remote sensing

MDPI

Article

An Optimization Method for Collaborative Radar Antijamming Based on Multi-Agent Reinforcement Learning

Cheng Feng [1], Xiongjun Fu [1,2,*], Ziyi Wang [1], Jian Dong [1], Zhichun Zhao [1] and Teng Pan [3]

[1] Beijing Institute of Technology, Beijing 100081, China
[2] Tangshan Research Institute of BIT, Tangshan 063007, China
[3] Beijing Institute of Space Systems Engineering, Beijing 100094, China
* Correspondence: fuxiongjun@bit.edu.cn

Abstract: Attacking a naval vessel with multiple missiles is an important way to improve the hit rate of missiles. Missile-borne radars need to complete detection and antijamming tasks to guide missiles, but communication between these radars is often difficult. In this paper, an optimization method based on multi-agent reinforcement learning is proposed for the collaborative detection and antijamming tasks of multiple radars against one naval vessel. We consider the collaborative radars as one player to make their confrontation with the naval vessel a two-person zero-sum game. With temporal constraints of the radar's and jammer's recognition and preparation interval, the game focuses on taking a favorable position at the end of the confrontation. It is assumed the total jamming capability of a shipborne jammer is constant and limited, and the shipborne jammer allocates the jamming capability in the radar's direction according to the radar threat assessment result and its probability of successful detection. The radars work collaboratively through prior centralized training and obtain a good performance by decentralized execution. The proposed method can make radars collaborate to detect the naval vessel, rather than only considering the detection result of each radar itself. Experimental results show that the proposed method in this paper is effective, improving the winning probability to 10% and 25% in the two-radar and four-radar scenarios, respectively.

Keywords: radar antijamming; multi-agent reinforcement learning; game theory

Citation: Feng, C.; Fu, X.; Wang, Z.; Dong, J.; Zhao, Z.; Pan, T. An Optimization Method for Collaborative Radar Antijamming Based on Multi-Agent Reinforcement Learning. *Remote Sens.* **2023**, *15*, 2893. https://doi.org/10.3390/rs15112893

Academic Editors: Eugin Hyun and Inoh Choi

Received: 5 April 2023
Revised: 18 May 2023
Accepted: 23 May 2023
Published: 1 June 2023

1. Introduction

Missile-borne radar is an important sensor for leading the missile to attack an enemy. When the radar detects non-cooperative targets, it often confronts various jamming from the defense system of the target. For example, a missile-borne radar will face a complex three-dimensional shipborne jamming system when detecting naval vessels [1]. Due to the platform, the power and volume of the shipborne jammer are almost unlimited, resulting in a solid jamming capability. However, the power and volume of the missile-borne radar are limited, resulting in relatively weak antijamming capability. Therefore, it is very difficult for one single radar to complete the detection and antijamming task when detecting a naval vessel, and it is also difficult to ensure the reliability and accuracy of the detection results. To improve the success probability of the detection task, multiple collaborative radars can be used to solve the problem [2]. Multiple radars can collaborate to form a detection network, effectively improving the antijamming capability [3], promoting the recognition ability of the target [4], and achieving the integration of target detection, reconnaissance, guidance, and evaluation [5].

These radars can collaborate in distributed and centralized ways, which both need continuous communication during the flight. In the distributed way, each radar carries out preliminary echo processing and obtains initial features before conducting information fusion. Targets may be located first and then validated by cross-calculation and checking [6–8]. In the centralized way, each radar submits the original echo directly to the fusion terminal

for processing, which provides the target location in the end [9–11]. The differences mainly lie at the signal processing level. Centralized fusion processing has a large workload, which places high requirements on communication between radars. Distributed collaboration may result in the loss of some information before fusion, adversely affecting its effectiveness. These collaboration methods optimize detection performance but when multiple radars are transmitting and receiving signals simultaneously, there can be problems with space–time alignment, clutter suppression, coherent processing, and other aspects, which make it difficult to apply in practice [12,13]. When the communication is not stable, the radars will take action by only considering their benefits, which will result in the deterioration of detection performance.

Radar functions as a typical agent that detects, senses, processes information, and responds, which makes the theory of multi-agent a possible solution to enhance radar capabilities in detection and antijamming. Many studies focus on multi-agent cooperation involving multi-weapon and multi-UAV, which can be classified into two aspects. The first aspect is based on spatial analysis, path planning, and so on. It aims to improve radar accuracy by addressing the complex geometric relationship between each agent and the detection target, eliminating false targets, and then solving the problem of target tracking. In [14], a diffusion Kalman filtering algorithm based on the covariance intersection method was proposed under the condition of multi-sensors, resulting in a stable estimate for each agent regardless of whether the system is uniformly observable locally by the measurements of its neighbors. In a study of [15] the decentralized detection problem with N sensors and a central processor, it was concluded that using decentralized detection, as opposed to centralized detection with CFAR processing, may lead to better performance in a homogeneous background. A general scheme of SRC-based centralized target detection in multistatic radar was proposed by [16], which examined the symbiosis relationship of neighboring SRCs in exceeding the test threshold during centralized target detection. These research topics often involve filtering in signal processing environments and predicting the future state of the target using current state information, which requires a relatively stable target state.

The second is to optimize the detection performance of collaborative radar by allocating resources more reasonably and taking target detection probability (Pd) and the echo signal-to-noise ratio (SNR) as optimization goals. Ref. [17] investigated a game theoretic power allocation scheme based on the estimate of the signal-to-disturbance ratio (SDR) and performed a Nash equilibrium analysis for a multistatic MIMO radar network. Ref. [18] proposed a game theoretic waveform allocation algorithm of a MIMO radar network, which used potential games to optimize the performance of radars in the clusters. Ref. [19] proposed a hybrid Bayesian filter that operated by partitioning the state space into smaller subspaces and thereby reducing the complexity involved with high-dimensional state space, and jointly estimated the target state by comprising the positions and velocities of multiple targets. However, little consideration is given to the jamming behavior of non-cooperative targets, and the optimization condition assumes that the non-cooperative target will maintain its current state without altering its behavior, which is inconsistent with actual scenarios. Additionally, each collaborative sensor must communicate information or partial communication during the detection process, sharing observed data with the decision-making terminal, which then optimizes the information before commanding each agent. This is difficult to guarantee in practical confrontational environments.

It is a great challenge for radars to collaborate without communication. Multi-agent reinforcement learning (MARL) is an effective tool to complete the collaborative detection and antijamming task of multiple radars. MARL is a method applicable to multi-agent-based reinforcement learning (RL) [20,21]. Ref. [22] proposed a MARL-based method in target tracking, which analyzed static collaborative detection by a UAV swarm. However, it did not consider the antijamming process. RL is an important method of machine learning, which takes environmental feedback as the input target and uses a trial-and-error method to find the optimal behavior strategy [23]. RL has been widely used in handicraft manu-

facturing, robot control, optimization and scheduling, and other fields [24]. Collaborative detection among multiple radars is a process of interacting with the environment, making the evaluation, exchanging information, and making feedback. MARL can be used to optimize collaborative detection. Compared with RL, one agent's action will also become the state of other agents after it takes actions according to the strategy in MARL, which will lead to complex state space and convergence difficulty [25]. One of the important research contents of MARL is how to optimize the behavior of each agent without explicit communication among agents. By applying deep learning theory to MARL and using a deep neural network to simulate its value function or strategy function, the problem that the state space of RL is too large to calculate can be solved and the optimization result is easier to convergent.

