battlefield need to be considered simultaneously for the intelligent detection problem.

References

[1] Lu D ,Wang X ,Wu X , et al.Adaptive allocation strategy for cooperatively jamming netted radar system based on improved cuckoo search algorithm[J].Defence Technology,2023,24(06):285-297.

[2] SHENG H ,ZHANG J ,YAN Z , et al.New multi-UAV formation keeping method based on improved artificial potential field[J].Chinese Journal of Aeronautics,2023,36(11):249-270.

[3] Xirui X, Shucai H, Daozhi W, et al. Multiradar Joint Tracking of Cluster Targets Based on Graph-LSTMs[J].Journal of Sensors,2022.

[4] Lima Filho G M D, Medeiros F L L, Passaro A. Decision support system for unmanned combat air vehicle in beyond visual range air combat based on artificial neural networks [J]. Journal of Aerospace Technology and Management, 2021, 13: 3721.

[5] Zhen Z, Chen Y, Wen L, et al. An intelligent cooperative mission planning scheme of UAV swarm in uncertain dynamic environment [J]. Aerospace Science and Technology, 2020, 100: 105826.

[6] Sun Z, Piao H, Yang Z, et al. Multi-agent hierarchical policy gradient for Air Combat Tactics emergence via self-play [J]. Engineering Applications of Artificial Intelligence, 2021, 98: 104112.

[7] Yan J, Jiu B, Liu H, et al. Simultaneous multibeam resource allocation scheme for multiple target tracking [J]. IEEE Transactions on Signal Process, 2015, 63(12): 3110-3122.

[8] Zhang H, Xie J, Ge J, et al. A hybrid adaptively genetic algorithm for task scheduling problem in the phased array radar [J]. European Journal of Operational Research, 2019, 272: 868-878.

[9] Galinier P, Hertz A. Solution techniques for the large set covering problem[J]. Discrete Applied Mathematics, 2007, 155(3): 312-326.

[10] Pereira J ,Averbakh I .The Robust Set Covering Problem with interval data[J].Annals of Operations Research,2013,207(1):217-235.

[11] Yelbay S B, Birbil I, Bu¨Lbu¨L K. The set covering problem revisited: an empirical study of the value of dual information[J]. Journal of Industrial & Management Optimization, 2015, 11(2):575-594.

[12] Al-Shihabi S, Arafeh M, Barghash M. An improved hybrid algorithm for the set covering problem[J]. Computers & Industrial Engineering, 2015, 85: 328-334.

[13] Crawford B, Soto R, Cuesta R, et al. Application of the artificial bee colony algorithm for solving the set covering problem[J]. The Scientific World Journal, 2014, 2014:1-8.

[14] M. Zatman, "Radar resource management for UESA," Proceedings of the 2002 IEEE Radar Conference (IEEE Cat. No.02CH37322), Long Beach, CA, USA, 2002, pp. 73-76.

