Distributed interference resource optimization algorithm based on cooperative multi-agent reinforcement learning

1st Xiao Yan
State Key Laboratory of CEMEE, National University of Defense Technology
Changsha, China
1325427975@qq.com

2nd Qingping Wang*
State Key Laboratory of CEMEE, National University of Defense Technology
Changsha, China
Andywpq007@163.com

3rd Wenhong Liu
Teaching and Research Support Center, National University of Defense Technology
Changsha, China
Wenhongliu@163.com

4th Yujie Liu
State Key Laboratory of ATR, National University of Defense Technology
Changsha, China
liuyujienudt@nudt.edu.cn

***Abstract*—Due to the communication and sensing capability limitations of the unmanned aerial vehicle (UAV) formation, the performance of distributed jamming resources optimization tends to be unsatisfactory. To address the problem, this paper uses the collaborative multi-agent reinforcement learning method to significantly improve its performance. Meanwhile, the UAV has weak communication capabilities and is easily destroyed, which can lead to network disconnection and communication failure. Considering that most existing multi-agent reinforcement learning methods rely on neural network fitting of agent strategies and lack adaptability to changes in the number of agent nodes, this paper extends and optimizes the agent's policy network by leveraging the permutation equivariance of HPI and the permutation invariance of HPE to adapt to changes in UAV nodes and interference targets. Through simulations, it has been verified that the method proposed significantly improves the performance of distributed optimization and its adaptability to dynamic environments.**

Keywords—distributed optimization; collaborative multi-agent; extended multi-agent; multi-agent reinforcement learning

I. INTRODUCTION

In recent years, the intelligent jamming decision has gradually become a research focus of electronic warfare[1]. Based on the reconnaissance and perception of the battlefield, it forms a comprehensive situation analysis of the battlefield. At the same time, it employs optimization algorithms to form efficient distribution and utilization of the limited jamming resources, achieving effective and stable jamming capabilities, thus improving the perception and countermeasure capability in the complex electromagnetic environment[2], [3]. Aiming at the insufficient prior information on both sides in the actual combat environment, which leads to deficiency in the real-time performance of traditional intelligent optimization algorithms, the reinforcement learning method has gradually been widely applied in the field of intelligent electronic warfare with its characteristic of relying on no prior knowledge[4]. Zhang Bokai et al. proposed a multi-function radar cognitive jamming decision system, and analyzed the supportive application of reinforcement learning and deep reinforcement learning to the system through combining the circumstance of limited prior knowledge under battlefield confrontation[5]; Rao Ning et al. proposed a decentralized jamming resources allocation algorithm based on multi-agent deep reinforcement learning (MADRL) to improve the efficiency of the allocation of jamming resources in the electronic countermeasures[6].

In practice, the jamming system encounters multiple hostile radars, and it needs to quickly establish cognitive relationships with threat targets and dynamically schedule jamming resources. Meanwhile, as the scale of the UAV jamming formation expands, the overall communication resources of the system are in short supply as a result of the communication and sensing capability constraints, which requires optimized decisions to achieve efficient jamming and communication resources allocation. Based on this problem, this paper designs a distributed decision-making method under constrained communication to achieve the overall distributed interference performance of the system.

In addition, UAVs have weak survivability and are prone to network disconnection and communication failure in complex electromagnetic environments. At the same time, when the formation deepens its knowledge of the environment, it may perceive new jamming targets, resulting in a change in the number of decision nodes and mission targets, thereby posing challenges to the application of existing online learning algorithms in multi-agent networks. Multi-agent reinforcement learning fits agent strategies or value functions through neural networks that take environment states or agent observations as inputs and output agent actions or Q-values corresponding to the actions. In an adversarial scenario, the dynamic changes in the number of drone nodes and jamming target radar nodes in the cluster network will simultaneously cause changes in the observation dimension, action dimension, and environment state dimension of the agent. To adapt to these changes, not only should the neural network structure be adjusted, but the agent strategy also needs to be retrained.

