Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7726338, 10 pages
https://doi.org/10.1155/2022/7726338

Research Article

Optimization Design of Multi-UAV Communication Network Based on Reinforcement Learning

Zhengyang Cao [1,2,3]

[1]*State Key Laboratory of Strength and Vibration for Mechanic Structures, School of Aerospace Engineering, Xi'an Jiaotong University, Xi'an, 710049 Shaanxi, China*
[2]*Shaanxi Key Laboratory of Service Environment and Control for Flight Vehicles, Xi'an Jiaotong University, Xi'an, 710049 Shaanxi, China*
[3]*Xi'an ASN UAV Technology Co., Ltd., Xi'an, 710065 Shaanxi, China*

Correspondence should be addressed to Zhengyang Cao; caozhengyang@stu.xjtu.edu.cn

Received 19 July 2022; Revised 19 August 2022; Accepted 26 August 2022; Published 9 September 2022

Academic Editor: Kapil Sharma

In recent years, due to the application of high-definition video codec technology, high-precision satellite navigation technology, mobile base station positioning technology, and broadband technology, the performance of UAVs has been greatly improved. In the military field, drones have become an important weapon alongside missiles on the battlefield. In the future, military drones will perform strategic missions such as battlefield reconnaissance and long-range destruction. Outside the military field, DJI's Zenmuse series drones are used for filming, MG series drones are used for pesticide spraying, and Beijing Zhonghangzhi unmanned helicopters are used for geological surveys, precise inspection of power lines, and maritime law enforcement. With the continuous improvement of technical specifications, UAV communication technology requires further research and development. This paper has conducted research experiments on the optimization of multi-UAV communication network based on reinforcement learning. The experimental data show that it is marked as the AoI value corresponding to the completion of a certain self-task. It can be seen that the final AoI of the communication trajectory of reinforcement learning is 115, and the AoI greedy strategy finally obtains AoI of 140 seconds, achieving about 18% of the total AoI reduce, which effectively improve the performance of the system. From the above data, the research of reinforcement learning method has great benefits for the development of UAV communication.

1. Introduction

In the 100 years since the first drone was invented in 1917, the drone industry has grown rapidly, from a single flight to a multibillion dollar output. With the continuous development of military high-tech, unmanned aerial vehicles (UAVs) have emerged in modern warfare due to their low cost, easy maintenance, reduced casualties, and the ability to perform tasks in a variety of complex and harsh environments and gradually attracted the attention of the world's military powers. In future wars, there will be less and less direct human participation and more confrontation between unmanned military equipment. In order to adapt to the situation of informatized warfare and quickly collect and process real-time and accurate intelligence information, UAVs have been widely used in modern combat command for reconnaissance, surveillance, and other tasks, especially the reconnaissance of enemy areas and important targets. In the networked environment, the mode of warfare has undergone tremendous changes. As an important part of the Internet era, "sharing" also plays an important role in networked warfare. In a networked environment, the "distance" between troops is approximately zero, and information resources and services can be easily shared. Under the traditional platform-centric combat mode, in order to obtain the information they need, each unit needs to configure corresponding equipment, such as unmanned aerial vehicles and ground stations. Under the condition of underdeveloped intelligence sharing, various combat units often have the problem of duplication of resource allocation and

duplication of information collection. For the UAV troops, the repeated deployment of a large number of UAVs will also lead to problems in airspace security, electromagnetics, and communication management, which will greatly limit the effective use of UAV system resources.

With the rapid development of UAV technology, higher requirements are put forward for the simultaneous use of multiple types and numbers of UAVs in the same war. Under the traditional use of UAVs, the managers and users are the same unit. The result is that the tasks between different units cannot be efficiently coordinated and unified, and the information obtained cannot be shared in time, which is extremely unfavorable for joint operations and high-level commanders' war decisions. In addition, under the complex informationized battlefield conditions in the future, due to the different coverages of battlefield communication links, different mission requirements, and different mission capabilities of various UAVs, a single type of UAV system obviously cannot meet the needs of the battlefield. By establishing a generalized monitoring and control network, the UAVs, ground stations, decision-making systems, and control systems scattered in the battlefield environment are connected into an organic whole. Through this network, personnel at all levels can timely understand the battlefield situation, share intelligence information, and achieve seamless command, control, and communication, which is obviously beneficial to improving efficiency.

After the optimization of multi-UAV communication network based on reinforcement learning in this paper, the data shows that the AoI of the communication trajectory of reinforcement learning is 118 and 157 when the number of ground nodes is 6 and 10, respectively; the AoI of the greedy algorithm is 139 and 199, respectively; that is, the AoI that can be obtained by the reinforcement learning algorithm is smaller, and with the increase of the number of ground nodes, its advantages are more obvious. From the above data, it can be seen that the multi-UAV communication network optimization research experiment of reinforcement learning is of great significance for promoting the development of the current multi-UAV communication network.

2. Related Work

This paper studies some technologies of multi-UAV communication network, which can be fully applied to the research in this field. The main research goal of Amorim et al. is to obtain the path loss index and shadow model of wireless channels between airborne UAVs and cellular networks [1]. Fawaz et al. improve the performance of existing relay-assisted FSO systems by relaxing these two highly restrictive assumptions by integrating UAVs as buffer-assisted mobile relays into traditional relay-assisted FSO systems [2]. Mamaghani and Hong studied the problem of maximizing the average secrecy rate for UAV wireless communication systems, where UAVs are used to transmit confidential information to ground destinations in the presence of ground passive eavesdroppers [3]. Wang et al. studied the average packet error probability and effective throughput of control links in UAV communication, where a ground central station sends control signals to UAVs that require ultra-reliable low-latency communication [4]. Liu et al. designed a recurrent neural network based on long short-term memory for UAV position prediction [5]. These methods provide some references for our research, but due to the short time and small sample size of the relevant research, they have not been recognized by the public.

