

Article

Autonomous Task Planning of Intelligent Unmanned Aerial Vehicle Swarm Based on Deep Deterministic Policy Gradient

Qiang Jiang^{1,2}, Yongzhao Yan^{1,2}, Yinxing Dai², Zequan Yang³, Huazhen Cao¹, Bo Wang^{1,4,*}  and Xiaoping Ma^{1,2}

¹ National Key Laboratory of Science and Technology on Advanced Light-Duty Gas-Turbine, Institute of Engineering Thermophysics, Chinese Academy of Sciences, Beijing 100190, China; jiangqiang23@mails.ucas.ac.cn (Q.J.); yanyongzhao19@mails.ucas.ac.cn (Y.Y.); caohuazhen@iet.cn (H.C.); maxiaoping@iet.cn (X.M.)

² University of Chinese Academy of Sciences, Beijing 100049, China; daiyinxing22@mails.ucas.ac.cn

³ Wasion Group Co., Ltd., Changsha 410000, China; 17810679436@163.com

⁴ Qingdao Institute of Aeronautical Technology, Qingdao 266400, China

* Correspondence: wangboc@iet.cn

Abstract: Intelligent swarm is a powerful tool for targeting high-value objectives. Within the Anti-Access/Area Denial (A2/AD) context, an unmanned aerial vehicle (UAV) swarm must leverage its autonomous decision-making capability to execute tasks with independence. This paper focuses on the Suppression of Enemy Air Defenses (SEAD) mission for intelligent stealth UAV swarms. The current research field mainly faces challenges in fully simulating the complexity of real-world scenarios and in insufficient autonomous task planning capabilities. To address these issues, this paper develops a representative problem model, establishes a six-tier standardized simulation environment, and selects the Deep Deterministic Policy Gradient (DDPG) algorithm as the core intelligent algorithm to enhance the autonomous task planning capabilities of UAV swarms. At the algorithm level, this paper designs reward functions corresponding to UAV swarm behaviors, aiming to motivate UAV swarms to adopt more effective action strategies, thereby achieving autonomous task planning. Simulation results demonstrate that the scenario and architectural design are feasible and that artificial intelligence algorithms can enable the UAV swarm to show a higher level of intelligence.



Academic Editor: Pablo Rodríguez-González

Received: 30 January 2025

Revised: 4 March 2025

Accepted: 24 March 2025

Published: 3 April 2025

Citation: Jiang, Q.; Yan, Y.; Dai, Y.; Yang, Z.; Cao, H.; Wang, B.; Ma, X. Autonomous Task Planning of Intelligent Unmanned Aerial Vehicle Swarm Based on Deep Deterministic Policy Gradient. *Drones* **2025**, *9*, 272. <https://doi.org/10.3390/drones9040272>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: swarm; autonomous decision-making; simulation environment; intelligent algorithm

1. Introduction

Due to its distributed and highly flexible nature, UAV swarms can effectively leverage their numerical advantages to conduct reconnaissance and strike missions across multiple regions and targets. However, the experience of modern warfare shows that high-value targets are often hidden behind the front line and in a relatively deep position in order to protect themselves from direct enemy attacks or to better fulfill their strategic functions [1]. Employing intelligent stealth UAV swarms in penetration and search tasks is an effective approach. However, real battlefield environments are highly dynamic, characterized by constantly changing target behaviors, electromagnetic conditions, and weather patterns. The current simulation environment makes it difficult to update these dynamic changes in real time, which may lead to adaptability issues for the trained algorithms in practical applications.

In the SEAD mission, the UAV swarm enters the target reconnaissance and strike area, executes the final search task, and completes the closed loop of the whole task. High-value

enemy targets are usually located in the deep area of the enemy. Compared with the conventional mission environment, the SEAD mission environment is in a strong refusal state. On the one hand, due to the distance between the fleet and the rear command center, it is difficult for rear command personnel to obtain real-time dynamic information in a timely manner. Without considering satellite communication support, it is also challenging to exert real-time control over the UAV swarm. At the same time, the UAV swarm will face various enemy countermeasures in the process of performing established combat missions. Typical countermeasures, like electromagnetic interference or air defense firepower, are difficult challenges of swarm's autonomous task ability, and a practical problem that restricts the application of swarms [2–5].

The traditional air combat confrontation theory is based on the principle of “win by energy”, with both sides' aircraft aiming to achieve higher relative altitude, better attack angles, and more favorable relative battle positions [6]. However, the swarm is mainly in the form of “group-attack verse group-defense”, and the comparative advantages between single machines are difficult to be transform into the combat effectiveness of the group. Swarms should aim to “win by intelligence”, and realize “gathering of intelligence” through orderly division of labor and coordinated operations [7]. Facing the complex rejection environment, the formation, battle position and strategy of each UAV in the swarm are very complex, which would exceed people's conceptual design ability and is difficult to achieve directly. It is an effective method to use artificial intelligence [8,9].

In recent years, researchers have increasingly applied multi-agent deep reinforcement learning (MARL) to various applications in UAV swarms, such as combat, formation, obstacle avoidance, and navigation [10–27]. For example, to achieve efficient swarm coordination and combat strategies through intelligent algorithms, Yang et al. (2024) proposed a Decomposed Prioritized Experience Replay-based Multi-Agent Deep Deterministic Policy Gradient (DP-MADDPG) algorithm for drone swarm combat [28]. This algorithm integrates decomposition mechanisms and Prioritized Experience Replay (PER) into the traditional MADDPG framework, overcoming the technical challenges of converging to local optima and dominant strategies, thereby improving the success rate of drone combat. Deng et al. (2024) introduced an Evolutionary Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (E-MATD3) algorithm, which combines evolutionary algorithms with multi-agent reinforcement learning to effectively address the issue of strategy cycling in traditional self-play training [29]. The E-MATD3 algorithm employs an attention-based network architecture to adapt to dynamic changes in the number of drones in aerial combat and uses death masking to avoid weight bias caused by drone crashes. Aschu et al. (2024) proposed a multi-agent deep reinforcement learning (MADRL) approach for local path planning in a UAV swarm, specifically targeting the precise landing task in a 3D environment [30]. This method addresses the challenges faced by traditional control and planning methods in drone swarm landing scenarios. Kumari et al. (2024) developed a method for reliable identification and tracking of individual drones within a swarm using stereo vision cameras and multi-agent reinforcement learning [31]. This method effectively deals with the high density and dynamic behavior of UAV swarms, providing technical support for precise tracking within the swarm. Although these algorithms have their own advantages, the Deep Deterministic Policy Gradient (DDPG) algorithm better meets the task planning requirements of UAV swarms in complex battlefield environments, with its efficiency and accuracy in continuous action space [32–38]. In addition, the experience replay mechanism and target network design of the DDPG algorithm enable it to have higher sample efficiency and convergence speed when dealing with complex tasks. These characteristics make the DDPG algorithm well-suited for autonomous task planning of UAV swarms, effectively improving task success rates and swarm survivability.

