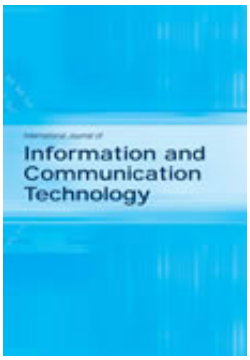
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Trajectory planning for enhanced multi-agent deep deterministic policy gradient-based multi-UAV assisted maritime communication

Zhenyu Xu

Article History:

Received:	18 December 2024
Last revised:	28 February 2025
Accepted:	28 February 2025
Published online:	28 April 2025

Trajectory planning for enhanced multi-agent deep deterministic policy gradient-based multi-UAV assisted maritime communication

Zhenyu Xu

School of Information Engineering,
Shanghai Maritime University,
Shanghai, Shanghai, China
Email: xzyzz123@163.com

Abstract: A multi-UAV flight-path planning method is developed to provide communication services to user ships in blind maritime communication zones. The developed approach considers several limitations, such as the maximum flight speed and flight range. Owing to the limited energy and transmission range of UAVs, communication resources may be distributed unevenly, which could result in communication inequality. To address these problems, an optimisation problem is created to maximise the combined metrics of communication fairness and UAV energy efficiency. Considering the complexity of solving this optimisation problem, a deep-learning algorithm, DRL-1, is proposed. DRL-1 optimises the multi-UAV flight path-planning problem by utilising Ape-X and RNN based on the MADDPG method. Simulation result 1 demonstrates that the proposed optimisation algorithm effectively enhances the UAVs' energy efficiency and communication fairness. Simulation result 2 shows a significant improvement in UAVs' energy efficiency and communication fairness as the number of UAVs increases.

Keywords: maritime communication; unmanned aerial vehicle; UAV; trajectory planning; Markov decision process; MDP; deep learning; multi-agent deep deterministic policy gradient; DDPG.

Reference to this paper should be made as follows: Xu, Z. (2025) 'Trajectory planning for enhanced multi-agent deep deterministic policy gradient-based multi-UAV assisted maritime communication', *Int. J. Information and Communication Technology*, Vol. 26, No. 9, pp.64–82.

Biographical notes: Zhenyu Xu is a graduate student majoring in Information and Communication Engineering at Shanghai Maritime University. His primary research focuses on information and communication technology and maritime internet.

1 Introduction

In recent years, the maritime industry has expanded rapidly. Offshore wireless communication services, especially in coastal areas, have seen an increase in demand with increasing activity in sea (Wei et al., 2021, 2019). Currently, wireless broadband communication services in coastal regions are primarily provided by shore-based ground stations. However, owing to their limited communication coverage, blind spots exist in

nearshore communication coverage. Therefore, the establishment of a nearshore maritime communication network that fulfils the increasing communication demands is required (Dong et al., 2022). Wireless broadband coverage is restricted, and building base stations for maritime communication networks is difficult because of the special characteristics of the maritime environment. Currently, the wide coverage of maritime wireless communications relies on satellite communication systems. However, owing to their high cost and latency, satellites cannot effectively fulfil the demands of maritime users in real time (Evans, 2014; Hadinger, 2015).

Unlike satellite communication, which has high latency and cost, unmanned aerial vehicles (UAVs) carrying communication access points offer several advantages, including high mobility, flexibility, low cost, and on-demand deployment. They can rapidly establish networks in a short time, acting as wireless relay nodes or aerial base stations to offer wireless communication services to ships in maritime blind spots and facilitate prompt and efficient communication services. UAVs have been proven indispensable in emergency communication, sea rescue, and other fields (Chen et al., 2022; Li et al., 2020).

Guezouli et al. (2018) proposed that the mobility of nodes in wireless sensor networks can optimise maximise the coverage radius of the base station and end-to-end data transmission latency. The high mobility of UAVs can serve as communication nodes in wireless networks. Existing research on UAVs as mobile aerial base stations for wireless communication services has been focused on terrestrial consumers. To maximise the minimum average achievable rate for users, Yang et al. (2023) examined a scenario of UAV-assisted downlink communication for ground users. They presented an optimisation problem involving UAV trajectory restrictions, power constraints, and user-access scheduling. Liu et al. (2018) examined the use of unmanned aerial base stations to improve the performance and coverage of communication networks under a variety of circumstances, including emergency communication and network access in remote locations. Lang et al. (2022) proposed a method for wireless resource allocation and trajectory optimisation in a UAV-assisted communication system based on user trajectories for UAV-assisted downlink mobile communication systems.

In contrast to terrestrial communication, the unique characteristics of maritime communication must be considered when planning deployment, trajectory optimisation, and resource scheduling of UAV-assisted maritime communication networks. For instance, line-of-sight communication can be applied to clear air-to-sea communication links and open-sea surfaces, where path loss is predominantly determined by the placement of ships and UAVs. The fundamental design, channel properties, use cases, opportunities, and difficulties of UAV-based maritime communication systems have been described in a previous study (Akhtar and Saeed, 2022). To provide communication services to ships in maritime blind spots, Tang et al. (2021) suggested a path planning method based on non-orthogonal multiple access for a single UAV. The goal of this method was to minimise the maximum ship throughput while optimising the joint power and transmission time allocation under airborne communication energy constraints. Nevertheless, they used only one UAV to serve as a transient aerial base station for communication with limited coverage (Tang et al., 2021). Owing to the wide range of blind areas in maritime communication, using a single UAV tends to result in ineffective communication.

For UAV route planning, existing reinforcement learning-based path planning techniques are generally superior to conventional techniques (Sun et al., 2021; Yan et al., 2021; Zhang et al., 2021). Through autonomous decision making, agents can interact with their surroundings in reinforcement learning approaches. They can then use feedback to optimise learning decisions, ultimately determining the most effective approaches to execute tasks based on experience. A method based on deep reinforcement learning (DRL) has been proposed in dynamic environments to overcome the problems of high relative error and long response time of traditional methods (Jing and Zhang, 2023; Ma and Hu, 2019). The network convergence speed of the algorithm and path-planning performance of agents have been improved by optimising the reward function and revising the Q-value calculation approach based on the deep Q network (Li and Geng, 2023).

As a single UAV can only cover a certain area for communication, using many UAVs as makeshift aerial base stations can increase the communication coverage. To optimise the deep deterministic policy gradient (DDPG) and achieve path planning for numerous UAVs in a positional environment, Qiao et al. (2022) used a prioritised experience replay mechanism. However, determining the best joint approach for UAV-assisted wireless communication networks is challenging because the problems of UAV trajectory design and power allocation are non-convex (Zhao et al., 2020). Such a problem can be modelled as a Markov decision process (MDP), and the trajectory can be optimised using DRL.