Based on reinforcement learning, we have reviewed the following related materials to optimize the research on multi-UAV communication networks. Gershman and Daw review the major advances in the psychology and neuroscience of reinforcement learning over the past two decades through comprehensive experimental studies on simple learning and decision-making tasks [6]. Li et al. tried to introduce qualitative rules into reinforcement learning and represented these rules through a cloud inference model [7]. Peng et al. showed that reinforcement learning methods can be adapted to learn robust control policies capable of imitating a wide range of example motion clips [8]. Sallab et al. proposed a framework for autonomous driving using deep reinforcement learning, gave a brief introduction to deep reinforcement learning, and then described the proposed framework [9]. Ying et al. proposed a new deep reinforcement learning method, an advanced reinforcement learning algorithm that uses a deep Q-network to approximate the Q-valued action function [10]. He et al. proposed a new deep reinforcement learning method, and the simulation results under different system parameters showed its effectiveness [11]. These methods provide sufficient literature support for our study of multi-UAV communication network optimization with reinforcement learning.

3. Overview of Reinforcement Learning and UAV Communication

The characteristics of UAV battery power supply make power consumption a factor that has to be considered in the UAV support network. Therefore, this paper takes the development of UAVs to a higher level by studying reinforcement learning to optimize the UAV communication network.

3.1. Overview of Reinforcement Learning. Reinforcement learning is an important machine learning method that has been at the forefront of intelligent control and artificial intelligence research in recent years [12]. Among various learning methods, reinforcement learning has the ability to adapt to complex systems and self-training. It approaches optimal control policies through trial-and-error learning that interacts with the environment, a learning mechanism that has been successfully applied to nonlinear control, artificial intelligence for solving complex problems, robot control, optimization, and planning.

Reinforcement learning (RL) first appeared in the 1950s as learning by trial and error in a dynamic environment. The agent method does not calculate the task performance of the agent but guides the agent through rewards and

punishments, which is now becoming an important branch of machine learning and artificial intelligence [13].

There are two strategies to solve the problem of reinforcement learning: first, searching the action space to find actions that work better in the environment, this approach has been used for example in genetic algorithms and genetic programming; second, using statistical methods and dynamic programming techniques to assess the utility of actions in the state of the world. RL is based on autonomous learning by exploring an unknown environment, whereby an agent acquires knowledge about the environment to optimize the course of action.

Reinforcement learning is a machine learning method that learns by interacting with the environment and taking feedback from the environment as input. The basic idea is to learn by trial and error, matching environmental states and actions, and the agent (learner or decider) interacts with the environment over time and assumes that the cumulative reward is maximized. In reinforcement learning, cues are provided by the environment with the purpose of providing some sort of evaluation of how good or bad the chosen action is, rather than telling the agent how to choose the right action. Since the external environment provides very little information for the agent, it must rely on the experience of interacting with the environment to learn independently. Therefore, the agent uses the evaluation signals from the external environment to optimize its decision and find the best behavior policy.

Reinforcement learning is a method that focuses more on learning through interaction and decision-making than other machine learning methods. In reinforcement learning, the agent must figure out, through trial and error, which activity brings the greatest immediate reward. The action not only affects the immediate reward but is also important for all subsequent rewards. With the deepening of research, reinforcement learning can be divided into the following branches: logic-based reinforcement learning, hierarchical learning, multiagent learning, POMDP learning, etc. [14]. The agent environment structure diagram is shown in Figure 1.

Reinforcement learning problems can be viewed as a framework for learning directly from interactions and goal achievement. Learners and decision-makers are called agents, and all other elements except agents are called environments. These interactions are continuous, the agent chooses actions, the environment reacts to these actions and generates new situations for the agent, and the environment returns reward values. Through the above process, the agent learns how to optimize its behavioral policy in the environment in order to maximize the cumulative reward over time.

In addition to the two components of the environment and the learning agent, a reinforcement learning system also needs four other main subelements: the strategy, the reward function (or cost function), the value function, and the optional environment model. The strategy is to define the learning mode or the action behavior mode displayed by the learning agent in a given time. The reward function, defining the goal of a reinforcement learning problem, is

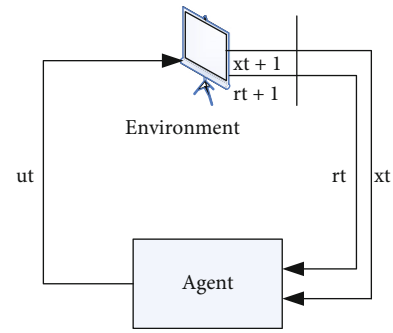

Figure 1: Agent environment structure diagram.

its main task. The value function refers to the accumulation of the expected rewards of the learning agent from the current state to the final state. The environment model (optional), before the learning agent actually has not experienced the future action, takes the possible future situation into account through the model and then makes planning decisions for future action selection [15]. The schematic diagram of the relationship between the four core parts of reinforcement learning is shown in Figure 2:

Early reinforcement learning systems were a trial-and-error learning method that was almost the opposite of planning decision-making methods. However, adding the planning method of the model and state space to the reinforcement learning makes the reinforcement learning closely related to the dynamic programming method, so that the reinforcement learning method gradually becomes clear. It can be seen that the main goal of reinforcement learning is to obtain the best strategy through continuous improvement of the strategy to achieve the final goal. The development route of reinforcement learning is shown in Figure 3.