This paper focuses on two key issues:

- (1) building a standard framework of simulation environments for intelligent stealth swarms in a rejection environment, and form a simulation environment with high standardization and generalization ability with scientific and reasonable logic and combined with swarm characteristics;
- (2) based on the existing simulation environment, aiming at typical SEAD tasks, use a reinforcement learning algorithm to realize autonomous tasks of intelligent swarms.

The rest of the paper is organized as follows. Section 2 introduces the scenario where a UAV swarm autonomously completes the task of searching and striking ground targets through reinforcement learning. Section 3 designs a simulation and verification platform of autonomous task confrontation empowered by artificial intelligence. Section 4 sets reward functions for mission UAVs and control rules for ground targets. Through analysis of simulation results, it is proven that a reinforcement learning algorithm enables the UAV swarm to display higher levels of intelligence. Section 5 presents the conclusions.

2. Problem Model in Mission Scenario

The mission scenario is shown in Figure 1. Our stealth UAV swarm has broken through the enemy's dense defense circle and entered the target search area. At this stage, the mission UAV obtains the real-time position and motion information of enemy high-value targets through radar or Electro-Optical (EO) load, and performs the search mission. Enemy ground vehicles adopt scattered avoidance mode, keep moving and scattered avoidance, which increases the difficulty of the mission. In highly dynamic and highly time-sensitive scenarios, maximizing the ability to strike all enemy mobile ground targets within a limited strike window time places extremely high demands on the autonomous task planning capability of our mission swarm.

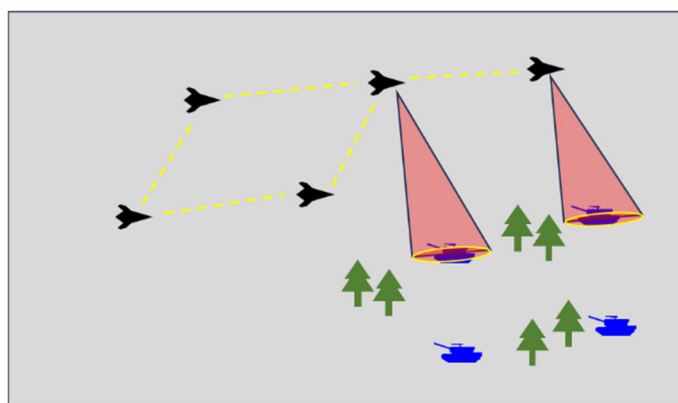


Figure 1. UAV swarm autonomous search.

The mission UAVs are controlled by a reinforcement learning algorithm, with each UAV acting as an independent agent. The internal limiting factor is the absence of centralized cooperative command and dispatch, while the swarm's intelligence relies on the emergent intelligence provided by the artificial intelligence algorithm. There are two external limiting factors. First, there is electromagnetic interference in the environment, and there are distance constraints and noise interference in the communication and observation between UAVs, that is, friendly UAVs can only obtain each other's state information within a certain communication range, and can only observe the state information of an enemy UAV swarm within a certain radar or EO visual distance, and random noise is added to the communication information and observation information. Second, the target has maneuverability, adopts expert strategy control, and has the ability of global observation

and spontaneous decentralized defense, which restricts the UAV swarm's ability to carry out the established tasks.

The goal of the control framework and simulation training is to enhance the intelligence level of the mission UAV swarm and develop its group awareness. When hitting ground targets, swarms can connect target allocation, mission planning and maneuver control, and quickly close the killing chain of “target-discovery to swarm-maneuver”. In this process, the task itself and the algorithm itself are deeply analyzed. Through the combination of control framework and related algorithms, the swarm has the ability of terminal autonomous confrontation. Through simulation experiments, the feasibility of the method is verified.

3. Construction of Simulation Model

According to the task requirements of stealth swarm, the simulation and verification platform of autonomous task confrontation empowered by artificial intelligence is designed. The simulation platform is structured into six levels, from bottom to top: mission level, kinematics model level, intelligent algorithm level, sensor information fusion level, command and control level, and effectiveness evaluation level. Figure 2 depicts the design of the adversarial simulation verification platform.

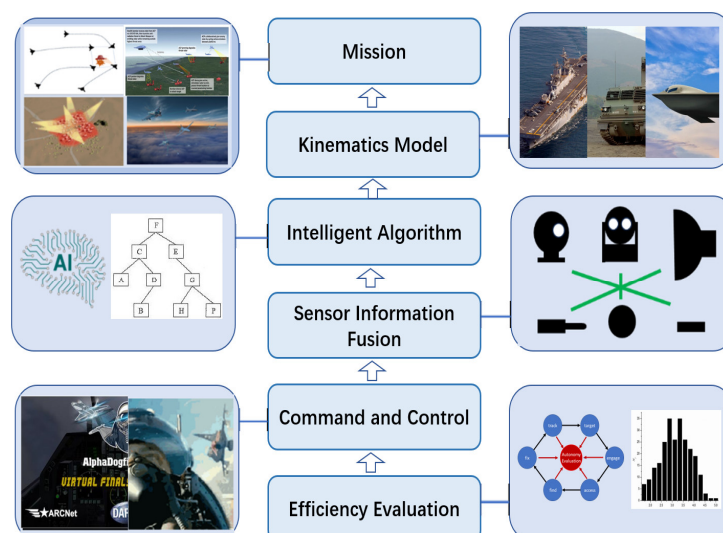


Figure 2. Design of simulation model.

3.1. Mission Level

In the mission level, confrontation scenes are set, and the characteristics of all parties can be refined and added to the simulation environment model according to different task backgrounds, so as to realize various task scenes such as air combat confrontation, air–sea confrontation or air–ground attack.

The mission level is the foundation of the UAV swarm countermeasure simulation platform, and its core purpose is to build a highly realistic virtual combat scene, simulating the elements and attributes in the actual battlefield. Researchers can use this level to customize various battlefield conditions.

Through careful design, the simulation platform can provide users with a highly realistic and customizable battlefield environment, so as to facilitate various tactical drills and strategy tests. This not only greatly enhances the practicability and educational significance of simulation, but also provides an ideal experimental platform for tactical research and technical development of UAV swarms.