MARL can be implemented in many ways. Ref. [26] is a method based on value function, which proposed the value decomposition networks (VDNs) method to centrally train a joint value network (Q network), which is obtained by the sum of the local Q networks of all agents. It can deal with the problems caused by the non-stationary environment through centralized training, decoupling the complex interrelationships between agents and realizing the decentralized execution of all agents. It follows the centralized training decentralized execution (CTDE) framework. Ref. [27] is an actor-critic method, which proposed multi-agent deep deterministic policy gradient (MADDPG) algorithm to extend the deep deterministic policy gradient (DDPG) to a multi-agent environment. The MADDPG algorithm assumes that each agent has its own independent critic network and actor network, and that each agent has its own independent utility function, which can solve the multi-agent problem of a cooperative environment, competitive environment, and mixed environment. Ref. [28] is a method based on experience replay (ER), which used ER to increase the stability of training a Q-function and break the correlation between data. It is based on Q-learning and also follows the framework of CTDE. The confrontation between collaborative radars and the jammer is a dynamic game process. Ref. [29] proposed a game confrontation model based on the non-real-time characteristic of radar and jammer behaviors, which can be used to make decisions during the confrontation. The collaborative detection and antijamming process of multiple radars for a naval vessel can be regarded as the process of multi-agent and target interaction, which can be optimized with MARL. By assuming the reasonable response of a naval vessel to the detection of multiple radars and setting the reward of radars for the detection results at different rounds, the CTDE framework MARL algorithm can be used to improve the tacit cooperation ability of various radars during training. When working in the actual scenario, the corresponding actions can be automatically executed by using the experience of training, and the behaviors can be directly optimized by skipping the information communication process to improve the collaborative detection and antijamming capability.

To solve the above problems, this paper proposes an antijamming strategy optimization method based on MARL to depict and model the dynamic confrontation between missile-borne radars and shipborne jammers. The proposed method improves the collaborative detection and antijamming capability of multiple radars without communications. The main contributions of this paper are as follows:

- Constructing a game model by considering multiple radars as one player rather than many players makes the confrontation between collaborative missile-borne radars and shipborne jammers be regarded as a two-person zero-sum game. The synthetic result of radars' detection of the naval vessel can be calculated by the jamming effect of shipborne jammers against each radar, which is easier to obtain. By using the game model with temporal constraints, the condition that the radars win the game is that at least one radar detects the naval vessel in tracking mode when the confrontation ends.
- In the game model, it is assumed the total jamming capability of the shipborne jammer is constant and limited and is allocated in each radar's direction by solving the optimization problem with the restraints of the radar threat assessment result and its probability of successful detection. The radars work collaboratively through prior cen-

tralized training and obtain a good performance by decentralized execution without communication.

- The simulation experiments contain the comparison with the DDPG algorithm, by which each radar only considers the result of itself, which proves that the success probability of radars' detection and antijamming optimized by the proposed MARL method is significantly improved. Our method achieves radar collaboration without communication in a confrontation scenario.

The remainder of this paper is organized as follows. Section 2 presents a model of the collaborative detection process of multiple radars, including the game confrontation model of multiple collaborative radars and a shipborne jammer, the settings for radar detection and antijamming capabilities, and the jamming capability settings for the shipborne jammer, as well as the calculation method of successful detection probability. In Section 3, the decision-making process of the shipborne jammer under the MARL model is considered, and appropriate rewards are set for the behavior of collaborative radars. Section 4 carries out simulation experiments and compares the results with the effect of radar without collaborative training. The results demonstrate that the success probability of radar collaborative detection and antijamming is greatly improved, verifying the effectiveness of the algorithm. The experimental results are discussed in Section 5, and the conclusions of this study are presented in Section 6.

2. Materials

The behavior of a single radar adopts the model of [28], in which the temporal constraints of both the radars and the jammer are considered as they consume time to recognize the other player's behavior and the preparation for their behavior. A round of the game is divided into four parts, including the recognition of jamming on the radar, preparation of antijamming actions, the recognition of radar actions by the jammer, and the preparation of jamming.

In the past, modeling the behavior of jammers was often relatively simple. The allocation method of jamming capability in the radars' direction is not seriously discussed, and the influence of the radar's operating mode on the allocation of jamming capability is not considered. We consider the decision-making of the shipborne jammer as an optimization problem, with the restraints of the radar threat assessment result and its probability of successful detection.

For a single radar, it is hoped to increase the proportion of radar dominance interval in a confrontation round as much as possible to win the game. For all collaborative radars, it is expected that at least one radar can complete the detection and antijamming task at the end of the game, without specific requirements for a single radar.

2.1. Game Model between Multiple Radars and a Single Jammer

When launching multiple missiles for coordinated attacks on the naval vessel at the same time, the radars often carry the same type and have the same capabilities. It is often difficult to communicate between these radars when collaborating to detect targets. To ensure that the research is consistent with the real scenario, it is assumed that there is no communication between the radars. Therefore, it can be considered that the collaborative detection and antijamming process of multiple radars is a composite of game confrontation between a single radar and a shipborne jammer.

The elements of game theory generally include players, strategy sets, and utility functions. Players are the participants in the game, the strategy sets are the sets of actions that can be taken by each player, and the utility function corresponds to the benefits of the actions taken by the radar and naval vessel, respectively [30].

In this game, multiple radars are considered as a single player, while shipborne jammers are considered as the other player. Real-time changes in detection and antijamming for each radar pose difficulty in calculating the synthetic result of their detection on the naval vessel. As the confrontation between collaborative missile-borne radars and the

shipborne jammer is a two-person zero-sum game, the synthetic result of radar detection on the naval vessel can be calculated based on the jamming effect of shipborne jammers on each radar, which is easier to obtain. The confrontation between missile-borne radars and the shipborne jammers is shown in Figure 1.

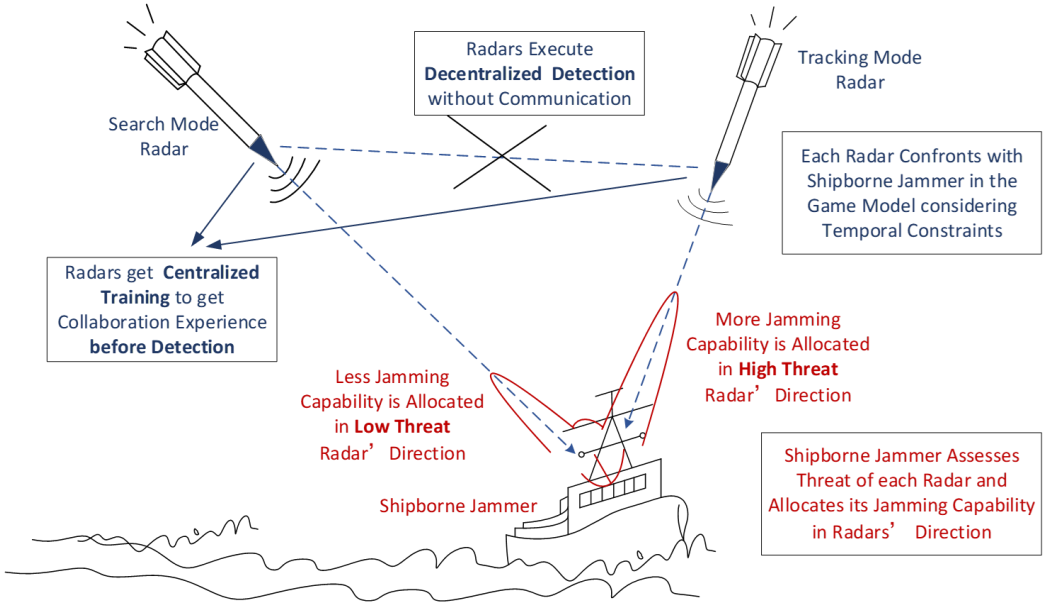


Figure 1. A schematic of the detection of a naval vessel by collaborative radars.