Various studies have primarily focused on terrestrial environments (Qiao et al., 2022; Zhao et al., 2020), which greatly differ from maritime environments in terms of available resources and communication scenarios. Nevertheless, effective solutions to the trajectory optimisation problem for multiple UAVs have been found in both environments to jointly enhance communication coverage. The existing multi-UAV trajectory optimisation schemes for terrestrial scenarios are often inapplicable to practical maritime scenarios. In maritime environments, it is essential to consider variables such as UAV energy consumption and available communication range and optimise the flight trajectories of several UAVs accordingly.

Furthermore, resources are often unfairly distributed in studies that have attempted to improve communication coverage using several UAVs to boost the total system throughput. In fact, more resources are allocated to locations with better transmission environments, which results in unfair communication services. We focus on creating a multi-UAV-assisted near-shore maritime wireless communication system considering the restricted resources available for maritime communication and the need for stable communication services. By optimising the trajectory of multi-UAV-assisted communication with limited resources, the proposed system aims to balance equitable coverage across service areas with the maximum overall system energy efficiency.

We present a multi-agent DDPG (MADDPG) algorithm that integrates a distributed architecture (Ape-X) and recurrent neural networks (RNNs) for optimising a reward function based on reinforcement learning to control the flight trajectories of multiple UAVs. This is performed by considering various constraints and major environmental uncertainties. The primary contributions of this study are as follows:

1 A multi-UAV-aided communication-coverage enhancement system is developed by considering user service demands and near-shore wireless communication service resources. Considering the unique requirements of the maritime model, several

charging stations are deployed along a shoreline. As UAVs cannot recharge their energy during service, a minimum safe energy threshold for UAVs is defined, and the UAVs return to recharge when their remaining energy reaches the threshold. A multi-UAV collaborative covering approach for maritime communication blind spots is proposed based on the established model. To improve the overall energy economy and fairness under limited resources, a joint flight trajectory optimisation problem for numerous UAVs is explored under limitations such as UAV flight range, maximum UAV flight speed, and minimum safe distance between multiple UAVs.

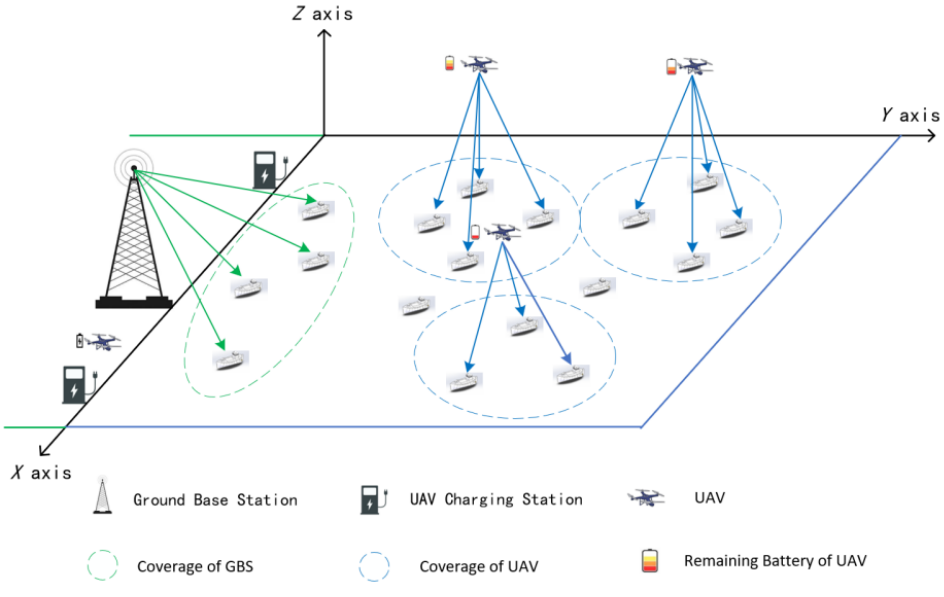
2 The problem model is formulated as an MDP because of the multiple restrictions and high computational complexity of this model considering a multi-UAV scenario. Given that this scenario uses a multi-UAV model, safety distance restrictions between UAVs need to be considered. Consequently, the MDP reward function is formed by using the optimisation function for the total energy efficiency and considering fairness as a positive reward, and using the UAV flying range, distance between UAVs, and other factors as negative rewards. Accordingly, a deep learning method, DRL-1, is presented after optimisation using the Ape-X distributed architecture and RNN modules. The MADDPG algorithm is suggested to optimise the reward function.

3 The proposed optimisation algorithm, DRL-1, can make corresponding decisions quickly based on dynamic changes in the environment through comparisons of the simulation experiment results, demonstrating a stronger learning ability, better long-term dependency modelling ability, and better generalisation ability that other algorithms. In comparing the learning efficiency and algorithm performance over the same period as the pre-optimisation algorithm, DRL-1 exhibits a considerable improvement. The accuracy of the findings in this study is confirmed by comparing the experimental results with varying numbers of UAVs, demonstrating that the learning effect of the algorithm becomes more notable as the number of serving UAVs increases.

2 System model

A scenario in which ships within a specific distance range of the coastline can receive communication service coverage from shore-based communication base stations is illustrated in Figure 1. However, ships located farther from the coastline exceed the communication range covered by shore-based base stations. The area in which these ships are located is referred to as the offshore maritime communication blind zone. Assume that K ships situated in the offshore maritime communication blind zone have wireless communication coverage from N UAVs, with $n = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, K$. The UAVs are launched at random locations and fly at constant altitude h in a single flight. They fly to designated spots to assist one or more ships with communication. The UAV returns to the land-based charging station to replenish its battery before the next round of deployment when its remaining energy reaches the minimum safe energy threshold. The UAV flying time is split into M equal time slots, λ, with $m = 1, 2, \ldots, M$.