However, this series of adjustments significantly increases the computational burden of the algorithm, making it difficult to meet the real-time requirements of practical adversarial applications. Currently, research on expandability mainly focuses on network structures such as Deep Sets, Self-Attention, and GNN. These networks independently extract input features through shared modules and aggregate them in a specific way to satisfy permutation invariance while adapting to input dimension changes. Permutation invariance captures the characteristics that the state and observation information are independent of the order of entities, which improves the

learning efficiency of the algorithm to a certain extent.However, in most real scenarios, changes in the number of entities will change the dimensions of both the neural network inputs (state and observation) and outputs (actions). Therefore, to deal with variability issues such as the variable length input and output problem and permutation isotropy, expandability optimization is carried out in this paper. Specifically, by introducing the HPI layer and the HPE layer, and embedding adaptive variable-length input modules with permutation isotropy and adaptive variable-length output modules with permutation invariance in the QMIX algorithm, the agent strategies can effectively adapt to dynamic changes in the number of entities. The specific contributions are as follows:

1. A multi-factor jamming resources allocation model is constructed based on confrontation scenarios. Combining communication resource constraints, a collaborative multi-agent strategy learning method based on Expandability-QMIX （E-QMIX） is proposed. By decomposing the global hybrid Q-function into local Q-functions for each UAV node, the jamming decision problem in the large interference action space and radar state space is effectively solved, and the efficiency of jamming decision is improved.

2. Two types of neural network modules are designed to adapt to variable-length inputs and variable-length outputs. QMIX is improved to enhance the algorithm's ability to adapt to states, observations, and actions in variable-length dimensions, so that the overall network structure of the algorithm is independent of the number of nodes, and there is no need to change the structure of the network for retraining when the number of nodes changes. At the same time, the permutation isotropy of the inputs and the permutation invariance of the outputs are ensured to improve the efficiency of the algorithm for sample training in fixed-size tasks.

II. COOPERATIVE-JAMMING MODEL

As shown in figure 1, this paper investigates the scenario of flight formation consisting of UAV clusters and manned aircraft to break through ground-based group network radars. The whole breakout period is uniformly divided into multiple equal-length time slots, which is convenient for discrete processing and analysis. The UAV detects the radar detection beam and schedules the UAV in its communication range to cooperatively transmit the interference beam in real time to reduce the effective detection beam of the radar. Therefore, the goal of this paper is to maximize the overall interference benefit of UAV swarm under communication constraints through interference decision optimization.

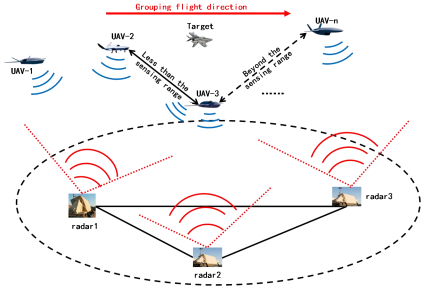

Fig. 1. Diagram of UAV cluster cooperative jamming network radar

In this scenario, it is assumed that all radars in the M unit are in search mode and the detection capability of the radar system is directly limited by the performance of its target detectors. Therefore, the UAV only needs to form effective jamming on the target detector by detecting the signals emitted by the radar and allocating the jamming resources for modulation and forwarding, then it can be considered as an effective jamming on the radar system as a whole. In this paper, the number of radars effectively jamming m^{jst} is chosen as an efficiency index to assess the effectiveness of jamming decision[7]. Firstly, referring to the radar equation, it is known that the target echo power received by the single station radar m is:

$$S_m = \frac{P_t G_t^2 \sigma \lambda^2}{(4\pi)^3 R_m^4 L_r} e^{-0.46\delta R_m} \tag{1}$$

In the formula, R_m is the distance from the radar m to the target n. From the reconnaissance equation, the jamming power of the UAV n jamming signal received by the radar m is[8]:

$$J_{mn} = K(\theta_{mn}) \frac{P_{jn} G_{jn} G_r \lambda^2}{(4\pi)^2 R_{mn}^2 L_j} e^{-0.23\delta R_{mn}} \tag{2}$$

