Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8805416, 13 pages
https://doi.org/10.1155/2022/8805416

Research Article

Deep Reinforcement Learning-Based UAV Data Collection and Offloading in NOMA-Enabled Marine IoT Systems

Yanpeng Dai, Ziyi Liang, Ling Lyu, and Bin Lin

School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

Correspondence should be addressed to Ling Lyu; jidalvling@126.com

Received 29 January 2022; Revised 22 April 2022; Accepted 26 April 2022; Published 12 May 2022

Academic Editor: Qiang Ye

The rapid growth of maritime wireless communication demand and the complex offshore wireless communication environment have brought challenges to ensure the real-time and reliability of data transmission in the marine Internet of Things (MIoT). Unmanned aerial vehicles (UAVs) have great advantages in enhancing coverage and channel quality. Hence, we investigate a UAV-assisted data collection and data offloading system based on nonorthogonal multiple access (NOMA) technology in this paper. We jointly optimize the buoy-UAV association relationship, transmit powers, and the UAV trajectory to minimize the total mission completion time while ensuring data transmission requirements. We first propose a UAV trajectory optimization algorithm based on deep reinforcement learning (DRL). Then, we design a heuristic algorithm to effectively solve the subproblem of power control and the association relationship. Finally, we propose a joint optimization scheme to solve the minimization problem. Simulation results show the effectiveness of the proposed scheme.

1. Introduction

Marine environmental monitoring is indispensable with the continuous increase of human marine activities. A large amount of meteorological and hydrological data leads to an increase in the demand for maritime wireless communication [1, 2]. Buoys are widely deployed in the ocean due to their low cost and flexible deployment. With the development of technology, buoys can be used for marine environment monitoring with a variety of sensors and communication equipment and can also be powered by power supply methods such as lithium-ion batteries and solar energy [3, 4]. However, the transmit power of the buoy is limited. Traditional maritime wireless communication methods, such as land base stations and satellites, have disadvantages such as limited coverage and long transmission distance, which seriously affect the real-time and reliability of information transmission [5]. For the current five-generation (5G) and the upcoming six-generation (6G) era, it is of great significance to build an efficient and dynamic maritime communication network [6]. Therefore, the UAV-assisted wireless communication system (UWCS) has received widespread attention.

Unmanned aerial vehicles (UAVs) have the advantages of maneuverability and easy manipulation, which can be deployed on demand and enlarge coverage [7]. It is easier to establish a line-of-sight (LoS) channel and a stronger communication link with the target device, which can better deal with the variable ocean environment [8, 9]. In the marine Internet of Things (MIoT), aiming at the problem of the large number and wide distribution of buoys, UAV can act as a mobile base station, collecting data collected by buoys from the target area and offloading the data to the OBS [10, 11]. Furthermore, the limited spectrum resources of MIoT also pose a challenge to the reliability and efficiency of data transmission. Nonorthogonal multiple access (NOMA) technology is considered a promising technology in the 5G era [12]. Compared with orthogonal multiple access (OMA) technology, NOMA greatly improves the spectrum efficiency in the presence of limited spectrum resources by allowing multiple users to access simultaneously in the same channel and relying on power domain multiplexing and successive interference cancellation (SIC) decoding technology [13–15].

Recently, much research has applied NOMA technology to UAV-assisted wireless communication system. Zhao et al. in [13] investigated a NOMA-assisted UAV large-scale IoT data collection system and proposed a data collection optimization algorithm. The results show that, compared with the traditional UWCS, the NOMA-based UWCS has better performance in data collection. W. Chen et al. in [16] maximized the sum rate of the UAV-assisted uplink NOMA system by jointly optimizing the UAV location, buoy sensor grouping, and power control. The simulation results show the performance gain of NOMA in the sum rate of the system. Tang et al. in [17] investigated the scenario of a UAV-assisted marine wireless communication downlink in which the UAV hovers continuously to provide services to multiple groups of ships. Obviously, the NOMA-based UWCS has great advantages in enhancing coverage and strengthening communication links.

In the above work, the optimization problem is usually formulated as a mixed-integer nonconvex problem. They can usually be divided into several subproblems, which can be solved by traditional optimization techniques and iterative algorithms [18]. However, the above solutions may have high computational complexity. Furthermore, the buoys associated with the UAV and NOMA cochannel interference vary with UAV position. The complex dynamic changes bring great challenges to the traditional convex optimization technology as well. With the development of machine learning technology, reinforcement learning (RL) is considered to be an effective solution to the high-dynamic environment [19–21]. Deep reinforcement learning (DRL) solves the continuous state space problem that RL cannot solve by introducing deep neural network (DNN), such as deep Q -learning (DQN) and deep deterministic policy gradient (DDPG) [22]. At present, a lot of work has focused on the research of UWCS based on DRL. L. Wang et al. in [23] minimized the energy consumption of all user equipment by jointly optimizing UAV trajectories, user associations, and resource allocation. Two algorithms are proposed to effectively solve the minimization problem based on convex optimization and DRL technology, respectively. The results show that the DRL-based method is better than the convex optimization method. Zhang et al. in [24] studied the UAV lineup and user distribution change scenarios and developed a DDPG-based proactive self-regulation method for UAV networks, which is based on the proposed asynchronous parallel computing architecture. Wang et al. in [25] studied a UAV-assisted mobile edge computing system. They minimized the maximum processing delay and proposed a DDPG-based algorithm to solve the high-dimensional state space and continuous action space. However, the flight action space of actual UAV is continuous and high dimensional, which may bring dimensional disaster to traditional reinforcement learning methods (such as DQN) [23, 26]. DDPG exists the problem of overestimation. To solve the above problems, Fujimoto et al. in [27] proposed the twin-delayed deep deterministic (TD3) algorithm based on DDPG. In [28], Sun et al. considered the age of information (AoI) and energy consumption and proposed an AoI-energy-aware UAV trajectory optimization algorithm based on TD3.

In this paper, we investigate a UAV-assisted data collection and data offloading system in MIoT. Specifically, buoys are used to collect marine environment information in the sensing layer. The UAV collects the sensing information from buoys based on NOMA technology and offloads the collected data to the OBS. Our goal is to minimize the total mission completion time of the UAV by jointly optimizing the UAV trajectory, the buoy-UAV association relationship, the UAV transmit power, and the buoys transmit power. The main contributions of our paper are listed as follows.