Figure 1 UAV-network-assisted maritime communication model (see online version for colours)



The environment three-dimensional coordinate system is depicted in Figure 1, where the positive z axis represents the height above sea level, y axis represents the horizontal distance from the coastline, and x axis represents the coastline. The coordinates of the n^{th} UAV in the m^{th} time slot are expressed as $l_n[m] = (x_n[m], y_n[m], h_U)$, and those of the k^{th} ship in the m^{th} time slot are expressed as $l_k = (x_k[m], y_k[m], 0)$. Thus, distance $d_{n,k}[m]$ between the n^{th} UAV and k^{th} ship is expressed in the m^{th} time slot as

$$d_{n,k}[m] = \sqrt{\left(x_n[m] - x_k[m]\right)^2 + \left(y_n[m] - y_k[m]\right)^2 + h_U^2} \tag{1}$$

Controlling the flight range and maximum speed of a UAV and minimum safe distance between UAVs is essential in a multi-UAV collaboration scenario, in which UAVs have limited energy and cannot recharge while flying over the sea. These constraints are expressed as

$$l_k[m] = \left(x_n[m], y_n[m]\right) \in L_{\max} \tag{2}$$

where $L_{\max}$ denotes the maximum horizontal movement range of the UAVs. In addition,

$$v_n[m] = \frac{\sqrt{\left(x_n[m] - x_n[m-1]\right)^2 + \left(y_n[m] - y_n[m-1]\right)^2}}{\tau} \tag{3}$$

Here, $v_n[m]$ is the flight speed of the n^{th} UAV at the m^{th} time slot, with $v_n[m] \le V_{\max}$, and $V_{\max}$ denotes the maximum flight speed of the UAVs. Moreover,

$$d_{n,i}[m] = \sqrt{\left(x_n[m] - x_i[m]\right)^2 + \left(y_n[m] - y_i[m]\right)^2} \tag{4}$$

Here, $d_{n,i}[m]$ is the distance between the n^{th} and i^{th} UAVs for $n \neq i$, and $D_{\min}$ is the minimum allowable distance between UAVs, with $d_{n,i}[m] \geq D_{\min}$.

The onboard batteries of UAVs provide most of their energy, which is split between flight and communication tasks. The UAVs provide communication services exclusively to user vessels, and no inter-UAV communication is required. The communication energy consumption can be disregarded because it is much lower than the energy required for flight. The flight energy consumption of a UAV, which increases linearly with flight distance, can be described as follows:

$$w_{fn}[m] = 0.1 d_{fn}[m] \tag{5}$$

$$W_{fn} - \sum_{m=1}^{M} w_{fn}[m] \geq W_s \tag{6}$$

where $d_{fn}[m]$ is the flight distance of the n^{th} UAV in the m^{th} timeslot, $w_{fn}[m] \in W_{fn}$, W_{fn} is the maximum energy of the UAV, and W_s is the minimum safe energy threshold of the UAV. When the remaining energy of the UAV falls below W_s, it is recalled for charging. The communication channel between a UAV and ship can be regarded as a line-of-sight link, and the channel quality is primarily affected by the communication distance between the transmitting and receiving ends (Mozaffari et al., 2019). The channel gain follows the free-space path loss model, which can be expressed as

$$g_{n,k}[m] = a_0 d_{n,k}^{-2}[m] = \frac{a_0}{\left(x_n[m] - x_k\right)^2 + \left(y_n[m] - y_k\right)^2 + h_U^2} \tag{7}$$

where a_0 represents the fixed-channel transmission power. As the model is ideal for multi-UAV joint optimisation, the received signal of the user must incorporate interference from other UAVs. Assuming that there is more than one UAV, $p_{n,k}[m]$ denotes the communication power of the n^{th} UAV to the k^{th} ship at the m^{th} time slot. Within the m^{th} time slot, the signal to interference plus noise ratio between the n^{th} UAV and k^{th} ship can be expressed as

$$SINR[m] = \frac{p_{n,k}[m] g_{n,k}[m]}{\sum_{i=1, i \neq n}^{N} p_{i,k}[m] g_{i,k}[m] + n_f} \tag{8}$$

where n_f represents additive white Gaussian noise. Considering that the transmission power of a UAV is constrained by its transmission energy, let $P_{\max}$ denote the maximum communication transmission power of the UAV, and $0 \leq p_{n,k}[m] \leq P_{\max}$ indicates that the communication power of the UAV in each time slot does not exceed its maximum transmission power constraint.

A binary scheduling variable c is introduced to represent the utilisation of communication services provided by the UAVs and the receipt of these services by user ships. If $c_{n,k}[m] = 1$, within the m^{th} time slot, the n^{th} UAV can provide communication for the k^{th} ship; otherwise, $c_{n,k}[m] = 0$. At this point, the information transmission rate from the n^{th} UAV to the k^{th} ship in the m^{th} time slot can be expressed as

$$R_{n,k}[m] = c_{n,k}[m]\log_2\left(1+SINR[m]\right) = c_{n,k}[m]\log_2\left(1+\frac{p_{n,k}[m]g_{n,k}[m]}{\sum_{i=1,i\neq n}^{N} p_{i,k}[m]g_{i,k}[m]+n_f}\right) \tag{9}$$

Then, the average information transmission rate provided by the n^{th} UAV to the k^{th} ship in the m^{th} time slot is represented as

$$R_n[m] = \frac{1}{K}\sum_{k=1}^{K} R_{n,k}[m] \tag{10}$$

The average information transmission rate of all UAVs in the m^{th} time slot can be represented as

$$R[m] = \frac{1}{N}\sum_{n=1}^{N} R_n[m] \tag{11}$$

We aim to optimise UAV flight trajectories, control the total energy consumption of UAVs, and maximise the minimum average information transmission rate for user ships. Nevertheless, considering that this optimisation strategy can cause one ship to be covered for an extended period, while others are not, it is impossible to guarantee equitable communication possibilities for every user. Hence, we introduce the Jain fairness index to represent the fairness of the information transmission rate to provide equitable communication among all ships.

$$f = \frac{\left(\sum_{m=1}^{M} R[m]\right)^2}{M\left(\sum_{m=1}^{M} R[m]^2\right)} \tag{12}$$

Here, $f \in (0, 1)$, and a larger fairness index indicates more equitable communication provided by the UAVs.

By combining the above problems, the optimisation goal function can be expressed as

$$\delta[m] = f\frac{R[m]}{w_{fn}[m]} \tag{13}$$

We maximise objective function $\delta[m]$ by optimising the UAV flight trajectories based on the difficulties provided. The optimisation problem (P1) is formulated as follows:

$$
\begin{aligned}
&\textbf{max} \quad \textbf{min}\delta[m] \\
&\text{s.t.} \quad \delta[m] = f\frac{R[m]}{w_{fn}[m]} \\
&C1: \quad l_k[m] = \left(x_n[m], y_n[m]\right) \in L_{\max} \\
&C2: \quad v_n[m] \le V_{\max} \\
&C3: \quad d_{n,i}[m] \ge D_{\min} \\
&C4: \quad w_{fn[m]} \in W_{fn} \\
&C5: \quad 0 \le p_{n,k}[m] \le P_{\max} \\
&C6: \quad c_{n,k}[m] \in [0,1]
\end{aligned}
\tag{14}
$$

where $L_{\max}$, $V_{\max}$, and $D_{\min}$ are the UAV flight range, maximum flight speed, and minimum safe distance between UAVs, respectively. C1 represents the UAV flight range constraint; C2 constrains the UAV flight speed not to exceed $V_{\max}$; C3 indicates that the minimum safety distance between UAVs must be greater than $D_{\min}$; C4 states that the total energy consumed by the UAVs must not exceed the maximum onboard energy constraint, W_{fn}; C5 is the maximum communication transmission power constraint of the UAV; and in C6, $c_{n,k}[m]$ represents the access indication constraint for UAV-provided communication.

