Computer Communications 209 (2023) 283–301

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Autonomous confrontation strategy learning evolution mechanism of unmanned system group under actual combat in the loop

Wang Zhenhua [a,1], Guo Yan [a,*,2], Li Ning [a], Yuan Hao [a], Hu Shiguang [b], Lei Binghan [a], Wei Jianyu [a]

[a] College of Communications Engineering, Army Engineering University of PLA, Jiangsu, Nanjing, 210007, China
[b] College of Equipment Management and UAV Engineering, Air Force Engineering University of PLA, Shanxi, Xian, 710051, China

ARTICLE INFO

Keywords:
Unmanned system group
Actual combat in the loop
Autonomous confrontation strategy
ACS-ACL algorithm
Learning evolution mechanism

ABSTRACT

Confrontation of unmanned system group (USG) is an important combat pattern in future aerial combat, and autonomous confrontation strategy learning evolution is the pre-foundation of USG for actual combat application, researching multiple problems concerning the realization of USG autonomous confrontation strategy active learning evolution in high-dynamic actual combat scenario through continuous interaction with commander, an Autonomous Confrontation Strategy learning evolution mechanism of USG under Actual Combat in the Loop (ACS-ACL) was thence proposed. Select the Multi Agent Deep Deterministic Policy Gradient (MADDPG) algorithm as the baseline algorithm, introduce the Parallel Decoupling Reward Mechanism (PDRM) to make applicability improvement on MADDPG algorithm, establish the generation model of USG autonomous confrontation strategy; after generating the initial autonomous confrontation strategy, USG proactive initiation the Continuous Interaction (CI) with the commander of actual combat in the loop, and uploads the perception information of recessive battlefield situation, whereas commander makes proofreading supplement for the information of battlefield situation, and transmits them back to USG with combat intention together; USG updates replay experience pool, updates autonomous confrontation strategy according to the combat intention, and updates simultaneously interaction strategy with actual combat commander in the loop, and then establishes the autonomous confrontation strategy benign closed-loop learning evolution mechanism of USG. Assume USG execution of collaborative search moving target mission against the enemy, and a visual USG autonomous collaborative search dynamic confrontation game environment is constructed, and carry out a series of simulation validation experiments. By observation and comparison, the convergence efficiency and execution quality of autonomous confrontation strategy driven by combat intention are improved significantly, the autonomous confrontation strategy learning has the benign evolution trend, and further improves the credibility of actual combat application of autonomous confrontation strategy of USG.

Introduction

Future war pattern has significant attributes such as strong confrontation, strong game and high dynamics, information of ourselves and the enemy in battlefield is asymmetrical and incomplete, and both parties of ourselves and the enemy introduce plenty of unmanned combat equipment, thus future war combat pattern evolves gradually into the USG confrontation between ourselves and the enemy, making battlefield progress and rhythm expedite sharply [1], and thereby raising higher requirements on the abilities of front-line commander for battlefield situation evaluation and combat intention decision. Benefited from the wide and deep application of Artificial Intelligence Technology (AIT) such as Deep Reinforcement Learning (DRL) etc. in the field of USG autonomous collaborative control, USG represented by Unmanned Aerial Vehicle (UAV) has been increasingly applied in the military fields of collaborative reconnaissance, saturation attack, collaborative detection, collaborative round-up, collaborative search, and precision strike in recent years [2], and has already become an important support of establishing the precise and collaborative coordination system of multi-mission unmanned combat system. In which, using USG to make collaborative search on multiple moving time-sensitive target in target area is the important actual combat application direction of USG, and has important reference significance [3] for the construction of Autonomous Confrontation Strategy Learning Evolution (ACSLE) mechanism of USG. In the course that USG makes collaborative search on moving time-sensitive target, moving time-sensitive target, with a high

* Corresponding author.
E-mail addresses: wangzhenhua@aeu.edu.cn (Z. Wang), gongyou_3000@sina.com (Y. Guo).

[1] (1990—), Doctoral student, Main research directions: unmanned system group collaborative combat, multi-agent deep reinforcement learning, UAV group behavior autonomous decision, etc.

[2] (1971—), Professor, doctoral supervisor, Main research directions: signal processing, multi-agent deep reinforcement learning, compressive sensing, etc.

https://doi.org/10.1016/j.comcom.2023.07.006
Received 19 April 2023; Received in revised form 5 July 2023; Accepted 5 July 2023
Available online 17 July 2023

probability, might move alone or by group to the searched areas [4]. In order to improve collaborative search efficiency, USG is required to realize accurate prediction on the moving tendency of moving time-sensitive target by learning and evolving continuously autonomous confrontation strategy. Under actual combat scenario in future, battlefield situation will be complicated and varying, and information of ourselves and the enemy in battlefield will be unsymmetrical and incomplete, USG will be required to totally have stronger online learning evolution attribute [5], utilizes continuous exploration and identification in the interaction process of highly dynamic and complicated battlefield environment and information of ourselves and the enemy in the combat process to realize autonomous confrontation strategy learning evolution, realize the complexity degree of the environment in which USG is located always matches the behavioral autonomy capability of USG, and provide technical supports for carrying out large-scale application duplication of actual combat application in a short time.

In this paper, the research motivation is to further improve the actual combat application credibility of USG autonomous confrontation strategy, strengthen the guiding position of actual combat commander in the loop in the USG autonomous confrontation strategy learning evolution course, multiple problems concerning or related to the realization of USG autonomous confrontation strategy active learning evolution in high-dynamic actual combat scenario through continuous interaction with commander are researched, and an autonomous confrontation strategy learning evolution mechanism (ACS-ACL) of USG under actual combat in the loop is proposed. In which, USG internally adopts fully connected communication network, and its members are homogeneous and isomorphic individuals, and have the completely consistent kinematics and maneuvering characteristics parameters. Select MADDPG algorithm as the baseline algorithm, introduce PDRM to make applicability improvement on MADDPG algorithm, establish the generation model of USG autonomous confrontation strategy; after generating the initial autonomous confrontation strategy, USG makes proactive initiation on continuous interaction (CI) with actual combat commander in the loop, and uploads the perception information of recessive battlefield situation, whereas commander makes proofreading supplement for the information of battlefield situation, and transits them and combat intention back to USG; USG updates replay experience pool, updates autonomous confrontation strategy according to the combat intention, updates simultaneously interaction strategy with actual combat commander in the loop, and then establishes the autonomous confrontation strategy benign closed-loop learning evolution mechanism of USG. In order to make multidimensional verification on the performance of the proposed ACS-ACL algorithm, a visible USG autonomous collaborative search dynamic confrontation game environment was constructed, and a series of simulation validation experiments were carried out, finally multiple valuable conclusions were obtained.

The main contributions of the ACS-ACL algorithm are summarized as follows: (1) Firstly, the Continuous Interaction (CI) mechanism was proposed, continuously fill in self-experience gap through the continuous interaction (CI) with actual combat commander in the loop, expedite the experience learning process of USG, generate the USG autonomous confrontation strategy coupled with volition of commander, and further improve the actual combat application credibility of USG autonomous confrontation strategy, strengthen the guiding position of actual combat commander in the loop in the USG autonomous confrontation strategy learning evolution course. (2) Secondly, the autonomous confrontation strategy Active Learning Evolution(ALE) mechanism was proposed, under the action of active learning evolution (ALE) mechanism, USG autonomous confrontation strategy coupled with volition of commander triggers the targeted active learning evolution, on the one hand, it guides USG behavioral autonomy capability to always adapt to the strong-gaming, high mobility, and high-intensity complicated actual combat battlefield confrontation environment; on the other hand, it guides USG optimal joint strategy to constantly match with the USG optimal joint strategy of confrontation party to realize the continuous benign evolution of autonomous confrontation strategy learning under the driving of combat intention. (3) Thirdly, the Indirect Communication Channel (ICC) mechanism oriented to interior communication of USG is proposed, USG members utilize volition share buffer pool as the indirect communication channel (ICC), the convergence efficiency and quality of ACS-ACL algorithm are effectively optimized.

All work contents in this paper are organized as follows: Section 1 describes simply the related works; Section 2 provides the unmanned aerial vehicle cluster collaborative search moving target model and UAV kinematics description model; Section 3 provides USG autonomous confrontation strategy learning evolution mechanism (ACSLE of USG); Section 4 presents contrastive simulation experiment and discussions; Section 5 provides a comprehensive summary of this paper and presents directions for future work.

1. Related work

Scholars in China and abroad, aiming at the problems concerning Autonomous Confrontation Strategy Learning Evolution (ACSLE) of USG, have carried out abundant and useful exploratory works, and have yielded multiple representative achievements; the entry point of all of these works is USG autonomous confrontation strategy learning and evolution, the research scenario therein specifically involve in area coverage, status estimation, relative positioning, collaborative search, mission planning, autonomous obstacle avoidance, formation reconstruction, consistent tracking, communication relay, collaborative defense, collaborative confrontation, collaborative navigation, collaborative round-up etc., and the research methods can be roughly divided into two categories of model-based strategy driving, model-free-based strategy driving, briefly described as follows:

1.1. Model-based strategy driving

The core thinking of the model-based strategy driving is to make optimal maneuvering decision strategy search of USG members by establishing the descriptive model of USG elements such as position, gesture and energy, and by using one or multiple intelligent optimization algorithms, the representative achievements therefrom are described briefly as follows: Literature [6] presents one intelligent decision process framework for multi-to-multi UAV aerial combat, it utilizes the improved hybrid particle swarm optimization algorithm to search optimal target assignment of multi-to-multi UAV aerial combat on the basis of constructing multiple UAV moving elements advantage evaluation model, and further realize multiple UAV maneuvering collaborative decision by relying on Double Q-learning algorithm; Literature [7] presents one multiple UAV collaborative reconnaissance path planning method under multiple bases scenario, it establishes the global path planning model of multiple UAV under high dimensional constraint by relying on the graph theoretic approach, further uses the improved ant colony algorithm to solve the global path planning model of multiple UAV, and generates in the real-time way the collaborative reconnaissance path with battlefield application vale; Literature [8] presents one collaborative flight path planning method of multiple UAV oriented to dynamic and complicated aerial combat scenarios, it conducts dimensionality reduction for constraint planning computation complexity through hierarchical planning thought, uses Tent chaotic mapping factor to improve genetic algorithm, and designs a Chaos Elite Adapts-Genetic Algorithm (CEA-GA) to seek solutions for the three-dimensional collaborative curve flight path planning problem of multiple UAVs, and verify the effectiveness of CEA-GA method via simulation experiment; Literature [9] presents one multi-task planning framework oriented to multiple UAV collaborative strike against ground, in which the double- layer mutual coupling mechanism was introduced to optimize multi-task planning sequence, and then the Simulated Annealing - Picking-uniform-points Algorithm that can jump out

of local optimum was designed to realize the global optimal assignment for multi-task of multiple UAV; Literature [10] presents one intelligent collaborative aerial combat algorithm of multiple UAV groups based on predictive game tree, and capsulate a package of tactical maneuvering action based on existing aerial combat knowledge, and further designs a complete of aerial combat situation evaluation function to realize multiple UAV role assignment and real-time maneuvering decision, utilizes Unity3D to set up a set of simulation environment closing to real aerial combat scenario, and then carry out a series of comparative simulation verification experiments to verify the effectiveness of algorithm; Literature [11] presents one multiple UAV collaborative target defense algorithm based on exponentially averaged momentum pigeon-inspired optimization algorithm, it firstly makes modeling in three-dimensional space for the collaborative target defense system of multiple UAV to obtain the optimal control input amount of UAVs of ourselves and the enemy under confrontation, then constructs the target function of optimization algorithm with the multilevel penalty function method, and uses the Exponential average Momentum-Pigeon Inspired Optimization algorithm (EM-PIO) to solve the optimal target points, and finally verifies the effectiveness of EM-PIO algorithm through a series of experiments.

1.2. Model-free-based strategy driving

The core thinking of the model-free-based strategy driving is to utilize the model-free reinforcement learning to realize the learning confrontation strategy and autonomous evolution of USG in the process of making continuous interaction (CI) with environment, its basic thinking is to make applicability improvement and extension on deep reinforcement learning algorithms such as Deep Q Network (DQN), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), MADDPG, give full play to the overall perceptibility of deep learning and the real-time decision-making ability of reinforcement learning, trained USG members emerge actively good exploration competence, generate autonomously in the model-free pattern the USG autonomous confrontation strategy, and utilize the situation information of ourselves and the enemy in combat process to realize autonomous confrontation strategy learning evolution. The representative achievements are described briefly as follows: Literature [12] presents one decision flow framework oriented to collaborative aerial combat of multiple UAV, in which PPO algorithm was made multi-angle optimization and multiple optimization mechanisms such as self-adaption weight and preferential sampling were integrated to effectively improve the efficiency and stability of model training, and the feasibility of framework was verified through the Wargame Platform; Literature [13] presents one UAV autonomous maneuver decision method in the short-distance aerial combat scenario under highly dynamic and uncertain maneuver constraint, in which DQN algorithm was made extension and a staged training method called as “basic confrontation” was designed to help to reduce the training time and acquire the suboptimal, but effective results in the high-dimensional condition and action space; Literature [14] presents one DDPG algorithm with the course learning for the confrontation game of UAVs of both sides of ourselves and enemy involved in the conflict of interest in the complicated dynamic battlefield environment, in which the intelligence degree of confrontation policy of enemy’s UAV was progressively improved in the progressive form in the confrontation process, thus the convergence efficiency and scenario generalization ability of the DDPG algorithm training was improved hugely; Literature [15] presents one isomerous UAV group multi-task assignment model based on the Half-Random Q-learning (HR Q-learning) algorithm against the problem of optimal assignment of combat task of isomerous UAV group in the dynamic and complicated battlefield environment, in which the exploration process of Q learning algorithm was improved, the probability of obtaining invalid actions in the random scenario was reduced, thereby the success rate of combat tasks executed by isomerous UAV group was significantly improved; Literature [16] presents one MADDPG algorithm integrated with the Mixed Experience policy (ME-MADDPG) against the problem of the high-efficiency motion planning of decentralized multi-agent under the complicated dynamic constraint, in which the artificial potential field method was used to improve the generation quality of specimen at the stage of early training, the source of training data was expanded by utilizing the dynamic hybrid sampling policy, the stability of model training process was reinforced by adopting the delayed learning mechanism, and the feasibility of ME-MADDPG algorithm was verified through a series of experiments.