Early reinforcement learning algorithms learned through trial-and-error learning to achieve their goals. With the development of reinforcement learning, dynamic programming and optimal control algorithms and time difference (TD) learning algorithms appear in turn. These three main lines eventually constitute the main framework of modern reinforcement learning algorithms. From a modeling perspective, reinforcement learning falls into two categories: model-based learning algorithms and model-free methods. The former extracts empirical knowledge from the environment to establish a learning model and then determines the optimal strategy according to the model; the latter selects strategies through direct interaction with the environment. Commonly used model-free learning algorithms are as follows: AHC, temporal difference (TD), and Q-learning. The model-free method has the characteristics of iterative calculation, and its calculation amount is small, and because it cannot make full use of the prior knowledge, it is not as good as the model method in terms of convergence speed. The following mainly introduces several typical RL methods [16].

Dynamic programming methods use a value function to search for good policies and are suitable for solving large problems. If the environment is a finite Markov set, and

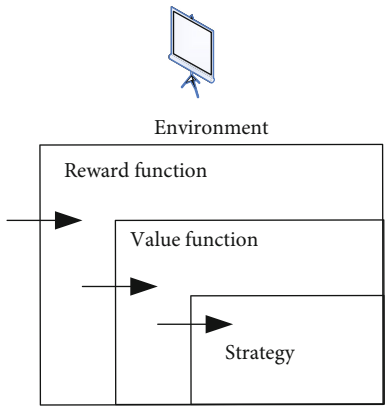

Figure 2: Schematic diagram of the relationship between the four core parts of reinforcement learning.

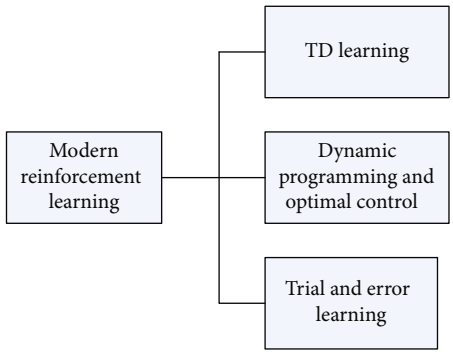

Figure 3: The development route of reinforcement learning.

for each policy, information about the dynamic environment is fully known, then the value function is given by

$$V^*(s) = \max_b \left\{ r_s(b) + \gamma \sum_{b \in B} P_{ss'}[b] V^{\pi^*}\left(s'\right) \right\}. \quad (1)$$

The dynamic programming algorithm requires an immediate reward value r and a state transition function P. In other words, this problem can only be solved if the environment model is known.

Monte Carlo is a model-free algorithm that iteratively learns by computing state, action, and value functions. The algorithm differs from traditional Q-learning in that it is based on a multistage reward averaging mechanism. Therefore, the algorithm converges too slowly and is less used in modern artificial intelligence. Its iterative formula is as follows:

$$V(s_t) = V(s_{t+1}) + \alpha[R_t - V(s_t)]. \quad (2)$$

Unlike dynamic programming, this algorithm does not require modeling of the environment or limited information about the environment, but it also has a slower convergence rate because it features learning average rewards.

The TD learning algorithm is one of the most important algorithms in the reinforcement learning method. It is a combination of the above two methods. The iterative formula of the TD (0) algorithm is

$$V(s_{t+1}) = V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)). \quad (3)$$

The TD algorithm was proposed in 1998, and it was proved that the TD algorithm must converge when the learning rate satisfies certain conditions. However, the convergence rate of the TD algorithm is slow because the agent only changes the value function estimates of neighboring states in each iteration [17]. The effective method is that when the agent obtains the instantaneous reward value, it can take any step backwards, which is the so-called multistep TD learning algorithm. The convergence rate of the TD (μ) learning algorithm is significantly improved by the following iterative formula:

$$V(s_t) = V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))e(s). \quad (4)$$

Among them, $e(s)$ is defined as the qualification trace of state s, which can be calculated in the following ways:

$$e(s) = \begin{cases} \gamma\mu e(s) + 1, \text{if } s = s_t, \\ \gamma\mu e(s), \text{otherwise}. \end{cases} \quad (5)$$

The Q-learning algorithm is a model-independent reinforcement learning algorithm proposed in 1989. Its essence is an off-policy TD learning algorithm. Different from the TD algorithm, the state is used in the Q-learning update iteration, and the reward value $Q(s, a)$ of the action pair is used as the estimation function, instead of the reward value $V(s)$ of the state in the TD as the estimation function. The iterative form of the Q-learning algorithm is as follows:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha_t \left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right). \quad (6)$$

Among them, a_t is the learning rate of the agent at time t, and γ is the discount factor. Under the nondeterministic Mahalanobis decision process, the learning rate of Q-learning also satisfies the following two conditions:

$$\sum_{l=1}^{\infty} \alpha_l = \infty, \quad (7)$$

$$\sum_{l=1}^{\infty} \alpha_l^2 < \infty. \quad (8)$$

When $1 \longrightarrow \infty$, Qk (s, a) will receive $Q^*(s, a)$ with probability 1.