3.2. Kinematics Model Level

Setting the dynamic models of the two sides can change the maneuverability of the two sides, realize the confrontation tests with different speed ratios and overload ratios, and simulate the multi-domain manned or unmanned platforms such as air, sea, and land by setting different types of dynamic kinematics models.

A three-dimensional global coordinate system is established with the ground in the center of the simulation area as the origin, and the position of each UAV in the UAV swarm in space is represented by (X, Y, Z) . In the simulation, in order to truly reflect the UAV motion state, according to the real state of the pilot operating the aircraft, the pilot's throttle lever and control lever operation mode are simulated, and a throttle channel and two heading angle channels are set up to establish a second-order three-degrees-of-freedom control model. Compared with the three-dimensional six-degrees-of-freedom equation, using the second-order three-degrees-of-freedom control model can reduce the calculation dimension of three degrees of freedom, save the computing consumption in the emergence of swarm intelligence, and pay more attention to the research problem itself.

$$\begin{aligned} x_{t+1} &= x_t + udt & u &= V \cos \gamma \cos \psi & V_{t+1} &= V_t + adt \\ y_{t+1} &= y_t + vdt & v &= V \cos \gamma \sin \psi & \gamma_{t+1} &= \gamma_t + \dot{\gamma}dt \\ z_{t+1} &= z_t + wdt & w &= V \sin \gamma & \psi_{t+1} &= \psi_t + \dot{\psi}dt \end{aligned} \quad (1)$$

The throttle channel controls the flight speed V to change the running speed of the aircraft by changing the acceleration a . The two angle channels control the pitching angle γ and yaw angle Ψ , respectively, and change the running direction of the aircraft by changing the pitching angular acceleration $\dot{\gamma}$ and yaw angular acceleration $\dot{\psi}$. In order to limit aircraft overload, boundary constraints are set for three control quantities.

The action space of the mission UAV is three-dimensional, and the output control quantities are $\begin{bmatrix} a \\ \dot{\gamma} \\ \dot{\psi} \end{bmatrix}$, $a \in [a_{\min}, a_{\max}]$, $\dot{\gamma} \in [\dot{\gamma}_{\min}, \dot{\gamma}_{\max}]$, $\dot{\psi} \in [\dot{\psi}_{\min}, \dot{\psi}_{\max}]$. The attributes of the relevant parameters in the kinematic model are given in Table 1.

Table 1. Parameter settings.

Parameter	Symbol	Unit	Value
position	(X, Y, Z)	km	$[0, 2000]$
speed	V	km/h	$[40, 100]$
acceleration	a	m/s^2	$[-10, 10]$
pitch angle	γ	$^\circ$	$[-\pi/3, \pi/3]$
yaw angle	Ψ	$^\circ$	$[-\pi/3, \pi/3]$
pitch angular acceleration	$\dot{\gamma}$	m/s^2	$[-\pi/3, \pi/3]$
yaw angular acceleration	$\dot{\psi}$	m/s^2	$[-\pi/3, \pi/3]$

3.3. Intelligent Algorithm Level

Researchers can determine the kernel-driven algorithm of the agent, preset the DDPG algorithm to control task swarm, use expert rules to control defensive swarm, and have algorithm secondary development interface on this level, which can improve the established algorithm according to simulation results and expectations.

The intelligent algorithm level is the core component of the UAV swarm countermeasure simulation platform, in which UAVs are endowed with an advanced autonomous decision-making ability and behavior control. The purpose of this level is to simulate the

intelligent behavior of UAVs in a complex battlefield environment, so that they can independently complete key tasks such as path planning, obstacle avoidance, target recognition, and tracking.

In terms of function realization, the intelligent algorithm level uses advanced artificial intelligence technologies such as reinforcement learning to provide powerful data processing and decision support for UAVs. Through machine learning algorithms, UAVs can learn and optimize their behavior pattern from historical data. Reinforcement learning can continuously improve the decision-making quality of UAVs through interactive learning with the environment, so that they can make more reasonable choices when facing unknown challenges.

Furthermore, the role of intelligent algorithms must account for the cooperative work among UAVs. It is essential to achieve information sharing and coordinated task execution within the swarm through decentralized intelligent decision-making algorithms, thereby enhancing the overall operational efficacy of the UAV swarm.

3.4. Sensor Information Fusion Level

At this level, the communication within the UAV swarm and the observation of local targets are considered, corresponding to the state space of UAV action, communication interaction, and radar/EO observation information fusion that are carried out; multiple parameters are reserved; and the limited degree of the environment relative to the task swarm is regulated.

Figure 3 illustrates the state space composition of the mission UAV. The state space of mission UAV i consists of three parts, namely, its own state information O_i^A , communication information $O_{i,k}^A$, and radar observation information R_i , which are expressed as follows:

$$O_i = O_i^A + O_{i,k}^A + R_i, (i, k \in N^A, i \neq k; j \in N^D) \quad (2)$$

Self-state information O_i^A includes three spatial positional quantities and three speed quantities, which are expressed as follows:

$$O_i^A = \{x_i^A, y_i^A, z_i^A, u_i^A, v_i^A, w_i^A\} \quad (3)$$

The communication information $O_{i,k}^A$ includes three spatial position quantities and three speed quantities of all friendly UAVs except itself, and the communication radius is d_O . When the friendly side is not within the communication radius, the information would be empty, which is represented as follows:

$$O_{i,k}^A = \begin{cases} \{x_k^A, y_k^A, z_k^A, u_k^A, v_k^A, w_k^A\} + O_{noise}, k \neq i, d_{Oik} \leq d_O \\ \{0, 0, 0, 0, 0, 0\}, k \neq i, d_{Oik} > d_O \end{cases} \quad (4)$$

Radar observation information R_i includes three spatial position quantities and three speed quantities of defensive targets, and the observation radius is d_R . When the enemy is not within the observation radius limit, the information would be empty, which is expressed as follows:

$$R_i = \begin{cases} \{x_j^D, y_j^D, z_j^D, u_j^D, v_j^D, w_j^D\} + R_{noise}, d_{Rij} \leq d_R \\ \{0, 0, 0, 0, 0, 0\} + R_{noise}, d_{Rij} > d_R \end{cases} \quad (5)$$