To sum up, many scholars in China and abroad have made in-depth study on ACSLE problems of USG and have achieved multiple gratifying results, but there are certain limitations to all the model-based strategy driving and model-free-based strategy driving research methods. Therein, the model-based strategy driving research methods can realize real-time assessment on battlefield situation and have dynamic attributes in a certain sense, but rely highly on expert experience, have difficulties in designing situation evaluation function [17], have no way to meet battlefield real-time decision demands, and are weak in the generalization ability of battlefield environment, so they apply generally to USG autonomous confrontation strategy learning and evolution scenario with relatively stable battlefield environment and moderate requirements on real-time capability; the model-free-based strategy driving research methods can realize the one that USG learns confrontation strategy and makes autonomous evolution in the process of making CI with environment, and have good scenario adaption and generalization capability; they do not need to acquire the dynamical model of environment and special communication demands at the stages of training and execution, but they are always of low efficiency in the utilization rate of samples when dealing with ACSLE problems of relatively large-scale USG, resultantly the strategy convergence time extends, the interpretability becomes worse, and the actual combat application credibility of USG autonomous confrontation strategy is worse.

2. Problem description and modeling

USG execution of collaborative search moving-target mission against the enemy is the important combat pattern of cross-domain collaborative combat. Select multiple UAV as the USG entities, assume that multiple UAV of ourselves are distributed disorderly in a certain unknown continuous dynamic limited combat airspace, multiple Unmanned Ground Vehicle(UGV) of the enemy are distributed disorderly on the ground mapped by such combat airspace, UAVs of ourselves are taking with visual reconnaissance equipment to realize the effective search and identification on ground targets, UGVs of the enemy are taking with counterreconnaissance equipment [18] to effectively elude the search and identification of the UAVs of ourselves. UAVs of ourselves and UGVs of the enemy have opposite tactical purpose, and significant confrontation attributes, UAVs of ourselves are expected to search and identify the UGVs of the enemy within the shortest time and realize stable tracking, and UGVs of the enemy are expected to effectively elude from search and identification or maximize the delay of the time being searched and identified. In order to increase problem research difficulty and focus on problem essence, consider the combat scenario in which six UAVs of ourselves make collaborative search on two UGVs of the enemy, UAVs of ourselves and UGVs of the enemy have different maneuvering characteristics, and UAVs of ourselves are superior in numbers, UGVs of the enemy have faster maneuvering speed. In the above-mentioned combat scenario, UAVs of ourselves and UGVs of the enemy would conduct tactical confrontation, and six UAVs of ourselves need to make interactive collaboration internally, it belongs to the typical competition scenario on the confrontation level between ourselves and the enemy, and the typical cooperation scenario on the level of interactive collaboration [19], the ACSLE problem of USG turn into the autonomous collaborative control problem of multiple UAV under the co-existing scenario of cooperation and competition.

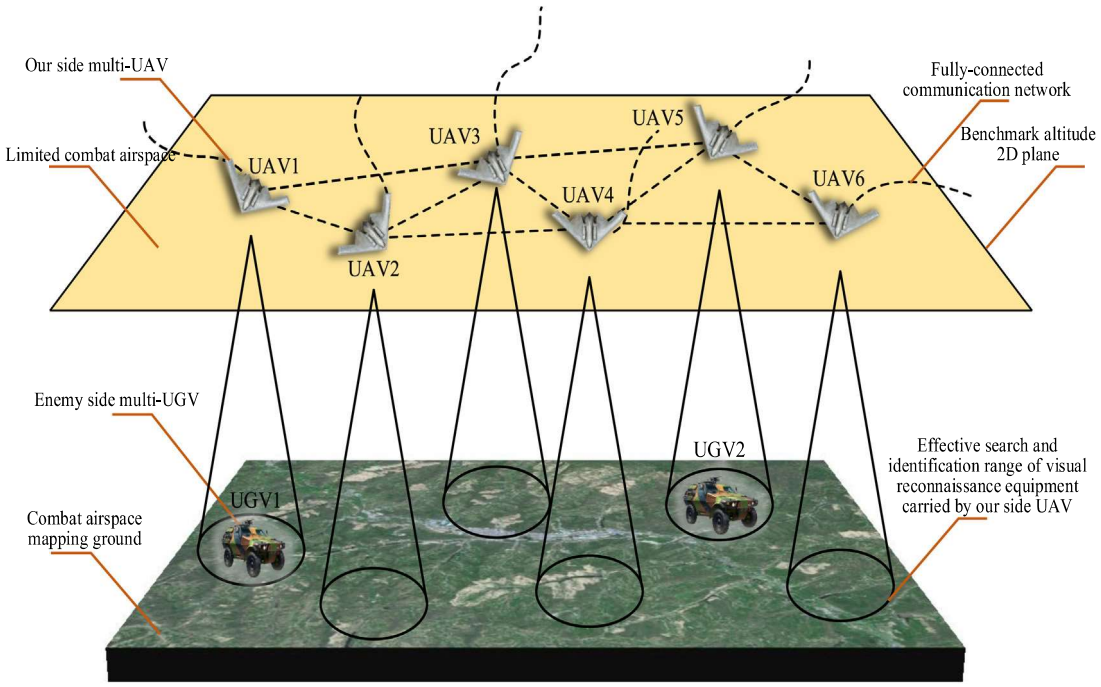


Fig. 1. Description schematic diagram for ACSLE problem of USG.

In the paper, assume the search and identification airspace of UAVs of ourselves and the elusion ground of UGVs of the enemy are of the limited state, six UAVs of ourselves execute search mission on the same flight plane, and neglect the surface relief factors of UGVs of the enemy, USG adopts the fully connected communication network internally, and USG members are all homogeneous and isomorphic individuals, and have completely consistent kinematic performance parameters, therefore the ACSLE problem of USG description schematic diagram is shown in Fig. 1.

Select the fixed-wing UAV as the kinematic modeling object, assume all search UAVs of ourselves execute search mission on the 2D plane with benchmark altitude, assume UAVs is a rigid body of constant quality and with equally distributed mass, neglect the impact of earth curvature, the gravitational acceleration of UAVs in flight process would not be affected by flight height, neglect the flight attitude description of UAVs, and consider only the maneuvering description. Hypothesize the position coordinates of the search UAVs of ourselves in the search airspace at the current moment is $\left[x_S^n, y_S^n\right]^{\mathrm{T}} (n = 1, 2, \ldots, 6)$, so the nonlinear kinematic model description of the search UAVs of ourselves is Formula (1).

$$\begin{cases} \dot{x}_S^n = v_S^n \cos \psi_S^n \\ \dot{y}_S^n = v_S^n \sin \psi_S^n \\ \dot{\psi}_S^n = \omega_S^n \end{cases} \quad (n = 1, 2, \ldots, 6) \tag{1}$$

Therein, v_S^n represents the velocity magnitude of the searching UAVs of ourselves, it is a constant and remains unchanged within the whole combat cycle; ω_S^n represents the angular velocity magnitude of the search UAVs of ourselves; $\omega_S^n \in \left[-Max\omega_S^n, Max\omega_S^n\right]$ represents the motion control variable constraint of the search UAVs of ourselves; $Max\omega_S^n$ represents the maximum angular velocity of the search UAVs of ourselves. All search UAVs of ourselves are homogeneous and isomorphic, define the simulation time step of a decision cycle as ΔT, define the maximum change value of course angle as $\Delta\psi_{max}$, define the turning radius as r, define the minimum turning radius as r_{min}, define the maximum lateral overload as n_{max}, so $Max\omega_S^n$ can be determined by Formulas (2) and (3).

$$\begin{cases} Max\omega_S^n \Delta T = \Delta\psi_{max} \\ n_{max} g r_{min} = \left(v_S^n\right)^2 (n = 1, 2, \ldots, 6) \\ 2r \sin \Delta\psi_{max} \approx v_S^n \Delta T \end{cases} \tag{2}$$

$$Max\omega_S^n \Delta T = \arcsin\left(\Delta T n_{max} g \Big/ \left(2v_S^n\right)\right) (n = 1, 2, \ldots, 6) \tag{3}$$

3. Autonomous confrontation strategy learning evolution mechanism

3.1. Autonomous confrontation strategy generate model

In 2017, Ryan Lowe et al. of Open AI, on the basis of DDPG algorithm, integrated organically with Actor–Critic thought and proposed the MADDPG algorithm to effectively solve the cutthroat competition of resources among USG members, it provides new thought for solving autonomous collaborative control problem of USG under continuous dynamic environment. MADDPG algorithm, through the cooperation and competition relationship award function structure that is designed rationally, allows to co-exist the cooperation and competition relationship simultaneously among multi-agents; by estimating all agent strategies, utilizing sufficiently global information when making centralized training, and relying only on local information when making decentralized execution, the system environment instability problem of multi-agent can be well soothed. In the MADDPG algorithm, each agent follows the DDPG algorithm architecture, and the strategy parameter set of multi-agent system is defined as $\theta = \left(\theta_1, \theta_2, \ldots, \theta_N\right)$, so the joint strategy space of multi-agent system is determined by Formula (4). MADDPG algorithm solves the problem of environmental instability of multi-agent system and the problem [20] of outdated experience in the Experience Replay (ER) mechanism through its architecture of "centralized training and decentralized execution", see Fig. 2 for details. where the motivation that the architecture of "centralized training and decentralized execution" can solve the problem of environmental instability of multi-agent system is determined by the Formula (5).

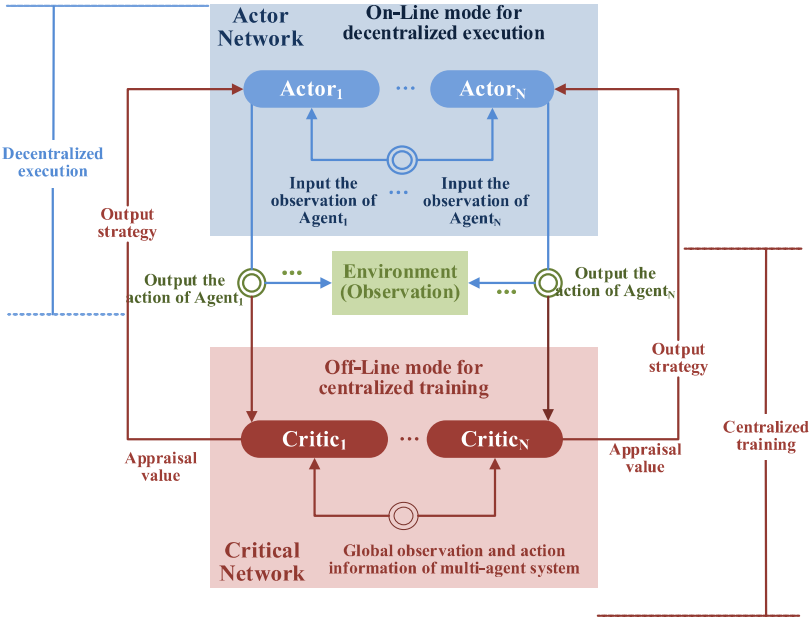


Fig. 2. Schematic mode diagram for centralized training and decentralized execution in MADDPG algorithm.

When $\pi_i \neq \pi_i'$, and if all multi-agent policies are already known, even though the policies of agents change, the environmental instability thereof can still be deemed as "unaffected", so the Formula (5) at the time is still true.

$$H : \pi = \left(\pi\left(\theta_1\right), \pi\left(\theta_2\right), \ldots, \pi\left(\theta_N\right)\right) \tag{4}$$

$$\begin{cases} P\left(s'|s, a_1, \ldots, a_N, \pi_1, \ldots, \pi_N\right) = P\left(s'|s, a_1, \ldots, a_N\right) \\ P\left(s'|s, a_1, \ldots, a_N, \pi_1', \ldots, \pi_N'\right) = P\left(s'|s, a_1, \ldots, a_N\right) \end{cases} \quad \pi_i = \pi_i' or \pi_i \neq \pi_i' \tag{5}$$

In order to maximize whole reward and member reward of USG in the time parallel, improve the convergence efficiency of MADDPG algorithm, integrate the thought of parallel decoupling (PDRM-MADDPG) on the basis of MADDPG algorithm, use the mutually independent centralized Parallel Benchmark Critic network (PB-Critic) and personalized Parallel Decoupling Critic network (PD-Critic) to simultaneously maximize the whole reward and the member reward of USG in the form of parallel decoupling, the run logic diagram of the PDRM-MADDPG algorithm is shown in Fig. 3. For the centralized Parallel Benchmark Critic network (PB-Critic), define the joint state of continuous dynamic environment as $s = \left(o_1, o_2, \ldots, o_N\right)$, and use the joint action of all agents to estimate the Q value of state–action value of the agent i under the action of joint state s, define its expression as $Q_\psi^g\left(s, a_1, a_2, \ldots, a_N\right)$, and then express the expectation symbol in E, the experience pool in D, and define the joint policy of all agents as ψ, determine the whole policy gradient formula of the agent i by the formula (6); the target policy defined by the network parameter $\theta' = \left\{\theta_1', \theta_2', \ldots, \theta_N'\right\}$ is $\pi' = \left\{\pi_1', \pi_2', \ldots, \pi_N'\right\}$, so the loss function corresponded by the centralized Parallel Benchmark Critic network (PB-Critic) can be determined by the Formula (7).

$$\nabla J\left(\theta_i\right)_{PB} = E_{s,a\sim D}\left[\nabla_{\theta_i}\pi_i\left(a_i|o_i\right)\nabla_{a_i}Q_\psi^g\left(s, a_1, a_2, \ldots, a_N\right)|_{a_i=\pi_i(o_i)}\right] \tag{6}$$

$$\begin{cases} L\left(\psi\right) = E_{s,a,r,s'}\left[\left(Q_\psi^g\left(s, a_1, a_2, \ldots, a_N\right) - y_z\right)^2\right] \\ y_z = r_z + \gamma Q_{\psi'}^g\left(s', a_1', a_2', \ldots, a_N'\right)|_{a_i'=\pi_i'(o_i')} \end{cases} \tag{7}$$