(i) We jointly consider the UAV trajectory, buoy-UAV association relationship, and transmit powers to investigate the UAV's total mission completion time minimization problem. The above minimization problem is a mixed-integer nonconvex problem. Accordingly, we divide the total mission process of UAV into data collection stage and data offloading stage for analysis

(ii) We propose a UAV trajectory optimization algorithm based on TD3 to solve the UAV trajectory coupling of data collection and data offloading since the minimization problem is a mixed-integer nonconvex problem. Furthermore, we design a heuristic algorithm to effectively solve the above problem due to the coupling between the buoy-UAV association relationship and the buoys transmit power

(iii) We propose a joint TD3-based trajectory optimization, power control, and buoy-UAV association relationship scheme that effectively solves the mixed-integer nonconvex problem. The simulation results show that the proposed scheme can effectively shorten the UAV's total mission completion time while ensuring that the data transmission requirements are met

The remainder of this paper is organized as follows. Section 2 presents the system model and the problem formulation. Section 3 briefly introduces the TD3 algorithm and proposes the TD3-based UAV trajectory optimization algorithm. Then, we design a heuristic algorithm to solve the subproblem of power control and buoy-UAV association relationship. Finally, we propose a joint optimization scheme for the minimization problem. Simulation results and conclusion are given in Sections 4 and 5.

2. System Model and Problem Formulation

2.1. Network Model. We consider a UAV-assisted MIoT system as shown in Figure 1, which includes a UAV base station, M buoys, and an OBS. Each buoy senses and stores hydrometeorological data and is powered by a lithium-ion battery to ensure that it has sufficient energy to transmit data. The total mission time of the UAV is denoted as T_{total} and divided into K time slots. The time slot length is

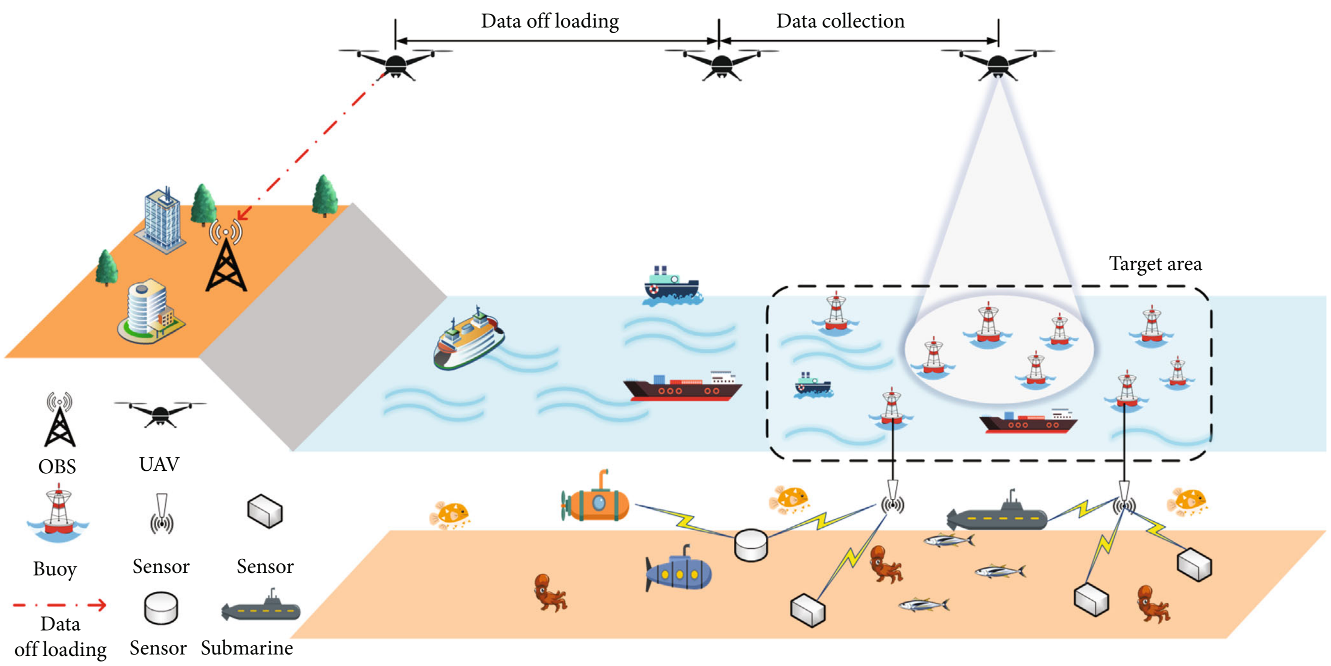

Figure 1: UAV-assisted marine IoT system.

δ. The total mission process of UAV consists of two stages. The first stage is that the UAV utilizes NOMA to collect data from M buoys, and the number of time slots of this stage is K_{co}. The UAV is allowed to collect data from at most U buoys in each time slot where $U \leq M$. The second stage is that the UAV offloads all the collected data to the OBS after completing the first stage, and the number of time slots of this stage is K_{of}. Therefore, T_{total} can be expressed as

$$T_{\text{total}} = K\delta = (K_{\text{co}} + K_{\text{of}})\delta. \tag{1}$$

Let $\mathscr{M} = \{1, 2, \cdots, M\}$ denote the set of all buoys. We denote that the horizontal coordinate of the UAV in the k-th time slot is $q_k = (x_k, y_k)$, where $k \in \mathscr{K} = \{1, \cdots, K_{\text{co}}, K_{\text{co}} + 1, \cdots, K\}$. Let $\mathscr{K}_{\text{co}} = \{1, 2, \cdots, K_{\text{co}}\}$ denote the set of the data collection time, and $\mathscr{K}_{\text{of}} = \{K_{\text{co}} + 1, \cdots, K\}$ denote the set of the data offloading time.

The fixed flight height of UAV is H, and the flight velocity of UAV in the k-th time slot is V_k. Then, the UAV should follow the maximum flight velocity constraints, which are expressed as

$$\begin{aligned}
&V_{k+1} = V_k + \Delta_k\delta, \forall k \in \mathscr{K},\\
&V_{\min} \leq \|V_k\| \leq V_{\max}, \forall k \in \mathscr{K},\\
&\|\Delta_k\| \leq \Delta_{\max}, \forall k \in \mathscr{K},\\
&q_{k+1} = q_k + V_k\delta + \frac{1}{2}\Delta_k\delta^2, \forall k \in \mathscr{K},
\end{aligned} \tag{2}$$

where $V_{\max}$ and $\Delta_{\max}$ are the maximum flight velocity and acceleration, respectively, and $V_{\min}$ is the minimum flight velocity.