The Sarsa learning algorithm is proposed as an improved network form of the Q-learning algorithm, which still uses Q -value iteration. The iterative calculation formula for the

value function of the Markov decision process in the Sarsa learning algorithm is

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)). \tag{9}$$

The main difference between Sarsa algorithm and Q algorithm is that the Q algorithm uses the maximum operator of the next action value function to adjust the action value function estimate of the current step, while the Sarsa algorithm only uses the actual trajectory data in the Mahalanobis decision process to adjust the action value function estimation [18].

The Markov decision chain is regarded as a basic multistep prediction model, which is the mathematical basis for reinforcement learning research. For the Markov decision chain, setting the state space as S, then

$$P_{ij}^{y} = P\{M_y = j | M_0 = i\}, \forall i, j \in S, y \in Y. \tag{10}$$

It is called the y-step transition probability of the Markov decision chain starting from state i and transferring to j after y-steps. Letting P be a matrix composed of all elements, there are

$$P^{(y)} = \left(P_{ij}^{y}\right). \tag{11}$$

The above formula is the y-step transition probability matrix of the Markov decision chain, and the one-step transition probability matrix is P.

Reinforcement learning methods first mathematically model a class of stochastic problems with discrete states and discrete time. In practice, the Markov decision model is the most commonly used [19].

The discrete-time finite Markov decision process can be expressed as

$$\forall i, j \in S, a \in A, y \geq 0, \tag{12}$$

$$P\left(M_{y+1} = j \middle| M_y = i, A_y = a, M_{y-1}, A_{y-1}, \cdots, M_0, A_0\right), \tag{13}$$

$$P\left(M_{y+1} = j \middle| M_y = i, A_y = a\right) = P(i, a, j). \tag{14}$$

J is the objective function of decision optimization. The state transition probability P satisfies

$$\sum_{j \in S} P(i, s, j) = 1. \tag{15}$$

There are two main types of decision optimization objective function J of the Markov decision process. That is, the discounted total return target and the average expected return target, respectively, were shown in the following two equations:

$$J_d = E\left(\sum_{t=0}^{\infty} r_t\right), 0 < \gamma < 1, \tag{16}$$

$$J_a = \limsup_{Y \longrightarrow \infty} \frac{1}{Y} E\left(\sum_{t=0}^{Y-1} r_t\right). \tag{17}$$

These two decision optimization objective functions have been widely studied and applied in the field of dynamic programming, and a lot of research has also been done in reinforcement learning theory and algorithms, mainly for the total return discount objective function.

3.2. Overview of UAV Communication. The unmanned aerial vehicle (UAV) is a powered, radio-controlled, or autonomously programmed aircraft operated by an unmanned pilot. The first drones were developed by the British in the 1970s and were primarily used as target drones in the initial stages; after entering the 1960s, the research on drones focused on reconnaissance; since then, the development of drones has entered the era of demand traction; since the 1980s, the miniaturization of UAVs has become another main direction for the development of UAVs. Many small UAVs have been used in civilian applications due to their advantages of light weight, good concealment, and low price [20, 21].

Currently, the development and use of UAVs is on the rise worldwide, mainly due to modern military and civilian needs and technological developments. The use of military UAVs has expanded from traditional aerial surveillance, battlefield monitoring, and battlefield assessment to combat, ground attack, missile interception, and even air combat. UAVs not only support manned combat aircraft but also replace manned aircraft in many situations. Currently, UAV research focuses on high-altitude and unmanned combat aircraft.

In terms of military use, UAVs such as the Global Hawk and Predator in the United States and Heron and Hunter in Israel are all UAVs with relatively successful research and development and excellent parameter performance. UAVs for military purposes can be divided into the following: target drones, mainly to identify the flight status and attack process of various aircraft; unmanned reconnaissance aircraft, used to monitor the battlefield and provide various intelligence information for combat troops; decoy drones are mainly used to induce the enemy to turn on the radar to obtain radio wave information and attract enemy firepower; signal jamming drones are used for electromagnetic interference and electronic detection of the enemy; and unmanned combat drones are a combination of fighter jets and drones. They can usually carry small precision weapons and can attack and intercept missiles to achieve combat purposes. UAVs are mainly used for civilian use in forest fire fighting, communication relay, pesticide spraying, aerial photography of competitions, and meteorological detection.

Compared with manned aircraft, UAV has the following characteristics and advantages: simple structure, no traditional cockpit, and UAV is much smaller than manned aircraft; this safety feature is great, and there will be no accidents; the performance is good, and the pilot factor does not need to be considered when developing the UAV; concealment, compared to manned aircraft, the size and mirror

surface of drones are much smaller, coupled with unique and complex designs and concealment materials, making them much more stealthy and survivable; the cost is low, the cost of UAV is only one-tenth or even a few percent of manned aircraft; and it is convenient and flexible to take off and land and has a short fuselage [22, 23].

In order to maximize the role of a single UAV, expand the application field of UAVs and make UAVs safer and more reliable when performing various tasks such as aerial surveillance, reconnaissance, and combat, and a multi-UAV system is proposed. Several research activities related to UAVs have been carried out in many fields. In the field of joint control of multiple UAVs, the main aspects being studied are the command and control of multiple UAVs, trajectory planning, and multimission [24].