Personalized Parallel Decoupling Critic network (PD-Critic) aims at maximizing member reward, so define the private observation in the continuous dynamic environment as o_i, then use the private action of agent to estimate the Q value of state–action value of the agent i in the action of private observation o_i [21], and define its expression as $Q_i^\pi\left(o_i, a_i\right)$, express the expectation symbol in E, define the private policy of the agent i as ψ_i, so the member policy gradient formula of the agent i is determined by the Formula (8), the loss function corresponded by the personalized Parallel Decoupling Critic network (PD-Critic) is determined by the Formula (9).

$$\nabla J\left(\theta_i\right)_{PD} = E_{s_i,a_i\sim D}\left[\nabla_{\theta_i}\pi_i\left(a_i|o_i\right)\nabla_{a_i}Q_{\psi_i}^\pi\left(o_i, a_i\right)|_{a_i=\pi_i(o_i)}\right] \tag{8}$$

$$\begin{cases} L\left(\psi_i\right) = E_{o,a,r,o'}\left[\left(Q_{\psi_i}^\pi\left(o_i, a_i\right) - y_u\right)^2\right] \\ y_u = r_u^i + \gamma Q_{\psi_i'}^{\pi'}\left(o_i', a_i'\right)|_{a_i'=\pi_i'(o_i')} \end{cases} \tag{9}$$

Under the action of PDRM-MADDPG algorithm, the value Q of state–action value of the agent i includes two parts: whole parallel decoupling and member parallel decoupling. In order to maximize the whole reward and the member reward of USG simultaneously, the policy gradient formula of the agent i is the linear superposition sum of the Formula (6) and (8), specially determined by the Formula (10). When conducting the model training, set the weight of whole reward and member reward rationally according to the scenario to be faced. For the pure cooperative scenario, the common view is that both weights of whole reward and member reward are consistent, at the moment $\alpha_1 = \alpha_2 = 1$ is selected.

$$\begin{cases} \nabla J\left(\theta_i\right) = \alpha_1\nabla J\left(\theta_i\right)_{PB} + \alpha_2\nabla J\left(\theta_i\right)_{PD} \\ \alpha_1 \in [0, 1], \alpha_2 \in [0, 1] \end{cases} \tag{10}$$

3.2. Continuous interaction mechanism of actual combat commander in the loop

In order to further improve the actual combat application credibility of USG autonomous confrontation strategy, strengthen the guiding position of actual combat commander in the loop in the autonomous confrontation strategy learning evolution process of USG, and guide the experience learning process of USG to coupling with the combat intention of actual combat commander in the loop, the Continuous Interaction (CI) mechanism is thence introduced. Under the action

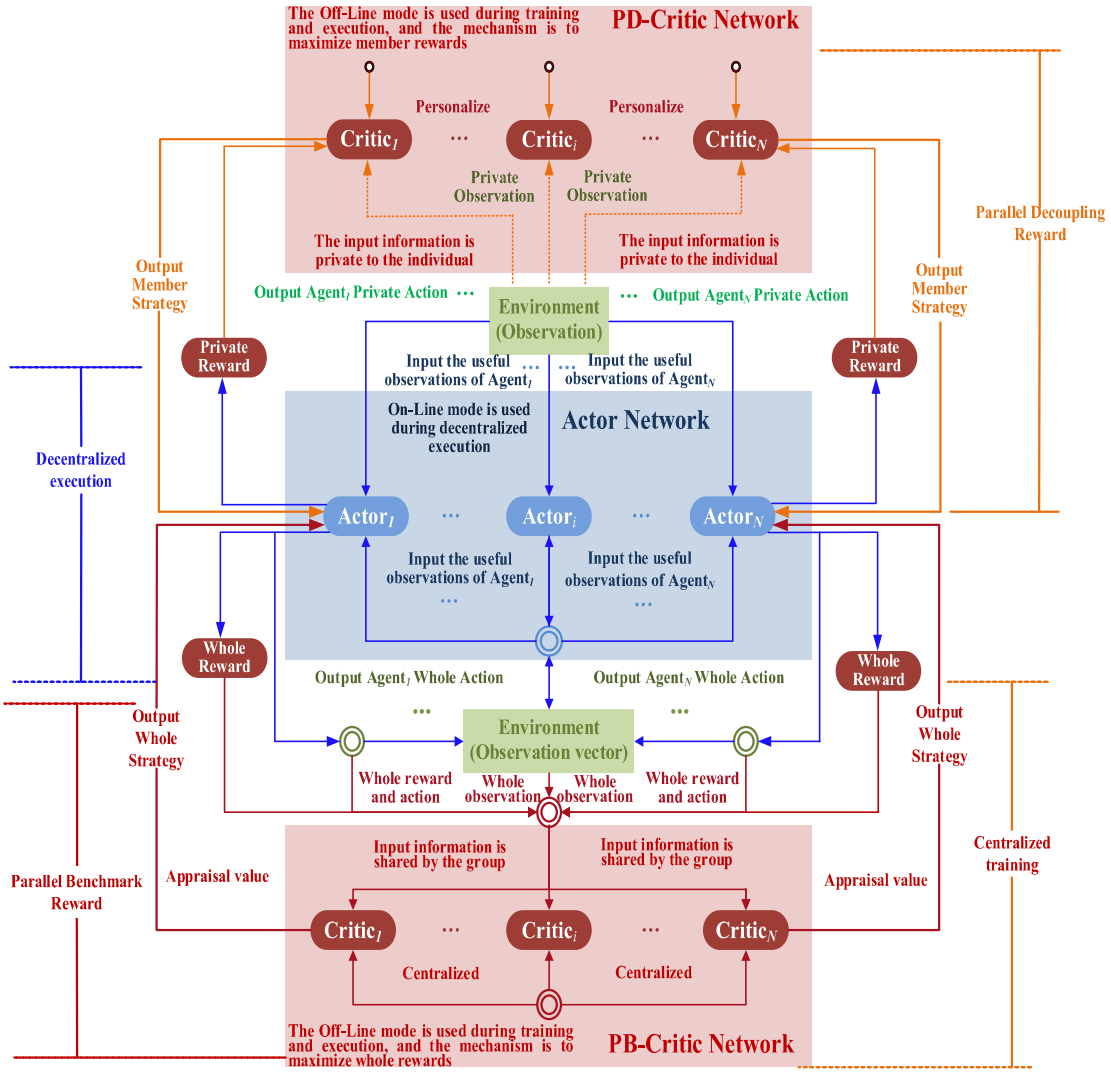


Fig. 3. The run logic diagram of the PDRM-MADDPG algorithm.

of continuous interaction (CI) mechanism, after generating the initial autonomous confrontation strategy, USG makes proactive initiation on the continuous interaction (CI) with actual combat commander in the loop, and then uploads the perception information of recessive battlefield situation, whereas commander makes proofreading supplement for the information of battlefield situation, and transits them and combat intention back to USG. USG continues to fill in self-experience gap through the continuous interaction (CI) with the actual combat commander in the loop, and provides experience driving for USG autonomous confrontation strategy active learning evolution. The continuous interaction (CI) mechanism of actual combat commander in the loop runs through the entire combat cycle of the USG. Firstly, accelerate self-experience learning process by coupling the combat intention of actual combat commander in the loop; Secondly, conduct characteristics identification on the historical data set of continuous interaction (CI) by using expected value function, extract the battlefield situation information interested by actual combat commander in the loop; Finally, update the continuous interaction (CI) strategy by predicating the battlefield situation information interested by the actual combat commander in the loop in future. The continuous interaction (CI) mechanism of actual combat commander in the loop needs to balance USG interactive learning strategy update and learning interaction strategy update, the interactive learning strategy update aims to guide continuous interaction (CI) between USG and actual combat commander in the loop, expedite the experience learning process of USG; the learning interaction strategy update aims at leading USG to continuously optimize the continuous interaction (CI) strategy of actual combat commander in the loop, and maps the combat intention of actual combat commander in the loop to USG autonomous confrontation strategy in the end-to-end form. The logic diagram of continuous interaction (CI) mechanism of actual combat commander in the loop is shown by Fig. 4, and the pseudo code description is shown in Algorithm 1.

Assume the USG is placed in the high-dynamic actual combat scenario $\xi = (S_i, A_i, P_i, P_{i0})$, the Intrinsic Motivation (IM) of USG is to perceive recessive battlefield situation as complete as possible, continuously fill in self-experience gap through the continuous interaction (CI) with actual combat commander in the loop, and continuously optimize the continuous interaction (CI) strategy of actual combat commander in the loop. In ξ, S_i represents the instantaneous battlefield situation, A_i represents the instantaneous interaction space, $P_i : S_i \times A_i \to S_i$ represents the instantaneous transition probability of battlefield situation, P_{i0} represents the probability measure of initial situation distribution. Since the state space of USG is continuous and the action space is dispersed [22], USG autonomous decision process can be formalized into the Infinite Horizon-Markov Decision Process (IH-MDP), and it is

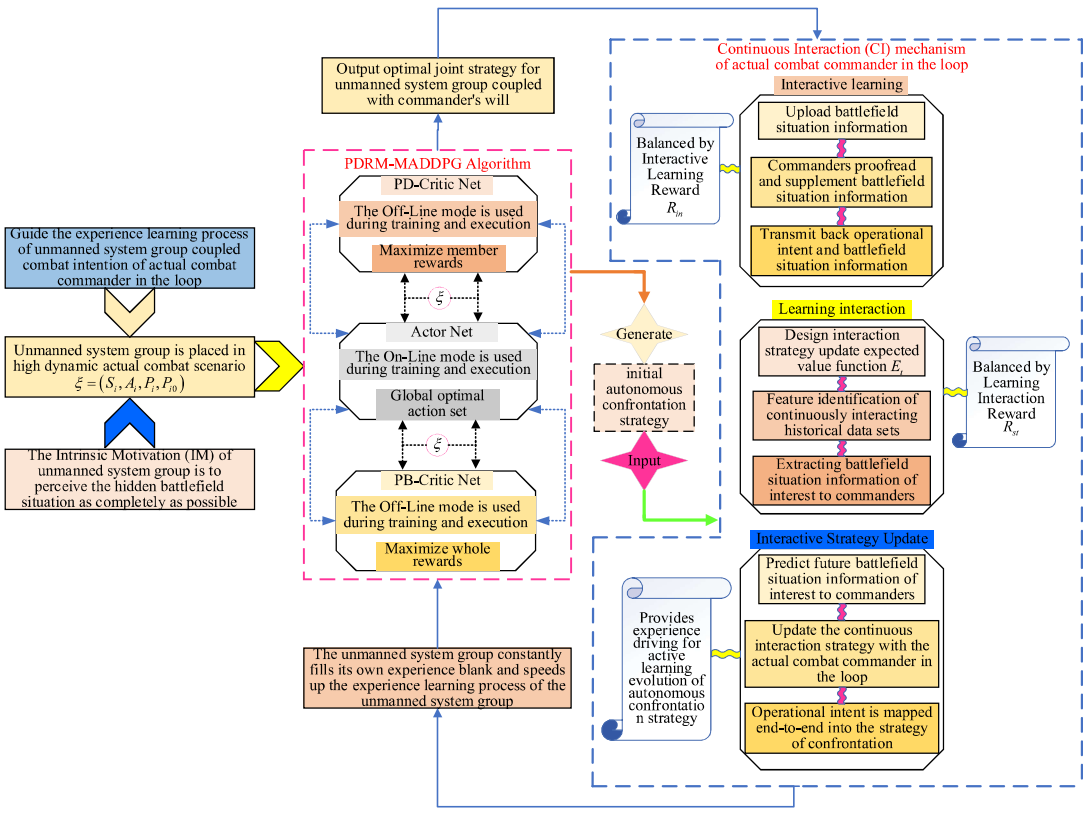


Fig. 4. Logic diagram of continuous interaction (CI) mechanism of actual combat commander in the loop.

determined by Formula (11).

$$M_{IH}=\left(\overline{S}_i,A_i,\overline{P}_i,\overline{P}_{i0},R,\mu_{CI}\right) \tag{11}$$

Define USG historical interaction data set as D_{CI}, μ_{CI} represents the interaction strategy, R represents the reward function, $\overline{S}_i=S_i\times D_{CI}\times\mu_{CI}$ represents the transition state of battlefield situation. For the instantaneous t moment in the interaction process, The historical interaction datasets D_{CI}^{t} and D_{CI}^{t+1} of the USG are determined by Formula (12) and (13), the transition state of the battlefield situation is determined by Formula (14), and μ_{CI}^{t+1} is determined by Formula (15).

$$D_{CI}^{t}=\left\{s_0,a_0,r_0,\ldots,s_{t-1},a_{t-1},r_{t-1},s_t\right\}\in D_{CI} \tag{12}$$

$$D_{CI}^{t+1}=D_{CI}^{t}\cup\left\{a_t,s_{t+1}\right\} \tag{13}$$

$$\overline{s}_t=\left(s_t,D_{CI}^{t},\mu_{CI}^{t}\right)\in\overline{S}_i \tag{14}$$

$$\mu_{CI}^{t+1}=E_t\left(D_{CI}^{t+1}\right) \tag{15}$$

In Formula (15), E_t represents the expected value of training, updating continuously the continuous interaction (CI) strategy with actual combat commander in the loop. In order to lead USG to interact the battlefield situation information [23] that interested by actual combat commander in the loop as far as possible, the continuous interaction (CI) reward R_{CI} is designed, R_{CI} is composed of interactive learning reward R_{in} and learning interaction reward R_{st} with linear coupling relationship, R_{in} aims to guide USG to proactive initiation the continuous interaction (CI) with actual combat commander in the loop, R_{st} aims to guide USG to update the interaction strategy with actual combat commander in the loop, R_{CI} is specifically determined by Formula (16).

$$\begin{cases} R:\overline{S}_i\rightarrow R_{CI}=\sigma\cdot R_{in}+(1-\sigma)\cdot R_{st} \\ 0\le\sigma\le 1 \end{cases} \tag{16}$$