The horizontal coordinate of the m-th buoy is $D_m = (x_m, y_m)$. The horizontal coordinate of OBS is $D_0 = (x_0, y_0)$. If the time slot δ is small enough, the motion of the UAV in each time slot can be regarded as static approximately. Hence, the distance between UAV and the m-th buoy at the k-th time slot is $d_{m,k} = \sqrt{\|q_k - D_m\|^2 + (H - H_m)^2}$, $\forall k \in \mathscr{K}_{\text{co}}$, $\forall m \in \mathscr{M}$. In the stage of data offloading, the distance between UAV and OBS at the k-th time slot is $d_{0,k} = \sqrt{\|q_k - D_0\|^2 + (H - H_0)^2}$, $\forall k \in \mathscr{K}_{\text{of}}$. H_m and H_0 denote the antenna heights of the m-th buoy and OBS, respectively.

2.2. Transmission Model

2.2.1. Channel Model. We adopt the model of the air-to-ground channel and the two-ray path loss model [29] and give the LoS and NLoS path loss models of the buoy-UAV and UAV-OBS links, respectively.

Specifically, the channel gain of buoy-UAV and UAV-OBS links is expressed as $h_{i,k} = 1/L_{i,k}, \forall k \in \mathscr{K}, \forall i \in \{0, 1, \cdots, M\}$, where $L_{i,k}$ denotes the average path loss in the k-th time slot, expressed as

$$\begin{aligned}
L_{i,k} &= P_{i,k}^{\text{LoS}} L_{i,k}^{\text{LoS}} + \left(1 - P_{i,k}^{\text{LoS}}\right) L_{i,k}^{\text{NLoS}},\\
L_{i,k}^{\text{LoS}} &= \left(\frac{4\pi d_{i,k}}{\lambda}\right)^2 \mu_{i,k}\xi_{\text{LoS}},\\
L_{i,k}^{\text{NLoS}} &= \left(\frac{4\pi d_{i,k}}{\lambda}\right)^2 \mu_{i,k}\xi_{\text{NLoS}},
\end{aligned} \tag{3}$$

where $L_{i,k}^{\text{LoS}}$ and $L_{i,k}^{\text{NLoS}}$ are the average path loss for LoS and NLoS, $\mu_{i,k} = 1$. ξ_{LoS} and ξ_{NLoS} are the excessive path loss for LoS and NLoS paths, respectively, λ is the wavelength, and $P_{i,k}^{\text{LoS}}$ denotes the probability of LoS link which is expressed as

$$P_{i,k}^{\mathrm{LoS}} = \frac{1}{1 + a \exp\left(-b\left(\psi_{i,k} - a\right)\right)}, \tag{4}$$

where a and b are two constant values depending on the environment and $\psi_{i,k}$ denotes the elevation angle between the m-th buoy (or OBS) and UAV, which is given by $\psi_{i,k} = (180^\circ/\pi) \times \arcsin\left(H - H_i/\sqrt{\|q_k - D_{i,k}\|^2}\right)$.

2.2.2. UAV Data Collection from Buoys. In this stage that $\forall k \in \mathscr{K}_{\mathrm{co}}$, let $\alpha_{m,k}$ denote the association indicator between m-th buoy and UAV. $\alpha_{m,k} = 1$ means that the m-th buoy is associated with the UAV in the k-th time slot. Otherwise, $\alpha_{m,k} = 0$.

In uplink NOMA system, the UAV is regarded as a receiver to receive signals from multiple buoys at the same time and allows multiple buoys to share the same channel. The SIC decoding technique is used to demodulate the received signals with different received power levels. The successfully demodulated signal is deleted from all received signals, and the later decoded signal receives less cochannel interference. Therefore, the buoy with high channel gain is usually demodulated first, and its interference comes from the buoys with worse channel gain [30, 31]. The cochannel interference of uplink transmission between the m-th buoy and UAV in the k-th time slot can be given by

$$I_{m,k} = \sum_{i \in \mathscr{M}_k} \alpha_{i,k} P_{i,k} h_{i,k}, \forall k \in \mathscr{K}_{\mathrm{co}}, \tag{5}$$

where $\mathscr{M}_k = \{i | i \in \mathscr{M}, h_{m,k} > h_{i,k}\}$ is the set of the buoy whose channel gain is worse than the m-th buoy in the k-th time slot.

Hence, the signal-to-interference-noise-ratio (SINR) between the m-th buoy and UAV at the k-th time slot is expressed as

$$g_{m,k} = \frac{P_{m,k} h_{m,k}}{\sigma^2 + I_{m,k}}, \forall m \in \mathscr{M}, \forall k \in \mathscr{K}_{\mathrm{co}}, \tag{6}$$

where $P_{m,k}$ is the transmit power of the m-th buoy in the k-th time slot and σ^2 is the noise power.

The transmission rate of m-th buoy in the k-th time slot is expressed as

$$R_{m,k} = \alpha_{m,k} B \log_2\left(1 + g_{m,k}\right), \forall m \in \mathscr{M}, \forall k \in \mathscr{K}_{\mathrm{co}}, \tag{7}$$

where B is the spectrum. In order for the received signal to be demodulated successfully, the SIC demodulation condition that SINR needs to meet is as [32]

$$\frac{\alpha_{m,k} P_{m,k} h_{m,k}}{\sum_{i \in \mathscr{M}_k} \alpha_{i,k} P_{i,k} h_{i,k} + \sigma^2} \geq \eta_{\mathrm{SIC}}, \forall m \in \mathscr{M}, \tag{8}$$

where η_{SIC} denotes the SIC threshold.

2.2.3. UAV Data Offloading to OBS. In this stage that $\forall k \in \mathscr{K}_{\mathrm{of}}$, let β_k denote the association indicator between UAV and OBS. $\beta_k = 1$ means that the UAV can offload the data to the OBS at the k-th time slot. Otherwise, $\beta_k = 0$.

The signal-to-noise-ratio (SNR) between the UAV and OBS at the k-th time slot needs to satisfy the following condition:

$$g_{0,k} = \frac{P_k h_{0,k}}{\sigma^2} \geq \bar{g}_0, \forall k \in \mathscr{K}_{\mathrm{of}}, \tag{9}$$

where P_k is the transmit power of the UAV and $\bar{g}_0$ is the SNR threshold.