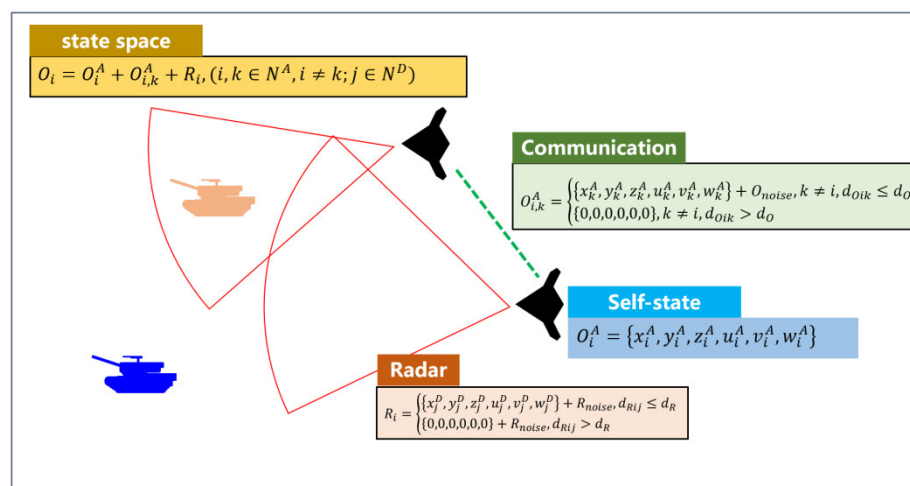


Figure 3. State space of mission UAVs.

3.5. Command and Control Level

In the command and control level of the UAV swarm countermeasure simulation platform, the setting of intelligent control attributes is crucial for realizing tactical decision-making and action coordination. This level primarily serves to establish the control strategies for UAVs in adversarial engagements, which are mainly divided into two categories: reinforcement learning control and expert decision control.

The purpose of the Reinforcement Learning Agent (RLA) is to enable the mission UAV swarm to autonomously learn and optimize its behavior strategy in the countermeasure environment. The reinforcement learning agent learns the optimal strategy by interacting with the environment, that is, by executing actions, receiving feedback, and updating its decision model, it gradually improves its performance. Reinforcement learning is especially suitable for dynamic and uncertain environments, which can adapt to the enemy's strategic changes and realize adaptive tactical decision-making.

The expert strategy manipulates the defensive swarm by using an expert system or preset tactical rules, thereby simulating the decision-making process of human experts. Expert strategy is based on a set of fixed rules or heuristic methods to guide UAV behavior, and these rules are usually formulated by domain experts according to experience. Expert strategy can simulate the decision-making of human commanders, provide a predictable and interpretable control mode, and facilitate the analysis and evaluation of tactical effectiveness.

3.6. Efficiency Evaluation Level

In the UAV swarm countermeasure simulation platform, the core function of the efficiency evaluation level is to judge the outcome. It is based on a series of key indicators to determine the tactical results of opposing sides. Mission completion is the primary index to judge the outcome, which measures the efficiency and success rate of UAV swarm in performing tasks such as target strike, reconnaissance, and surveillance. Survival rate is also very important. It evaluates the durability of swarms by the damage rate of UAVs and the survival time of battlefields. Resource consumption is also a factor that cannot be ignored, which relates to the efficiency of the use of key materials such as fuel and ammunition, and reflects the continuous combat capability of swarm. Reaction time is used to assess the response speed of UAV swarms to enemy threats. This directly affects the timeliness of tactics and the reduction of enemy attack opportunities, so it is also a key component of judging the outcome. In addition, the judgment of winning or losing should follow a series of preset rules and conditions, such as specific mission objectives and

battlefield rules of engagement, which provide clear standards for judgment. A real-time monitoring and feedback mechanism ensures the dynamic and adaptive nature of victory and defeat judgments, allowing commanders to adjust tactics and strategies based on real-time data. In the end, judging the outcome is a process of comprehensively considering various factors. By comprehensively analyzing task completion, survival rate, resource consumption, reaction time, etc., scientific and objective evaluation results are obtained, which provide support for tactical decision-making and strategy optimization.

The operating logic of the complete simulation system is as follows:

Firstly, the mission level defines the mission scenario and objectives, providing initial conditions for the kinematics model level. Then, the kinematics model level generates the dynamic behavior of the UAV and target based on mission requirements, providing data for the sensor information fusion level. The sensor information fusion level processes sensor data to provide environmental perception information for the intelligent algorithm level. The intelligent algorithm level generates decision instructions based on sensor information, providing decision support for the command and control level. The command and control level generates specific control commands based on the instructions of intelligent algorithms, and feeds them back to the dynamic model level to update the behavior of the UAV. Finally, the efficiency evaluation level assesses the completion status of missions and system efficiency, providing feedback to the mission level for adjusting subsequent mission settings.

The simulation environment is built on the local host, the software environment is compiled by python, the GPU is GeForce RTX 4070Ti 12G, the CPU is Intel 13,900 K, and the memory is DDR5 32G. The parameter settings related to the simulation environment are shown in Table 2, where the red side refers to the mission UAV and the blue side refers to the ground target.

Table 2. Simulation environment parameter settings.

Parameter Type	Camp	Name	Symbol	Value
environmental parameters	/	range of 3D scene area	/	$2000 \times 2000 \times 1000$
		training round	n	20000
		number of round steps	t	200
		time step	dt	0.2
refusal factor parameters	red	effective communication interaction distance	d_{RO}^R	700
		effective radar observation distance	d_{RR}^R	500
	blue	effective communication interaction distance	d_{RO}^B	∞
		effective radar observation distance	d_{RB}^B	∞
Ground reconnaissance and strike mission	red	minimum safe flight altitude	h_{R-G}	200
		flight altitude above the ground	h_{safe}	/
		distance between UAV and target	d_{R-T}	/
		effective detection range of reconnaissance payload	d_R	500
		strike decision time steps	t_d	5
		effective strike distance	d_{kill}	200
	blue	air defense identification radius	d_0	1000
		distance between targets	d_{B-B}	/

4. Implementation of Autonomous Task Planning

Unmanned aerial vehicle (UAV) task planning refers to the process of formulating flight routes for UAVs and conducting task allocation and overall management based on the tasks that the UAVs need to complete, the number of UAVs, and the types of mission payloads they carry. UAV swarm task planning, in addition to considering the requirements of the task itself, must also take into account the constraints of coordination and consistency

among UAVs in executing the task together. It is necessary to design collaborative flight routes for the UAV swarm according to the task planning indicators, so as to achieve the optimal or nearly optimal overall combat effectiveness. In this paper, the tasks that the mission UAVs need to complete include collision avoidance among UAVs, search and strike of ground targets, and maintaining a suitable flight altitude.