In the formula, θ_{mn} is the tensor angle between the protected target and the UAV relative to the radar m, $K(\theta_{mn})$ is the jamming direction mismatch loss. Suppose that the effects of noise and clutter can be ignored. Combining equation (1) and (2), the total average interference-signal-ratio jointly generated by blanket jammers to the radar m is:

$$j_{sm}(\theta_m) = \frac{4\pi R_m^4 L_r}{P_t G_t \overline{\sigma}} e^{0.46\delta R_m} \sum_{n=1}^{N} [K(\theta_{mn}) \frac{P_{jn} G_{jn}}{L_j R_{mn}^2} e^{-0.23\delta R_{mn}}] \tag{3}$$

Assume that the interference-signal-ratio to meet effective interference requirement, namely the blanket coefficient, is K_j and that the jamming effectiveness is the probability that the total average interference-signal-ratio is greater than or equal to K_j. Thus, the probability of jamming effectiveness of blanket jamming on radar is:

$$E_{jsm} = 1 - \exp(-\frac{j_{sm}}{K_j}) \tag{4}$$

Combined with the radar network structure, the probability of effectively jamming all radars can be obtained as

$$P_{jam} = \prod_{m=1}^{M} E_{jsm} \tag{5}$$

where M is the total number of radars to be jammed. The number of effectively jammed radars can be obtained as

$$m_j = \sum_{m=1}^{M} E_{jsm} \tag{6}$$

In the UAVs and multi-radar confrontation mode, the UAV may belong to the one-to-one mode as well as the many-to-one

or many-to-many mode under a certain time slot. Therefore, the average probability of effectively jamming any radar is selected as a quantitative index to evaluate the jamming effectiveness and is expressed as:

$$P_j = \frac{m_j}{M} \tag{7}$$

Referring to Equation (5)-(7), the assessment function of jamming effectiveness designed in this paper requires the probability that the number of effectively jammed radars is larger than or equal to the required number n_{jst} and is expresses as:

$$E_{jam} = \sum_{k=n_{jst}}^{N} C_n^k P_j^k (1-P_j)^{N-k} \tag{8}$$

III. Distributed Optimization Algorithm Based on Cooperative Multi-Agent Reinforcement Learning

A. Distributed locally observable markov process modeling

The agents established in this paper are fully collaborative[9] and thus can be represented as $(S,U,\{O^n\}_{n\in N},A,R,P)$: for each time step t , the environment information is characterized by the state $s_t \in S$ and each agent $n \in N$ gets an observation o_t^n . Subsequently, according to the policy $\dot{\pi}^n(a_t^n \mid o_t^n)$: $O^n \times A^n \to [0,1]$, the agents select to get the action $a_t^n \in A^n$. All the agent actions form a joint action which denotes as $a_t = [a_t^1, a_t^2, ..., a_t^N]$. The environment transition probability function $P(s_{t+1} \mid s_t, a_t)$ transfers to the state s_{t+1} at the next moment. At the same time, the multi-agent in a fully cooperative state will receive a common reward r_t . The goal of multi-agent reinforcement learning is to obtain the joint optimal policy of all the agents, maximizing the expected long-term cumulative reward $\mathbb{E}_t R(\dot{\boldsymbol{\pi}}) = \sum_{\tau=t}^{\infty} \gamma^t r_t$. Where $\dot{\boldsymbol{\pi}} = \{\dot{\pi}^n\}_{n\in N}$ represents the set of strategies of all the agents and represents the reward discount factor.

B. Collaborative multi-agent decision-making based on E-QMIX algorithm

In the actual confrontation, the number of UAV nodes and the number of interference targets in the cluster are easy to change, which puts forward higher requirements for the scalability of the distributed optimization algorithm. By introducing an automatic weight generation mechanism on the basis of the FC layer, the HPI layer and the HPE layer realize the dynamic adjustment of the dimension of the weight matrix to adapt to variable-length input and variable-length output.