3 UAV trajectory planning using DRL-1 based on MADDPG

The optimisation of multi-UAV flight trajectories is a challenging problem with significant state complexity. The optimisation in this study focuses on the variables of UAV flight trajectories considering elements such as the UAV information transmission rate, power, energy consumption, and fairness index of the provided communication services, which add to the problem complexity. Therefore, the optimisation problem is formulated as an MDP. An efficient approach, DRL-1, is devised to solve this problem using MADDPG and incorporating the Ape-X distributed architecture with RNNs.

3.1 MDP

An MDP arises from the interaction between an agent and its environment, encompassing states S, actions A, reward function R, and state transition probabilities P. The creation of an MDP for optimisation problem P1 is outlined below.

The collection of environmental states S is represented by the state space. It describes the state of the UAV at a specific timeslot, including details such as the UAV position, ship position, and the separation between UAVs based on the optimisation problem.

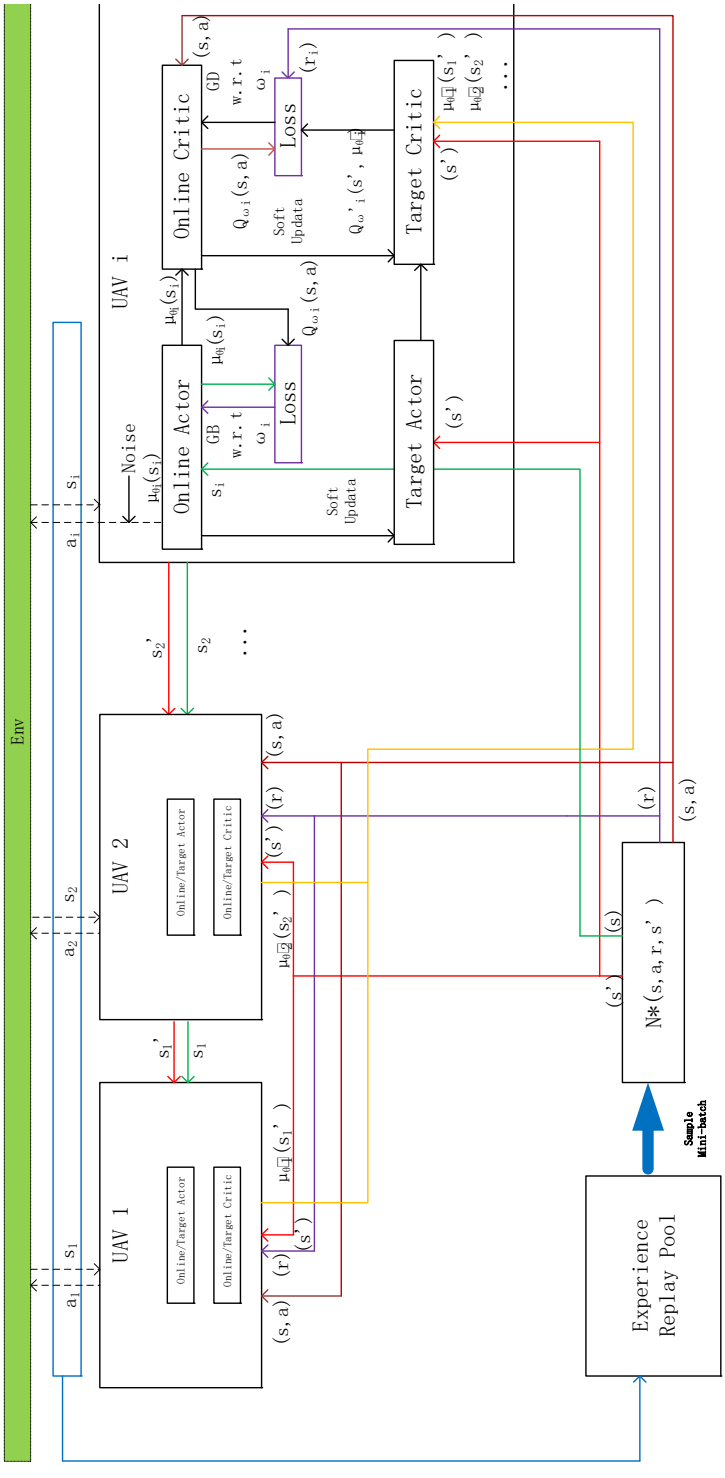
$$s[m] = \left\{l_n[m], l_k[m], d_{n,k}[m], D_{\min}\right\} \tag{15}$$

Here, $s[m] \in S$ denotes a subset of the state space.

Action space A represents the set of possible actions, referring to the set of actions that the UAV can take to change its flight status in response to the state space.

$$a[m] = \left\{l_n[m+1], c_{n,k}[m], p_{n,k}[m]\right\} \tag{16}$$

Figure 2 Diagram of MADDPG (see online version for colours)



Here, $a[m] \in A$ denotes a subset of the action space of the UAV.

Reward function R represents the immediate reward that the UAV receives from a set of actions performed in response to the state space. The reward function is related to the objective function. In this study, the reward function is denoted as $r[m] = r_c[m] + r_f[m] + r_b[m]$, where $R = \sum_{m=1}^{M} r[m]$. Here, $r_c[m]$ is defined by three components: fairness, information transmission rate, and energy consumption. The numerator of the reward represents the benefit, and the denominator represents cost $r_f[m]$. Moreover, $r_b[m]$ denotes the penalty for violating the UAV flight range constraints, and $r_f[m]$ and $r_b[m]$ denote the penalties for violating collision constraints. Both are negative values.