To achieve autonomous task planning for the mission UAVs, it is necessary to set up a reward function in the DDPG algorithm to control the mission UAVs and use expert rules to control ground targets after establishing the simulation platform. By analyzing the simulation results, the impact of the reinforcement learning algorithm on mission UAV swarm can be obtained. In this section, we introduce the principle of the DDPG algorithm and set reward functions for mission UAVs and expert rules for ground targets. Finally, we analyze the simulation results.

4.1. Principle of DDPG

In the intelligent algorithm level, in order to test the effectiveness of the overall simulation framework, this paper chooses the widely used deep deterministic strategy gradient algorithm (DDPG). The DDPG algorithm is an advanced reinforcement learning technology that is specially designed to deal with complex problems with continuous action space. The DDPG algorithm combines deep learning and the deterministic strategy gradient method, and solves the limitation of traditional reinforcement learning methods in continuous action space by using neural network to approximate strategy and value function.

The core principle of the DDPG algorithm is based on the strategy gradient method in reinforcement learning, which optimizes the strategy by directly increasing the gradient of the strategy function. Different from traditional value-based methods, the strategy gradient method directly models the decision-making process of actions, allowing effective exploration and utilization in continuous action space. The DDPG algorithm uses empirical playback mechanism and stores historical interaction data in playback buffer. The learning process randomly selects samples from this buffer, which helps the algorithm learn more efficiently from experience and reduces the time series dependence of data.

The DDPG algorithm adopts a deterministic policy, which means that in a given state, the policy will produce a deterministic action instead of a random action. This simplifies the strategy optimization process and allows more precise action control. The DDPG algorithm includes two neural networks: actor network and critic network. The actor network is responsible for generating strategies, that is, selecting actions. The critic network evaluates the quality of the current strategy and provides guidance on the direction of strategy improvement.

The DDPG algorithm, based on the Actor–Critic framework, is classified as an off-policy algorithm. The actions in the DDPG algorithm are deterministic, which is more conducive to decision-making in continuous action spaces. The DDPG algorithm features a replay buffer to store samples generated from the agent's interactions with the environment. In the experiment, the size of the buffer is set to 10^6 , the corresponding discount factor is set to 0.95, and the network model will begin to randomly sample and update when the number of interactions between the agent and the environment reaches 10^3 . The policy and value networks in the DDPG algorithm correspond to the actor and critic, respectively. In the experiment, both the policy and value networks consist of hidden layers with five fully connected layers, each with 128 neurons, and the activation function is ReLU. The learning rate for network training is set to 3×10^{-4} , and the batch size for stochastic gradient descent is 1024. When setting the hyperparameters, the target network update interval is $d = 1$ s, and the soft update rate $\tau = 0.5$ s.

4.2. Reinforcement Learning Reward Function

A well-designed reward function can motivate UAV swarms to adopt more effective action strategies, thereby achieving autonomous task planning. The reward for the mission UAV swarm consists of two parts.

Within time T , after the mission UAV swarm hits all targets, it is judged that the mission swarm wins the round, and after the round ends, the swarm gains the shared sparse reward R_{a1} . If the mission UAV swarm does not destroy all targets within time T , the round ends and the swarm gains a shared reward of 0. This is the reward for the first part.

The second part refers to the mission UAV obtaining dense rewards through the relative motion with the ground target, which are divided into collision avoidance reward R_{b1} , search reward R_{b2} , and flight altitude reward R_{b3} in simulation. For the swarm SEAD mission, the core command and control points are patrolling, detection, locking, striking, and safe collision avoidance. The collision avoidance reward is R_{b1} , which is primarily for the safe progression of the swarm and also to promote dispersion, thereby increasing search efficiency. Detection, locking, and striking are combined into Search Reward R_{b2} ; R_{b2} is a time-dependent quantity. To encourage the swarm to adopt a more aggressive strategy, these three actions are linked in sequence. The reward is only obtained after completing the three consecutive operations. Set the low-altitude patrolling reward as Flight Altitude Reward R_{b3} .

The reward R_b of intelligent UAV swarms consists of three parts,

$$R_b = w_1 R_{b1} + w_2 R_{b2} + w_3 R_{b3} \quad (6)$$

Collision Avoidance Reward R_{b1} . In order to avoid the internal collision of the swarm in complex confrontation, set the collision avoidance reward R_{b1} . The formula is as follows:

$$R_{b1} = \begin{cases} 1 - \frac{d_{A-A}}{d_r}, d_{A-A} < d_{safe} \\ 0, d_{A-A} \geq d_{safe} \end{cases} \quad (7)$$

The safe distance d_{safe} , is set here. When the relative distance d_{A-A} between the mission UAV and the adjacent UAV is less than d_{safe} , $d_{safe} = 200$, it will gain a continuous negative reward, and when the relative distance between the mission UAV is greater than d_{safe} , it will gain a reward of 0. R_{b1} is set up to guide mission drones to keep a safe distance between them.

Search Reward R_{b2} . In the scenario setting, the UAV swarm uses radar to search the ground. When the relative distance between the target and the UAV d_t^{A-D} is less than the UAV's radar detection range d_{find} , $d_{find} = 1500$, the search and positioning process starts, and the UAV starts to calculate the reward. If the target remains found for five time steps, the positioning is judged to be effective, and the UAV successfully searches for the target, obtaining the cumulative reward. If the target escapes, the location fails, restart the search process, and the reward is empty and recalculated.

$$R_{b2} = \begin{cases} \sum \frac{(d_{find} - d_t^{A-D})}{d_{find}}, d_t^{A-D} \leq d_{find} \\ 0, d_t^{A-D} > d_{find} \end{cases} \quad (8)$$

Flight Altitude Reward. The reward R_{b3} for fixed height patrol is limited by safe height. Because the detection distance of UAVs in a preset mission is limited, when this item is not added, a UAV will reduce its flight height to the limit in exchange for the maximum ground detection coverage. Therefore, this item is added to limit the flying height h_{A-G} of UAVs, $h_{A-G} = 800$, so as to encourage UAVs to keep patrolling and searching more actively.

$$R_{b3} = \begin{cases} 1 - \frac{h_{A-G}}{h_{safe}}, h_{A-G} < h_{safe} \\ 0, h_{A-G} \geq h_{safe} \end{cases} \quad (9)$$

4.3. Expert Rule Model

In order to fully stimulate the intelligent emergence of the mission UAV swarm, expert strategy is considered to control ground targets. Similar to the design of the reward function, the expert rules were also crafted with battlefield rules in mind. We employed the artificial potential field method to control the movement of ground targets. In this method, ground targets are treated as points within a potential field, and the relationships between targets and targets, as well as among targets themselves, are translated into potential field forces based on the battlefield situation.