Among them, the HPI layer satisfies the permutation invariance[10]. Its specific implementation is to introduce a supernetwork $\mathrm{HN}_{in}(\cdot)$ composed of a single FC layer to generate the weight value $\mathbf{W}_{in}^j = \mathrm{HN}_{in}(\mathbf{x}^j)$ of each input element, and then obtain the output of the entire network :

$$\mathbf{y} = \sum_{j=1}^{J} \mathbf{W}_{in}^j \cdot \mathbf{x}^j = \sum_{j=1}^{J} \mathrm{HN}_{in}(\mathbf{x}^j) \cdot \mathbf{x}^j \tag{9}$$

The HPE layer satisfies the replacement equivariant[11].A similar approach to the HPI layer is used to introduce a supernetwork $\mathrm{HN}_{out}(\cdot)$ to generate a weight value $\mathbf{W}_{out}^j = \mathrm{HN}_{out}(\mathbf{x}^j)$ that matches the number of output elements, so each output element $\mathbf{y}^j = \mathrm{HN}_{out}(\mathbf{x}^j)\mathrm{HPI}(\mathbf{x}^j)$ is obtained.

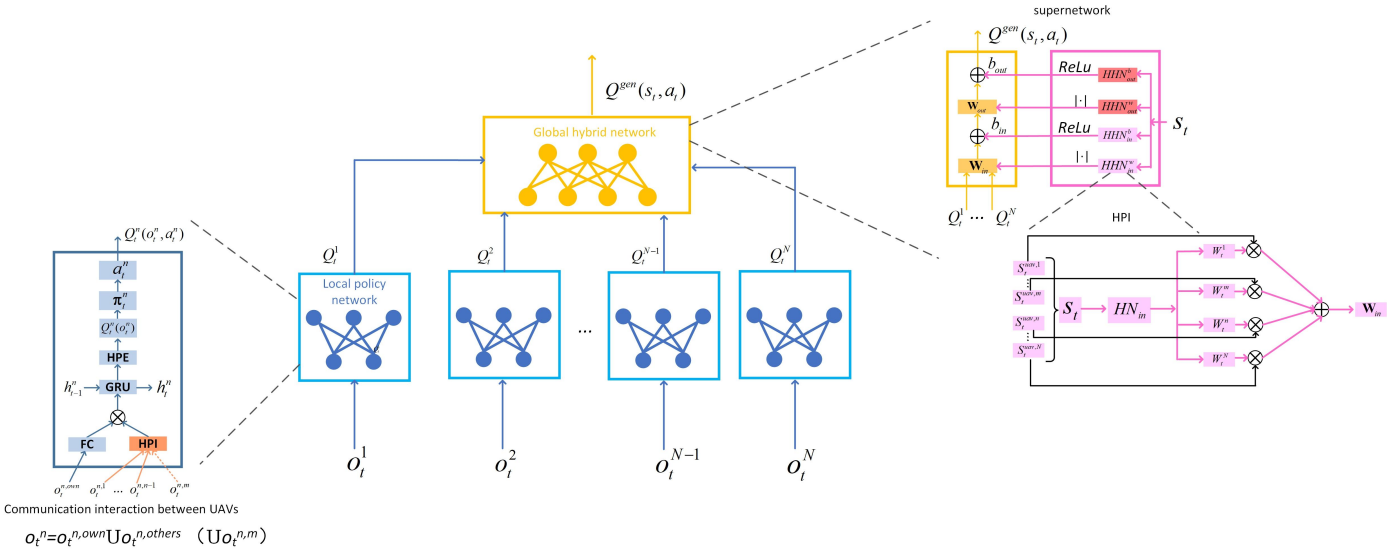

Fig. 2. E-QMIX algorithm framework diagram

Based on the QMIX algorithm[12], this paper introduces HPI and HPE modules for extensive improvement, and obtains the overall framework of the E-QMIX algorithm as shown in Figure 2. Its core idea is to transform the training of the global Q function into the local Q function training of each UAV node agent by means of value function decomposition.