[15] ZHANG Hua-rui, YANG Hong-wen, YU Wenxian. Design of Optimal Search Operation Parameters for Phased Array Radar [J]. ACTA ARMAMENTARII ,2012,33(09): 1062-1065.
[16] WU Qi-hua, LIU Jin, AI Xiao-feng , et al. Method of Search Parameters Optimization of Phased Array Radar for Antimissile Mission[J]. Modern Defence Technology, 2016,44(02):165-170.
[17] Haowei Z ,Weijian L ,Xiao Y .Resource saving based dwell time allocation and detection threshold optimization in an asynchronous distributed phased array radar network[J].Chinese Journal of Aeronautics,2023,36(11):311-327.
[18] Shally G ,Nanhay S .Toward intelligent resource management in dynamic Fog Computing- based Internet of Things environment with Deep Reinforcement Learning: A survey[J].International Journal of Communication Systems,2022,36(4):
[19] Xu W ,Sen W ,Xingxing L , et al.Deep Reinforcement Learning: A Survey.[J].IEEE transactions on neural networks and learning systems,2022,PP
[20] Feng C ,Fu X ,Wang Z , et al.An Optimization Method for Collaborative Radar Antijamming Based on Multi-Agent Reinforcement Learning[J].Remote Sensing,2023,15(11):
[21]Wen J ,Yihui R ,Yanping W .Improving anti-jamming decision-making strategies for cognitive radar via multi-agent deep reinforcement learning[J].Digital Signal Processing,2023,135
[22] Yang Q , Han Z , Wang H ,et al. Radar Waveform Design Based on Multi-Agent Reinforcement Learning[J].International Journal of Pattern Recognition and Artificial Intelligence, 2021.
[23] Zheng, Z.; Li, W.; Zou, K. Airborne Radar Anti-Jamming Waveform Design Based on Deep Reinforcement Learning. Sensors 2022, 22.
[24] Ahmed A M , Ahmad A A , Fortunati S ,et al.A Reinforcement Learning based approach for Multi-target Detection in Massive MIMO radar[J]. 2020.
[25] Haowei Zhang, Junwei Xie, Binfeng Zong. Bi-objective particle swarm optimization algorithm for the search and track tasks in the distributed multiple-input and multiple-output radar[J]. Applied Soft Computing, 2021, 1568-4946.
[26] Payam H, Mahmoud K, Reza F M. The optimal search for multi-function phased array radar[C]. Antennas & Propagation Conference, Loughborough, UK, 2009: 609-612.
[27] Jiejiang C ,Shaowei C ,Yiyuan W , et al. Improved local search for the minimum weight dominating set problem in massive graphs by using a deep optimization mechanism[J].Artificial Intelligence, 2023, 314
[28] V. Cacchiani, V.C. Hemmelmayr, F. Tricoire, A set-covering based heuristic algorithm for the periodic vehicle routing problem, Discrete Applied Mathematics, 2014, 53-64.
[29] CHEN JIANYONG, WANG JIAN, SHAN ZHICHAO. Optimal search algorithm for a random invariant speed target detecting in discrete time[J]. Systems Engineering and Electronics, 2013, 35(8): 1627-1630. (in Chinese)
[30] Zhe Gao, Zhaochen Sun, Shuxiu Liang, Probability density function for wave elevation based on Gaussian mixture models, Ocean Engineering, Volume 213, 2020.
[31] ZHAO Feng, BI Li, ZHOU Ying, et al. Optimal Search Order of Tracking and Guiding Radars in Ballistic Missile Defense Based on Search Data Rate[J]. ACTA ELECTRONICA SINICA, 2009, 37(12).
[32] Xiao L, Xie Y, Gao S, Li J, Wu P. Generalized Radar Range Equation Applied to the Whole Field Region. Sensors (Basel). 2022 Jun 18;22(12):4608.
[33] ZHANG Yong-jie, LI Shao-hong, ZHU Hai-bing. Research on Modeling and Simulation of Optimal Search of Multifunction Phased Array Radar[J]. Journal of System Simulation, 2008, 20(16): 4248-4251.
[34] DENG Gui-fu, LIU Hua-lin, XU Lei. Optimization of Search Parameters of Long Range Phased Array Radar[J]. Radar Science and Technology, 2012, 10(1): 32-36.
[35] Abdelaziz B, Fouad, Mir, et al. An Optimization Model and Tabu Search Heuristic for Scheduling of Tasks on a Radar Sensor[J].IEEE sensors journal,2016,16(17).
[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need[C]. 31st Conference on Neural Information Processing Systems. 2017.
[37] Li D ,Wang D ,Li J .Large Range of a High-Precision, Independent, Sub-Mirror Three-Dimensional Co-Phase Error Sensing and Correction Method via a Mask and Population Algorithm [J]. Sensors, 2024, 24(1):
[38] JIANG Jian-lin, CHENG Kun, WANG Can-can, et al. Improved Genetic Algorithm for Set Covering Problem[J]. MATHEMATICS IN PRACTICE AND THEORY, 2012, Vol.42, No.5.
[39] Zomaya A Y , Yee H .The Observations on Using Genetic Algorithm for Dynamic Load-Balancing[J].IEEE Trans on Parallel and Distributed Systems, 2001,12(9): 899-911.
[40] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[J]. 2017.
[41] Haarnoja T , Zhou A , Hartikainen K ,et al. Soft Actor-Critic Algorithms and Applications[J]. 2018.
[42] Fujimoto S , Van Hoof H , Meger D .Addressing Function Approximation Error in Actor-Critic Methods[J]. 2018.
[43] Lillicrap T. P., Hunt, J. J., et al. Continuous control with deep reinforcement learning. Journal of Machine Learning Research, 17(1), 834-839.