2.2. Radar Detection and Antijamming Capability Setting

In the confrontation between radars and shipborne jammers, the strategy set and utility function are unknown generally, but the specific revenue value can be estimated and evaluated by prior information.

The strategy set and utility functions of one player are generally unknown to the other. However, in the confrontation between radar and naval vessels, they can be estimated by prior knowledge.

The utility matrix is the matrix of benefits that the players can obtain when they take different actions. It is assumed that the radar and the naval vessel both know the utility matrix of the opponent. The actions of the naval vessel include barrage noise (BN), responsive spot noise (RSN), doppler noise (DN), range false targets (RFT), and velocity gate pull-off (VGPO) or other jamming attacks. The actions of the radar are simply regarded as anti-BN, anti-RSN, etc. The numbers of actions of radar and naval vessels are m and n. The utility matrix of the radar is expressed by:

$$E = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & & e_{2n} \\ \vdots & & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{pmatrix} \tag{1}$$

where the rows represent the radar's actions and the columns represent the naval vessel's actions, and e_{ij} is the benefit of the i_{th} action of the radar to the j_{th} action of the naval vessel. It represents the success probability of antijamming. A large value e_{ij} means radar antijamming actions have a higher probability of effectiveness. The game between a single missile-borne radar and a shipborne jammer adopts the model in [28].

2.3. Jamming Capability Setting of a Shipborne Jammer

When missiles attack a naval vessel, they often launch attacks in as many directions as possible simultaneously, and missile-borne radars detect the naval vessel in corresponding

directions. Therefore, the naval vessel also needs to allocate jamming capabilities accordingly. It is assumed that the total jamming capability of the jammer is fixed, and it can be allocated in accordance with the demand. This means the jamming capability integration between 0° and 360° is constant. The integration of the allocated jamming capability in each radar direction is:

$$J = \int_0^{2\pi} L(\theta) d\theta \tag{2}$$

where J is the total jamming capability of the shipborne jammer and $L(\theta)$ is the distribution function of jamming capability.

To describe the jamming effect of shipborne jammers, the following definitions need to be made.

Threshold 1: If the jamming capability allocated by the shipborne jammer at a specific direction exceeds threshold 1 (Th_1), the missile-borne radar in that direction cannot perform effective detection. In the nearby area $L_0(\alpha)$, the jamming capability is depicted as:

$$L_0(\alpha) = A_1 \mathrm{sinc}(\frac{\alpha}{A_2}) \tag{3}$$

where α is the angle of deviation from the direction of the missile-borne radar and the shipborne jammer, A_1 is the peak value of the jamming capability allocated in the direction, and A_2 is the adjustment coefficient for the peak decline speed. Here, it is assumed that the jamming capability at the connection between the missile-borne radar and the shipborne jammer is at a peak, and the other angles away from the peak are reduced according to the *sinc* function, and only the first zero point is considered.

If the radar's perceived jamming capability is lower than Th_1, the probability of jamming against the current direction radar decreases exponentially, which is expressed as:

$$P_k = \begin{cases} 1, & when\ A_1 \geq Th_1 \\ e^{A_3(A_1 - Th_1)}, & when\ A_1 < Th_1 \end{cases} \tag{4}$$

where A_3 is the adjustment coefficient for the descending speed of the jamming success probability.

Threshold 2: If the radar's perceived jamming capability exceeds Th_1 and the duration in this condition exceeds threshold 2 (Th_2), the radar will lose its detection and tracking ability for a while and change its state from tracking mode to searching mode. Th_2 is a limiting condition for optimizing radar antijamming strategies.

Threshold 3: Threshold 3 (Th_3) is the period in which the radar cannot track and detect after switching to search mode. During the duration of Th_3, the shipborne jammer can allocate the jamming capability from one radar direction to another.

If the antijamming performance of radar collaborative detection is insufficient, shipborne jammers can use a strategy of sequential jamming. They can concentrate resources to first jam the radar with the greatest threat, followed by the radar with the next greatest threat. However, this could limit the effectiveness of multi-radar collaborative detection. To improve detection ability, radars can intentionally decrease their threat level so as not to become the primary target of shipborne jammers. Alternatively, considering the constant total jamming capability of the radar, one radar can attract the main attention of shipborne jammers, while allowing for detection in other directions.

The success of antijamming toward one missile-borne radar is related to the jamming capability allocated by the shipborne jammer in the radar direction, and the benefits of antijamming actions. After the shipborne jammer allocates a jamming capability in a certain direction that exceeds Th_1, the radar will be defeated in that confrontation round. If the missile-borne radar does not take reasonable antijamming actions, it is difficult for the missile-borne radar to implement effective detection. The formula for the success probability of missile-borne radar antijamming P_{r_k} is as follows:

$$P_{r_k} = 1 - P_k\left(1 - e_{ij}\right) \tag{5}$$

where e_{ij} indicates the benefits obtained from Equation (1) when the current radar takes the i_{th} action and the shipborne jammer takes the j_{th} action, P_k represents the jamming capability allocated by the shipborne jammer to the allocation of the k_{th} missile-borne radar.

2.4. Decision Process Setting for the Shipborne Jammer

The decision-making process for setting shipborne jammers includes four steps: data collection, threat assessment, decision-making, and jamming capability allocation.

Step 1: Data collection. The objects of data collection include information such as the distance between the missile-borne radar and the naval vessel, the velocity of the radar, the state of the missile-borne radar (search mode or tracking mode), and the pulse descriptor word (PDW) of the missile-borne radar.

Step 2: Threat assessment. It aims to assess the threat level of each radar against the naval vessel based on the collected data. The result of threat assessment to the radar is obtained by weighting based on the distance between the radar and the naval vessel and the radar operating mode as follows:

$$D_k = m_k \frac{v_k}{d_k} \tag{6}$$

where d_k is the distance between the missile-borne radar and the naval vessel, v_k represents the velocity of the radar, and m_k is the coefficient which is corresponding to the radar operating mode.

Step 3: Decision making. It aims to solve the optimization problem with the constraints of radar threat assessment results and successful detection probability to determine how to allocate the jamming capability to each radar based on the result of the threat assessment of each radar, while the total jamming capability of the shipborne jammer is constant.

Step 4: Jamming capability allocation. In this step, the shipborne allocates the jamming capability of the shipborne jammer in the radars' direction according to the optimization result.

2.5. Setting the Success Probability of Composite Jamming for the Shipborne Jammer

After obtaining the distribution of jamming capabilities of shipborne jammers in one radar direction, the corresponding jamming success probability at this time can be obtained according to Equation (5).

When allocating total jamming capability, it is generally difficult to fully jam all radars. At this time, the probability of successful detection of each radar and their threat assessment result to the shipborne jammer are the constraints for optimization. Therefore, the overall optimization goal for shipborne jammers is expressed as:

$$\begin{aligned} &\max_{A_{1k}} \left(\sum_{k=1}^{n} TA_k * Pr_k \right) \\ &subject\ to\ J = \sum_{k}^{n} \left(\int_{\theta_k - A_2\pi}^{\theta_k + A_2\pi} A_{1k} \mathrm{sinc}\left(\frac{\theta}{A_2}\right) d\theta \right) \end{aligned} \tag{7}$$

where TA_k is the threat assessment result of the k_{th} radar.

First, based on the current action of the radar and jammer, the utility is obtained by Equation (1). Then, compared with the threshold Th_1, the jamming capability of the shipborne jammer allocated in the current missile-borne radar direction is calculated according to Equation (5).