Identify applicable funding agency here. If none, delete this text box.

In multi-agent reinforcement learning, the global optimal Q function defines that when all agents follow the global optimal strategy $\dot{\pi}^*$ under state s_t, the expected long-term cumulative reward obtained by taking joint action a_t is :

$$Q^{gen}(s_t, a_t) = \mathbb{E}_{\dot{\pi}^*}[\bar{R}(\dot{\pi}^*) \mid s_t, a_t] = \mathbb{E}_{\dot{\pi}^*}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_t, a_t] \tag{10}$$

Therefore, the optimal joint action is expressed as the action that maximizes $Q^{gen}(s_t, a_t)$. Given the global optimal Q function, the optimal global strategy $\dot{\pi}^*$ can be expressed as:

$$a_t^* = \arg\max_{a_t} Q^{gen}(s_t, a_t) \tag{11}$$

The global optimal Q function needs to follow the Bellman optimality principle, so it satisfies the expression :

$$Q^{gen}(s_t, a_t) = \mathbb{E}_{s_{t+1} \in P}[r_t + \gamma \max_{a_{t+1}} Q^{gen}(s_{t+1}, a_{t+1})] \tag{12}$$

The local optimal Q function defines the local long-term expected cumulative reward value obtained by the agent n according to the local optimal strategy $\dot{\pi}^{n*}$ after taking action a_t^n under local observation o_t^n , which is expressed as follows :

$$Q^n(o_t^n, a_t^n) = \mathbb{E}_{\dot{\pi}^*}[\sum_{t=0}^{\infty} \gamma^t r_t^n \mid o_t^n, a_t^n] \tag{13}$$

where r_t^n represents the local reward value obtained by the action of the agent n under the time slot t . Referring to the definition of the global optimal strategy in formula (11), it is not difficult to obtain the expression of the local optimal strategy :

$$a_t^{n*} = \arg\max_{a_t^n} Q^n(o_t^n, a_t^n) \tag{14}$$

The local optimal Q function also follows the optimal Bellman equation, which is expressed as :

$$Q^n(o_t^n, a_t^n) = \mathbb{E}_{o_t^n \in P}[r_t^n + \gamma \max_{a_{t+1}^n} Q^n(o_{t+1}^n, a_{t+1}^n)] \tag{15}$$

In the scenario discussed in this paper, all UAV nodes have the same optimization objectives, so they share global rewards. On the one hand, according to the global environment state, the joint action of the agent and the global reward, the global Q function can be learned by centralized training. However, the global Q function is difficult to be deployed to each agent in a distributed manner. On the other hand, the local Q function can be used to generate the local action of the agent, but it is difficult to learn the local Q function directly because the local reward of each agent cannot be obtained explicitly.

In order to solve the above problems, a global Q function can be trained and decomposed into a local Q function of each UAV by means of value function decomposition. The specific implementation method is to use a decomposition function f^{decom} to decompose the global Q value into multiple local Q values. At the same time, in order to ensure the consistency between the global optimal strategy and the local strategy of a single agent, the optimal joint action taken by all agents according to the global optimal Q function should be equivalent to the local optimal action taken by each UAV node according to the local Q function. Expressed as :

$$\arg\max_{a_t} Q^{gen}(s_t, a_t) = \begin{pmatrix} \arg\max a_t^1 Q^1(o_t^1, a_t^1) \\ \vdots \\ \arg\max a_t^n Q^n(o_t^n, a_t^n) \end{pmatrix} \tag{16}$$