Probability P is the state transition probability, which represents the probability that the UAV will transition to the next state after executing an action in a given state. It is determined jointly by the state space of the timeslot in which the UAV is located and the actions of the UAV:

$$P_{s[m+1]s[m]}^{a[m]} = P\left\{s[m+1] \middle| s[m], a[m]\right\} \tag{17}$$

3.2 MADDPG

The MADDPG algorithm is a DRL algorithm for solving multi-agent collaboration problems and originates from DDPG. Liu et al. (2019) showed that the actor-critic network serves as the fundamental building block of the DDPG algorithm, which uses a dual neural network architecture, namely, the current and target networks, for both the policy and value functions. The stochastic gradient approach is used to train the parameters in the actor-critic network. This architecture accelerates convergence and enhances the algorithm learning stability. The following elements are found within its framework:

- Current-actor network μ: primarily responsible for updating θ and selecting actions a based on state s.
- Target actor network μ': copies θ updated to θ' and selects a' through s'.
- current critic network Q: calculates $Q(s, a, \omega)$ and $y = r + \gamma Q'(s', a', \omega')$ based on Q. Here, r represents the reward.
- Target critic network Q': responsible for calculating $Q'(s', a', \omega')$ for target value Q. Network parameters ω' are updated periodically using ω.

To enhance the learning efficiency, Gaussian noise n_f is introduced, and the initial action of each UAV at each stage is selected as follows:

$$a[m] = \mu\left(s[m] \middle| \theta\right) + n_f \tag{18}$$

At this point, the following decay function is introduced:

$$L = \frac{1}{N_b} \sum_{1}^{N_b} \left[y[m] - Q\left(s[m], a[m] \middle| \omega\right) \right]^2 \tag{19}$$

Here, N_b represents the sample size. The current critic network, Q, is updated through backpropagation based on the loss function.

Based on loss function L, policy gradient $\nabla_\theta J(\theta)$ for updating the parameters of network Q can be obtained to determine optimised parameters θ.

For parameters ω of current actor network μ, the state of the UAV in the current time slot is $s[m]$, and the next action output in this state is $a[m]$. Policy gradient $\nabla_\theta J(\theta)$ for updating parameters θ is given by

$$\nabla_\theta J(\theta) = \frac{1}{N_b}\sum_{m=1}^{M}\left[\nabla_a Q(s,a|\omega|_{s=s[m],a=\mu(s[m])}\nabla_\theta \mu(s|\theta)\Big|_{s=s[m]}\right] \tag{20}$$

In the DDPG algorithm, a soft update method is adopted. Learning rate τ is introduced, which is much smaller than 1. Moreover, θ' and ω' are updated through τ weighted averaging with 1, and the results are assigned to the target networks as follows:

$$\theta' = \tau\theta + (1-\tau)\theta' \tag{21}$$

$$\omega' = \tau\omega + (1-\tau)\omega' \tag{22}$$

Compared with the DDPG algorithm, the MADDPG algorithm can handle multi-agent problems. The MADDPG model comprises multiple DDPG networks, as illustrated in Figure 3. In this model, each agent learns independently and stores all its learning experiences in a shared experience buffer. When upgrading the target networks, it is necessary to access the experiences gained by all agents, and the parameter updates consider the experiences of all agents. In MADDPG, data sampling and collection are performed separately for each agent, while training and learning are conducted jointly. Therefore, MADDPG exhibits better performance than DDPG when dealing with multi-agent environments.

Algorithm 1: Maritime UAV trajectory planning based on MADDPG

1 Initialise actor and critic networks
2 Copy parameters of current actor and critic networks to their corresponding target networks: $\theta \to \theta'$, $\omega \to \omega'$
3 Clear the experience replay buffer
4 For each episode e = 1, 2, 3, …, E
  a Initialise position and velocity of UAV
  b Initialise Gaussian noise n_f and state s
  c For each time slot m = 1, 2, 3, …, M
    i Select action based on noise and current policy $a[m] = \mu(s[m] \mid \theta) + n_f$, execute action $a[m]$ with UAV, and obtain reward $r[m]$ and next state $s[m + 1]$
    ii Store the current state, action, reward, and next state ($s[m]$, $a[m]$, $r[m]$, $s[m + 1]$) in experience replay buffer
    iii Sample data from experience replay buffer using a sample minibatch
    iv Upgrade critic network by minimising decay function L
    v Calculate policy gradient $\nabla_\theta J(\theta)$ to upgrade actor network
    vi Target network parameters are updated using equations (21) and (22).
  end for
end for

In each time slot, the UAV follows the policy to select an action and flies to a location in the area to provide communication services. Meanwhile, after the UAV selects an action, it obtains $s[m+1]$ and $a[m]$ from the environment based on action $a[m]$, and stores $s[m]$, $a[m]$, $r[m]$, and $s[m+1]$ in the experience buffer. Subsequently, based on the data stored in the experience buffer, parameters θ and ω of the current and target networks are upgraded based on decay function L and policy gradient $\nabla_\theta J(\theta)$. At this point, an action is completed, and the experience gained from this action is used iteratively to find the flight route with the best reward.

To make the learning process more efficient and the learning outcomes more accurate, we introduce the Ape-X architecture and RNNs to enhance the algorithm performance.

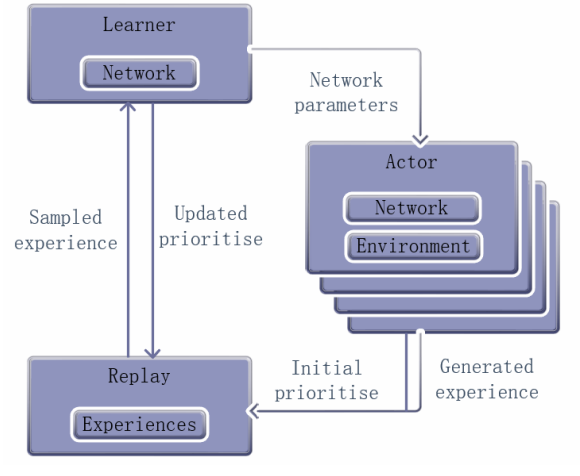
3.3 Computational complexity analysis of MADDPG

In this paper, the algorithmic complexity of MADDPG is divided into two components: the complexity of the actor network and the complexity of the critic network. The actor network has a complexity of $O(N)$, whereas the critic network exhibits a complexity of $O(N^2)$. It is evident that the actor network's complexity grows linearly with the number of UAVs, while the critic network's complexity scales quadratically with the UAV count. Therefore, the overall algorithmic complexity of MADDPG is primarily dominated by the critic network's processing of multi-UAV joint actions. As the number of UAVs increases, the algorithmic complexity of MADDPG continues to rise.

3.4 Ape-X

Ape-X is distributed replay memory architecture, as shown in Figure 3. It incorporates multiple actors that are distributed to generate data with the environment and store those data in a shared experience replay buffer. A learner can sample data generated by multiple actors from the replay buffer, update the priority of experiences, and complete the update of network parameters.