The transmission rate between the UAV and OBS at the k-th time slot is given by

$$R_{0,k} = \beta_k B \log_2\left(1 + g_{0,k}\right), \forall k \in \mathscr{K}_{\mathrm{of}}. \tag{10}$$

2.3. Problem Formulation. Our goal is to minimize the UAV's total mission completion time by jointing optimization of the buoy-UAV association relationship, UAV transmit power, buoys transmit power, and UAV trajectory. Let $A_m = \{\alpha_{m,k}, \forall m \in \mathscr{M}, \forall k \in \mathscr{K}_{\mathrm{co}}\}$ denote the buoy-UAV-associated variables, $P = \{P_k, \forall k \in \mathscr{K}_{\mathrm{of}}\}$ denote the UAV transmit power during data offloading, $P_m = \{P_{m,k}, \forall m \in \mathscr{M}, \forall k \in \mathscr{K}_{\mathrm{co}}\}$ denote the buoys transmit power, and $Q = \{q_k, \forall k \in \mathscr{K}\}$ denote the UAV trajectory. The total mission completion time minimization problem can be formulated as

$$\begin{aligned}
(P1):\ & \min_{\mathbf{A}_m, \mathbf{P}, \mathbf{P}_m, \mathbf{Q}} T_{\mathrm{total}}, \\
\text{s.t.} C1 &: 0 \leq P_{m,k} \leq P_{m_{\max}}, \\
C2 &: 0 \leq P_k \leq P_{\max}, \\
C3 &: \alpha_{m,k} \in \{0, 1\}, \forall m \in \mathscr{M}, \forall k \in \mathscr{K}_{\mathrm{co}}, \\
C4 &: \beta_k \in \{0, 1\}, \forall k \in \mathscr{K}_{\mathrm{of}}, \\
C5 &: \sum_{m=1}^{M} \alpha_{m,k} \leq U, \forall k \in \mathscr{K}_{\mathrm{co}}, \\
C6 &: \sum_{k=K_{\mathrm{co}}+1}^{K} R_{0,k}\delta \geq \sum_{m=1}^{M} C_m, \\
C7 &: \sum_{k=1}^{K_{\mathrm{co}}} R_{m,k}\delta \geq C_m, \forall m \in \mathscr{M}, \\
C8 - C11 &: (3), (4), (13), (14)
\end{aligned} \tag{11}$$

In problem $P1$, $C1$, and $C2$ restrict the maximum transmit power of UAV and buoy, respectively. $C5$ limits the maximum number of buoys that can be associated with the UAV in each time slot. Let C_m denote the data size that needs to be collected in m-th buoy. $C6$ ensures that all data collected by the UAV is offloaded to OBS. $C7$ ensures that the data collection requirements of each buoy are met. $C8$ and $C9$ are the UAV maximum velocity and maximum acceleration constraints, respectively. $C10$ is the SIC demodulation constraint that SINR needs to meet in the data collection stage. $C11$ is the SNR constraint in the data offloading. Problem $P1$ is a mixed-integer nonconvex problem since it contains binary relational variables, which makes it difficult to be solved effectively.

3. Proposed Scheme

In order to solve problem $P1$, we first propose a TD3-based UAV trajectory optimization algorithm (TTO). Then, we design a heuristic algorithm to solve the power control problem while determining the buoy-UAV association relationship (PCAR). Finally, the above two algorithms are combined to effectively solve the problem $P1$.

3.1. TD3-Based UAV Trajectory Optimization. In our system, the start position of data offloading stage is similar to the end position of data collection, which is named as the transition position (TP) between two stages. Therefore, the UAV trajectories of these two stages are coupled, and the TP cannot be determined in advance. Furthermore, the UAV trajectory changes dynamically according to the requirements of data collection and data offloading. The traditional deterministic optimization method is difficult to solve the above problems [19]. Therefore, in this paper, we use an advanced DRL method, TD3, to solve the UAV trajectory optimization subproblem with the given transmit powers and association relationships. In the following, we first give the definitions of state, action, and reward and then briefly introduce the TD3.

3.1.1. State Definition. In the data collection stage, the UAV's action is closely related to the remaining data size of buoys and the buoy-UAV association relationship. Similarly, in the data offloading stage, the UAV's action is related to the remaining data size of UAV and the UAV-OBS association relationship. Furthermore, the UAV only collects data from buoys in the target area. Hence, we define the state space as

$$S_k = \{\alpha_k, c_k, x_k, y_k, \beta_k, c_k^{\mathrm{uav}}, \rho_k\}, \tag{12}$$

where the variables contained in the above expression are defined as

(i) $\alpha_k = \{\alpha_{1,k}, \alpha_{2,k}, \cdots, \alpha_{M,k}\}$ denotes the set of the buoy-UAV association relationship in the k-th time slot

(ii) $c_k = \{c_{1,k}, c_{2,k}, \cdots, c_{M,k}\}$ denotes the remaining data size of each buoy in the k-th time slot

$$c_{m,k} = C_m - \sum_{j=1}^{k-1} R_{m,j}\delta \tag{13}$$

(iii) c_k^{uav} denotes the remaining data size of UAV

$$c_k^{\mathrm{uav}} = \sum_{k=1}^{K_{\mathrm{co}}} \sum_{m=1}^{M} R_{m,k}\delta - \sum_{j=K_{\mathrm{co}}+1}^{k-1} R_{0,j}\delta \tag{14}$$

(iv) ρ_k denotes the boundary penalty information of the UAV to judge whether the UAV position exceeds the target area in the k-th time slot

3.1.2. Action Definition. Based on the above state and environment information, the UAV's action is defined as

$$A_k = \{\varphi_k, V_k\}, \tag{15}$$

where $\varphi_k \in (0, 2\pi]$ denotes the flight angle of UAV in the k -th time slot and $V_k \in [0, V_{\max}]$.

3.1.3. Reward Definition. Let $K_{\max}$ denote the number of maximum total mission completion time slots. $K^* = K_{\max} - K$ if the UAV completes the mission in K time slots. Then, we design the reward function as

$$r_k = \begin{cases} R_k\delta + \rho_k + K^*, & \text{if C6 and C7 are satisfied,} \\ C_k(\text{or } R_k\delta) + \rho_k, & \text{otherwise.} \end{cases} \tag{16}$$

It can be seen from (16) that the shorter the time it takes for the UAV to complete the mission, the greater the reward it will eventually obtain.

Note that for better performance, we reduce the order of magnitude of B and C_m by n_1 orders of magnitude to be less than or equal to the order of magnitude of $K_{\max}$ when calculating the state and reward.