In Formula (16), interactive learning reward R_{in} belongs to variable dominant reward signal, USG can acquire initial interactive learning reward without experiencing learning and training, and automatically release interactive learning reward signal after initially generating autonomous confrontation strategy based on PDRM-MADDPG algorithm; such signal guides USG to proactive initiation the continuous interaction (CI) with actual combat commander in the loop, and interactive learning reward R_{in} gets dynamic change as interactive learning process deepens. Define the evaluation expectation value function of USG historical interaction dataset at the instantaneous moment t as E_v^t, define the initial interactive learning reward as B, define the random adjustment coefficient as ρ_{in}, so the specific interactive learning reward R_{in} is determined by formula (17); the learning interaction reward R_{st} belongs to variable recessive reward signal, USG needs to firstly conduct interactive strategy learning [24], and then generate the corresponding learning interaction reward according to interactive strategy update results. Define the evaluation expectation value function of interactive strategy at the instantaneous t moment as E_e^t, define the random adjustment coefficient as ρ_{st}, so the specific learning interaction reward R_{st} is determined by Formula (18).

$$R_{in}=\rho_{in}\cdot E_v^t\left(D_{CI}^{t}\right)+B \tag{17}$$

$$R_{st}=\rho_{st}\cdot E_e^t\left(\mu_{CI}\right) \tag{18}$$

In order to maintain the dynamic balance of interactive learning reward R_{in} and learning interaction reward R_{st} under time parallel, the continuous interaction (CI) mechanism balancing strategy π_B is designed, and the process that USG is mapped to the instantaneous interaction space from tradition state of battlefield situation is interpreted in the way of dynamic balance [25], so the specific π_B is determined by Formula (19). Define the expectation value function of continuous interaction (CI) reward as E_π, the optimal continuous interaction (CI) mechanism balancing strategy π_B^* corresponds to the maximized continuous interaction (CI) reward, so the specific π_B^* is determined by Formula (20).

$$\pi_B:\overline{S}_i=S_i\times D_{CI}\times\mu_{CI}\rightarrow A_i \tag{19}$$

Algorithm 1: Continuous Interaction (CI) Mechanism

Input: Initial autonomous confrontation strategy of USG(Based on PDRM-MADDPG algorithm)

1. Initialize high-dynamic actual combat scenario $\xi = \left(S_i, A_i, P_i, P_{i0}\right)$

2. Initialize initial interactive learning reward B

3. Initialize random adjustment coefficient ρ_{in} and ρ_{st}

4. Initialize reward function coefficient σ

5.For agent *i=1* to *N* do

6. Perceive of recessive battlefield situation as completely as possible under the action of Intrinsic Motivation (IM)

7. Proactive initiation the Continuous Interaction (CI) with actual combat commander in the loop

8. Generate historical interaction data set D_{CI}^{t}

9. Update $R_{in} = \rho_{in} \cdot E_v^t\left(D_{CI}^t\right) + B$

10. Update $D_{CI}^{t+1} = D_{CI}^t \cup \left\{a_t, s_{t+1}\right\}$

11.For episode=1 to MaxEpisode do

12. Update interaction strategy $\mu_{CI}^{t+1} = E_t\left(D_{CI}^{t+1}\right)$

13. Update $R_{st} = \rho_{st} \cdot E_e^t\left(\mu_{CI}\right)$

14. Calculate $R : \overline{S}_i \rightarrow R_{CI} = \sigma \cdot R_{in} + \left(1-\sigma\right) \cdot R_{st}$

15. Calculate $\pi_B^* = \arg\max_\pi E_\pi\left[\sum_t R_{CI}\left(\overline{s}_t\right)\right]$

16.End For

17.End For

Output: Autonomous confrontation strategy of USG coupled with volition of commander(Based on Continuous Interaction (CI) Mechanism)

$$\pi_B^* = \arg\max_\pi E_\pi\left[\sum_t R_{CI}\left(\overline{s}_t\right)\right] \tag{20}$$

3.3. Autonomous confrontation strategy active learning evolution mechanism

In order to guide the commander volition-coupled USG autonomous confrontation strategy active learning evolution generated through continuous interaction (CI) mechanism, utilize sufficiently the volition information of actual combat commander in the loop released in the process of continuous interaction (CI) to update the Replay Experience Buffer Pool (REBP) of USG, and map the combat intention of actual combat commander in the loop to USG autonomous confrontation strategy learning evolution process in the end-to-end form, the Active Learning Evolution (ALE) mechanism of confrontation strategy is introduced. Under the action of active learning evolution (ALE) mechanism of confrontation strategy, the USG autonomous confrontation strategy coupled with the volition of commander triggers the targeted active learning evolution (ALE). On the one hand, it guides USG behavioral autonomy capability to always adapt to the strong-gaming, high mobility, and high-intensity complicated actual combat battlefield confrontation environment; on the other hand, it guides USG optimal joint strategy to constantly match with the USG optimal joint strategy of confrontation party to realize the continuous benign evolution of autonomous confrontation strategy learning under the driving of combat intention. The active learning evolution (ALE) mechanism of confrontation strategy runs through the entire combat cycle of USG. Firstly, it utilizes sufficiently the volition information of actual combat commander in the loop to update USG Replay Experience Buffer Pool to form the Volition Share Buffer Pool (VSBP) of actual combat commander in the loop oriented to USG, serving as the Indirect Communication Channel (ICC) of USG interior communication; afterwards, in order to guide USG behavioral autonomy capability to constantly adapt to the strong-gaming, high-maneuvering, and high-intensity complicated actual combat battlefield confrontation environment, the environmental adaption evolution mechanism is introduced; in order to guide USG optimal joint strategy to constantly match with the USG optimal joint strategy of confrontation party, the strategy matching evolution mechanism is introduced; finally, in order to maintain the dynamic balance between environmental adaption evolution mechanism and strategy matching evolution mechanism under time parallel, the active learning evolution (ALE) mechanism balancing strategy is designed, the process

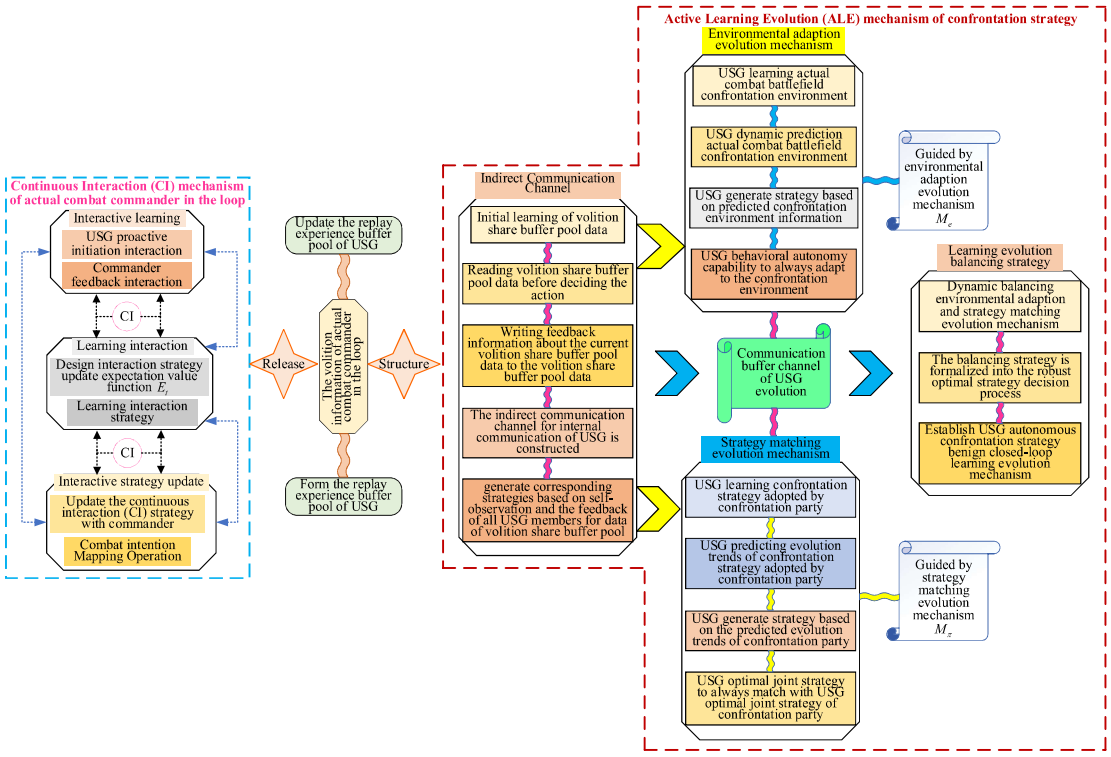


Fig. 5. Logic diagram of active learning evolution (ALE) mechanism of confrontation strategy.

that USG autonomous confrontation strategy learning coupled with the volition of commander evolves into the USG autonomous confrontation strategy coupled with benign closed-loop learning evolution attributes is restored in the way of dynamic balance, and the solving process of balancing strategy is formalized into the robust optimal strategy decision process. The logic diagram of active learning evolution (ALE) mechanism of confrontation strategy is shown by Fig. 5, and the pseudo code description is shown in Algorithm 2.

In order to further strengthen the stability and collaboration of USG autonomous confrontation strategy active learning evolution (ALE) process, expedite USG experience learning process, improve USG learning and training convergence efficiency and quality, the Indirect Communication Channel (ICC) concept oriented to interior communication of USG is introduced, USG members utilize volition share buffer pool as the indirect communication channel (ICC). Before making decision on execution action, USG members firstly conduct traversal read on data of volition share buffer pool, and then write the feedback information of current data in the volition share buffer pool, finally generate corresponding strategies based on self-observation and the feedback of all USG members for data of volition share buffer pool, and simultaneously updates volition share buffer pool. Define the volition share buffer pool of actual combat commander in the loop oriented to USG as G, the instantaneous autonomous confrontation strategy of USG member i is defined as μ_{θ_i}, so the autonomous confrontation strategy of USG member i is determined by Formula (21).

$$\mu_{\theta_i} : O_i \times G \mapsto A_i \tag{21}$$

After acquiring the instantaneous private observation o_i, USG member i map it into instantaneous expectation state s_i^e through coding function $\mathbb{Z}$, where the coding function $\mathbb{Z}$ is the deep fully connected neural network [26], the network parameter is θ_i^e, the instantaneous expectation state s_i^e is the determining factor of USG member i for deciding execution action, and simultaneously it plays important guiding effect in the process that USG members i make reading and feedback on data of volition share buffer pool, and specifically it is determined by Formula (22).

$$s_i^e = \mathbb{Z}_{\theta_i^e}\left(o_i\right), \quad s_i^e \in S_i^E \tag{22}$$

After coding the instantaneous private observation o_i into the instantaneous expectation state s_i^e, USG members i start executing the operation of reading data of volition share buffer pool, and USG members i are allowed to conduct Memory Learning (ML) on the data of volition share buffer pool in the process of reading the data of volition share buffer pool. Under the action of memory learning (ML) mechanism, USG members i can not only learn the information of volition of actual combat commander in the loop, but also access to the experience information learned by other USG members, define the memory learning (ML) vector of USG member i as V_i^{ML}, the input of V_i^{ML} is instantaneous expectation state s_i^e, the output of V_i^{ML} is memory learning (ML) information [27], and the learnable weight is W_i^{ML}, so the specific V_i^{ML} is determined by Formula (23). Define the instantaneous data of volition share buffer pool as g, express the activation function in $sf(\cdot)$; define the reading necessity weight as W_i^{Re}, so the reading weight of data of volition share buffer pool after activation ϑ_i is determined by Formula (24).

$$V_i^{ML} = W_i^{ML} s_i^e \tag{23}$$

$$\vartheta_i = sf\left(W_i^{Re}\left[s_i^e, V_i^{ML}, g\right]\right), \quad \vartheta_i \in [0,1] \tag{24}$$

In Formula (24), $\left[s_i^e, V_i^{ML}, g\right]$ represents the performing tandem operations for three vectors; ϑ_i represents the necessity degree of USG member i memory learning (ML) on instantaneous data g of volition share buffer pool. memory learning (ML) allows each member of USG to make autonomous learning adjustment W_i^{ML} and W_i^{Re}, signifying that all USG members can read and interpret the data in volition share buffer pool G in their own particular ways. Formalize the reading operation of USG member i into the deep fully connected neural network χ, the network parameter is θ_i^χ, so the reading operation Re_i of USG member i is determined by Formula (25).

$$\begin{cases} \theta_i^\chi = \left\{W_i^{ML}, W_i^{Re}\right\} \\ \mathrm{Re}_i = \chi_{\theta_i^\chi}\left(s_i^e, g\right) \sim \chi_{\theta_i^\chi}\left(o_i, g\right) \end{cases} \tag{25}$$

USG members autonomously decides the experience information to be shared by reading and interpreting the data information in volition

share buffer pool, write then in the volition share buffer pool in the form of data feedback, and thereby realize the indirect communication among USG members. USG member i read and interpret the data information in the volition share buffer pool, and then generate the initial feedback information set f_i, f_i is generated [28] on the basis of the own private observation of USG member i and the feedback information of other USG members for current volition share buffer pool. Define the write necessity weight as W_i^{Wr}, express the activation function in $sf(\cdot)$, so the initial feedback information set f_i is determined by Formula (26), define the open sharing gating function as κ_o, the corresponding weight is $W_i^{\kappa_o}$, the prohibit sharing gating function is κ_p, the corresponding weight is $W_i^{\kappa_p}$, so the expected feedback information set f_i^E is determined by Formula (27).

$$f_i = sf\left(W_i^{Wr}\left[s_i^e, g\right]\right) \sim sf\left(W_i^{Wr}\left[o_i, g\right]\right) \tag{26}$$

$$f_i^E = \kappa_o \odot f_i + \kappa_p \odot f \tag{27}$$

In Formula (27), $\odot$ represents the Hadamard product, USG members utilizes the gating effect of Hadamard product to autonomously decide the experience information [29] to be shared. Formalize the writing operation of USG member i into the deep fully connected neural network η, the network parameter is θ_i^{η}, so the writing operation Wr_i of USG member i is determined by Formula (28).

$$\begin{cases} \theta_i^{\eta} = \left\{W_i^{Wr}, W_i^{\kappa_o}, W_i^{\kappa_p}\right\} \\ \mathrm{Wr}_i = \eta_{\theta_i^{\eta}}\left(s_i^e, g\right) \sim \eta_{\theta_i^{\eta}}\left(o_i, g\right) \end{cases} \tag{28}$$