It is assumed that P_k represents the jamming probability to the k_{th} missile-borne radar, and then the success probability of shipborne jammer composite jamming to all radars is:

$$P = \prod_{k=1}^{n} P_k \tag{8}$$

where P is the success probability of shipborne jammer composite jamming to all radars, which means the shipborne jammers must ensure that the success probability of jamming to each missile-borne radar is very high to ensure the overall low probability of detection of the naval vessel.

3. Methods

RL faces two challenges in a multi-agent environment. The first is that the strategies of each agent are constantly changing in the training process, causing the environment observed by each agent to be unstable, and unstable policies cannot be applied during execution. Second, the multi-agent system should consider not only the individual intelligence degree but also the autonomy and sociality of the whole system. The combination of the optimal strategy of each agent may not necessarily be the optimal strategy of the whole system.

To obtain the optimal strategy of multi-radar collaborative antijamming, we express the multi-radar collaborative detection model as a Markov decision process (MDP) and use the MADDPG algorithm to choose the optimal strategy.

3.1. Markov Decision Process

In the process of MARL, each agent collects environmental information, makes action decisions, and evaluates the effectiveness of action decisions by calculating corresponding rewards in the next round. Through MARL, agents gradually learn optimal strategies through past experiences.

In order to use MARL to solve the optimal strategy, the model established in the above chapter needs to be restated in the Markov process. The typical MDP is a tuple $\{S, A, P, R, \gamma\}$, which contains the necessary elements in the MDP. S is the environment observed by the agent, including missile-borne radar antijamming method, the jamming style and jamming capability of the jammer, the current radar operating mode, and the distance between the radar and the jammer. A is the antijamming action set and its operating modes of the radar. R is the action reward of the radar. The collaborative goal of all radars is that at least one radar can complete the detection and antijamming task. Although one radar's probability of successful detection may be low when it attracts the main jamming capability of a shipborne jammer, its behavior is meaningful. Therefore, no additional reward is needed for the single radar. If a missile-borne radar is in tracking mode when the game comes to an end, it is considered that the radar accurately locates the naval vessel and completes the detection task, and then all radars are rewarded. γ is the discount factor. The smaller γ is, the more attention is paid to instant rewards; the larger γ is, the more attention is paid to the possible future rewards.

The Markov decision process requires determining actions before performing state transitions. The state transition process is as follows. At the time step t, assuming the environment is in the state S_t, the RL agent observes the state S_t and performs action a_t according to its strategy. This action causes the environment to move to the next state $s_{t+1} \sim P(.|s_t, a_t)$ according to the transition probability function and returns a scalar reward value $R(s_t, a_t, s_{t+1})$ to the agent according to the reward function. The objective of RL is to find an optimal strategy to maximize the cumulative reward of the expected discount obtained by the agent during the above interaction with the environment.

Based on the above optimization objective, given a strategy π, we can define the state-action function (i.e., Q-function, representing the expected discount cumulative reward that the agent can obtain if it continues to follow the given strategy π after taking action α under state s), and the value function (i.e., V-function, representing that under state s, the cumulative reward of the expected discount that an agent can obtain by following the given strategy π) is:

$$Q^{\beta}(s, a) = \mathbb{E}^{\beta}\left[\sum_{t \geq 0} \gamma^t R(s_t, a_t, s_{t+1}) \mid a_0 = a, s_0 = s\right], \forall s \in S, a \in A, \tag{9}$$

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t\geq 0} \gamma^{t} R(s_t, a_t, s_{t+1}) \mid s_0 = s\right], \forall s \in S \tag{10}$$

In Equation (9), the expected reward is based on the state transition function P_{π} composed of an infinitely long state-action trajectory locus $T = (s_0, a_0, s_1, a_1)$, and P_{π} is derived from the state transition probability P, strategy π, initial state s, and initial action α (only for the Q- function). The best strategy can be expressed as:

$$\pi^{*}(s, a) = \underset{\pi}{\operatorname{argmax}}\, Q^{\pi}(s, a). \tag{11}$$

3.2. MADDPG Solution Procedure

The MADDPG algorithm is an extension of the DDPG in multi-agent systems, which makes the traditional RL more suitable for multi-agent environments. Different from DQN, which uses a greedy strategy to select actions, the DDPG algorithm uses the actor-critic network to select actions. Specifically, it selects agent actions for actor networks. The critic network scores the actions [31], and the two networks work together to learn the optimal strategy. In the multi-agent environment, each agent is trained by the DDPG. Changes in agent strategies during training lead to unstable environments, making it difficult for the critic network and actor network to converge. To solve this problem, [26] puts forward the MADDPG network, the core part of which is to adopt the framework of centralized training and decentralized execution. During training, strategies are allowed to use the information from other agents to simplify the process, but the information is not used during testing, and each agent performs decentralized execution. The central idea of the MADDPG algorithm is as follows. For a multi-agent system, the strategy set of all agents is $\pi = \{\pi_{1,\ldots,}\pi_N\}$, and the gradient of the expected reward of agent k can be written as the following formula:

$$\nabla_{\theta_k} J(\theta_k) = \mathrm{E}_{\mathrm{s}\sim \mathrm{p}^{\mathrm{u}}, \mathrm{a}_k \sim \pi_k}[\nabla \theta_k \log \pi_k(\mathrm{a}_k \mid \mathrm{o}_k) \mathrm{Q}_k^{\pi}(\mathrm{x}, \mathrm{a}_1, \ldots\ldots, \mathrm{a}_{\mathrm{N}})] \tag{12}$$

where Q is a value function in a multi-agent set and x is composed of observed values of all agents $\mathbf{x} = (O_1, \ldots, O_N)$.

Extend the above idea to deterministic strategies and consider N consecutive strategies μ_{θ_k}, and then the gradient is written as:

$$\nabla_{\theta_k} J(\mu_k) = \mathrm{E}_{\mathrm{x,a}\sim \mathrm{D}}\Big[\nabla \theta_k \mu_k(\mathrm{a}_k \mid \mathrm{o}_k) \nabla_{\mathrm{a}_k} \mathrm{Q}_k^{\mu}(\mathrm{x}, \mathrm{a}_1, \ldots\ldots, \mathrm{a}_{\mathrm{N}}) \mid \mathrm{a}_k = \mu_k(\mathrm{o}_k)\Big] \tag{13}$$

The main motivation in the MADDPG is that if we know the actions taken by all the agents, the environment is still stable under the condition that their strategies change. Because for any strategy $\pi_k \neq \pi_k'$, there is:

$$\mathrm{P}(\mathrm{s}' \mid \mathrm{s}, \mathrm{a}_1, \ldots, \mathrm{a}_{\mathrm{N}}) = \mathrm{P}(\mathrm{s}' \mid \mathrm{s}, \mathrm{a}_1, \ldots, \mathrm{a}_{\mathrm{N}}, \pi_1', \ldots, \pi_{\mathrm{N}}') \tag{14}$$

According to the principle of the MADDPG, we train the collaborative antijamming strategy of multi-missile-borne radars. By taking the modeling process in Section 3.1, each missile-borne radar chooses the corresponding antijamming strategy and operation mode transition strategy based on its observed information. After accumulating certain training experiences, the experience replay is started and the parameters of the actor network and critic network are updated. Different from other RL scenarios, multi-missile-borne radar coordination does not reset the environment with a fixed step size but chooses the moment when the missile hits the naval vessel and calculates whether the missile-borne radar detects the naval vessel in the game. If it detects the naval vessel, it is a radar victory, and then it resets the environment and starts a new round of the game.