Due to the characteristics of strong network dynamics and individual control autonomy, the multi-UAV system requires the multi-UAV network to have automatic networking and adapt to the rapid changes of network topology. Mobile ad hoc network (MANET) becomes the network technology for multi-UAV systems. The "UAV Roadmap" and "UAV System Integration Roadmap" issued by the US military put forward the important role of UAVs in the future global information network and pointed out that UAV self-organizing network will be the main content of multi-UAV network research in the future. Based on the above research and analysis, it can be concluded that the usefulness of the information transmitted between multiple UAVs is highly dependent on the communication performance; so, communication plays a key role in UAV dynamics.

Due to its fast flight speed, cooperative autonomy, limited energy, and irregular topology changes, the multi-UAV system also puts forward higher requirements for the network technology it adopts. The characteristics of the multi-UAV network can be summarized as follows: without a central node, in order to increase the robustness of the UAV system, the overall structure of the multi-UAV network should be equivalent to a peer-to-peer network; self-organizing, multi-UAV systems should have the characteristics of rapid deployment and rapid combat; when some drones in the network fail and cannot continue to perform tasks, the network topology needs to be rebuilt to maintain normal network communication; so, the network needs to be self-healing; dynamically changing network topology, the external environment of the UAV is complex during the execution of the mission, and its movement track is generally executed according to the preset route; good QOS and high security, due to the characteristics of wireless communication and the influence of the unknown environment in which the drone performs the task, the data transmission between the drones needs to have high security.

Mobile autonomous network (MANET) is not a new technology, it has been used for more than 40 years, and the idea and concept of MANET was first proposed in the United States in 1968.

The protocol of MANET network consists of physical layer, link layer, network layer, and application layer. The operating environment of the MANET network is very different from that of the wired network; so, the technology chosen for the network is also very different, especially in the lower three layers of the network: the physical layer, the link layer, and the network layer. The corresponding relationship between the MANET protocol stack model and the OSI model is shown in Figure 4.

The physical layer is responsible for modulation, coding, transmission, and reception of wireless data. The communication layer is divided into a medium access layer (MAC) and a logical link control layer (LLC), which are responsible for regulating access to shared wireless channels and control of logical links. The network layer is a key feature of the metropolitan area network technology, which distinguishes it from other networks by its basic features. The network layer provides transport protocols, mobile communication algorithms, and dynamic single-path and multipath routing algorithms for the metropolitan area network. Network layer routing protocols usually meet the following requirements: distributed operation, loop-free routing, demand-driven routing, high security, and support for one-way communication.

UAV MANET has three typical applications: battlefield coverage, battlefield reinforcement, and extended applications. Battlefield coverage refers to deploying multiple UAVs to form a metropolitan area network in order to expand the reconnaissance capabilities of UAVs to achieve coverage of the entire battlefield; battlefield reinforcements are mainly used for long-range target reconnaissance. The distance from the base station to the target exceeds the communication range of a single UAV; so, multiple UAVs must be deployed to transmit target reconnaissance signals; the extended application diagram of the UAV MANET network is shown in Figure 5.

The UAV network is connected to the global information network through relay satellites or ground UAV control stations and can be used as a channel to receive and transmit information required for cyber warfare by transmitting reconnaissance signals and forwarding control instructions to the ground combat network.

4. Multi-UAV Communication

4.1. Multi-UAV for Reinforcement Learning. This section simulates the multi-UAV communication trajectory and analyzes the results to verify its effectiveness. The TensorFlow framework is used to build a reinforcement learning network, in which both the actor and critic networks are two-layer fully connected networks, and the rectified linear unit (ReLU) and the sigmoid function are used as activation functions.

Since there is no work to solve the joint problem of continuous communication and linking of multiple UAVs, the communication trajectory obtained by the strategy proposed in this section is compared with the communication trajectory obtained by the basic greedy strategy to verify its feasibility and effectiveness. The purpose of the compared greedy algorithm is to minimize the AoI, and each time the ground node with the smallest AoI is selected for service, the

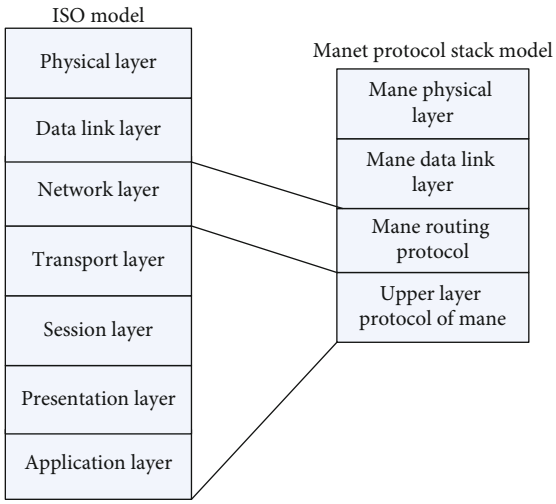

Figure 4: Correspondence between MANET protocol stack model and OSI model.

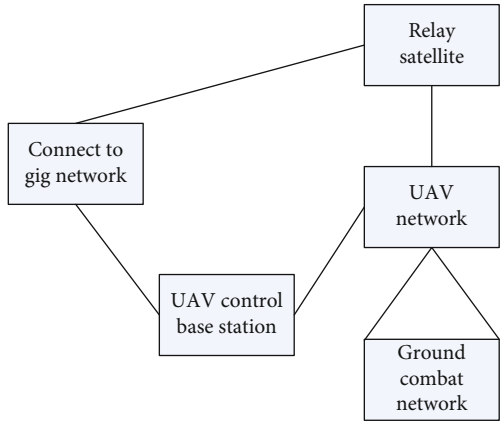

Figure 5: Schematic diagram of extended application of UAV MANET network.

communication trajectory is to fly to the selected node until the information transmission is completed.