In order to guide USG behavioral autonomy capability to always adapt to the strong-gaming, high mobility, and high-intensity complicated actual combat battlefield confrontation environment, the environmental adaption evolution mechanism M_{env} is introduced; the environmental adaption evolution mechanism M_{env} is the deep fully connected neural network ϖ, the network parameter is θ_i^{ϖ}, define o_i^{env} to represent the observational components of USG member i for environment [30], express the expectation symbol in E, so the environmental adaption evolution mechanism M_{env} is determined by Formula (29). Under the action of environmental adaption evolution mechanism M_{env}, the execution action a_i^{env} of USG member i is jointly decided by environment observational components o_i^{env}, volition share buffer pool reading feedback Re_i, volition share buffer pool writing feedback Wr_i, it is specifically determined by Formula (30).

$$M_{env} = E\left[\varpi_{\theta_i^{\varpi}}\left(o_i^{env}\right)\right] \tag{29}$$

$$a_i^{env} = M_{env}\left[s_i^e, \mathrm{Re}_i, \mathrm{Wr}_i\right] \sim M_{env}\left[o_i^{env}, \mathrm{Re}_i, \mathrm{Wr}_i\right] \tag{30}$$

In order to guide USG optimal joint strategy to always match with the USG optimal joint strategy of confrontation party, the strategy matching evolution mechanism M_{pol} is introduced, the strategy matching evolution mechanism M_{pol} is the deep fully connected neural network ζ, the network parameter is θ_i^{ζ}, and definition o_i^{pol} represents the observational components [31] of USG member i against the strategy of the enemy, express the expectation symbol in E, so the strategy matching evolution mechanism M_{pol} is determined [32] by Formula (31). Under the action of strategy matching evolution mechanism M_{pol}, the execution action a_i^{pol} of USG member i is jointly decided by strategy observational components o_i^{pol}, volition share buffer pool reading feedback Re_i, volition share buffer pool writing feedback Wr_i, it is specifically determined by Formula (32).

$$M_{pol} = E\left[\zeta_{\theta_i^{\zeta}}\left(o_i^{pol}\right)\right] \tag{31}$$

$$a_i^{pol} = M_{pol}\left[s_i^e, \mathrm{Re}_i, \mathrm{Wr}_i\right] \sim M_{pol}\left[o_i^{pol}, \mathrm{Re}_i, \mathrm{Wr}_i\right] \tag{32}$$

In order to maintain the dynamic balance between environmental adaption evolution mechanism and strategy matching evolution mechanism under time parallel, the active learning evolution (ALE) mechanism balancing strategy P_{bal} is designed, the balancing strategy P_{bal} is an input–output asymmetric multilayer neural network ξ, the network parameter is θ_i^{ξ}, express the expectation symbol in E, it is specifically determined [33] by Formula (33). The input of balancing strategy P_{bal} is based on the action a_i^{env} generated by environmental adaption evolution and the action a_i^{pol} generated by strategy matching evolution, the output of balancing strategy P_{bal} is based on the optimal action a_i^* generated by balancing strategy, the optimal action a_i^* is specifically determined by Formula (34).

$$P_{bal} = E\left[\xi_{\theta_i^{\xi}}\left(a_i^{env}, a_i^{pol}\right)\right] \tag{33}$$

$$a_i^* = P_{bal}\left[\left(a_i^{env}, a_i^{pol}\right), \mathrm{Re}_i, \mathrm{Wr}_i\right] \tag{34}$$

4. Simulation verification experiment

4.1. Experimental setting

Simulation verification experiment codes are all written based on Python language, and construct a visual USG autonomous collaborative search dynamic confrontation game environment, integrated development environment adopts PyCharm Professional.2021.3 for Linux, and tool platform is based on Anaconda.3, deep learning training framework is based on Pytorch.1.11, visual tool adopts TensorBoard, model training is based on the rented Aliyun server, and server's instance configuration is GPU computation-type GN7-8 core processor, the memory is 64 GB, and software's host operating system Ubuntu Server 18.04 LTS 64 bits. The ACS-ACL algorithm has built-in centralized Parallel Benchmark Critic (PB-Critic), personalized Parallel Decoupling Critic (PD-Critic), Actor network, interactive strategy network, learning evolution strategy network and other neural networks,all neural networks aforesaid, on structure, belong to the fully connected four-layer neural network, the specific structure diagram is shown in Fig. 6.In which, the PB-Critic network contains two hidden layers, the number of hidden units in each layer is respectively 128 and 64, activation function adopts ReLU, and adopts Linear layer connectivity output layer [34], its complete structure is expressed in $27 \times 128 \times 64 \times 1$; the PD-Critic network contains two hidden layers, the number of hidden units in each layer is respectively 96 and 64, activation function adopts ReLU, and adopts Linear layer connectivity output layer, its complete structure is expressed in $19 \times 96 \times 64 \times 1$; every Actor network corresponds to independent PB-Critic network and PD-Critic network; Actor network contains two hidden layers, the number of hidden units in each layer is respectively 150 and 50, activation function adopts ReLU, and adopts activation function Tanh connectivity output layer, its complete structure is expressed in $15 \times 150 \times 50 \times 1$; interaction strategy network contains two hidden layers, the number of hidden units in each layer is respectively 64 and 32, activation function adopts Leaky ReLU, and adopts activation function Softplus connectivity output layer, its complete structure is expressed in $9 \times 64 \times 32 \times 1$; learning evolution strategy network contains two hidden layers, the number of hidden units in each layer is respectively 64 and 32, activation function adopts Leaky ReLU, and adopts activation function Tanh connectivity output layer, its complete structure is expressed in $15 \times 64 \times 32 \times 1$.

ACS-ACL algorithm involves in multiple hyper-parameters such as initial interactive learning reward, random adjustment coefficient, attenuation step size, attenuation base, experience pool capacity, batch learning sample number, reward function coefficient and so on in the process of learning and training, for the specific assignments in learning and training process is shown in Table 1. Construct a visual USG autonomous collaborative search dynamic confrontation gaming environment, contain 6 UAVs of ourselves and 2 UGVs of the enemy in the environment; all UAVs and UGVs are marked the two-dimensional trajectory in different colors, UAVs of ourselves execute the collaborative search mission on UGVs of the enemy, UAVs of ourselves are taking with visual reconnaissance equipment to realize the effective search and identification on ground targets, UGVs of the enemy can make

Algorithm 2: Active Learning Evolution(ALE) Mechanism

Input: Autonomous confrontation strategy of USG coupled with volition of commander(Based on Continuous Interaction (CI) Mechanism)

1. Initialize the volition share buffer pool G of actual combat commander in the loop

2. Initialize Memory Learning (ML) related weights W_i^{ML} 、 W_i^{Re}

3. Initialize initial feedback information set f_i

4. Initialize writing process related weights W_i^{Wr} 、 $W_i^{\kappa_o}$ 、 $W_i^{\kappa_p}$

5. Initialize the Memory Learning (ML) vector V_i^{ML} of USG member i

6.For agent *i=1* to *N* do

7. Generate autonomous confrontation strategy $\mu_{\theta_i}: O_i \times G \mapsto A_i$ of USG member i

8. USG member i acquires instantaneous private observation o_i

9. Calculate the instantaneous expectation state $s_i^e = \mathbb{Z}_{\theta_i^e}(o_i), \quad s_i^e \in S_i^E$

10. Calculate the reading weight $\vartheta_i = sf\left(W_i^{Re}\left[s_i^e, V_i^{ML}, g\right]\right), \quad \vartheta_i \in [0,1]$ of data of volition share buffer pool

11.For episode=1 to MaxEpisode do

12. Calculate the network parameter $\theta_i^{\chi} = \left\{W_i^{ML}, W_i^{Re}\right\}$ for reading operation

13. Calculate reading operation $\mathrm{Re}_i = \chi_{\theta_i^{\chi}}\left(s_i^e, g\right) \sim \chi_{\theta_i^{\chi}}\left(o_i, g\right)$

14. Calculate initial feedback information set $f_i = sf\left(W_i^{Wr}\left[s_i^e, g\right]\right) \sim sf\left(W_i^{Wr}\left[o_i, g\right]\right)$

15. Calculate expectation feedback information set $f_i^E = \kappa_o \odot f_i + \kappa_p \odot f$

16. Calculate the network parameter $\theta_i^{\eta} = \left\{W_i^{Wr}, W_i^{\kappa_o}, W_i^{\kappa_p}\right\}$ for writing operation

17. Calculate writing operation $\mathrm{Wr}_i = \eta_{\theta_i^{\eta}}\left(s_i^e, g\right) \sim \eta_{\theta_i^{\eta}}\left(o_i, g\right)$

18.End For

19.For agent *i=1* to *N* do

20. Calculate the environmental adaption evolution mechanism $M_{env} = E\left[\varpi_{\theta_i^{\varpi}}\left(o_i^{env}\right)\right]$

21. Calculate the action $a_i^{env} = M_{env}\left[s_i^e, \mathrm{Re}_i, \mathrm{Wr}_i\right] \sim M_{env}\left[o_i^{env}, \mathrm{Re}_i, \mathrm{Wr}_i\right]$ generated based on environmental adaption evolution mechanism

22. Calculate the strategy matching evolution mechanism $M_{pol} = E\left[\zeta_{\theta_i^{\zeta}}\left(o_i^{pol}\right)\right]$

23. Calculate the action $a_i^{pol} = M_{pol}\left[s_i^e, \mathrm{Re}_i, \mathrm{Wr}_i\right] \sim M_{pol}\left[o_i^{pol}, \mathrm{Re}_i, \mathrm{Wr}_i\right]$ generated based on strategy matching evolution mechanism

24. Calculate the balancing strategy $P_{bal} = E\left[\xi_{\theta_i^{\xi}}\left(a_i^{env}, a_i^{pol}\right)\right]$ of Active Learning Evolution (ALE) mechanism

25. Calculate the optimal action $a_i^* = P_{bal}\left[\left(a_i^{env}, a_i^{pol}\right), \mathrm{Re}_i, \mathrm{Wr}_i\right]$ generated based on balancing strategy

26.End For

27.End For

Output: Autonomous confrontation strategy of USG coupled with benign closed-loop learning evolution attribute(Based on Active Learning Evolution(ALE) Mechanism)

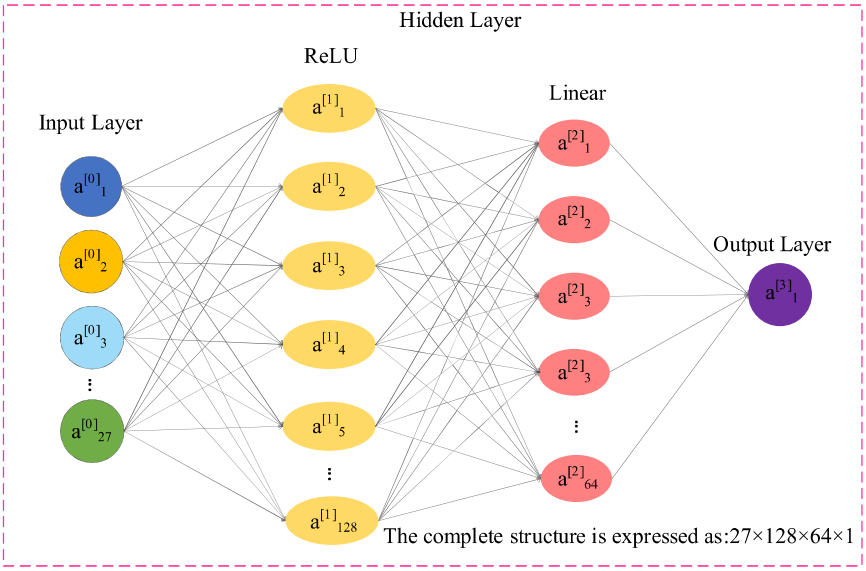


Fig. 6. Schematic diagram for neural network structure of ACS-ACL algorithm.

Table 1
Training hyper-parameter assignment table of ACS-ACL algorithm.

Hyper-parameters	Assignment	Hyper-parameters	Assignment
Initial interactive learning reward	2600	Reward function coefficient	0.3000
Random adjustment coefficient -in	0.1600	Random adjustment coefficient -st	0.2200
Attenuation step size	1×10^{-4}	Actor network learning rate	0.001
Attenuation base	60000	Threshold for training episode	6000
Experience pool capacity	36000	Time step threshold for episode	1800
Number of batch learning specimen	128	Simulation time step	0.100
Discount attenuation factor	0.980	Inertial update rate	0.010
Weight-adjusting factor	0.960	Critical network learning rate	0.0001

Table 2
Parameter table for autonomous collaborative search dynamic confrontation game environment.

Environmental parameters	Assignment	Environmental parameters	Assignment
Initial speed of UAVs of ourselves	8.6 m/s	Initial course of UAVs of ourselves	$[0,2\pi]$
Initial speed of UGVs of enemy	9.2 m/s	Initial course of UGVs of enemy	$[0,2\pi]$
Angular velocity threshold of UAVs	3 rad/s	Initial position of UAVs of ourselves	random
Angular velocity threshold of UGVs	1.8 rad/s	Initial position of UGVs of enemy	random
Mutual exclusivity distance of UAVs	32 m	Lock radius of UGVs of enemy	1.6 m
Mutual exclusivity distance of UGVs	12 m	Reconnaissance altitude of UAVs	800 m

omni-directional motion and are taking with counterreconnaissance equipment to effectively elude from the search and identification of UAVs of ourselves. For the visualization interface is shown in Fig. 7. UAVs of ourselves and UGVs of the enemy have opposite tactical objectives, with significant confrontation attributes, UAVs of ourselves are expected to search and identify the UGVs of the enemy within the shortest time and realize stable tracking, and UGVs of the enemy are expected to effectively elude from search and identification or maximize the delay of the time being searched and identified. Observe the distinctive combat intention, strategies, tactics, and methods of operation emerged in the process that UAV groups of ourselves search UGVs of the enemy, make comprehensive analysis the improvement effect on the autonomous confrontation strategy of convergence efficiency and execution quality under the driving of combat intention. In order to accelerate the convergence process of the ACS-ACL algorithm, definite the initial position and maneuvering characteristics of UAVs of ourselves and UGVs of the enemy, and set the dynamic confrontation game environment parameters, for the specific assignments in the process of learning and training is shown in Table 2.