Ideally, multi-radar should collaborate to complete the detection task in the time domain. Time collaboration means that the detection is carried out in steps from a similar

direction so that when a shipborne jammer jams with the missile-borne radar in track mode with a high threat level, other missile-borne radars can turn to track mode to locate the naval vessel when it is approaching so that the shipborne jammer has no time to implement the jamming. Then, the radars complete the detection task and take over the guided missile to attack the naval vessel. Through the MARL method, the missile-borne radar can learn the ideal collaborative strategy and accomplish the detection task.

By using the framework of CTDE, the critic part of each agent can obtain the action information of all the other agents during the training and introduce a critic that can observe the global situation to guide actor training. While engineering, we can only take action by using the environmental information observed locally by a single agent. Each agent maintains a separate critic network and actor network. In the application, only the parameters of the actor network of the agent are needed, which conforms to the condition that is difficult for the missile-borne radars to communicate when performing tasks. The training pseudo-code is shown in Algorithm 1.

Algorithm 1: Multi-agent deep deterministic policy gradient for N agents.

for episode = 1 to M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial state x
  for t = 1 to the End of the game **do**
    for each agent i, select action $a_i = \boldsymbol{\mu}_{\theta_i}(o_i) + \mathcal{N}_t$ w.r.t. the current policy and exploration
    Execute actions $a = (a_1, \ldots, a_N)$ and observe reward r and new state x'
    Store (x, a, r, x') in replay buffer $\mathcal{D}$
    x $\leftarrow x'$
      for agent i = 1 to N **do**
Sample a random minibatch of S samples $\left(x^j, a^j, r^j, x'^j\right)$ from $\mathcal{D}$
Set $y^j = r_i^j + \gamma Q_i^{\mu'}\left(x'^j, a'_1, \ldots, a'_N\right)\Big|_{a'_k = \mu'_k(o_k^j)}$
Update critic by minimizing the loss
$\mathcal{L}(\theta_i) = \frac{1}{S}\sum_j \left(y^j - Q_i^{\mu}\left(x^j, a_1^j, \ldots, a_N^j\right)\right)^2$
        Update actor using the sampled policy gradient:
        $$\nabla_{\theta_i} J \approx \frac{1}{S}\sum_j \nabla_{\theta_i}\mu_i\left(o_i^j\right)\nabla_{a_i} Q_i^{\mu}\left(x^j, a_1^j, \ldots, a_i, \ldots, a_N^j\right)\Big|_{a_i = \boldsymbol{\mu}_i(o_i^j)}$$
      end for
      Update target network parameters for each agent i:
      $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$
    end for
end for

To improve the efficiency of missile-borne radar collaborative detection and antijamming, it is necessary to optimize the collaborative mode with the algorithm. The ideal condition of radar coordination is that real-time communication can be carried out in the process of detection and antijamming. In this way, real-time action adjustments can be made in the process of detection and antijamming in flight to improve the probability of radar successful detection. However, it is often difficult for the missile-borne radar to communicate during the implementation of detection and antijamming. Therefore, the absence of communication during collaborative antijamming can be regarded as the working condition of multi-radar cooperation. Based on the characteristics of the MADDPG algorithm for CTDE, the optimal strategy can be obtained through centralized training, and the optimal action can be given by using only local information in the application. The flowchart of the proposed algorithm is shown in Figure 2. The left part shows the radars performing centralized training to learn appropriate strategies for choosing operating modes and antijamming actions. They can obtain status information on other radars during the training stage. The right part shows the decentralized execution stage. The radars

perform the jamming recognition and then choose antijamming actions and operating modes. Then, it comes to the jammer round. It collects radar information and assesses the radar's threat to the jammer. Then, it selects jamming actions, obtains the radar's detection success probability and allocates jamming capabilities to each radar. We should judge if the confrontation has come to an end. If so, the experiment shows the confrontation result; otherwise, the experiment goes on to jamming recognition. Repeat this loop until the confrontation comes to the end of the game.

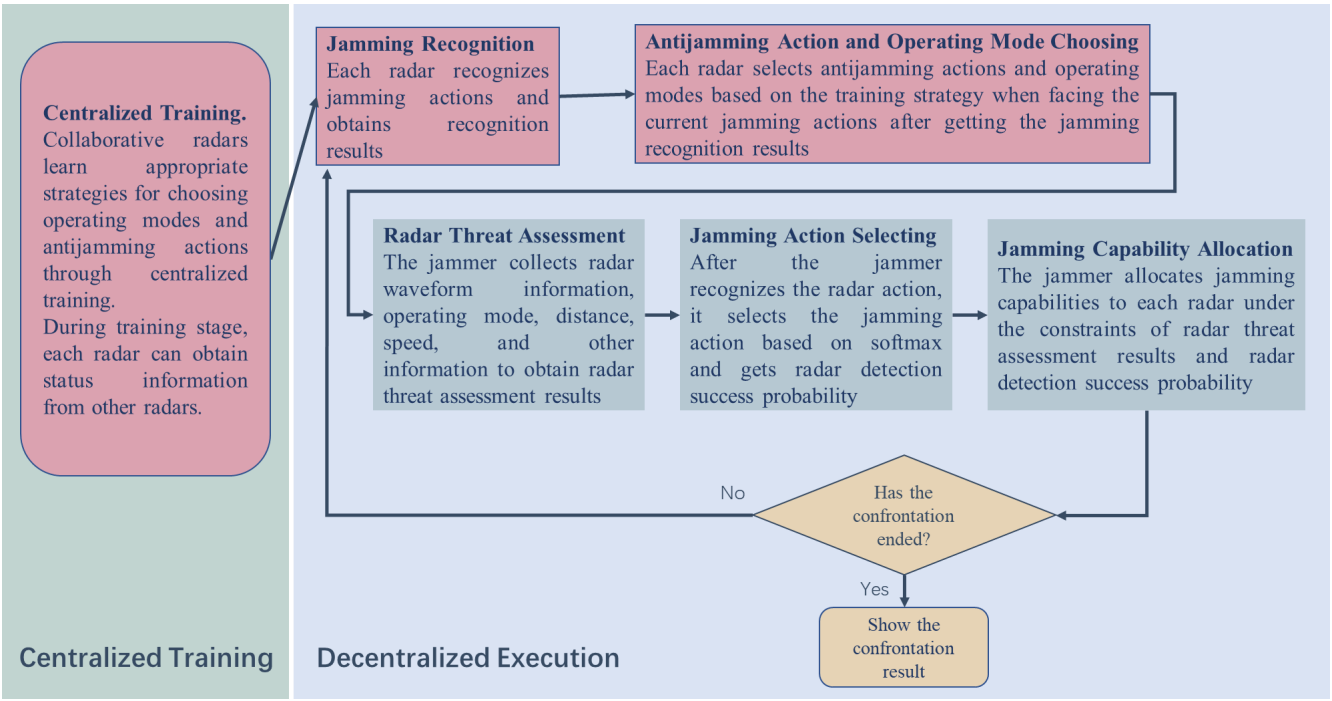


Figure 2. Flowchart of the MADDPG algorithm in multi-missile-borne radars vs. the shipborne jammer.

4. Results

In this section, we evaluated the performance of the MADDPG in the constructed game model between radars and jammers with the DDPG (each agent individually adopts the DDPG algorithm in training), regular strategy (each agent adopts the strategy with the best antijamming effect), and random strategy (each agent selects its strategy randomly). To show the advantage of the MADDPG in the selection of a multi-agent collaborative strategy, we carried out the simulations with two radars and four radars, respectively.