3.1.4. TD3. The TD3 has the following advantages [27]:

(i) Clipped double Q-learning for actor-critic: TD3 contains two critic networks. For the two target Q-values generated by the two critic target networks, the minimum of them is selected to suppress the overestimation problem caused by high variance, expressed as

$$y = r_k + \gamma \min_{i=1,2} Q_{\theta_{i'}}\left(S', \tilde{A}\right) \tag{17}$$

where y is the target value which is used to update the two critic networks, r is reward, γ is the reward discount factor, $\tilde{A}$ is the target policy, $\theta_{i'}$ is the target network parameter, and S' is the next state

(ii) Target networks and delayed policy updates: the TD3 algorithm updates the actor and its target network after a fixed number of updates to the critic network, expressed as

$$\theta_{i'} \longleftarrow \tau\theta_i + (1-\tau)\theta_{i'}, i = 1, 2 \tag{18}$$

where τ is the update parameter

(iii) Target policy smoothing regularization: TD3 smoothes the estimate and reduces the error by adding a small amount of random noise to the target actor network and averaging over minibatch, expressed as

$$\tilde{A} \longleftarrow \pi_{\phi'}(S_k) + \varepsilon, \varepsilon \sim \text{clip}(\mathcal{N}(0, \tilde{\omega}), -c, c) \tag{19}$$

It can be seen from (12) that most of the state dimensions are related to buoys. Only two dimensions are related to the UAV's position, two dimensions are related to OBS, and one dimension is related to UAV boundary penalty information. Therefore, there is a problem of dimension imbalance. Dimension spread technology can effectively solve this problem. We spread the above state dimensions. For example, we connect the position state dimension of UAV to a spread network composed of M neurons and spread its dimension to M [33, 34]. Furthermore, we set a termination flag l to indicate whether UAV has completed its mission. l is applied to the target value function. Hence, the Q-value of the target value function is 0 after the UAV completes the mission, so as to make the critic learning performance more stable [34].

In summary, our proposed algorithm TTO is shown in Algorithm 1. The update process of TD3 is shown in Figure 2. In lines 1-3, we first initialize the network parameters and the experience replay buffer $\mathscr{B}$. In lines 5-17, we initialize the environment, obtain the initial state information, and set a time variable k to represent the time spent by UAV to perform the mission. Moreover, a termination mark l is set. $l = 0$ indicates that UAV has not completed the mission. Then, UAV makes an action selection according to the observed state and environmental information. Specifically, UAV constantly interacts with the environment and updates the actor and critic networks. The actor network outputs the action A_k to be executed by the UAV according to the state information S_k. Then, the transition information (S_k, A_k, r_k, S', l) is stored in $\mathscr{B}$. If the UAV flies beyond the target area, the action will be canceled and the boundary penalty information will be given. In the data collection stage, we remain $\beta_k = 0$. The data offloading stage begins when UAV completes the data collection mission. Then, we remain $\alpha_k = 0$. $l = 1$ if the UAV completes the data offloading mission. The episode is terminated when $l = 1$ or $k = K_{\max}$.

In lines 18-28, N transition information is randomly selected from $\mathscr{B}$ to form a minibatch, which is input into the actor and critic network. The actor network calculates the corresponding $\tilde{A}$ according to S'. After selecting the target Q -value and smoothing the target policy according to S' and $\tilde{A}$, the critic network minimizes the loss function and updates the critical network by the following way:

$$\theta_i \longleftarrow \arg \min_{\theta_i} N^{-1} \sum \left(y - Q_{\theta_i}(S_k, A_k)\right)^2. \tag{20}$$

Then, the actor network is updated in the way of delayed update by the deterministic policy gradient, expressed as

$$\nabla_\phi J(\phi) = N^{-1} \sum \nabla_{A_k} Q_{\theta_1}(S_k, A_k) \Big|_{A_k = \pi_\phi(S_k)} \nabla_\phi \pi_\phi(S_k). \tag{21}$$

Finally, the optimal trajectory of UAV is obtained by cyclic iteration until the maximum number of episodes $E_{\max}$.

3.2. Power Control and Buoy-UAV Association Relationship. Given the UAV trajectory, the problem $P1$ can be written as

$$\begin{aligned} (\text{P2}): & \min_{\mathbf{A}_m, \mathbf{P}, \mathbf{P}_m, \mathbf{Q}} \quad T_{\text{total}}, \\ \text{s.t.} & \quad \text{C1} - \text{C7}, \text{C10}, C11 \end{aligned} \tag{22}$$

Obviously, the problem $P2$ is still a mixed-integer nonconvex problem. Given the mission completion time and the association relationships of buoy-UAV and UAV-OBS, $P2$ can be transformed into a problem of maximizing the total transmission data size in the k-th time slot, which can be divided into two parts.

First, in the data collection stage, it can be seen from problem $P2$ that the SINR between UAV and buoys is related not only to the transmit power of buoys but also to the buoy-UAV association relationship. Therefore, in order to determine the buoy-UAV association relationship, we first introduce Lemmas 1 and 2.

Lemma 1. *The UAV must be associated with the first U^* $(U^* \le U)$ buoys with the larger channel gain in the k-th time slot.*

Proof. As can be seen from $P2a$, the total transmission data size C_k depends on the summation term $\sum_{m=1}^{U} P_m h_m$. Therefore, we might as well assume that the transmit power of all buoys is the maximum transmission power $P_{m_{\max}}$. The channel gain between M buoys and UAV is expressed in descending order as $\{h_1 \longrightarrow h_2 \longrightarrow \cdots \longrightarrow h_M\}$. Obviously, selecting the first U^* buoys with the largest channel gain can maximize the total transmission data size. □

Lemma 2. *The transmit power of the buoy with the largest channel gain among the buoys associated with the UAV in the k-th time slot must be $P_{m_{\max}}$.*

Proof. Except for the buoys transmit power, other assumptions are the same as Proof. If the UAV is associated with U buoys, for the buoy with the largest channel gain, the SINR between it and the UAV is expressed as