4.2. Contrast experiment

In order to verify the learning and training effects of ACS-ACL algorithm on the basis of visual USG autonomous collaborative search dynamic confrontation game environment, two confrontation scenarios with increasing difficulties are designed. In every confrontation scenario, both ACS-ACL algorithm and PDRM-MADDPG algorithm are compared, then they are respectively used to control multiple UAV of ourselves to realize the effective collaborative search on UGVs of the enemy, finally multidimensional observations and comparisons are made for the training results of both algorithms. Aiming at the confrontation scenario 1, the autonomous confrontation strategy of UGVs of the enemy adopts Independent Learning Deep Deterministic Policy Gradient (IL-DDPG) algorithm, it is the adaptive application of DDPG algorithm in multi-agent field, every UGV of the enemy adopts respectively independent DDPG algorithm in the process of strategy learning and action execution, and the information vision of every UGV of the enemy is only limited to itself and searching UAV, UAVs of ourselves respectively and simultaneously adopt independent two

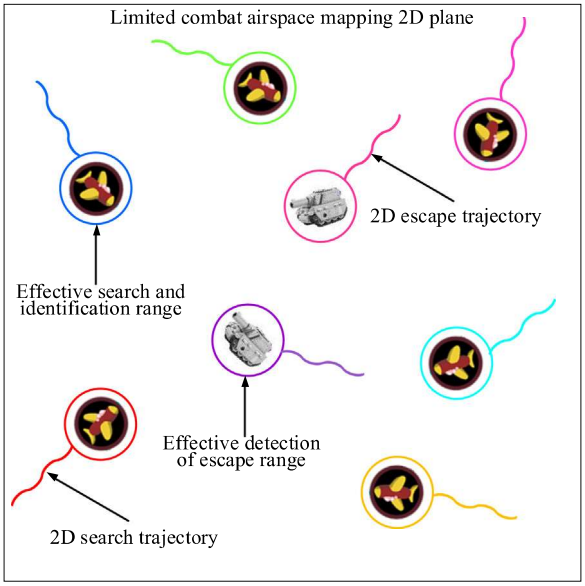


Fig. 7. Visualization interface of collaborative search dynamic confrontation game environment.

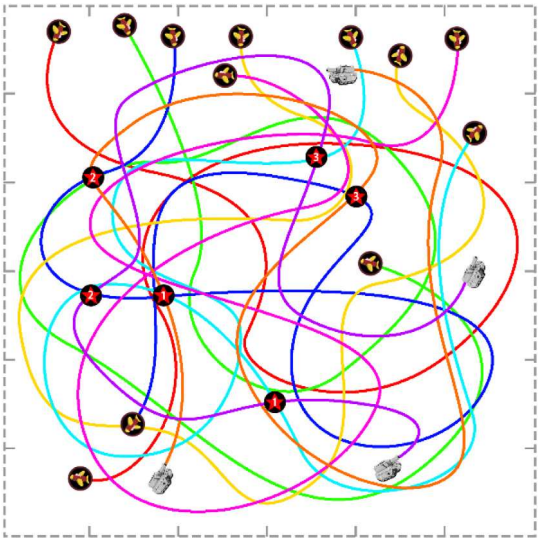


Fig. 8. Search trajectory diagram of UAVs of ourselves when adopting PDRM -MADDPG algorithm.

autonomous confrontation strategies such as PDRM -MADDPG, ACS-ACL, so the two-dimensional search trajectory diagram of confrontation parties and the average reward curves of UAVs of ourselves under the action of confrontation scenario 1 are respectively shown in Figs. 8–10.

Under the action of confrontation scenario 1, UAVs of ourselves locked UGVs of the enemy simultaneously three times at the following time of 15.06 s, 30.24 s, 45.35 s when adopting PDRM-MADDPG algorithm as the autonomous confrontation strategy, and at the following time of 8.34 s, 15.03 s, 35.51 s when adopting ACS-ACL algorithm as the autonomous confrontation strategy.

Aiming at the confrontation scenario 2, UAVs of ourselves respectively and simultaneously adopt independent two autonomous confrontation strategies such as PDRM-MADDPG algorithm and ACS-ACL algorithm, the autonomous confrontation strategy of UGVs of the enemy adopts MADDPG algorithm, so the two-dimensional search trajectory diagram of confrontation parties and the average reward curves of UAVs of ourselves under the action of confrontation scenario 2 are respectively shown in Figs. 11–13.

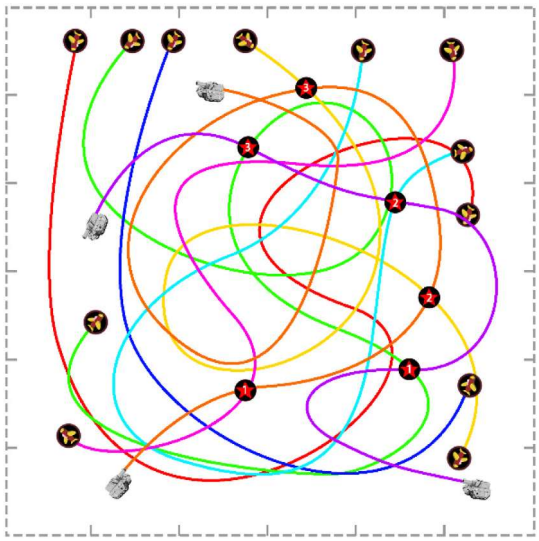


Fig. 9. Search trajectory diagram of UAVs of ourselves when adopting ACS-ACL algorithm.

Under the action of confrontation scenario 2, UAVs of ourselves locked UGVs of the enemy simultaneously three times at the following time of 18.94 s, 32.61s , 42.74 s when adopting PDRM-MADDPG algorithm as the autonomous confrontation strategy, and at the following time of 10.83 s, 22.92 s, 32.12 s when adopting ACS-ACL algorithm as the autonomous confrontation strategy.

4.3. Ablation experiment

In order to research the independent influences of continuous interaction (CI) mechanism, active learning evolution (ALE) mechanism on USG autonomous confrontation strategy learning evolution performance, ablation experiment is respectively carried out against continuous interaction (CI) mechanism, active learning evolution (ALE) mechanism. In order to evaluate the independent influences of continuous interaction (CI) mechanism on USG autonomous confrontation strategy learning evolution performance, continuous interaction (CI) mechanism is frozen, active learning evolution (ALE) mechanism is activated, and updates of USG Replay Experience Buffer Pool no longer relies on the volition information of actual combat commander in the loop any more, at the time the algorithm is recorded as FRE-CI. In order to guarantee the programming realization completeness, reserve the volition share buffer pool of actual combat commander in the loop, utilize periodicity to make sampling from experience buffer pool and then fill in the volition share buffer pool of actual combat commander in the loop with the data acquired therefrom, the indirect communication channel (ICC) among USG members exist in name only, the other elements of ablation experiment follow the flows of Algorithm 2 to execute, therefore the ablation experiment box plot under continuous interaction (CI) mechanism is shown in Fig. 14. In which, the horizontal axis represents the adopted confrontation strategy, the longitudinal axis represents the mean value reward with standard deviation, Fig. (a) shows the Confrontation Scenario 1, and Fig. (b) shows the Confrontation Scenario 2.

In order to evaluate the independent influences of active learning evolution (ALE) mechanism on USG autonomous confrontation strategy learning evolution performance, active learning evolution (ALE) mechanism is frozen, and continuous interaction (CI) mechanism is activated, at the time the algorithm is recorded as FRE-ALE. In order to guarantee the programming realization completeness, reserve still the volition share buffer pool of actual combat commander in the loop, utilize the volition information of actual combat commander in

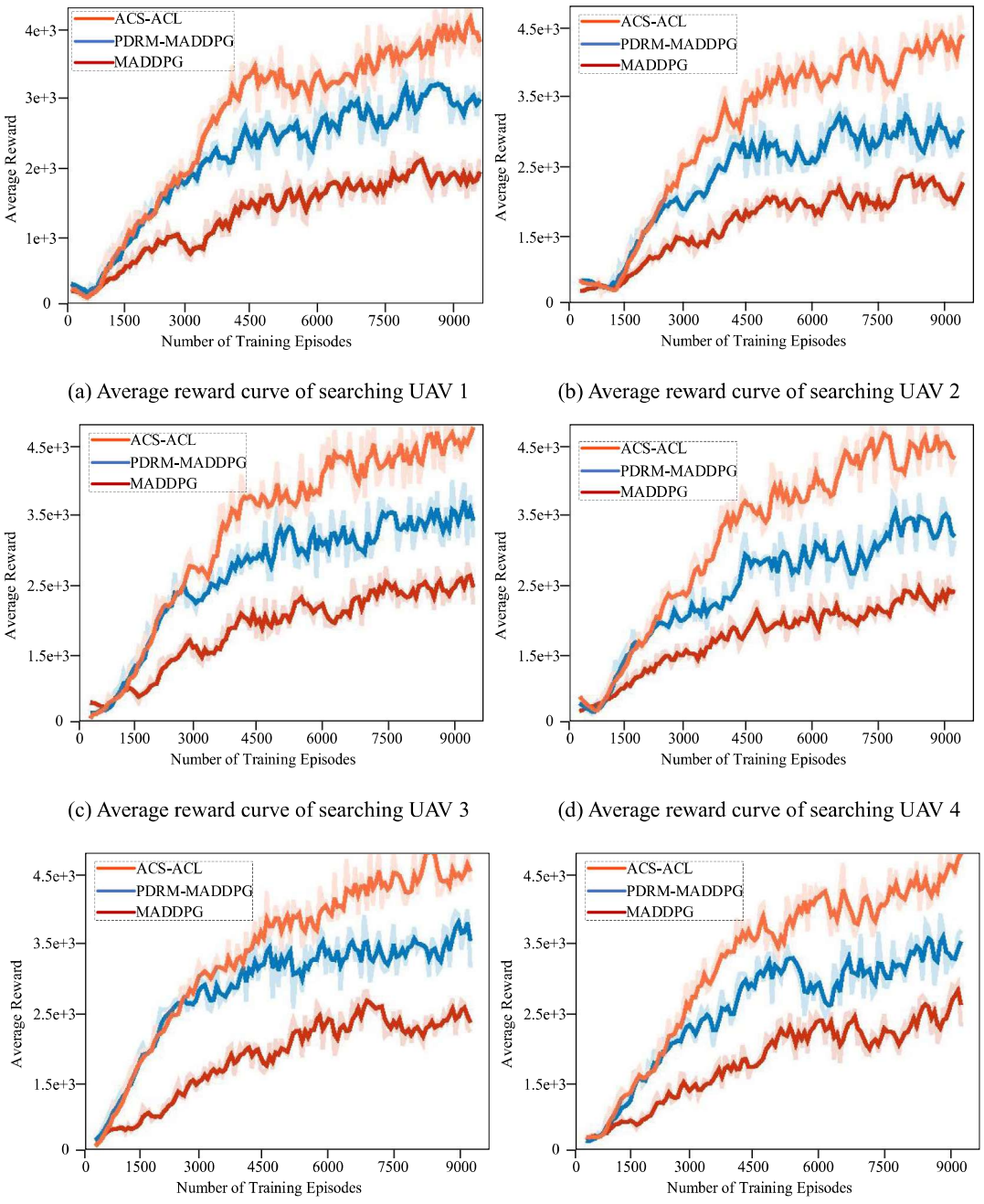


Fig. 10. Average reward curve of searching UAV when confrontation party adopts IL-DDPG algorithm.

the loop released in the process of continuous interaction (CI) to fill in the volition share buffer pool of actual combat commander in the loop, the indirect communication channel (ICC) among USG members is activated normally, export the USG autonomous confrontation strategy coupled with commander's volition, the other elements of ablation experiment follow the flows of Algorithm 1 to execute, therefore the ablation experiment box plot under active learning evolution (ALE) mechanism is shown in Fig. 15. In which, the horizontal axis represents the adopted confrontation strategy, the longitudinal axis represents the mean value reward with standard deviation, Fig. (a) shows the Confrontation Scenario 1, and Fig. (b) shows the Confrontation Scenario 2.

4.4. Experimental results analysis

By observing experimental data provided by group Figs. 8 and 9, and group Figs. 11 and 12, we can know that, under two confrontation scenarios with increasing difficulties and compared with the PDRM - MADDPG algorithm, the ACS-ACL algorithm has significant advantages on the qualitative and quantitative level. On the qualitative level, the complexity of the search trajectory diagram when UAVs of ourselves adopt ACS-ACL algorithm reduces significantly, and the reduction in complexity of search trajectory diagram is more significant as the difficulty of confrontation scenario increases; on the quantitative level, the time spent on locking UGVs of the enemy simultaneously three times

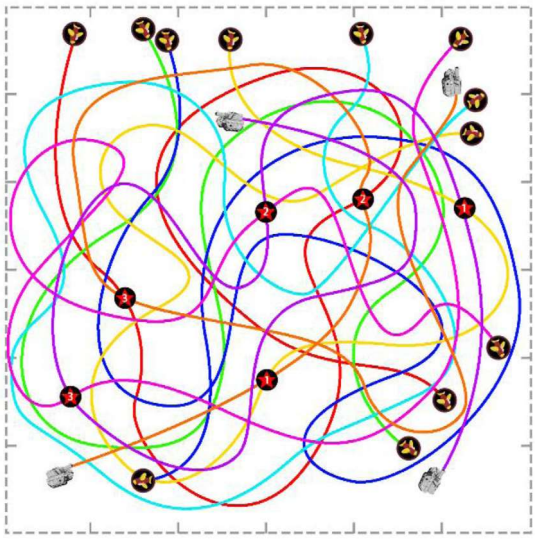


Fig. 11. Search trajectory diagram of UAVs of ourselves when adopting PDRM -MADDPG algorithm.