4.1. Two-Radar Experiment

There are five strategies for setting up a radar and four strategies for jammers to choose from. The initial distance of the missile is 80 km, and the radar velocity is 1 km/s. When the missile reaches the naval vessel, the game ends. According to Section 2.2, the radar antijamming probability matrix in the simulation is as follows:

$$E = \begin{pmatrix} 1 & 0.1 & 0.2 & 0.2 & 0.9 \\ 0.2 & 1 & 0.1 & 0.1 & 0.8 \\ 0.2 & 0.3 & 1 & 0.1 & 0.8 \\ 0.2 & 0.2 & 0 & 1 & 0.9 \end{pmatrix}$$

Due to the existence of the preparation interval and jammer recognition interval for radar actions, the actions with the best antijamming effect may not necessarily be the best. For example, when radar takes the first antijamming against the first jamming style, it has achieved a good antijamming effect at the current time. However, when the jammer

changes its jamming strategy, the antijamming effect will decrease when the radar is in recognition time. In this situation, adopting other strategies may bring greater rewards and it requires radars to learn to obtain the optimal combination strategy.

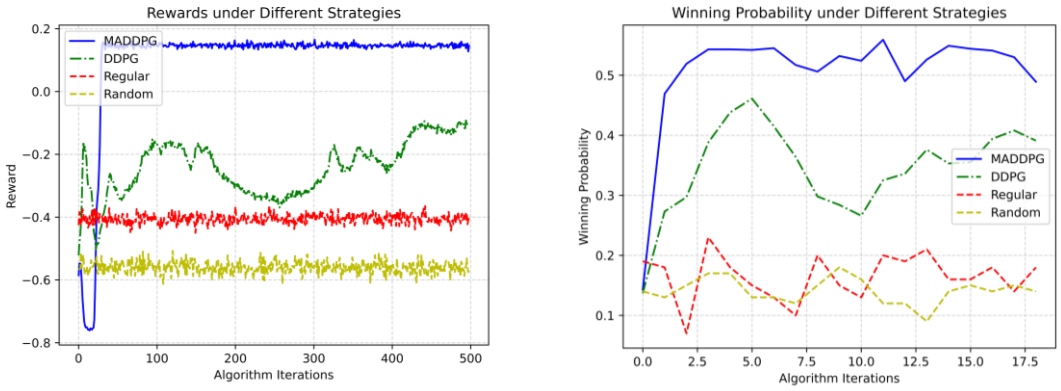
The collaboration between the two radars is mainly reflected in the collaboration of operating modes, in which each radar can choose whether to switch to tracking mode from search mode. In the case of multi-agent cooperation, the missile-borne radar can approach the naval vessel and switch to tracking mode when the jammer focuses on the jamming capability to attack another radar, accurately positioning the naval vessel's position and breaking through defense.

The network parameters involved in the simulation are shown in Table 1.

Table 1. Simulation parameter table.

Parameter	Value
Total time steps	16,000,000
Actor network learning rate	0.0001
Critic network learning rate	0.001
Batch size	256
Reward discount factor	0.95
Motion exploration rate	0.9
Reduced exploration rate	0.000005
Initial distance of agent	80 km
Agent speed	1 km/s
Radar operating mode	Search mode, track mode
Search mode weight	10
Tracking mode weight	20
Radar acceptance jamming threshold (Th1)	0.5
Tracking mode jamming interval threshold (Th2)	5
Minimum duration of search mode (Th3)	5

The simulation results are shown in Figure 3.



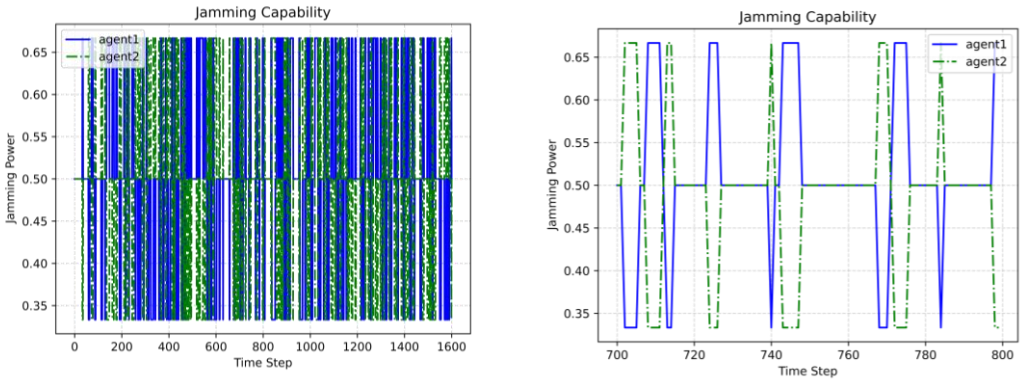
(**a**) (**b**)

Figure 3. A comparison of different strategies in the two-radar scenario. (**a**) Reward in the two-radar scenario. (**b**) Winning probability in the two-radar scenario.

As can be seen in Figure 3a, for the MARL problem, the reward value of the fixed optimal antijamming strategy is relatively constant and higher than the random strategy. The DDPG algorithm cannot learn the stable strategy due to the change in other agent strategies, and the reward value fluctuates in a large range. The highest reward can exceed the regular strategy, while the lowest reward is lower than the random strategy. There is no convergence trend in the simulation. When each radar optimizes its antijamming strategy without collaboration, the result will not be good enough. By learning other agent strategies, the MADDPG algorithm learns a stable strategy with a higher reward value than

the strategy trained by the DDPG algorithm and the fixed strategy. It reaches convergence soon after the training and has better performance than the DDPG. To remove the influence of some singular results on the experiment, 1000 rounds of the game are conducted for the models trained at different stages to eliminate the randomness. As can be seen in Figure 3b, the strategy trained with the MADDPG can bring a higher average winning probability, representing that the missile-borne radars have better performance, while the average winning probability of the other three strategies is lower than the MADDPG. The winning probability of the MADDPG algorithm can reach 50% in the confrontation, which is 10% higher than the DDPG algorithm, demonstrating the effectiveness of collaboration optimization. The winning probability of the regular strategy and random strategy is very low, and we cannot see any trend of better performance. When the radars win the game at a percentage of less than 20%, the task can hardly be conducted.

In Figure 4a, we can find the jamming capability allocated by the shipborne radar to the two radars changing during the game. In Figure 4b, the trend can be observed more intuitively. This proves that the two radars have realized collaborative detection by changing their operation mode to alternate the threats to the shipborne jammer.



(**a**) (**b**)

Figure 4. Jamming capability allocated to the two radars. (**a**) During the whole game. (**b**) In typical time steps.

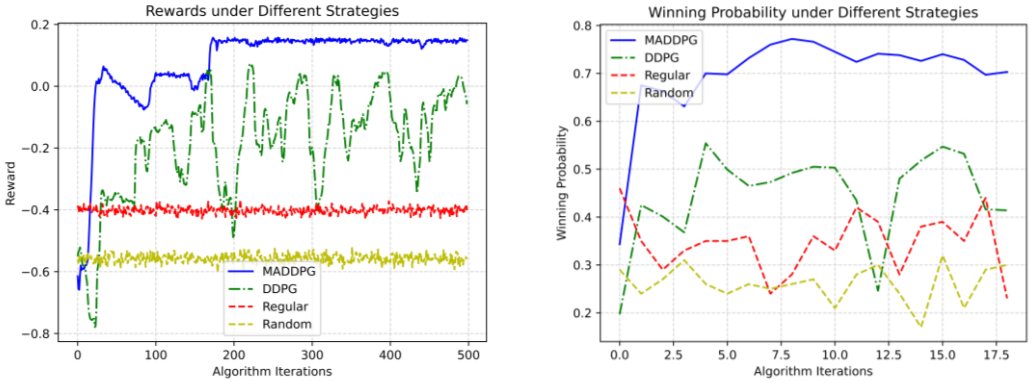
The experimental results show that the two missile-borne radars can collaborate by actively controlling their operating modes and antijamming actions, breaking through the jammer defense at weak points of the jammer's capability, and completing collaborative detection tasks.