In the simulation reinforcement learning communication strategy, when the task starts at $k = 0$, all nodes update the data once, then the ground nodes are updated twice when $k = 20$, the ground nodes are updated for the third time when $k = 40$, and the remaining nodes will not be updated again during the flight cycle. Since the flight time is required to reach the transmission range of the ground node, the communication trajectory obtained by this strategy has a trade-off between the flight time and the AoI. Jointly planning the trajectory of the drones working together and the drone link sequence make the AoI minimize.

The performance of the planned communication trajectory is analyzed. First, the simulation analysis of the time-dependent change process of the target performance AoI value during the task execution process is carried out. The comparison of the evolution process of AoI over time under different strategies is shown in Figure 6.

As can be seen from Figure 6, which is marked as the corresponding AoI value when a certain self-task is completed, it can be seen that the AoI finally reached by the communication trajectory of reinforcement learning is 115. The total AoI obtained by the AoI greedy strategy is 140 seconds, which reduces the total AoI by about 18% and effectively improves the performance of the system. This is because, compared with the greedy strategy, the communication trajectory through reinforcement learning can more reasonably plan the communication trajectory according to the information generation law of the century.

In order to further verify the performance of the proposed communication trajectory planning strategy, based on Monte Carlo simulation, the results obtained by 50 simulations were averaged, and the final Aol was further compared under different numbers of ground nodes. Assuming that the total amount of tasks remains unchanged, the total task amount is equally divided into different numbers of ground nodes to compare the final Aol value of the communication trajectory under the proposed reinforcement learning algorithm and the communication trajectory obtained by the Aol greedy algorithm. The final AoI comparison of different ground nodes is shown in Figure 7.

As can be seen from Figure 7, when the number of ground nodes is 6 and 10 for the communication trajectory of reinforcement learning, the AoI is 118 and 157, respectively; the greedy algorithm is 139 and 199; that is, the AoI that can be obtained by the reinforcement learning algorithm is smaller, and as the number of ground nodes increases, its advantages are more obvious. Therefore, in a more complicated situation with the increase of ground nodes, the strategy of simply selecting the nodes to be served according to the greedy strategy to determine the communication trajectory is not suitable. It is necessary to consider the collaborative work between UAVs as much as the reinforcement learning strategy and obtain better communication trajectories and link strategies through multiple offline learning.

4.2. Multi-UAV Route Planning. In this section, the communication route planning method based on the reinforcement learning algorithm is simulated and optimized. Two groups of simulation results are given before and after optimization. The parameters of the two groups of simulation experiments are different in the number of flight steps of the UAV, and other basic parameters are the same.

In simulation experiment 1, the initial speed direction of each UAV is the vertical boundary line pointing into the mission area. Assuming that the flying distance of the UAV after a fixed time interval is one step, when the number of flight steps in this experiment is 30 steps, the coverage rate of the six UAVs within 30 steps of flight changes as shown in Figure 8.

It can be seen from Figure 8 that with the increase of flight steps, the coverage of the mission area increases rapidly, the coverage of the task area is 0.76, 0.85, 0.97, 0.98, 0.98, 0.99, and 1, respectively, and the complete monitoring coverage of the task area is completed around the 15th step. It can be seen from the above simulation results that the

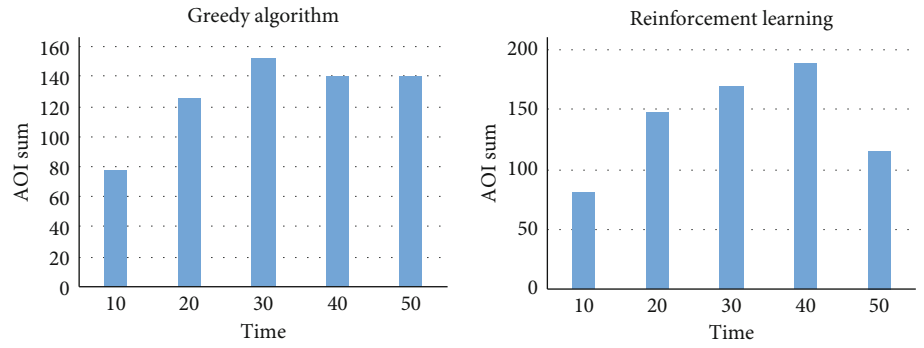

Figure 6: Comparison of the evolution of AoI over time under different strategies.

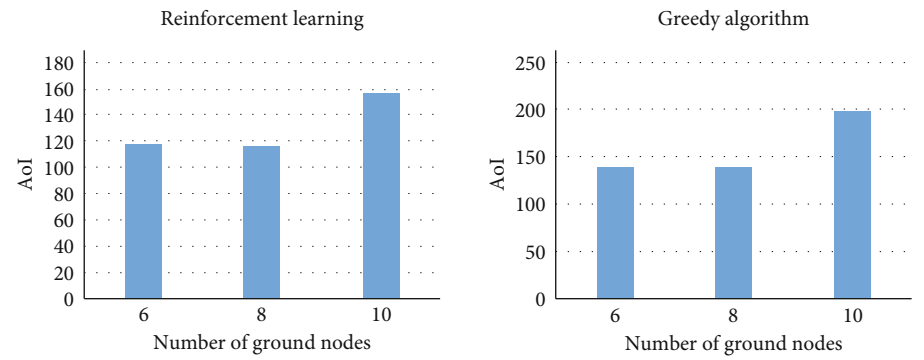

Figure 7: Final AoI comparison of different ground nodes.