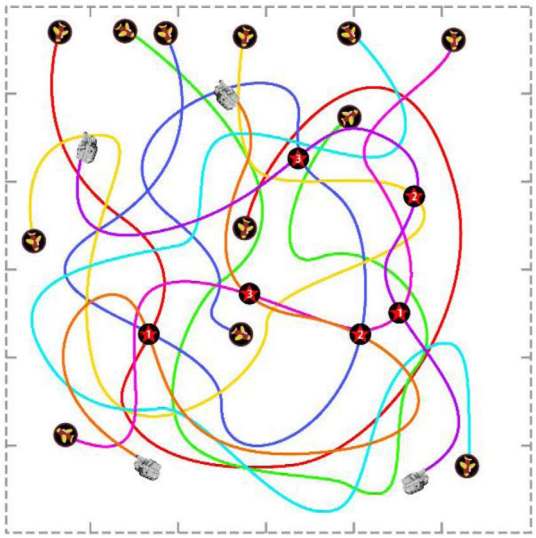


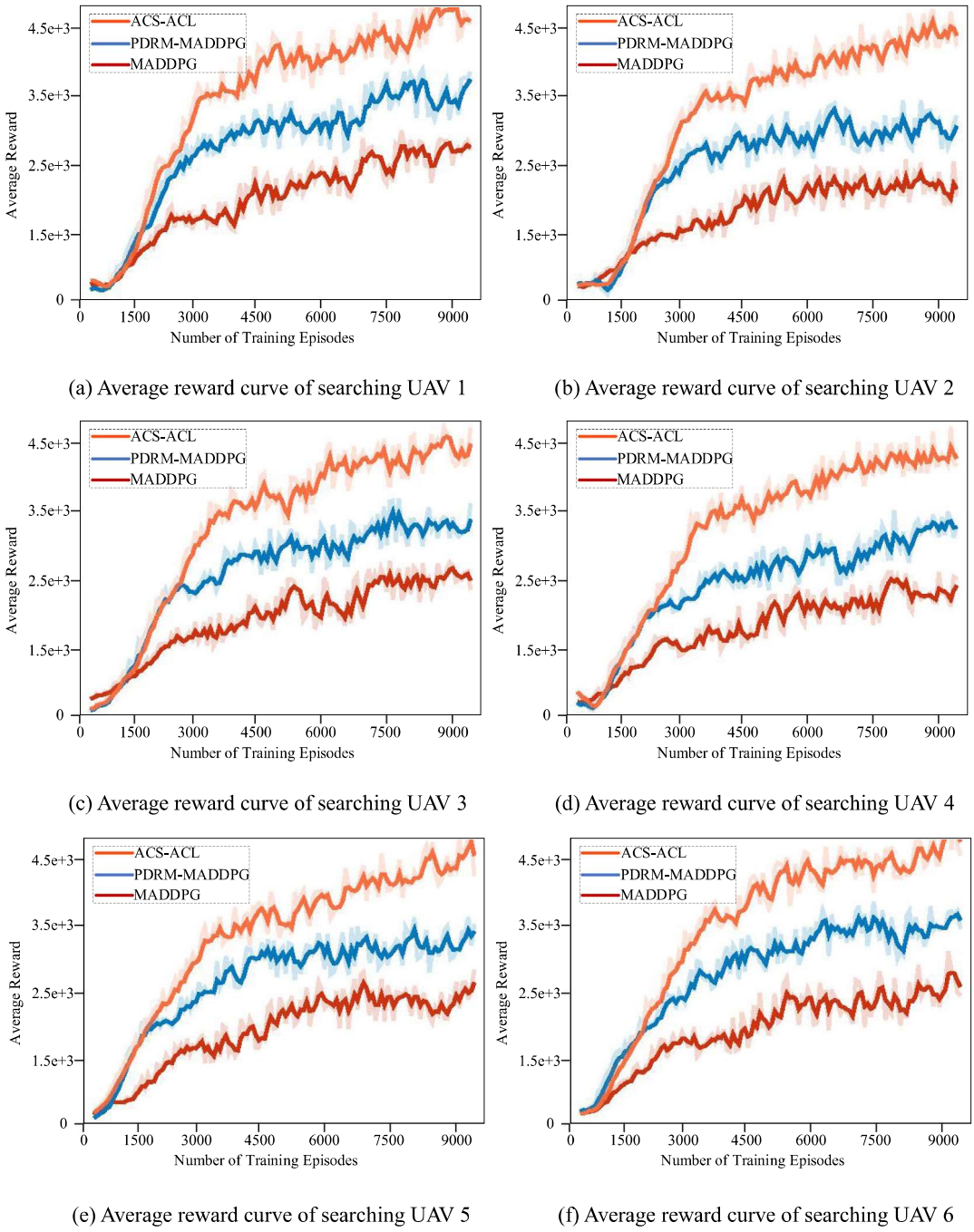
Fig. 12. Search trajectory diagram of UAVs of ourselves when adopting ACS-ACL algorithm.

when UAVs of ourselves adopt ACS-ACL algorithm as the autonomous confrontation strategy shortens significantly, and the reduction in the time spent on locking UGVs of the enemy simultaneously three times is more significant as the difficulty of confrontation scenario increases. By observing the experimental data as provided by group Figs. 10 and 13, we can know that, under two confrontation scenarios with increasing difficulties, the average reward of all searching UAVs will finally tend to get benign convergence, and the convergence tendency of average reward has benign advantage as the training episodes increase. When training Episodes are lower than 6000, the average reward of searching UAVs increase rapidly; when training Episodes are higher than 6000, the average reward of searching UAVs increase slowly, signifying that the autonomous collaborative search strategy has been fundamentally learned when the Episodes of multiple searching UAVs are about 6000 approximately.

Upon the multidimensional contrastive analysis on training and execution effects of ACS-ACL algorithm, PDRM-MADDPG algorithm, MADDPG algorithm, on the one hand, the algorithms of ACS-ACL and PDRM-MADDPG, under two confrontation scenarios, can all export effective collaborative search strategy, whereas the ACS-ACL algorithm, compared with the PDRM-MADDPG algorithm, gains the training and execution results with higher quality, and the MADDPG algorithm, under such two confrontation scenarios, cannot export effective collaborative search strategy, additionally there is the trend of getting worse gradually as the difficulty of confrontation scenario increases, at this time searching UAVs do not possess the autonomous collaborative search strategy; on the other hand, as the difficulty of confrontation scenario increases, and under the action of active learning evolution (ALE) mechanism, USG autonomous confrontation strategy coupled with volition of commander triggers the targeted active learning evolution, so the performance difference between ACS-ACL algorithm and PDRM-MADDPG algorithm enlarges further, and the convergence efficiency and quality of ACS-ACL algorithm has a tendency to become better as the training Episodes increase continuously, signifying that the ACS-ACL algorithm has the active learning evolution attribute, so the ACS-ACL algorithm can be guided to realize the autonomous confrontation strategy learning continuous benign evolution under the driving of combat intention by increasing the complexity of confrontation environment and strengthening the strategy intelligence degree of confrontation party.

By observing the data of ablation experiment as provided by group Figs. 14 and 15, we can know that continuous interaction (CI) mechanism has significant influence on the performance of ACS-ACL algorithm, especially on the convergence efficiency of ACS-ACL algorithm, it was just because that USG through continuous interaction (CI) with actual combat commander in the loop, continues to fill in self-experience gap, and expedites self-experience learning progress by coupling the combat intention of actual combat commander in the loop, and thereby optimizes the convergence efficiency of ACS-ACL algorithm; active learning evolution (ALE) mechanism has significant influence on the performance of ACS-ACL algorithm, especially on the convergence quality of ACS-ACL algorithm, it was just because that USG autonomous confrontation strategy coupled with volition of commander, under the action of active learning evolution (ALE) mechanism, triggers the targeted active learning evolution, and thereby optimizes the convergence quality of ACS-ACL algorithm. In order to further verify the applicability and robustness of ACS-ACL algorithm under the different confrontation scenarios, two algorithms such as ACS-ACL, PDRM-MADDPG were respectively made 26,000 times of Standard Monte Carlo (SMC) simulations in two confrontation scenarios with increasing difficulty, and counted statistically the final win rate, the details are shown in Fig. 16.

By making comprehensive observation on the statistical information about the win rate as provided in Fig. 16, the following basic conclusions could be made: the win rate of two algorithms such as ACS-ACL,PDRM-MADDPG declined as the difficulties in confrontation scenarios gradually increased, it was just because that, as the intelligence degree of autonomous confrontation strategy of UGVs of the enemy improved, and the complexity degree of confrontation scenario increases sharply, multiple searching UAVs were forced to adopt more advanced strategy to adapt to complicated confrontation scenario; under two confrontation scenarios with increasing difficulty, ACS-ACL algorithm obtains the win rate exceeding 70%, such win rate is significantly leading that of PDRM-MADDPG algorithm, and such win rate's reduction rate is far lower than PDRM-MADDPG algorithm, thus it further verifies the one that ACS-ACL algorithm has good multi-scenario applicability; The determinants of win rates of two algorithms such as ACS-ACL,PDRM-MADDPG, under the confrontation scenario with increasing difficulty include initial position of both confrontation parties (randomly generate within the limited area in this paper), quantity ratio of both confrontation parties (it is 6:2 in this paper), speed ratio of both confrontation parties (it is 86:92 in this paper), confrontation strategy adopted by both confrontation parties, initial interactive learning reward (it is 2600 in this paper), volition share buffer pool



(a) Average reward curve of searching UAV 1

(b) Average reward curve of searching UAV 2

(c) Average reward curve of searching UAV 3

(d) Average reward curve of searching UAV 4

(e) Average reward curve of searching UAV 5

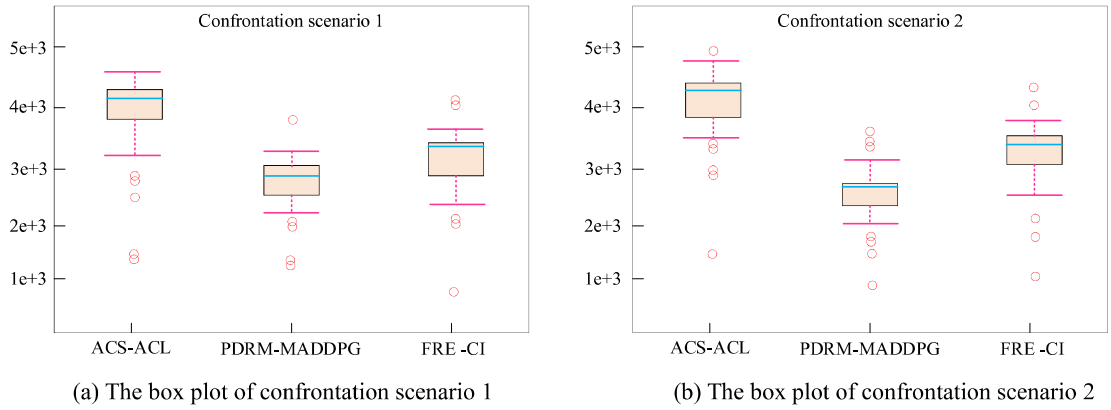
(f) Average reward curve of searching UAV 6

Fig. 13. Average reward curve of searching UAV when confrontation party adopts MADDPG algorithm.

capacity (it is 60,000 in this paper). When applying ACS-ACL algorithm to establish USG autonomous confrontation strategy benign closed-loop learning evolution mechanism, there is need to make concrete analysis as per specific scenario, set the initial interactive learning reward with good guidance, rationally select the reading–writing weights of volition share buffer pool; when activating continuous interaction (CI) mechanism, there is need to maintain the dynamic balance between interactive learning reward and learning interaction reward under time parallel; when activating active learning evolution (ALE) mechanism, there is need to maintain the dynamic balance between environmental adaption evolution mechanism and strategy matching evolution mechanism under time parallel, if so, there is the expectation to realize continuous benign evolution of autonomous confrontation strategy learning under the driving of combat intention.

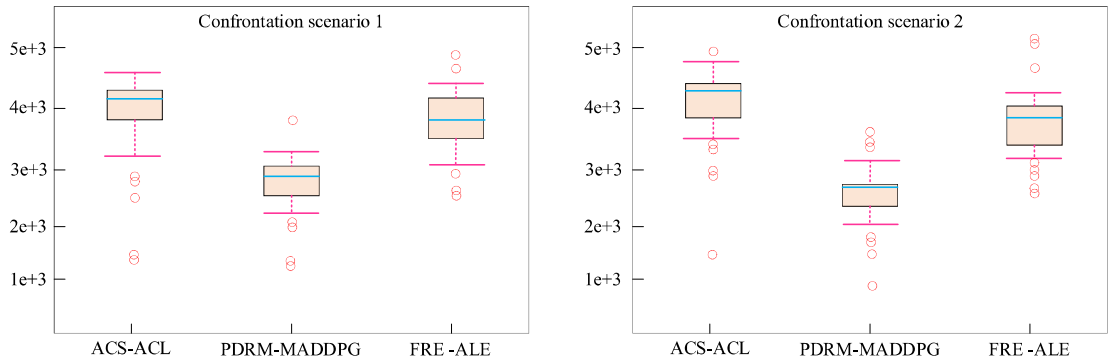
5. Conclusions

In order to strengthen the guiding position of actual combat commander in the loop in the process of USG autonomous confrontation strategy learning evolution, the combat intention of actual combat commander in the loop is mapped to USG autonomous confrontation strategy learning evolution process in the end-to-end form, thus an Autonomous Confrontation Strategy Learning Evolution Mechanism of Unmanned System Group Under Actual Combat in the Loop (ACS-ACL) is put forward. The contributions of ACS-ACL algorithm are mainly embodied on two levels: on the one hand, introduce continuous interaction (CI) mechanism, fill in continuously the experience gap of USG, expedite the experience learning process of USG, generate the USG autonomous confrontation strategy coupled with volition of commander, then provide experience driving to USG autonomous confrontation



(a) The box plot of confrontation scenario 1 (b) The box plot of confrontation scenario 2

Fig. 14. The ablation experiment box plot of continuous interaction (CI) mechanism.



(a) The box plot of confrontation scenario 1 (b) The box plot of confrontation scenario 2

Fig. 15. The ablation experiment box plot of active learning evolution (ALE) mechanism.