Figure 3 Diagram of Ape-X (see online version for colours)



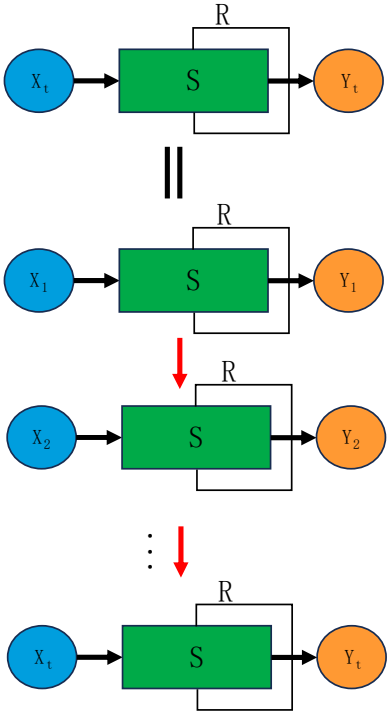
Every actor in Ape-X has a unique environment, while all actors use the same network model. Ape-X is more efficient than other approaches and greatly improves the learning efficiency because it can independently acquire experiences and store them in an experience replay buffer. Introducing the Ape-X architecture into the MADDPG algorithm effectively improves the convergence speed and performance.

3.5 RNN

As shown in Figure 4, RNNs are distinct from other neural networks in that they have cyclic connections inside their internal networks, which enable them to retain and analyse data from various points in a sequence.

In an RNN, a hidden layer exists between the input and output layers. Each time an RNN receives a new input at the input layer, it combines this input with the previous hidden layer state to generate a new hidden state that influences the output. This hidden state is then combined with the next new input, affecting the subsequent output. Therefore, RNNs can process time-series data of varying lengths and capture time-dependent relationships within a sequence.

Figure 4 RNN architecture (see online version for colours)



3.6 DRL-1

We propose an optimised DRL-1 algorithm that incorporates Ape-X and an RNN and is based on the MADDPG algorithm. UAVs can be trained by Ape-X using a vast amount

of empirical data. This improves their generalisation ability and enables simultaneous training of numerous UAVs, thereby accelerating learning. RNNs are essential for multi-UAV path-planning decision making because they can recognise long-term dependencies in time-series data, learn the dynamic features of the environment, and generalise to new scenarios. RNN integration can enhance the learning efficiency and help UAVs plan flight routes more effectively. The learning, long-term dependency modelling, and generalisation ability of the algorithm are improved when Ape-X and RNN are combined with MADDPG. This enables DRL1 to swiftly determine the best solutions for multi-UAV path planning.

4 Simulation experiment

To verify the effectiveness of the proposed solution, an environment was developed, and simulations were conducted using Python, followed by an analysis of the results from two simulation experiments.

- Results from simulation 1: effects on the maximum and average rewards before and after the Ape-X and RNN modules of the distributed architecture were added to the MADDPG algorithm when two UAVs were used.
- Results from simulation 2: effects of using the optimised MADDPG method with varying UAV counts on the average and maximum rewards.

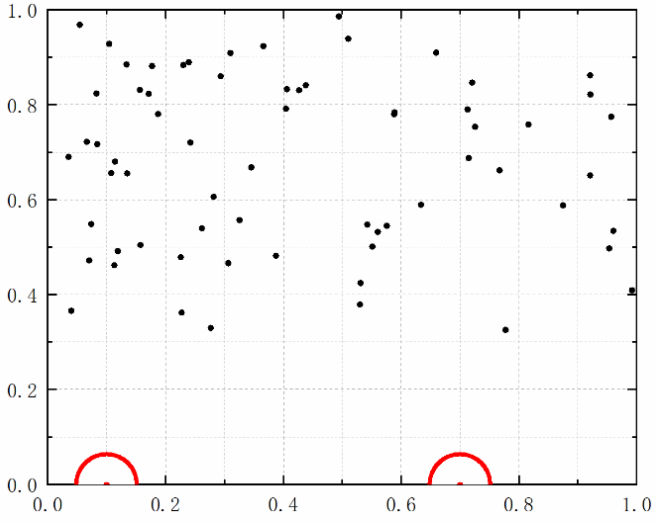
4.1 Simulation environment

Each time a UAV learns using the DRL-1 algorithm presented in this study, a flying cycle with 2,000 time slots is involved. The locations of the UAV land-based recharge stations and user ship distribution within the offshore communication blind zone in the m^{th} time slot are shown in Figure 5. In this case, the y axis goes into the ocean depth, while the x axis represents the coastline. The user ship coordinates, which are predetermined and randomly generated, are represented by black dots representing the locations within the offshore communication blind zone. The take-off position of the UAVs is [0, 0], and the red circles represent the UAV-land-based charging stations located at [0.1, 0] and [0.7, 0].

Table 1 Simulation parameters

Parameter	*Value*
Maximum coastline distance (X_{max})	1,000 m
Maximum ocean depth distance (Y_{max})	1,000 m
Altitude (H)	50 m
Maximum UAV flight speed (V_{max})	20 m/s
Minimum distance between UAVs (D_{min})	100 m
Maximum UAV transmission power (P_{max})	0.1 W
Fixed channel transmission power (a_0)	–60 dB
Gaussian white noise (n_f)	–110 dBm

Figure 5 Distribution of communication blind spots for user ships at sea in the m^{th} time slot and UAV charging stations along coastline (see online version for colours)



4.2 Simulation results

Path-planning experiments using two UAVs as examples are shown in Figures 6 and 7, which represent the trends in the maximum reward, average reward, and learning efficacy of the system as the number of learning iterations increases under various optimisation techniques. As illustrated in Figure 6, during the initial 0–100 learning iterations, the performance differences between the schemes are relatively small. Specifically, compared with the basic MADDPG algorithm, adding the RNNs increases the average reward by approximately 42%, adding the Ape-X module increases it by approximately 21%, and adding both Ape-X and RNN increases it by nearly 48%. By the 200th iteration, the non-optimised MADDPG method performs noticeably worse than the DRL-1 algorithm iteration, and the difference in the average rewards increases with the number of iterations. This suggests that the learning capacity of the MADDPG algorithm is constrained when solving more complicated models. The generalisation and learning capabilities of the DRL-1 algorithm are improved in complicated contexts by integrating Ape-X and RNN into the MADDPG algorithm. The modified DRL-1 method significantly outperforms the MADDPG algorithm after the 100th iteration, with an improvement of approximately 44%, as shown in Figure 7, which compares the maximum rewards during the learning process. Furthermore, compared with the non-optimised MADDPG algorithm, the performance of the algorithm increases by approximately 33% and 44%, respectively, upon the introduction of Ape-X or RNN, demonstrating notable improvements. The MADDPG algorithm tends to converge as the number of learning iterations increases gradually, and its difference from the other schemes increases. The DRL-1 algorithm performs noticeably better than the MADDPG algorithm for the same number of iterations after the addition of Ape-X and the RNN. Thus, the proposed DRL-1 algorithm performs better than the MADDPG algorithm as well as algorithms that solely use RNNs or Ape-X.