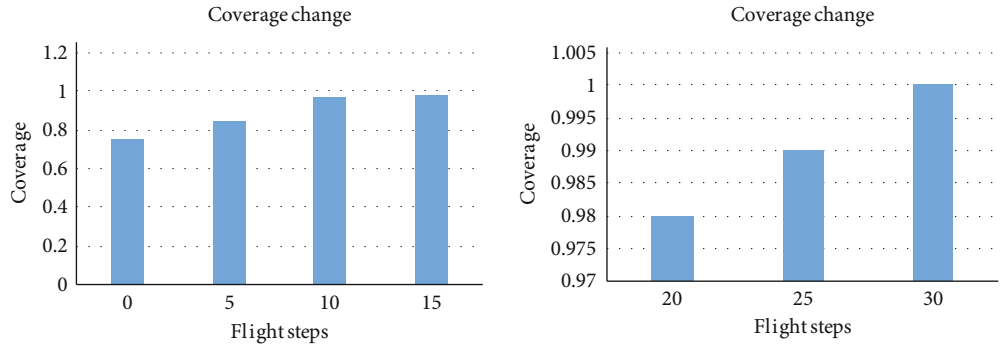

Figure 8: Coverage variation of 6 drones within 30 steps of flight.

route of multiple UAVs planned based on the reinforcement learning algorithm makes the percentage of surveillance coverage area of the UAV swarm sustainably maintained above 98% after convergence. It is proved that the flight of multiple UAVs along the route planned by this method can realize the maximum continuous monitoring of the designated target area by the UAV swarm.

In simulation experiment 2, the starting coordinates and initial speed directions of each UAV are the same as those in experiment 1, and the number of flight steps in this experiment is 300 steps. The coverage changes of 6 UAVs within 300 steps of flight are shown in Table 1.

It can be seen from Table 1 that the coverage rate of the six drones reached 0.970, 0.969, 0.970, 0.971, 0.972, and 0.973 at the lowest and 0.995, 0.996, 0.997, 0.998, 0.999, and 1 at the highest; when the drones fly longer and the number of flight steps is longer, the surveillance coverage of the target area by the drone group will fluctuate greatly. The onboard radar has an average coverage of the mission area of 98.1%; so, this will be improved on below.

Based on the above two simulation results, optimization analysis is carried out to demonstrate the effectiveness of the improved reinforcement learning method. The following two simulation experiments are introduced: the first

Table 1: Coverage change of 6 UAVs within 300 steps of flight.

Steps	50	100	150	200	250	300
1	0.977	0.970	0.990	0.973	0.968	0.995
2	0.978	0.971	0.991	0.974	0.969	0.996
3	0.979	0.972	0.992	0.975	0.970	0.997
4	0.980	0.973	0.993	0.976	0.971	0.998
5	0.981	0.974	0.994	0.978	0.972	0.999
6	0.982	0.975	0.995	0.979	0.973	1

Table 2: The optimized coverage of 6 UAVs within 300 steps of flight.

Steps	50	100	150	200	250	300
1	0.974	0.984	0.985	0.995	0.993	0.988
2	0.975	0.985	0.986	0.996	0.994	0.989
3	0.976	0.986	0.987	0.997	0.995	0.990
4	0.977	0.987	0.988	0.998	0.996	0.991
5	0.978	0.988	0.989	0.999	0.997	0.992
6	0.979	0.989	0.990	1	0.998	0.993

Table 3: Coverage of 6 UAVs within 300 steps of the flight with the reinforcement learning multistep method.

Steps	50	100	150	200	250	300
1	0.995	0.990	0.993	0.994	0.992	0.991
2	0.996	0.991	0.994	0.995	0.993	0.992
3	0.997	0.992	0.995	0.996	0.994	0.993
4	0.998	0.993	0.996	0.997	0.995	0.994
5	0.999	0.994	0.997	0.998	0.996	0.995
6	1	0.995	0.998	0.999	0.997	0.996

simulation experiment is mainly to compare with the method before optimization to test the effectiveness of the improved method; the second experiment is mainly to test whether the reinforcement learning method is still effective under different simulation parameters.

In simulation experiment one, because the main purpose is to compare with the experimental results before the improvement, the same basic parameters as the experiment before the improvement are used. The coverage changes of the six UAVs optimized by the reinforcement learning method within 300 steps of flight are shown in Table 2.

It can be seen from Table 2 that the optimized 6 UAVs fly within 300 steps, and the average coverage rate of all UAV airborne radars to the mission area during the entire flight process is 98.9%, which is higher than that before optimization. The coverage changes of the 6 UAVs in the reinforcement learning multistep method within 300 steps are shown in Table 3.

From Table 3, it can be concluded that the 6-plus UAV of the reinforcement learning multistep method has an average coverage rate of 99.5% of the mission area during the entire flight process within 300 steps.

Combining all the experimental data in this section, we can know that the two improved methods used later will improve the surveillance coverage of the target area by the UAV swarm, and the multistep method of reinforcement learning has the best effect.