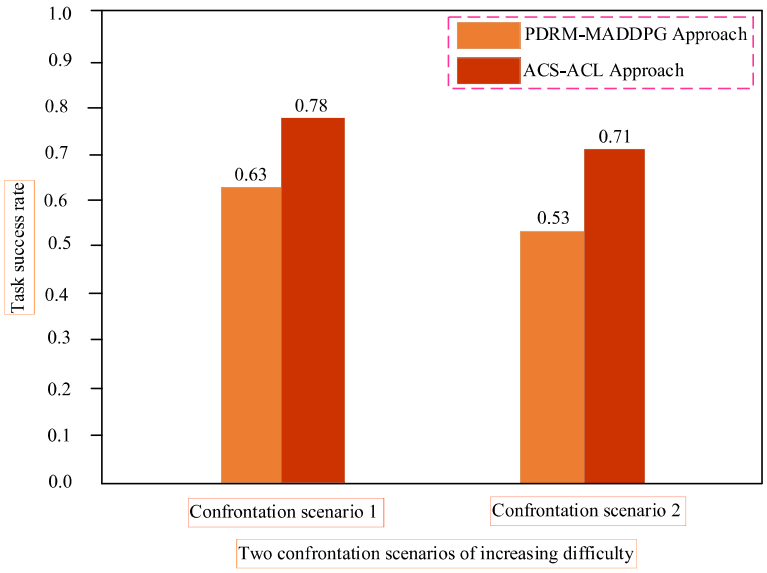


Fig. 16. Statistical diagram for win rate of two algorithms in different confrontation scenarios.

strategy active learning evolution; on the other hand, by introducing the active learning evolution (ALE) mechanism of confrontation strategy, guide the USG autonomous confrontation strategy coupled with volition of commander to trigger the targeted active learning evolution, realize USG behavioral autonomy capability to always adapt to the complicated actual combat battlefield confrontation environment, realize USG optimal joint strategy to always match with USG optimal joint strategy of confrontation party, realize continuous benign evolution of autonomous confrontation strategy learning for USG under the driving of combat intention.

A visual USG autonomous collaborative search dynamic confrontation game environment is constructed, and a series of comparison validation are carried out to make experiment. The experiment results therefrom show that ACS-ACL algorithm, compared with PDRM

-MADDPG algorithm, gains the training and execution effects of higher quality, and the convergence efficiency and quality of ACS-ACL algorithm have the trend of continuously becoming better as training Episodes continue to increase. By increasing the complexity of confrontation environment and strengthening the strategy intelligence degree of confrontation party, the performance difference between ACS-ACL algorithm and PDRM -MADDPG algorithm enlarges further, signifying that ACS-ACL algorithm has the active learning evolution attribute, can realize the continuous benign evolution of autonomous confrontation strategy under the driving of combat intention, and has considerable application prospect in unmanned confrontation field in future aerial combat. Through the comprehensive analysis on ablation experiment box plot, it can be known that continuous interaction (CI) mechanism has significant influence on the convergence efficiency of ACS-ACL algorithm, and active learning evolution (ALE) mechanism has significant influence on the convergence quality of ACS-ACL algorithm.

In the follow-up research works, in order to continuously strengthen the actual combat application credibility of USG autonomous confrontation strategy, further apply the ACS-ACL algorithm as mentioned in this paper to the confrontation scenario closer to actual combat, on the one hand, the Research Group will carry out the research of actual combat scenario simulation generation technology, utilize the generated combat scenario to continuously induce USG to learn collaborative combat experience, then drive the continuous benign evolution of USG autonomous confrontation strategy learning under the combat environment with increasing space complexity; on the other hand, the Research Group will carry out the research of multi-domain heterogeneous USG autonomous confrontation strategy learning evolution mechanism, extend USG attributes from single-domain, homogenesis and isomorphism to multi-domain, heterogeneity and isomerism, give full play to the resource complementarity of multi-domain heterogeneous USG, improve greatly the exploration precision, dimensionality, coverage of whole multi-domain heterogeneous USG and the cross-domain collaborative combat efficiency, and then adapt to the application demands of air-sea-ground multi-domain collaborative combat in future.

CRediT authorship contribution statement

Wang Zhenhua: Software, Data curation, Writing – original draft. **Guo Yan:** Conceptualization, Methodology. **Li Ning:** Supervision, Writing – reviewing & editing. **Yuan Hao:** Data curation. **Hu Shiguang:** Visualization, Investigation. **Lei Binghan:** Software, Validation. **Wei Jianyu:** Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to express their sincere gratitude to the Natural Science Foundation of Jiangsu Province of China (Grant No. BK20211227), the National Natural Science Foundation of China (Grant No. 61871400, 62273356) for providing funds to support this study. Also, We thank Home for Researchers editorial team (www.home-for-researchers.com) for language editing service.

Appendix A. Schedule 1 list of abbreviations

See Schedule A.1.

Appendix B

In order to make the paper more readable, Appendix B provides the dimensions and physical meanings information of state space and action space of UAVs of ourselves and UGVs of the enemy defined in the simulation verification process. The specific dimensions and physical

Schedule A.1
List of abbreviations.

Full name	Abbreviation
Unmanned System Group	USG
Autonomous Confrontation Strategy learning evolution mechanism of USG under Actual Combat in the Loop	ACS-ACL
Multi Agent Deep Deterministic Policy Gradient	MADDPG
Parallel Decoupling Reward Mechanism	PDRM
Continuous Interaction	CI
Artificial Intelligence Technology	AIT
Deep Reinforcement Learning	DRL
Unmanned Aerial Vehicle	UAV
Autonomous Confrontation Strategy Learning Evolution	ACSLE
Active Learning Evolution	ALE
Indirect Communication Channel	ICC
Chaos Elite Adapts-Genetic Algorithm	CEA-GA
Exponential average Momentum Pigeon-Inspired Optimization	EM-PIO
Deep Q Network	DQN
Proximal Policy Optimization	PPO
Deep Deterministic Policy Gradient	DDPG
Half-Random Q-learning	HR Q-learning
Mixed Experience-Multi Agent Deep Deterministic Policy Gradient	ME-MADDPG
Unmanned Ground Vehicle	UGV
Experience Replay	ER
Parallel Benchmark-Critic	PB-Critic
Parallel Decoupling-Critic	PD-Critic
Intrinsic Motivation	IM
Infinite Horizon-Markov Decision Process	IH-MDP
Replay Experience Buffer Pool	REBP
Volition Share Buffer Pool	VSBP
Memory Learning	ML
Independent Learning Deep Deterministic Policy Gradient	IL-DDPG
Standard Monte Carlo	SMC

meanings of the state space and action space of UAVs of ourselves and UGVs of the enemy as follows:

The action space of UAVs of ourselves is defined as follows (Six dimensions): flight course, flight altitude, flight speed, target locking, tracking activation, tracking sleep.

The state space of UAVs of ourselves is defined as follows (Eight dimensions): instantaneous longitude, instantaneous dimension, instantaneous speed, instantaneous course, instantaneous altitude, search state, target longitude, target dimension.

The action space of UGVs of the enemy is defined as follows (Five dimensions): movement direction, movement speed, locked escape, anti-tracking activation, anti-tracking sleep.

The state space of UGVs of the enemy is defined as follows (Five dimensions): instantaneous longitude, instantaneous dimension, instantaneous speed, instantaneous direction, escape state.

References

[1] Jin-wu Xiang, Xi-wang Dong, Wen-rui Ding, et al., Key technologies for autonomous cooperation of unmanned swarm systems in complex environments, Acta Aeronaut. Astron. Sinica 43 (10) (2022) 333–365.

[2] O. Abu Arqub, Z. Abo-Hammour, Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm, Inform. Sci. 279 (09) (2014) 396–415.

[3] Ting-ting Zhang, Yu-shi Lan, Ai-guo Song, Behavioral decision learning reward mechanism of unmanned swarm system, J. Beihang Univ. 47 (12) (2021) 2442–2451.

[4] H. Liu, K. Wu, K. Huang, et al., Optimization of large-scale UAV cluster confrontation game based on integrated evolution strategy, Cluster Comput. 35 (02) (2023) 147–169.

[5] S. Momani, B. Maayah, O. Abu Arqub, The reproducing kernel algorithm for numerical solution of Van der Pol damping model in view of the Atangana-Baleanu fractional approach, Fractals 28 (08) (2020) 1–12.

[6] Xing-yu Zhu, Jian-liang Ai, Research on intelligent decision making of many to many unmanned aerial vehicle air comba, J. Fudan Univ.(Nat. Sci.) 60 (04) (2021) 410–419.

[7] Yan Cao, Wan-yu Wei, Yu Bai, et al., Multi-base multi-UAV cooperative reconnaissance path planning with genetic algorithm, Cluster Comput. 22 (05) (2019) 5175–5184.

[8] Chao Wen, Wen-han Dong, Wu-jie Xie, et al., Multi-UAVs 3D cooperative curve path planning method based on CEA-GA, J. Beihang Univ. 1–19, [2022-03-18].

[9] Jing Yu, En-mi Yong, Han-yang Chen, et al., Bi-level mission planning framework for multi-cooperative UAV air-to-ground attack, Syst. Eng. Electron. 44 (09) (2022) 2849–2857.

[10] Wen qing Zhou, Ji-hong Zhu, Min-chi Kuang, et al., Multi-UAV cooperative swarm algorithm in air combat based on predictive game tree, Sci. Sinica Technol. 53 (02) (2023) 187–199.

[11] Hai-bin Duan, Bing-da Tong, Ji-Chuan Liu, et al., Coordinated target defense for multi-UAVs based on exponentially averaged momentum pigeon-inspired optimization, J. Beihang Univ. 48 (09) (2022) 1624–1629.

[12] Wei Shi, Yang-he Feng, Guang-quan Cheng, et al., Research on multi-aircraft cooperative air combat method based on deep reinforcement learning, Acta Automatica Sinica 47 (07) (2021) 1610–1623.

[13] Qi-ming Yang, Jian-dong Zhang, Guo-qing Shi, et al., Maneuver decision of UAV in short-range air combat based on deep reinforcement learning, IEEE Access 8 (12) (2019) 363–378.

[14] Xiao-wei Fu, Zhe Xu, Hui Wang, Generalization strategy design of UAVs pursuit evasion game based on DDPG, J. Northwest. Polytech. Univ. 40 (01) (2022) 47–55.

[15] Peng-xing Zhu, Xi Fang, Multi-UAV cooperative task assignment based on half random Q-learning, Symmetry-Basel 13 (12) (2021) 1–23.

[16] Kai-fang Wan, Ding-wei Wu, Bo Li, et al., ME-MADDPG: An efficient learning-based motion planning method for multiple agents in complex environments, Int. J. Intell. Syst. 37 (03) (2022) 2393–2427.

[17] Z. Abo-Hammour, O. Abu Arqub, S. Momani, et al., Optimization solution of Troesch's and Bratu's problems of ordinary type using novel continuous genetic algorithm, Discrete Dyn. Nat. Soc. 2014 (02) (2014) 1–15.

[18] Yu Sun, Jun Lai, Lei Cao, et al., A novel multi-agent parallel-critic network architecture for cooperative-competitive reinforcement learning, IEEE Access 8 (07) (2020) 135605–135616.

[19] T. Chu, J. Wang, L. Codecà, et al., Multi-agent deep reinforcement learning for large-scale traffic signal control, IEEE Trans. Intell. Transp. Syst. 21 (03) (2020) 1086–1095.

[20] Z. Wang, Y. Guo, N. Li, et al., Autonomous collaborative combat strategy of unmanned system group in continuous dynamic environment based on PD-MADDPG, Comput. Commun. 200 (02) (2023) 182–204.

[21] R. Krishna, D. Lee, L. Fei-Fei, et al., Socially situated artificial intelligence enables learning from human interaction, Proc. Natl. Acad. Sci. 119 (39) (2022) 1–8.

[22] Yang Jiachen, Zhang Jipeng, Wang Huihui, Urban traffic control in software defined internet of things via a multi-agent deep reinforcement learning approach, IEEE Trans. Intell. Transp. Syst. 22 (6) (2020) 3742–3754.

[23] Z. Movahedi, A. Bastanfard, Toward competitive multi-agents in Polo game based on reinforcement learning, Multimedia Tools Appl. 80 (17) (2021) 26773–26793.

[24] W. Qiu, C. Huang, Y. Chen, et al., A contract-based energy harvesting mechanism in UAV communication network, Comput. Commun. 199 (02) (2023) 50–61.

[25] R. Lowe, Y. Wu, A. Tamar, et al., Multi-agent actor-critic for mixed cooperative-competitive environments, Neural Inf. Process. Syst. (NIPS) 30 (06) (2017) 1–16.

[26] P.H. Leal, B. Kartal, M.E. Taylor, A survey and critique of multiagent deep reinforcement learning, Auton. Agents Multi-Agent Syst. 33 (06) (2019) 750–797.

[27] Hai-xia Peng, Xue-min Shen, Multi-agent reinforcement learning based resource management in MEC- and UAV-assisted vehicular networks, IEEE J. Sel. Areas Commun. 39 (01) (2021) 131–141.

[28] S. Momani, O. Abu Arqub, B. Maayah, Piecewise optimal fractional reproducing kernel solution and convergence analysis for the Atangana-Baleanu-Caputo model of the Lienard's equation, Fractals 28 (08) (2020) 1–13.

[29] S. Chigullapally, C.S.R. Murthy, Joint energy and throughput optimization for MEC-enabled multi-UAV IoRT networks, Comput. Commun. 201 (03) (2023) 1–19.

[30] J. Chen, L. Guo, J. Jia, et al., Resource allocation for IRS assisted SGF NOMA transmission: A MADRL approach, IEEE J. Sel. Areas Commun. 40 (04) (2022) 1302–1316.

[31] Kai Liu, Yu-yang Zhao, Gang Wang, et al., Self-attention-based multi-agent continuous control method in cooperative environments, Inform. Sci. 585 (03) (2022) 454–470.

[32] Zhu Kai, Zhang Tao, Deep reinforcement learning based mobile robot navigation: A review, Tsinghua Sci. Technol. 26 (5) (2021) 674–691.

[33] S. Wu, W. Xu, F. Wang, et al., Distributed federated deep reinforcement learning based trajectory optimization for air-ground cooperative emergency networks, IEEE Trans. Veh. Technol. 71 (8) (2022) 9107–9112.

[34] T.T. Nguyen, N.D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications, IEEE Trans. Cybern. 50 (09) (2020) 3826–3839.