There are two main expert strategies, namely, “nearby avoidance” and “collective collision avoidance”, which are transformed into potential field vectors to control the defensive UAV movement in turn. In Figure 4, there is a simple example of the evasion strategy for a ground target.

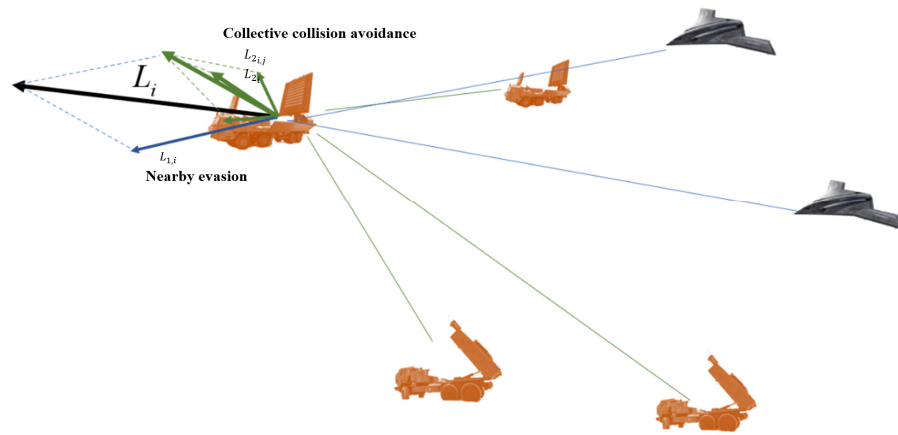


Figure 4. Target evasion strategy.

Nearby evasion L_1 means that when a ground target faces an invading swarm, it compares the relative position of each swarm with itself in sequence, aiming at removing the UAV with the strike path as quickly as possible. In the potential field, the potential field vector function is as follows:

$$L_{1,i} = \min \left[\frac{(|P_{i,j}| + \beta)^2 P_{i,j}}{|P_{i,j}|} \right], j \in 1, 2, \dots, N^A \quad (10)$$

where $L_{1,i}$ represents the escape vector of the i -th ground target, $P_{i,j}$ represents the relative position vector between the i -th ground target and the j -th mission UAV, and the relative position vector $P_{i,j}$ is transformed into the evasion vector $L_{1,i,j}$ of ground target i -th relative to UAV j -th, and N^{UAV} represents the number of mission UAV. The larger the value of β is, the more likely the ground target is to quickly escape from the mission UAV. In the simulation, the value of β was set to 0.8.

Collective collision avoidance L_2 means that when any target escapes, it keeps safe collision avoidance with all other ground targets. The potential field vector function is as follows:

$$L_{2,i,k} = - \frac{(|P_{i,k}| + \alpha)^2 P_{i,k}}{|P_{i,k}|}, i, k \in 1, 2, \dots, N^D, k \neq i \quad (11)$$

$$L_{2,i} = L_{2,i,1} + L_{2,i,2} + \dots + L_{2,i,k}$$

When the hit ground target moves, the relative position vector $P_{i,k}$ between itself and other targets is calculated in real time, and the relative position vector $P_{i,k}$ is transformed into the avoidance vector $L_{2,i,k}$ of ground target i -th relative to ground target k -th. N^{Target} represents the number of ground targets. Finally, all the collision avoidance vectors except i -th relative to itself form the collision avoidance control vector $L_{2,i}$; α is the collision avoidance inclination adjustment parameter of i -th. The larger the value of α is, the more conservative the ground targets are and the more inclined they are to maintain safe boundaries with each other. In the simulation, the value of α was set to 0.5.

The nearest-evasion potential field vector $L_{1,i}$ and the collective-collision-avoidance potential field vector $L_{2,i}$ together constitute the control vector L_i of the i -th target.

4.4. Simulation Results Analysis

Set up a UAV swarm to search and strike ground escape targets, with three UAVs and eight targets, respectively. The generation location and initial motion direction of ground escape targets are random. Within time T , after the mission UAV detects and eliminates all ground targets, it is determined that the mission swarm has won the turn.

Figure 5a illustrates the winning rate of the mission UAV in the 3v8 scenario, while Figure 5b shows the corresponding reward. Due to the uncertainty of the initial position, motion trajectory, and avoidance strategy of ground targets in each round of training, we used the results corresponding to every 100 rounds of training when drawing the win rate curve and reward value curve. They can demonstrate the effectiveness of the algorithm for UAV training. By analyzing Figure 5a, we can see that the winning rate increases with the number of rounds, reaching over 95% in 10,000 rounds. Although there are fluctuations, the independent adjusted winning rate is still high. The average reward also increases with the number of training rounds.

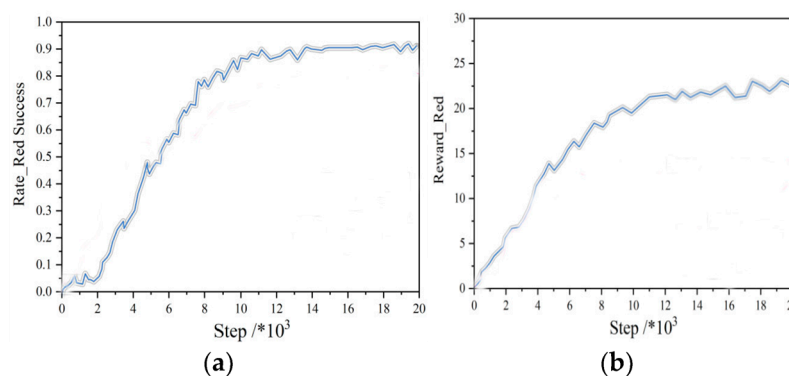


Figure 5. Winning rate and reward curve. (a) Winning rate; (b) Reward.

Figure 6 shows the strike process of the mission UAVs on the ground targets in the 3v8 scenario. The red dots represent the mission UAVs and their trajectories, and the blue dots represent the ground targets and their trajectories. When the mission UAV strikes the ground target, the blue dot will disappear and only the trajectory of the ground target will remain. According to the different ground target distribution and trajectory, the UAV swarm can independently complete the mission planning.

After training, the agent is demonstrated by the expert system. At this time, the UAV swarm shows a higher level of intelligence, and has independently evolved the patrol strike tactics. Three UAVs deploy their formation from the initial position, travel in a straight queue in the unknown area, and complete the locking strike after finding the target. Based on the new target position, plan the strike path. The three aircraft cooperated with each other to complete the attack on all eight targets within the specified time.