Figure 6 Average reward variation among different solutions (see online version for colours)

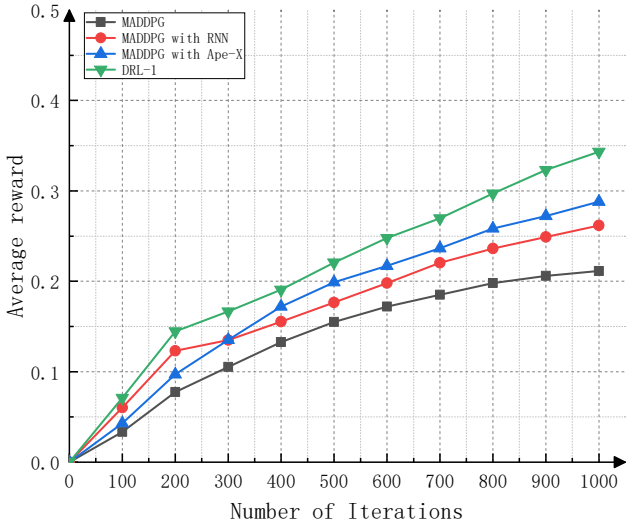
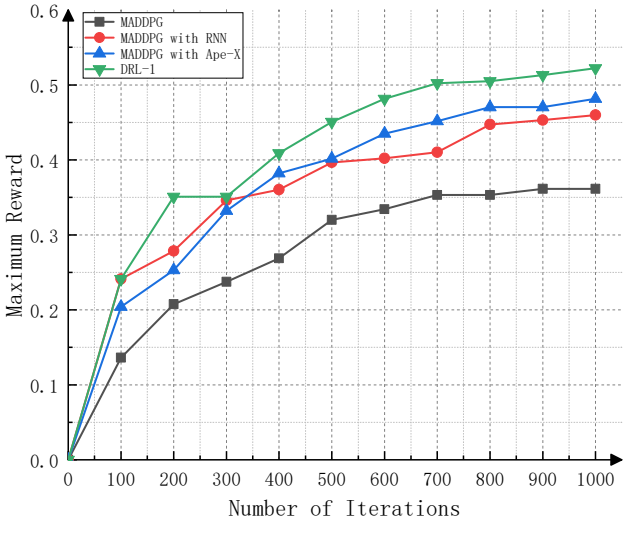


Figure 7 Maximum reward variation among different solutions (see online version for colours)



After 100 iterations, as shown in Figure 8, the average reward increases as the number of UAVs increases, and the difference increases as the number of learning iterations increases. After 400 repetitions, four UAVs receive an average reward that is approximately 67%, 54%, and 51% higher than that of one, two, and three UAVs, respectively. The average reward of fewer UAVs converges more quickly as the number of learning iterations increases, suggesting that fewer UAVs are insufficient to completely cover the service area. However, when the number of UAVs increases, this phenomenon progressively decreases. The effect of the number of UAVs on the maximum reward is illustrated in Figure 9. As the number of learning iterations

increases, it is clearly observed that when there is only one UAV, the highest reward obtained decreases and converges less than when there are more UAVs, indicating a lower performance with only one UAV. Furthermore, the maximum reward progressively increases with the number of UAVs after 100 repetitions. Four UAVs achieve the maximum rewards that are approximately 51%, 22%, and 33% higher than those of one, two, and three UAVs, respectively. The difference increases without exhibiting a discernible trend toward convergence. When the two numbers are combined, the magnitude and growth rate of the rewards improve dramatically with the number of UAVs, indicating higher performance and learning effects. This demonstrates the accuracy of the proposed method.

Figure 8 Impact of number of UAVs on average reward (see online version for colours)

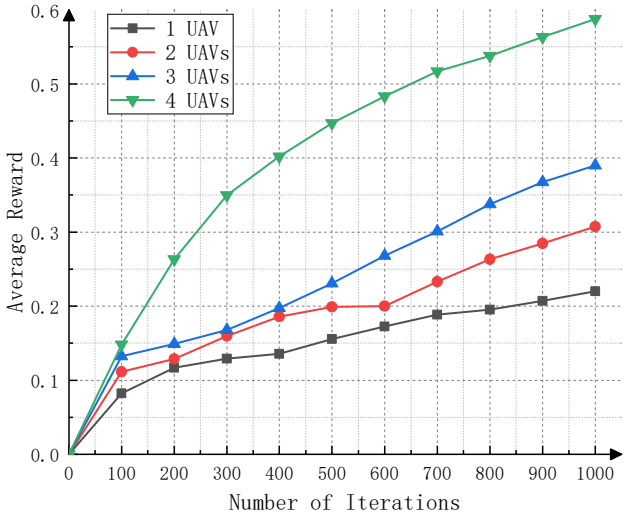
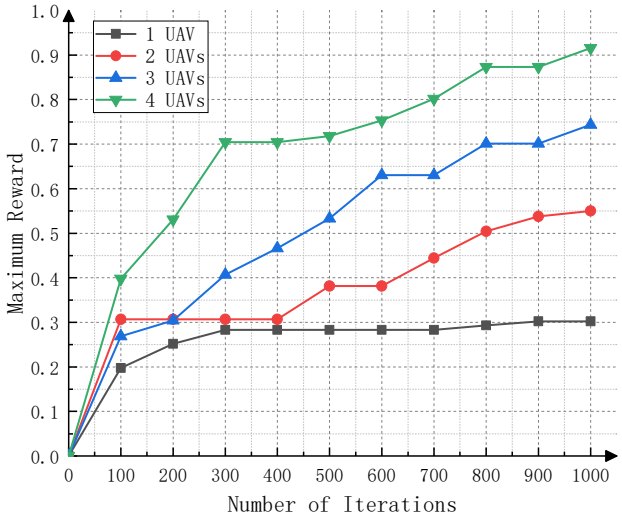


Figure 9 Impact of number of UAVs on maximum reward (see online version for colours)



5 Conclusions

This study examined the problem of cooperatively maximising the flight paths of several UAVs to offer temporary communication services to users in maritime communication dead zones. Accordingly, we examined the combined flight trajectory optimisation problem for numerous UAVs based on optimising the total energy efficiency and justice under limited resources, considering constraints such as UAV energy consumption, maximum flight speed, and minimum safety distance between UAVs. To optimise UAV flight trajectories, the problem was described as an MDP, and a DRL-1 algorithm was introduced. Ape-X and RNN modules improve the capacity of the algorithm for long-term dependency modelling and generalisation, and the proposed algorithm is based on MADDPG. Simulation results show that while dealing with multi-UAV scenarios and longer learning processes, the proposed DRL-1 algorithm substantially improved the performance compared with the pre-optimisation approach, thereby improving the algorithm learning ability. Higher reward values were acquired with an increasing number of UAVs, leading to better objective function results and enhanced learning efficiency, as confirmed by the second set of simulation results.