5. Conclusions

UAVs play an important role in intelligence and surveillance missions, electronic countermeasures, firepower, airborne early warning, target designation and communications, and important auxiliary information systems due to their flexibility and versatility, as well as the advantage of not having to worry about loss. At the same time, due to the information and intelligence requirements for high-speed communication, personal communication, and military covert/counter-secret communication, the demand for UAVs has greatly increased in recent years, and the demand for UAVs even exceeds the actual capabilities of existing systems. In the field of resource utilization and optimization, the field of optimizing UAV training has attracted extensive attention from international researchers. Therefore, it is very

important to study the use of reinforcement learning to optimize the UAV communication network, achieve higher throughput under the premise of ensuring communication quality, and avoid channel congestion and mutual interference in the UAV communication system. This paper puts forward practical suggestions for the development of multi-UAV communication through the research on the multi-UAV communication network of reinforcement learning, which has important theoretical and practical significance.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Conflicts of Interest

The author declares no potential competing interests in this paper.

Authors' Contributions

And all authors have seen the manuscript and approved to submit to your journal.

References

[1] R. Amorim, H. Nguyen, P. Mogensen, I. Z. Kovacs, J. Wigard, and T. B. Sorensen, "Radio channel modeling for UAV communication over cellular networks," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 514–517, 2017.

[2] W. Fawaz, C. Abou-Rjeily, and C. Assi, "UAV-aided cooperation for FSO communication systems," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 70–75, 2018.

[3] M. T. Mamaghani and Y. Hong, "Joint trajectory and power allocation design for secure artificial noise aided UAV communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2850–2855, 2021.

[4] K. Wang, C. Pan, H. Ren, W. Xu, L. Zhang, and A. Nallanathan, "Packet error probability and effective throughput for ultra-reliable and low-latency UAV communications," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 73–84, 2021.

[5] C. Liu, W. Yuan, Z. Wei, X. Liu, and D. W. K. Ng, "Location-aware predictive beamforming for UAV communications: a deep learning approach," *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 668–672, 2021.

[6] S. J. Gershman and N. D. Daw, "Reinforcement learning and episodic memory in humans and animals: an integrative framework," *Annual Review of Psychology*, vol. 68, no. 1, pp. 101–128, 2017.

[7] C. Li, L. Cao, and X. Chen, "Cloud reasoning model-based exploration for deep reinforcement learning," *Dianzi Yu Xinxi Xuebao/journal of Electronics & Information Technology*, vol. 40, no. 1, pp. 244–248, 2018.

[8] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "DeepMimic," *ACM Transactions on Graphics*, vol. 37, no. 4, p. 1, 2018.

[9] A. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 29, no. 19, pp. 70–76, 2017.

[10] H. Ying, Z. Zheng, and F. R. Yu, "Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10433–10445, 2017.

[11] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: a deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, 2018.

[12] G. Li, N. Cao, P. Zhu et al., "Towards smart transportation system," *Journal of Organizational and End User Computing*, vol. 33, no. 3, pp. 35–49, 2021.

[13] Y. He, F. R. Yu, N. Zhao, V. C. M. Leung, and H. Yin, "Software-defined networks with mobile edge computing and caching for smart cities: a big data deep reinforcement learning approach," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 31–37, 2017.

[14] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Transactions on Neural Networks*, vol. 29, no. 6, pp. 2063–2079, 2018.

[15] M. Cutler, T. J. Walsh, and J. P. How, "Real-world reinforcement learning via multifidelity simulators," *IEEE Transactions on Robotics*, vol. 31, no. 3, pp. 655–671, 2015.

[16] L. Lei, Z. Wang, and H. Zhang, "Adaptive fault-tolerant tracking control for MIMO discrete-time systems via reinforcement learning algorithm with less learning parameters," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 1, pp. 299–313, 2017.

[17] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter, "Control of a quadrotor with reinforcement learning," *IEEE Robotics & Automation Letters*, vol. 2, no. 4, pp. 2096–2103, 2017.

[18] T. Liu, X. Hu, S. E. Li, and D. Cao, "Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle," *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 4, pp. 1497–1507, 2017.

[19] S. S. Mousavi, M. Schukat, and E. Howley, "Traffic light control using deep policy-gradient and value-function-based reinforcement learning," *IET Intelligent Transport Systems*, vol. 11, no. 7, pp. 417–423, 2017.

[20] R. P. Selvam, "Earthworm optimization with deep transfer learning enabled aerial image classification model in IoT enabled UAV networks," *Fusion: Practice and Applications*, vol. 7, no. 1, pp. 41–52, 2022.

[21] M. R. Brust, G. Danoy, D. H. Stolfi, and P. Bouvry, "Swarm-based counter UAV defense system," *Discov Internet Things*, vol. 1, no. 1, p. 2, 2021.

[22] Y. Khosiawan, Y. Park, I. Moon, J. M. Nilakantan, and I. Nielsen, "Task scheduling system for UAV operations in indoor environment," *Neural Computing and Applications*, vol. 31, no. 9, pp. 5431–5459, 2019.

[23] K. Hossain, C. Mantel, and S. O. Forchhammer, "No-reference prediction of quality metrics for H.264-compressed infrared sequences for unmanned aerial vehicle applications," *Journal of Electronic Imaging*, vol. 28, no. 4, article 043012, 2019.

[24] E. Anderlini, D. Forehand, P. Stansell, Q. Xiao, and M. Abusara, "Control of a point absorber using reinforcement learning," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 4, pp. 1681–1690, 2016.