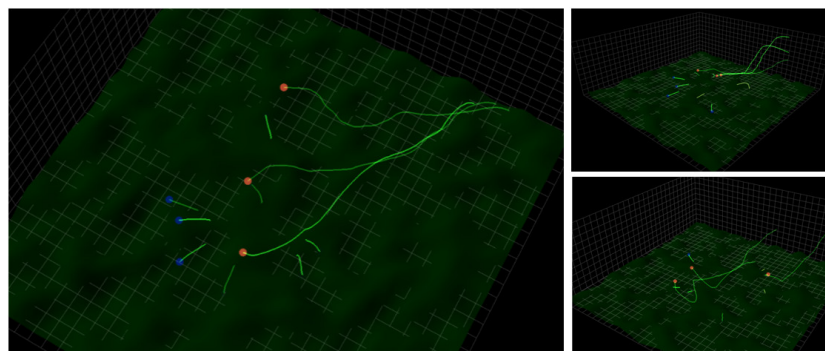


Figure 6. Confrontation trajectory.

5. Conclusions

This article successfully realizes the simulation scenario and architectural design for a UAV swarm executing SEAD tasks. In this process, the action strategy of the unmanned aerial vehicle fleet is dynamically generated by reinforcement learning algorithms, while the defense measures for ground targets are controlled by expert rule systems. The simulation results show that the proposed simulation scenario and system architecture are technically feasible, and the artificial intelligence algorithm has demonstrated significant effectiveness in SEAD tasks.

Although preliminary results have been achieved, there is still room for further expansions and improvements in the construction of simulation scenarios and the optimization of algorithms. For example, we can add more complex factors such as changing environmental conditions, more advanced enemy defense systems, and more complex collaboration mechanisms within the drone fleet to improve the realism and challenge of the simulation. In addition, reinforcement learning algorithms themselves also need further optimization to improve their adaptability and robustness in dynamic and uncertain environments. To further improve the adaptability and intelligence level of UAV swarms, Federated Learning and Multi Agent Reinforcement Learning (MARL) techniques can be introduced. Federated learning can achieve knowledge sharing and model updating among drones without sharing raw data, thereby improving the overall intelligence level of the swarm. Multi agent reinforcement learning can better handle the collaborative and competitive relationships between drones, achieving more efficient cluster decision-making.

Through these studies, we aim to further enhance the performance of UAV swarms in SEAD missions, thereby providing robust technical support for future unmanned combat systems.

Author Contributions: Conceptualization, Q.J.; Methodology, Q.J.; Software, Q.J., Y.Y. and Y.D.; Validation, Q.J., Y.Y. and Y.D.; Formal analysis, Y.Y. and Y.D.; Investigation, Z.Y.; Resources, Z.Y. and H.C.; Data curation, H.C.; Writing—original draft, B.W. and X.M.; Writing—review and editing, B.W. and X.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

DURC Statement: Current research is limited to the UAV Swarms control, which is beneficial and does not pose a threat to public health or national security. Authors acknowledge the dual-use potential of the research involving military application possibility and confirm that all necessary precautions have been taken to prevent potential misuse. As an ethical responsibility, authors strictly adhere to relevant national and international laws about DURC. Authors advocate for responsible

deployment, ethical considerations, regulatory compliance, and transparent reporting to mitigate misuse risks and foster beneficial outcomes.

Conflicts of Interest: The authors declare no conflicts of interest. Author Zequan Yang was employed by the company Wasion Group Co. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wu, Z.; Hu, B. Swarm rounding up method of UAV based on situation cognition. *J. Beijing Univ. Aeronaut. Astronaut.* **2021**, *47*, 424–430.
2. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; Mordatch, I. Multi-Agent Actor-Critic for mixed cooperative-competitive environments. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1706.02275.
3. Ma, S.; Zhang, H.; Yang, G. Target threat level assessment based on cloud model under fuzzy and uncertain conditions in air combat simulation. *Aerosp. Sci. Technol.* **2017**, *67*, 49–53.
4. McCune, R.; Purta, R.; Dobski, M.; Jaworski, A.; Madey, G.; Madey, A.; Wei, Y.; Blake, M.B. Investigations of DDDAS for command and control of UAV swarms with agent-based modeling. In Proceedings of the 2013 Winter Simulations Conference (WSC), Washington, DC, USA, 8–11 December 2013; pp. 1467–1478.
5. Qi, B.; Zhao, K.; Wang, X.; Rong, X. System Identification Method for Small Unmanned Helicopter Based on Improved Particle Swarm Optimization. *J. Bionic Eng.* **2016**, *13*, 504–514.
6. Radmanesh, M.; Kumar, M.; Sarim, M. Grey wolf optimization based sense and avoid algorithm in a Bayesian framework for multiple UAV path planning in an uncertain environment. *Aerosp. Sci. Technol.* **2018**, *77*, 168–179. [[CrossRef](#)]
7. Ren, Z.; Zhang, D.; Tang, S.; Xiong, W.; Yang, S.-h. Cooperative maneuver decision making for multi-UAV air combat based on incomplete information dynamic game. *Def. Technol.* **2022**, *27*, 308–317. [[CrossRef](#)]
8. Strickland, L.; Day, M.A.; DeMarco, K.J.; Squires, E.; Pippin, C. Responding to unmanned aerial swarm saturation attacks with autonomous counter-swarms. In Proceedings of the Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX, Orlando, FL, USA, 15–19 April 2018.
9. Vásárhelyi, G.; Virágh, C.; Somorjai, G.; Nepusz, T.; Eiben, A.E.; Vicsek, T. Optimized flocking of autonomous drones in confined environments. *Sci. Robot.* **2018**, *3*, eaat3536. [[CrossRef](#)]
10. Mao, H.; Zhang, Z.; Xiao, Z.; Gong, Z.; Ni, Y. Learning agent communication under limited bandwidth by message pruning. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 5142–5149.
11. Qu, C.; Gai, W.; Zhang, J.; Zhong, M. A novel hybrid grey wolf optimizer algorithm for unmanned aerial vehicle (UAV) path planning. *Knowl.-Based Syst.* **2020**, *194*, 105530. [[CrossRef](#)]
12. Qu, C.; Gai, W.; Zhong, M.; Zhang, J. A novel reinforcement learning based grey wolf optimizer algorithm for unmanned aerial vehicles (UAVs) path planning. *Appl. Soft Comput.* **2020**, *89*, 106099. [[CrossRef](#)]
13. Seong, M.; Jo, O.; Shin, K. Multi-UAV trajectory optimizer: A sustainable system for wireless data harvesting with deep reinforcement learning. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105891. [[CrossRef](#)]
14. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Hassabis, D. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv* **2017**, arXiv:1712.01815.
15. Soleyman, S.; Khosla, D. *Multi-Agent Mission Planning with Reinforcement Learning*; HRL Laboratories, LLC: Malibu, CA, USA, 2021.
16. Tang, C.; Lai, Y.C. Deep Reinforcement Learning Automatic Landing Control of Fixed-Wing Aircraft Using Deep Deterministic Policy Gradient. In Proceedings of the 2020 International Conference on Unmanned Aircraft Systems (ICUAS), Athens, Greece, 1–4 September 2020.
17. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Silver, D. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **2019**, *575*, 350–354. [[CrossRef](#)]
18. Wang, T.; Li, M.; Zhang, M.Y. Cooperative coverage reconnaissance of Multi-uav. In Proceedings of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 12–14 June 2020.
19. Wang, X.; Yang, Y.; Wang, D.; Zhang, Z. Mission-oriented cooperative 3D path planning for modular solar-powered aircraft with energy optimization. *Chin. J. Aeronaut.* **2022**, *35*, 98–109. [[CrossRef](#)]
20. Wang, X.-W.; Peng, H.-J.; Liu, J.; Dong, X.-Z.; Zhao, X.-D.; Lu, C. Optimal control based coordinated taxiing path planning and tracking for multiple carrier aircraft on flight deck. *Def. Technol.* **2022**, *18*, 238–248. [[CrossRef](#)]
21. Su, A.; Hou, F.; Hong, Y. Heterogeneous Policy Network Reinforcement Learning for UAV Swarm Confrontation. In Proceedings of the 2024 China Automation Congress (CAC), Qingdao, China, 1–3 November 2024; pp. 722–727. [[CrossRef](#)]