4.2. Four-Radar Experiment

In the simulation, except for the fact that Th_1 is 0.25, the radar parameters and training parameter settings are the same as in the simulation in Section 4.1. The simulation results are in Figure 5.

It can be seen in Figure 5a that, due to the increased environmental instability brought by the change in agent strategy, it is more difficult for the DDPG algorithm to learn the stable strategy; the reward value fluctuates in a large range and the highest point is smaller than the MADDPG algorithm. Similarly, the reward of the MADDPG algorithm also exceeds that of the fixed strategy and random strategy, and it is convergent. The simulation also carries out 1000 rounds of games on the model trained at different stages to eliminate randomness. As can be seen in Figure 5b, the strategy trained with the MADDPG algorithm can bring a higher average winning probability, which means that the missile-borne radar learns the collaborative detection method and completes the detection task better. The average winning probability of the other three strategies is lower than the MADDPG. The winning probability of the MADDPG algorithm can reach 70% in the confrontation, which is 20% higher than the DDPG algorithm, demonstrating the effectiveness of collaboration optimization. The winning probability of the regular strategy and random strategy is very

low, and we cannot see any trend of better performance. When the radars win the game at a percentage of less than 30%, the task can hardly be conducted.



(**a**) (**b**)

Figure 5. A comparison of the different strategies in the four-radar scenario. (**a**) Reward in the four-radar scenario. (**b**) Winning probability in the four-radar scenario.

5. Discussion

The stability of the rewards for regular and random strategies indicates that they have not been optimized and are fixed, which can be predicted before we really knew the result. Based on the previous two experiments, it can be observed that stable strategies can be learned through the MADDPG in our proposed multi-missile-borne radar collaborative detection model. During training, we observe that the reward initially increases and then stabilizes.

The MADDPG algorithm trains to obtain optimal multi-agent collaboration strategies that maximize overall rewards in the MDP. Experiments demonstrate that using the MADDPG enables radars to collaborate and generate operational synergy, simplifying the environment. Jamming capability allocation in the model shows that tracking mode radars receive more jamming capability. Experimental results indicate that the MADDPG-derived strategies actively weaken radar jamming in specific directions, allowing other radars to complete detection tasks more effectively. This collaborative approach outperforms individual intelligence strategies obtained via RL and traditional antijamming methods, increasing the probability of successful detection.

When more radars are used to complete the detection task, we can obtain a larger winning probability. Meanwhile, radars have a greater optimization space, as we can see from the experimental results that the gap between the MADDPG algorithm and the DDPG algorithm in the four-radar experiment is larger than the two-radar experiment.

We should compare this proposed method with other methods to see its performance. However, few methods have been applied in similar scenarios. That is why we set up the experiment by comparing it with the regular strategy, random strategy, and radars and jammers with the DDPG algorithm. MARL has been used in radar collaboration for planning the path [21] and other aspects. However, an antijamming strategy that considers the dynamics in the confrontation between radars and the jammer has not been widely studied.

In this study, the radars keep changing their actions and operating modes to win the game according to their strategy based on MARL. We need plenty of information on antijamming actions to obtain features that are used in the game if we want to verify the rationality of the proposed algorithm with real data, which is very difficult to achieve. Therefore, a simulation experiment that uses typical parameters is the best way to verify the rationality of the proposed algorithm, which has credibility to some degree.

Radars are guided when facing various jamming actions and detection probabilities through centralized training, and they know how to select their antijamming actions and

operating modes. By actively changing the operating mode of each radar and retaining the ability to quickly switch, the radar can change its threat level to the jammer, thereby causing the jammer to passively change the distribution of jamming capability. This mobilization process is continuous and dynamic, and each radar is constantly engaged in this process in the game, resulting in a constantly changing distribution of jammer capabilities. In the radar/jamming confrontation model, multiple radars can synergistically adjust the actions of jammers, finally resulting in weaknesses in the allocation of jamming capabilities, creating conditions for the radars to complete detection tasks.

The collaborative detection task of multiple radars is similar to the "hit and run" tactics in real-time strategy games. After the intelligent agent attacks, it leaves the attack range of the target and attracts the attention of the target. Other collaborative intelligent agents take the opportunity to attack and repeat the process until they win. The collaborative detection task of multiple radars needs to be completed without communication, making it much more difficult than "hit and run" tactics.

When applying this method in the confrontation between multi-missile radars and the shipborne jammer, the radars should change their operating mode actively. Then, they can easily change from search mode to track mode. Otherwise, they will have to search for the target and cannot obtain the target information within the interval of Th_3, which means the radar will lose the condition to collaborate with other radars. Then, other radars will withstand more jamming capability and will be easier to move to search mode passively. The situation will get worse and worse and finally lead to an unsuccessful collaboration.

In this article, we assume that the beam width of the radars is very narrow, and that the jammer uses a phased array antenna to form multiple beams adaptively in the direction of the radars, countering multiple radars simultaneously. However, if the missile-borne radars come from the same launch site, they will be jammed by one beam at the same time, rendering the proposed method ineffective. When analyzing the problem, we can separate it into two aspects. If the territory of the party launching the missiles is large and has many launch sites that are far away from each other, then the problem will not arise. However, if the territory of the launch party is not large enough, or even if all the missiles need to be launched from the same site, they can achieve the purpose of detecting from different directions through trajectory planning and launch time control. The missiles will not fly along a straight line between the launch site and the target but will maneuver in all directions separately and attack from different directions. By combining trajectory planning and launch time control, it is possible to achieve a focused strike when the missiles reach the target. This can avoid all missiles being jammed simultaneously, but it requires sacrificing a certain degree of timeliness. The flight interval of the missiles will become longer, leaving more reaction time for the target but greatly improving the synergy between the missiles. Therefore, the proposed method can work through reasonable planning in advance.

Multi-radar collaboration is often linked to multi-missile collaboration. In the past, research on multiple missiles attacking a naval vessel believed that as long as a large number of missiles simultaneously attacked ship targets from different directions, the detection task could be completed. However, insufficient consideration was given to the collaboration between missile-borne radars. We may determine the number of missiles to be launched based on prior knowledge, but we may also wonder if these missiles have a high enough probability to complete the attack mission. A sufficient number of missiles may indeed complete the mission, but this would result in a significant waste of cost. We may question whether even fewer missiles can also complete the mission. Alternatively, based on the method proposed in this article, we can conduct various simulation experiments in the model to obtain the probability of completing detection with different radar numbers, compare them with the expected minimum hit rate, determine the number of missiles and missile-borne radars used, and strike a balance between task completion rate and the number of missiles consumed. This can greatly improve the efficiency of a single radar in completing tasks.

6. Conclusions

This paper examines the optimization method of collaborative detection and anti-jamming of a naval vessel by multiple missile-borne radars. It constructs a two-person zero-sum game between the radars and a shipborne jammer and assumes a limited total jamming capability that is allocated to each radar direction by solving the optimization problem. The game focuses on securing a favorable position while considering temporal constraints. Using the MADDPG, a MARL method with CTDE framework, the study optimizes the collaborative detection and antijamming results without communication. The simulation confirms the effectiveness of the proposed method, which improves the winning probability of multiple radars.