$$g_1 = \frac{P_1 h_1}{\sum_{i=2}^{U} P_i h_i + \sigma^2}. \tag{23}$$

In order to meet the constraints $C10$ and $C12$ and maximize the total transmission data size, the value of

```
Initialize critic networks $Q_{\theta_1}$, $Q_{\theta_2}$, and actor network $\pi_\phi$ with random parameters $\theta_1$, $\theta_2$, and $\phi$.
Initialize target networks $\theta_1{}' \longleftarrow \theta_1$, $\theta_2{}' \longleftarrow \theta_2$, and $\phi' \longleftarrow \phi$.
Initialize experience replay buffer $\mathscr{B}$.
for episode =0 to $E_{\max}$ do
  Initialize the environment and state $S_0$, and the terminated flag $l = 0$.
  for epoch $k = 1$ to $K_{\max}$ do
    Select action $A_k = \pi_\phi(S_k) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \omega)$, and observe reward $r_k$ and next state $S'$.
    if the UAV flies beyond the target area then
      $\rho_k = 1$. Then cancel the UAV's action and update
      $r_k$, $S'$ based on the current state.
    end if
    if $\sum_{k=1}^{K_{co}} R_{m,k}\delta \ge C_m, \forall m \in \mathscr{M}$ then
      let $\alpha_k = 0$, and start the data offloading.
    else
      Let $\beta_k = 0$, and continue the data collection.
    end if
    if $\sum_{k=k^*}^{K} R_k\delta \ge \sum_{m=1}^{M} C_m$ then
      $r_k = R_k\delta + \rho_k + K^*$, and let $l = 1$.
    end if
    Store transition tuple $(S_k, A_k, r_k, S', l)$ in $\mathscr{B}$.
    if $\mathscr{B} > 2{,}000$ then
      Sample mini-batch of $N$ transitions $(S_k, A_k, r_k, S', l)$ from $\mathscr{B}$.
      $\tilde{A} \longleftarrow \pi_{\phi'}(S') + \varepsilon$, $\varepsilon \sim$ clip$(\mathcal{N}(0, \tilde{\omega}), -c, c)$.
      $y \longleftarrow r_k + (1 - l) \cdot \gamma \min_{i=1,2} Q_{\theta_i}{}'(S', \tilde{A})$.
      Update critics:
      $\theta_i \longleftarrow \arg\min_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(S_k, A_k))^2$.
      Update the actor policy $\phi$ by the deterministic policy gradient:
      $\nabla_\phi J(\phi) = N^{-1} \sum \nabla_{A_k} Q_{\theta_1}(S_k, A_k)|_{A_k = \pi_\phi(S_k)} \nabla_\phi \pi_\phi(S_k)$.
      Update target networks:
      $\theta_i{}' \longleftarrow \tau\theta_i + (1 - \tau)\theta_i{}'$.
      $\phi' \longleftarrow \tau\phi + (1 - \tau)\phi'$.
    end if
  end for
end for
return The UAV trajectory $\mathbf{Q}$
```

Algorithm 1: TD3-based trajectory optimization algorithm (TTO).

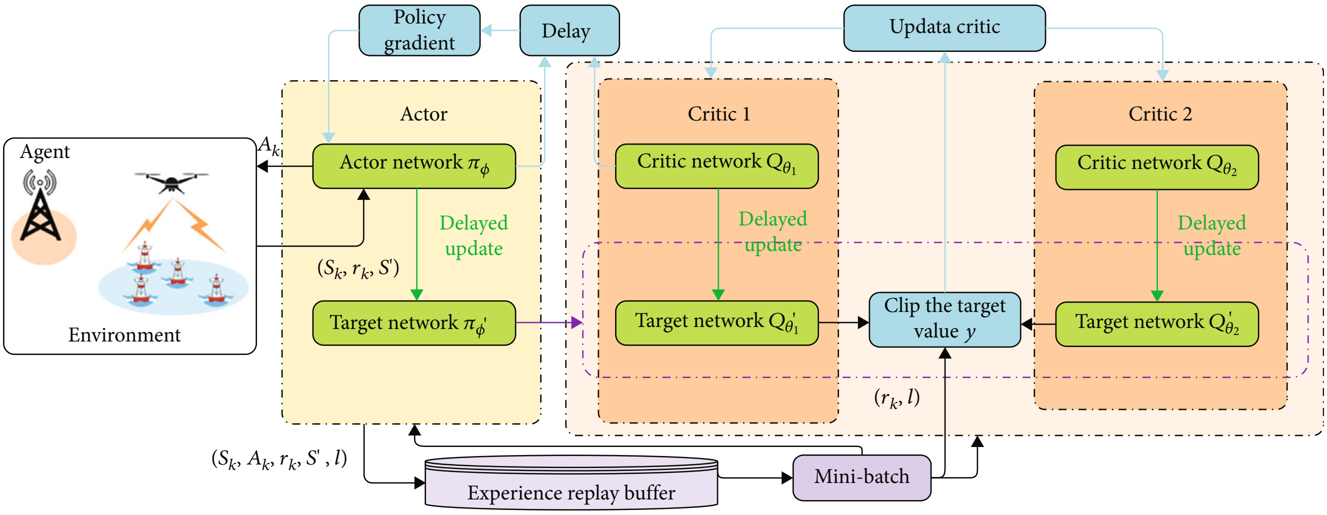

Figure 2: TD3-based UAV trajectory optimization.

```
Input the UAV's current position $q_k$, $\eta_{\mathrm{SIC}}$ and $U$.
Input current channel gain $g_{m,k}$, and sort it in descending order to obtain $g'_{m,k}$.
According to $U$, get initial NOMA group and the number of initial associated buoys $U^* = U$.
repeat
    Solve P2a to obtain the optimal solution $\mathbf{P}_m{}^*$ and the optimal value $C_k$.
    if the solution state is not optimal then
        Remove the one with the worst channel gain in the currently associated buoy.
        $U^* \longleftarrow U^* - 1$.
    end if
until $U^* = 1$.
return $\mathbf{P}_m{}^*$ and $\mathbf{A}_m$;
```

Algorithm 2: Power control and buoy-UAV association relationship algorithm (PCAR).