Declarations

All authors declare that they have no conflicts of interest.

References

Akhtar, M.W. and Saeed, N. (2022) *UAVs-enabled Maritime Communications: Opportunities and Challenges*, arXiv preprint arXiv:2206.03118, DOI: 10.1109/MSMC.2022.3231415.

Chen, X., Sheng, M., Li, B. and Zhao, N. (2022) 'Survey on unmanned aerial vehicle communications for 6G', *Journal of Electronics & Information Technology*, Vol. 44, No. 3, pp.781–789.

Dong, H., Song, L., Hua, C.Q., Liu, L.Y. and Tang, J.H. (2022) 'Survey of the research and development on the maritime communication technology', *Telecommun. Sci.*, Vol. 38, No. 5, pp.1–17.

Evans, B.G. (2014) 'The role of satellites in 5G', *2014 7th Advanced Satellite Multimedia Systems Conference and the 13th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, pp.197–202, IEEE, DOI: 10.1109/EUSIPCO.2015.7362886.

Guezouli, L., Barka, K., Bouam, S. and Zidani, A. (2018) 'A variant of random way point mobility model to improve routing in wireless sensor networks', *International Journal of Information and Communication Technology*, Vol. 13, No. 4, pp.407–423, DOI: 10.1504/IJICT.2018.095036.

Hadinger, P. (2015) 'Inmarsat Global Xpress the design, implementation, and activation of a global Ka-band network', *33rd AIAA International Communications Satellite Systems Conference and Exhibition*, Vol. 4303.

Jing, Z. and Zhang, H. (2023) 'Intelligent traffic assignment method of urban traffic network based on deep reinforcement learning', *International Journal of Information and Communication Technology*, Vol. 22, No. 1, pp.60–72, DOI: 10.1504/IJICT.2023.127683.

Lang, L., Wang, J., Wang, Y. and Zhao, Z. (2022) 'Radio resource and trajectory optimization for UAV assisted communication based on user route', *Journal on Communications*, Vol. 43, No. 3, pp.225–232.

Li, Q.R. and Geng, X. (2023) 'Robot path planning based on improved DQN algorithm', *Comput. Eng.*, Vol. 49, No. 12, pp.111–120, DOI: 10.19678/j.issn.1000-3428.0066348.

Li, X., Feng, W., Chen, Y., Wang, C.X. and Ge, N. (2020) 'Maritime coverage enhancement using UAVs coordinated with hybrid satellite-terrestrial networks', *IEEE Transactions on Communications*, Vol. 68, No. 4, pp.2355–2369, DOI: 10.1109/TCOMM.2020.2966715.

Liu, C.H., Chen, Z., Tang, J., Xu, J. and Piao, C. (2018) 'Energy-efficient UAV control for effective and fair communication coverage: a deep reinforcement learning approach', *IEEE Journal on Selected Areas in Communications*, Vol. 36, No. 9, pp.2059–2070, DOI: 10.1109/JSAC.2018.2864373.

Liu, J.W., Gao, F. and Luo, X.L. (2019) 'Survey of deep reinforcement learning based on value function and policy gradient', *Chin, J. Comput.*, Vol. 42, No. 6, pp.1406–1438.

Ma, M. and Hu, Y. (2019) 'Feature extraction algorithm for fast moving pedestrians with frame drop constraint based on deep learning', *International Journal of Information and Communication Technology*, Vol. 15, No. 4, pp.331–343, DOI: 10.1504/IJICT.2019.103199.

Mozaffari, M., Saad, W., Bennis, M., Nam, Y.H. and Debbah, M. (2019) 'A tutorial on UAVs for wireless networks: applications, challenges, and open problems', *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 3, pp.2334–2360, DOI: 10.1109/COMST.2019.2902862.

Qiao, Z., Li, S.L., Wang, J.Z. et al. (2022) 'UAV path planning based on PER-PDDPG', *Unmanned Syst. Technol.*, Vol. 5, No. 6, pp.12–23, DOI: 10.19942/j.issn.2096-5915.2022.6.055.

Sun, H., Hu, C. and Zhang, J. (2021) 'Deep reinforcement learning for motion planning of mobile robots', *Control and Decision*, Vol. 36, No. 6, pp.1281–1292, DOI: 10.13195/j.kzyjc.2020.0470.

Tang, R., Feng, W., Chen, Y. and Ge, N. (2021) 'NOMA-based UAV communications for maritime coverage enhancement', *China Communications*, Vol. 18, No. 4, pp.230–243.

Wei, T., Feng, W., Chen, Y., Wang, C.X., Ge, N. and Lu, J. (2021) 'Hybrid satellite-terrestrial communication networks for the maritime internet of things: key technologies, opportunities, and challenges', *IEEE Internet of Things Journal*, Vol. 8, No. 11, pp.8910–8934.

Wei, T., Feng, W., Wang, J., Ge, N. and Lu, J. (2019) 'Exploiting the shipping lane information for energy-efficient maritime communications', *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 7, pp.7204–7208.

Yan, J., Zhang, Q. and Hu, X. (2021) 'Review of path planning techniques based on reinforcement learning', *Jisuanji Gongcheng/Computer Engineering*, Vol. 47, No. 10, pp.16–25, DOI: 10.19678/j.issn.1000-3428.0060683.

Yang, Q., Chen, J. and Peng, Y. (2023) 'Unmanned aerial vehicle trajectory planning and power control algorithm based on deep deterministic policy gradient', *Journal of Beijing University of Posts and Telecommunications,* Vol. 46, No. 3, p.43, DOI: 10.13190/j.jbupt.2022-208.

Zhang, R., Wu, C., Sun, T. and Zhao, Z. (2021) 'Progress on deep reinforcement learning in path planning', *Computer Engineering and Applications*, Vol. 57, No. 19, pp.44–56.

Zhao, N., Cheng, Y., Pei, Y., Liang, Y.C. and Niyato, D. (2020) 'Deep reinforcement learning for trajectory design and power allocation in UAV networks', *ICC 2020–2020 IEEE International Conference on Communications (ICC)*, pp.1–6, IEEE, DOI: 10.1109/ICC40277.2020.9149196.