This paper models the decision-making process of the jammer, but there is no public data available for reference. Therefore, there is a possibility that the modeling may not be consistent with the actual decision-making process of the jammer, which may affect the optimization results of the method. However, this method is not specific to one decision-making process of the jammer. After obtaining enough information to update the model, this proposed method is still applicable.

Author Contributions: Conceptualization, C.F. and J.D.; methodology, C.F.; software, Z.W. and Z.Z.; validation, Z.W. and Z.Z.; formal analysis, C.F., Z.W. and Z.Z.; investigation, J.D.; resources, J.D.; data curation, Z.W. and Z.Z.; writing—original draft preparation, C.F., Z.W. and Z.Z.; writing—review and editing, C.F. and J.D.; visualization, Z.W. and Z.Z.; supervision, X.F. and T.P.; project administration, X.F.; funding acquisition, X.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by 111 Project of China, grant number B14010.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spezio, A.E. Electronic warfare systems. *IEEE Trans. Microw. Theory Tech.* **2002**, *50*, 633–644. [CrossRef]
2. Farina, A. Electronic Counter-Countermeasures. In *Radar Handbook*, 3rd ed.; Skolnik, M., Ed.; McGraw-Hill: New York, NY, USA, 2008; pp. 24.1–24.59.
3. Xiao, Z.; He, R.; Zhao, C. Cooperative Combat of Missile Formation: Concepts and Key Technologies. *Aerosp. Electron. Warf.* **2013**, *29*, 1–3.
4. Zhao, J.; Yang, S. Review of Multi-Missile Cooperative Guidance. *Acta Aeronaut. Et Astronaut. Sin.* **2017**, *38*, 22–34.
5. Wang, J.; Li, F.; Zhao, J.; Wang, C. Summary of Guidance Law based on Cooperative Attack of Multi-Missile method. *Flight Dyn.* **2011**, *29*, 6–10.
6. Taj, M.; Cavallaro, A. Distributed and Decentralized Multicamera Tracking. *IEEE Signal Process. Mag.* **2011**, *28*, 46–58. [CrossRef]
7. Liggins, M.E.; Chong, C.-Y.; Kadar, I.; Alford, M.G.; Vannicola, V.; Thomopoulos, S. Distributed fusion architectures and algorithms for target tracking. *Proc. IEEE* **1997**, *85*, 95–107. [CrossRef]
8. He, S.; Shin, H.; Xu, S.; Tsourdos, A. Distributed estimation over a low-cost sensor network: A Review of state-of-the-art. *Inf. Fusion* **2020**, *54*, 21–43. [CrossRef]
9. Akcakaya, M.; Nehorai, A. Adaptive MIMO Radar Design and Detection in Compound-Gaussian Clutter. *IEEE Trans. Aerosp. Electron. Syst.* **2011**, *47*, 2200–2207. [CrossRef]
10. Chong, C.Y.; Pascal, F.; Ovarlez, J.-P.; Lesturgie, M. Adaptive MIMO radar detection in non-Gaussian and heterogeneous clutter considering fluctuating targets. In Proceedings of the 2009 IEEE/SP 15th Workshop on Statistical Signal Processing, Cardiff, UK, 31 August–3 September 2009; pp. 9–12.
11. Wang, P.; Li, H.; Himed, B. A Parametric Moving Target Detector for Distributed MIMO Radar in Non-Homogeneous Environment. *IEEE Trans. Signal Process.* **2013**, *61*, 2282–2294. [CrossRef]
12. Yang, Y.; Blum, R.S. Phase Synchronization for Coherent MIMO Radar: Algorithms and Their Analysis. *IEEE Trans. Signal Process.* **2011**, *59*, 5538–5557. [CrossRef]
13. Yang, Y.; Su, H.; Hu, Q.; Zhou, S.; Huang, J. Centralized Adaptive CFAR Detection with Registration Errors in Multistatic Radar. *IEEE Trans. Aerosp. Electron. Syst.* **2018**, *54*, 2370–2382. [CrossRef]
14. Hu, J.; Xie, L.; Zhang, C. Diffusion Kalman Filtering Based on Covariance Intersection. *IEEE Trans. Signal Process.* **2012**, *60*, 891–902. [CrossRef]
15. Mathur, A.; Willett, P.K. Local SNR considerations in decentralized CFAR detection. *IEEE Trans. Aerosp. Electron. Syst.* **1998**, *34*, 13–22. [CrossRef]

16. Yang, Y.; Su, H.; Hu, Q.; Zhou, S.; Huang, J. Spatial Resolution Cell Based Centralized Target Detection in Multistatic Radar. *Signal Process.* **2018**, *152*, 238–246. [CrossRef]
17. Panoui, A.; Lambotharan, S.; Chambers, A. Game theoretic power allocation for a multistatic radar network in the presence of estimation error. In Proceedings of the Sensor Signal Processing for Defence, Edinburgh, UK, 8–9 September 2014; pp. 1–5.
18. Panoui, A.; Lambotharan, S.; Chambers, A. Waveform allocation for a MIMO radar network using potential games. In Proceedings of the Radar Conference, Arlington, VA, USA, 10–15 May 2015; pp. 751–754.
19. Chavali, P.; Nehorai, A. Scheduling and Power Allocation in a Cognitive Radar Network for Multiple-Target Tracking. *IEEE Trans. Signal Process.* **2012**, *60*, 715–729. [CrossRef]
20. Busoniu, L.; Babuska, R.; De Schutter, B. A Comprehensive Survey of Multiagent Reinforcement Learning. In *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*; IEEE: Piscataway, NJ, USA, 2008; Volume 38, pp. 156–172.
21. Wei, X.; Yang, L.; Cao, G.; Lu, T.; Wang, B. Recurrent MADDPG for Object Detection and Assignment in Combat Tasks. *IEEE Access* **2020**, *8*, 163334–163343. [CrossRef]
22. Xia, Z.; Du, J.; Wang, J.; Jiang, C.; Ren, Y.; Li, G.; Han, Z. Multi-Agent Reinforcement Learning Aided Intelligent UAV Swarm for Target Tracking. *IEEE Trans. Veh. Technol.* **2022**, *71*, 931–945. [CrossRef]
23. Sutton, R.S.; Barto, A.G. Reinforcement Learning: An Introduction. *IEEE Trans. Neural Netw.* **1998**, *9*, 1054. [CrossRef]
24. Gao, Y.; Chen, S.; Lu, X. Research on Reinforcement Learning Technology: A Review. *Acta Autom. Sin.* **2004**, *30*, 86–100.
25. Du, W.; Ding, S. Overview on Multi-Agent Reinforcement Learning. *Comput. Sci.* **2019**, *46*, 1–8.
26. Rashid, T.; Samvelyan, M.; De Witt, C.S.; Farquhar, G.; Foerster, J.; Whiteson, S. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv* **2018**, arXiv:1803.11485.
27. Lowe, R.; Wu, Y.I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
28. Foerster, J.; Nardelli, N.; Farquhar, G.; Afouras, T.; Torr, P.H.S.; Kohli, P.; Whiteson, S. Stabilising experience replay for deep multi-agent reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, NSW, Australia, 6–11 August 2017.
29. Feng, C.; Fu, X.; Lang, P.; Zhao, C.; Dong, J.; Pan, T. A Radar Anti-Jamming Strategy Based on Game Theory with Temporal Constraints. *IEEE Access* **2022**, *10*, 97429–97438. [CrossRef]
30. Neumann, J.V.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 1947.
31. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.