The formula (16) shows that there is a monotonic relationship between the global Q value and the local Q value, that is, when the local Q value rises, the global Q value will also rise. Therefore, the decomposition function f^{decom} needs to satisfy the constraints :

$$\frac{\partial f}{\partial Q^n} \geq 0, \forall n \in \mathcal{N} \tag{17}$$

C. Network parameter training update

In this paper, centralized training with decentralized execution architecture (CTDE) is adopted. The network parameters are trained by end-to-end method. Based on the Bellman equation of the global Q function, the loss function is obtained as follows :

$$L_{Q^{gen}}(\theta) = \mathbb{E}_{\mathcal{D}}[(y_t^{gen} - Q_\theta^{gen}(s_t, a_t))^2] \tag{18}$$

$$y_t^{gen} = r_t + \gamma \max_{a_{t+1}} Q_{\theta^-}^{gen}(s_{t+1}, a_{t+1}) \tag{19}$$

where $\mathcal{D}$ is the experience playback pool, y_t^{gen} is the Q value of the target network, and θ^- is the target network parameter of θ , which is used to calculate the loss function and the soft update of θ .

IV. Simulation and Analysis

In the simulation experiment, the confrontation scene between the UAV and the ground-based radar is set to a square area of 200 km x 200 km, and a 200 x 200 grid structure is constructed by a scaling ratio of 1:1000. The three radar positions are set to (200,200) (185,200) (200,185). In the initial state, the five UAVs approached the radar and interfered around the radar on the premise of ensuring the minimum safe distance between them. The simulation parameters are set as shown in Table 1.

In the hyperparameter setting of the E-QMIX algorithm, the embedding feature of the local Q network and the dimension of the hidden layer state are set to 128. In the training process, the maximum number of iterations is set to 3000, that is, up to 3000 time steps. The search probability ε of the agent gradually decreases from 0.9 to 0.05 with the increase of the number of iterations. The discount factor γ of the future reward is 0.9, and the network parameters are updated by soft update. The learning rate is set to 0.0025. The maximum storage of the experience playback pool is 2000 training rounds of experience samples, and 16 batches of samples will be randomly taken out for learning during each update. Set the system as a whole to train a total of 5000 rounds.

TABLE I. SIMULATION PARAMETER SETTING

Parameter	Numerical value
length of slot τ	1s
UAV flight altitude H	5km
The flight speed of UAV V	250m/s
Minimum distance between drones R_j	2km
Minimum distance between UAV and radar R_{tj}	20km
The threshold of communication perception γ_s	25dB
Radar transmit power P_t	150kw
Jammer maximum transmit power $P_{j\max}$	500w

Comparing the performance of distributed decision algorithm and centralized decision algorithm, this paper chooses CQL algorithm[13] and E-QMIX algorithm to compare. By comparing the communication overhead and decision effect of the two algorithms, the performance of the proposed algorithm in balancing communication overhead and decision effect is verified. Figure 4 and Figure 5 are the changes of radar network detection probability and communication rate during the training process of E-QMIX and CQL, respectively.

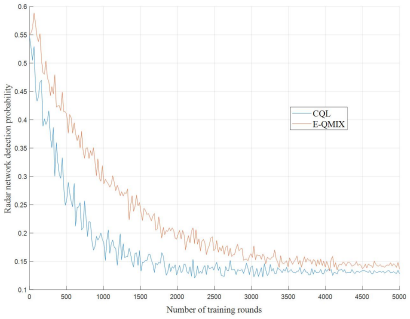

Fig. 3. The change curve of radar network detection probability during training process

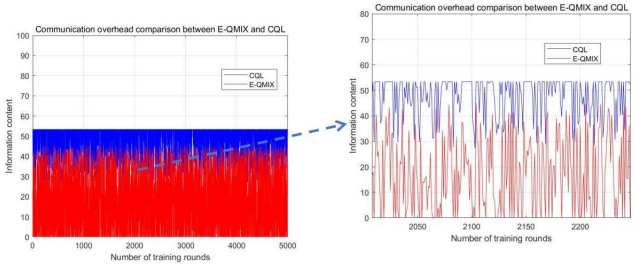

Fig. 4. Communication overhead comparison diagram(kbps)

By comparison, it can be seen that E-QMIX obtains a smaller state and action space through distributed suboptimal solution convergence decision, so its strategy is relatively conservative and less volatile in the optimization process. Finally, E-QMIX obtains a decision effect similar to CQL while reducing the demand for communication information rate.