22. Huo, W.; Li, J. Cooperative UAV Maneuver Decision-Making Based on Multi-Agent Reinforcement Learning. In Proceedings of the 2024 4th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Guangzhou, China, 8–10 November 2024; pp. 475–479. [\[CrossRef\]](#)
23. Feng, X.; Gao, R.; Tang, F.; Luo, J. Multi-UAV Collaborative Reconnaissance Based on Multi Agent Deep Reinforcement Learning. In Proceedings of the 2024 10th International Conference on Big Data and Information Analytics (BigDIA), Chiang Mai, Thailand, 25–28 October 2024; pp. 65–70. [\[CrossRef\]](#)
24. Cabra, K.; Delamer, J.-A.; Rabbath, C.-A.; Lechevin, N.; Williams, C.; Givigi, S. Improved Learning in Multi-Agent Pursuer-Evader UAV Scenarios via Mechanism Design and Deep Reinforcement Learning. In Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS), Chania, Crete, Greece, 4–7 June 2024; pp. 144–151. [\[CrossRef\]](#)
25. Iqbal, W.; Li, B.; Rouhbakhshmeghraz, A. Research on Informative Path Planning Using Deep Reinforcement learning. In Proceedings of the 2024 International Conference on Cyber-Physical Social Intelligence (ICCSI), Doha, Qatar, 8–12 November 2024; pp. 1–6. [\[CrossRef\]](#)
26. Xing, X.; Zhou, Z.; Li, Y.; Xiao, B.; Xun, Y. Multi-UAV Adaptive Cooperative Formation Trajectory Planning Based on an Improved MATD3 Algorithm of Deep Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2024**, *73*, 12484–12499. [\[CrossRef\]](#)
27. Zhang, C.; Wu, Z.; Li, Z.; Xu, H.; Xue, Z.; Qian, R. Multi-agent Reinforcement Learning-Based UAV Swarm Confrontation: Integrating QMIX Algorithm with Artificial Potential Field Method. In Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, 6–10 October 2024; pp. 161–166. [\[CrossRef\]](#)
28. Yang, J.; Yang, X.; Yu, T. Multi-Unmanned Aerial Vehicle Confrontation in Intelligent Air Combat: A Multi-Agent Deep Reinforcement Learning Approach. *Drones* **2024**, *8*, 382. [\[CrossRef\]](#)
29. Deng, X.; Dong, Z.; Ding, J. Uav Confrontation and Evolutionary Upgrade Based on Multi-Agent Reinforcement Learning. *Drones* **2024**, *8*, 368. [\[CrossRef\]](#)
30. Aschu, D.; Peter, R.; Karaf, S.; Fedoseev, A.; Tsetserukou, D. Marlander: A Local Path Planning for Drone Swarms Using Multiagent Deep Reinforcement Learning. In Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, 6–10 October 2024. [\[CrossRef\]](#)
31. Kumari, N.; Lee, K.; Barca, J.C.; Ranaweera, C. Towards Reliable Identification and Tracking of Drones Within a Swarm. *J. Intell. Robot. Syst.* **2024**, *110*, 84. [\[CrossRef\]](#)
32. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [\[CrossRef\]](#)
33. Kobayashi, T.; Ilboudo, W. T-soft update of target network for deep reinforcement learning. *Neural Netw.* **2021**, *136*, 63–71. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Liu, Y.; Liu, H.; Tian, Y.; Sun, C. Reinforcement learning based two-level control framework of UAV swarm for cooperative persistent surveillance in an unknown urban area. *Aerosp. Sci. Technol.* **2020**, *98*, 105671. [\[CrossRef\]](#)
35. Xuan, S.; Ke, L. Study on Attack-Defense countermeasure of UAV swarms based on multi-agent reinforcement learning. *Radio Eng.* **2021**, *51*, 360–366. (In Chinese)
36. Liu, F.; Wei, R.; Ding, C.; Jiang, L.; Li, T. Design of Att-MADDPG hunting control method for multi-UAV cooperation. *J. Air Force Eng. Univ. (Nat. Sci. Ed.)* **2021**, *22*, 9–14.
37. Liu, Z.; Li, Y.; Wu, Y. Multiple UAV formations delivery task planning based on a distributed adaptive algorithm. *J. Frankl. Inst.* **2023**, *360*, 3047–3076. [\[CrossRef\]](#)
38. Qie, H.; Shi, D.; Shen, T.; Xu, X.; Li, Y.; Wang, L. Joint Optimization of Multi-UAV Target Assignment and Path Planning Based on Multi-Agent Reinforcement Learning. *IEEE Access* **2019**, *7*, 146264–146272. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.