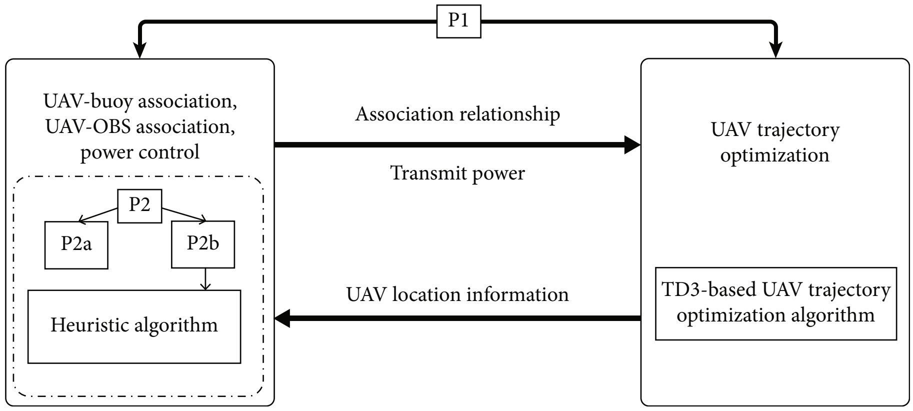

Figure 3: Joint TD3-based trajectory optimization, power control, and buoy-UAV association relationship scheme.

```
Input:  The UAV's initial position $q_1$, the buoys' position $D_m$, the OBS's position $D_0$;
Output:  $\mathbf{A}_m$, $\mathbf{P}$, $\mathbf{P}_m$, $\mathbf{Q}$;
for episode =0 to $E_{\max}$ do
    for epoch $k = 1$ to $K_{\max}$ do
        /* Lines 11-15 of Algorithm 1 */
        if $\sum_{k=1}^{K_{co}} R_{m,k}\delta \geq C_m, \forall m \in \mathcal{M}$ then
            Let $\alpha_k = 0$, and obtain current channel gain $g_{0,k}$.
            Set $\mathbf{P} = P_{\max}$.
            if $g_{0,k} \geq \bar{g}_0$ then
                Let UAV-OBS association relationship $\beta_k = 1$.
            end if
        else
            Let $\beta_k = 0$, and obtain current channel gain $g_{m,k}$.
            Update $P_{m,k}, \forall m \in \mathcal{M}$ and $\alpha_k$ with given $q_k$ by performing Algorithm 2.
            /* Lines 7-10 of Algorithm 1 */
            Update $q_k$ with given tranmist power and association relationship.
        end if
    end for
end for
```

Algorithm 3: Joint TD3-based trajectory optimization, power control, and buoy-UAV association relationship scheme (TTO-PCAR).

Table 1: Simulation parameters.

Parameter	Value
Maximum acceleration, $\Delta_{\max}$	25 m/s^2
Environment parameter, a, b	9.61, 0.16
Excessive path loss, ξ_{LoS}, ξ_{NLoS}	1, 20
Channel parameter, $\mu_{m,k}$	1
The antenna heights of buoys and OBS, H_m, H_0	0 m, 0 m
Wavelength, λ	0.15 m
Noise power, σ^2	-94 dBm
SINR and SNR threshold, $\bar{g}_{\text{co}}$, $\bar{g}_{\text{of}}$	10 dB, 3 dB
SIC threshold, η_{SIC}	10 dB
Order of magnitude parameter, n	6

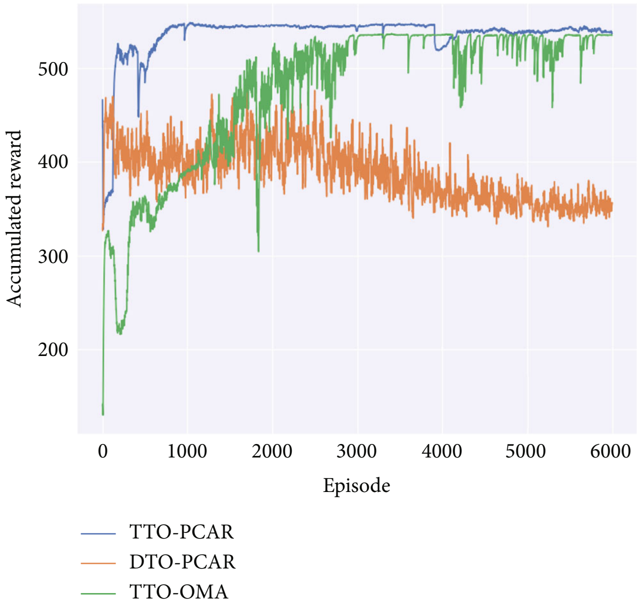

Figure 4: Accumulative reward.

denominator term P_1h_1 should be as large as possible, so as to maximize the value of molecular term and make more buoys connected to UAV. Therefore, P_1 should be $P_{m_{\max}}$.

The total transmission data size of U^* buoys in the k-th time slot is expressed as

$$C_k = \sum_{m=1}^{U^*} B \log_2\left(1 + \frac{P_{m,k}h_{m,k}}{\sum_{i\in\mathcal{M}_k} P_{i,k}h_{i,k} + \sigma^2}\right)\delta. \tag{24}$$

Therefore, problem $P2$ can be transformed into

$$\begin{aligned}
&(\text{P2a}):\ \max_{\mathbf{P}_m} C_k,\\
&\text{s.t.}\quad \text{C1}: 0 \le P_m \le P_{m_{\max}},\\
&\text{C10}: \frac{P_{m,k}h_{m,k}}{\sum_{i\in\mathcal{M}_k} P_{i,k}h_{i,k} + \sigma^2} \ge \eta_{SIC}
\end{aligned} \tag{25}$$

Due to the existence of cochannel interference between buoys, $P2a$ is still nonconvex. Therefore, we convert C_k into the following form:

$$C_k = B\delta \log_2\left(1 + \frac{\sum_{m=1}^{U^*} P_{m,k}h_{m,k}}{\sigma^2}\right). \tag{26}$$

Therefore, $P2a$ is a convex problem that can be solved by a standard convex optimization solver (such as cvxpy). □

The algorithm PCAR is shown in Algorithm 2. In Algorithm 2, the channel gains are first sorted in descending order. It is stipulated that the UAV is associated with U buoys at most, and the first U buoys with the largest channel gain are selected to form the initial NOMA group. Then, we solve $P2a$. If $P2a$ has an optimal solution, the optimal solution P_m and the optimal value C_k are obtained. If $P2a$ has no optimal solution, the buoy with the worst channel gain is removed from the current NOMA group to form a new NOMA group. The above process is repeated until $P2a$ has an optimal solution. Note that the UAV is associated with at least one buoy in each time slot.

Second, in the data offloading stage, the total transmission data size of UAV in the k-th time slot is expressed as

$$C_k^{\text{uav}} = B\delta \log_2\left(1 + \frac{P_k h_{0,k}}{\sigma^2}\right). \tag{27}$$

Then, $P2$ can be transformed into the following form:

$$\begin{aligned}
&(\text{P2b}):\ \max_{\mathbf{P}} C_k^{\text{uav}},\\
&\text{s.t.C2}: 0 \le P_k \le P_{\max},\\
&\text{C11}: \frac{P_k h_{0,k}}{\sigma^2} \ge \bar{g}_0
\end{aligned} \tag{28}$$

Problem $P2b$ is a standard convex problem, the optimal solution of which is $P_{\max}$.