V. CONCLUSION

In this paper, a distributed optimization algorithm based on cooperative multi-agent reinforcement learning is designed for the problem of weak communication ability of UAVs and the change of the number of individuals in the cluster, and its scalability is optimized to solve the optimization problem of maximizing the interference effect under the condition of constrained communication. Through simulation comparison, the proposed method reduces the demand for a large amount of communication information and achieves a similar decision-making effect compared with the CQL algorithm. At the same time, the scalability of the algorithm has been significantly improved, and the policy network can better adapt to the dimension changes of input and output. However, the proposed method still requires sufficient computational support to achieve online learning and requires knowledge-assisted pre-training. The next step still needs to explore the distributed decision-making method of real-time learning under the condition of no prior knowledge or small sample.

REFERENCES

[1] Z. Yin, J. Li and Z. Wang, "UAV Communication Against Intelligent Jamming: A Stackelberg Game Approach With Federated Reinforcement Learning," IEEE Trans. Green Communications and Networking, vol.8, no.1, pp. 1-8, July 2024.

[2] Y. Lin, M. Wang and X. Zhou, "Dynamic Spectrum Interaction of UAV Flight Formation Communication With Priority: A Deep Reinforcement Learning Approach," IEEE Trans. Cognitive Communications and Networking, vol.6, no.3, pp. 892-903, May 2020.

[3] H. Albinsaid, K. Singh and S. Biswas, "Multi-Agent Reinforcement Learning-Based Distributed Dynamic Spectrum Access," IEEE Trans. Cognitive Communications and Networking, vol.8, no.2, pp. 1174-1185, 2022.

[4] J. Xu, H. Lou and W. Zhang, "An Intelligent Anti-Jamming Scheme for Cognitive Radio Based on Deep Reinforcement Learning," IEEE Access, vol.8, no.1, pp. 202563-202572, 2020.

[5] B. Zhang and W. Zhu, "Research on Decision-making System of Cognitive Jamming against Multifunctional Radar," in IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC 2019), Dalian, China, March 2019.

[6] N. Rao, et al., "Joint Optimization of Jamming Link and Power Control in Communication Countermeasures: A Multiagent Deep Reinforcement Learning Approach," Wireless Communications and Mobile Computing, Vol.50, no.6, pp. 1319-1330, 2022.

[7] B. Cui, "Evaluation of radar countermeasure jamming effectiveness. Beijing," Publishing House of Electronics Industry, 2017.

[8] Q. Wang, X. Yan and F. Liu, "An Optimal Allocation Method for Cooperative Jamming Resources Based on Multi-agent Genetic Algorithm," in 2022 2nd International Conference on Frontiers of Electronics, Information and Computation Technologies (ICFEICT 2022)[C], Wuhan, China, Augest 2022.

[9] D. Maisto, F. Gregoretti and K. Friston, "Active inference tree search in large POMDPs," Neurocomputing, Vol.623, no.1, pp.0925-2312, 2025.

[10] G. Hou, T. Huang and F. Zheng, "A hierarchical reinforcement learning GPC for flexible operation of ultra-supercritical unit considering economy," Energy, Vol.289, no.5, pp. 129936, 2024.

[11] A. Karatzetzou, S. Apostolaki and E. Riga, "Hierarchical policy for seismic intervention of school buildings at urban scale,"Structures, Vol.48, pp. 669-680, 2023.

[12] W. Chen, S. Huang and J. Schneider, "Soft-QMIX: Integrating Maximum Entropy For Monotonic Value Function Factorization," 2024, arXiv:2406.13930v2. [Online]. Available: https://arxiv.org/pdf/2406.13930

[13] Y. Niu, B. Wan and C. Chen, "A Centralized Multi-User Anti-Composite Intelligent Interference Algorithm Based on Improved Q-Learning," Electronics, Vol.12, no.18, pp. 1803, 2023.