In summary, we propose a joint TTO and PCAR scheme (TTO-PCAR) to solve the problem $P1$. The TD3 agent is deployed in the OBS, and OBS maintains the communication with the UAV. During training, UAV collects data from buoys through traffic channel. Meanwhile, the UAV receives the states information of buoys through the control channel and feeds back the states information of itself and buoys to OBS. The OBS operates the proposed scheme with the above states information and sends the operation results to UAV in each time slot. Then, the UAV forwards relevant signalling (such as transmit power of buoys and buoy-UAV association relationship) to the buoy through the control channel. The specific process is shown in Figure 3 and Algorithm 3. Specifically, the UAV initial position is first given. In lines 3-12, P and A_m are obtained by Algorithm 2 according to the UAV current position q_k in the data collection stage. β_k is obtained according to q_k and $\bar{g}_0$ in the data offloading stage. Then, the UAV next position information is updated by lines 7-10 of Algorithm 1. Finally, the above process is repeated until $E_{\max}$.

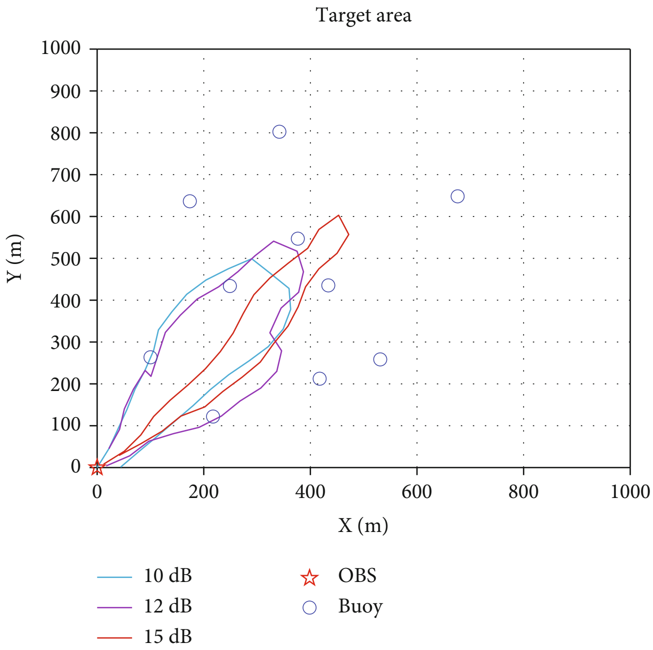

Figure 5: UAV trajectory with different SIC thresholds.

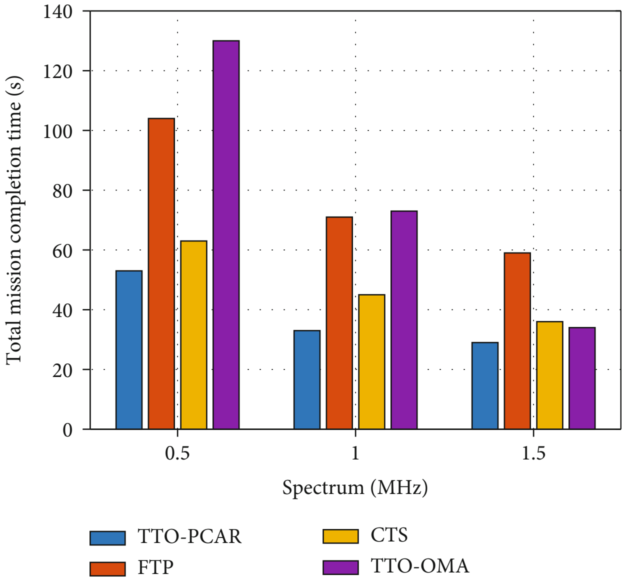

Figure 6: Total mission completion time with different schemes.

3.3. Complexity Analysis. TD3 contains two actor networks and four critic networks. Hence, the computational complexity of Algorithm 1 is $\mathcal{O}(2\sum_{e=1}^{E^{\mathrm{a}}} n^{\mathrm{a}}{}_{e} n^{\mathrm{a}}{}_{e-1} + 4\sum_{e=1}^{E^{\mathrm{c}}} n^{\mathrm{c}}{}_{e} n^{\mathrm{c}}{}_{e-1})$. E^{a} is the number of the fully connected layers of actor network. E^{c} is the number of the fully connected layers of critic network. n^{a} and n^{c} are the unit numbers in the e-th layer of the actor network and the critic network, respectively. Since the UAV is associated with at most U buoys in each time slot, the computational complexity of Algorithm 2 is $\mathcal{O}(U)$. Hence, the computational complexity of Algorithm 3 is $\mathcal{O}(E_{\max} K_{\max}(\sum_{e=1}^{E^{\mathrm{a}}} n^{\mathrm{a}}{}_{e} n^{\mathrm{a}}{}_{e-1} + \sum_{e=1}^{E^{\mathrm{c}}} n^{\mathrm{c}}{}_{e} n^{\mathrm{c}}{}_{e-1} + U))$.

4. Simulation Results

In simulation, the considered target area is 1000 m × 1000 m where $M = 10$ buoys are randomly distributed. A UAV is used to collect data with a fixed height $H = 100$ m from the target area. The flight velocity of UAV is $V_{\max} = 50$ m/s and $V_{\min} = 0$ m/s. The maximum transmit power of UAV and buoys is $P_{\max} = 0.1$ W and $P_{m_{\max}} = 24$ dBm, respectively. The position of OBS is $D_0 = (0, 0)$ m. The UAV is allowed to associate up to 3 buoys in each time slot, i.e., $U = 3$. The time slot length is $\delta = 1$ s. The data size range of each buoy is $C_m \in [10, 20]$ Mbits. The spectrum is $B = 1$ MHz. Furthermore, our proposed algorithm TTO is based on Pytorch. For actor and critic networks, we use a fully connected DNN with two hidden layers of 400 neurons. The learning rate is 0.0001. The experience memory buffer size is 100000. The minibatch size N is 256. The discount factor γ is 0.99. $\tilde{\omega} = 0.36$. $\tau = 0.005$. $K_{\max} = 300$. Other simulation parameters are shown in Table 1.

In order to compare performance, we use the following scheme as the comparison algorithm.

(i) UAV trajectory based on Fermat point (FTP) [35]: this scheme first regards each user as the vertex of a triangle to form multiple triangles. Then, the Fermat points of each triangle are taken as the hovering points of the UAV. The UAV hovers at the points in turn to collect data

(ii) UAV trajectory based on circle scheme (CTS): this scheme first finds the geometric center of all users as the center of the circle and then averages the distance from all users to the center of the circle to determine the radius of the UAV trajectory

(iii) UAV data collection based on OMA (TTO-OMA): this scheme refers to the UAV using OMA technology for data collection. The proposed TTO algorithm is still used to determine the UAV trajectory

(iv) DRL scheme based on DQN (DTO-PCAR): this scheme uses DQN instead of TD3 in our proposed algorithm

Figure 4 shows the comparison of accumulative reward for different schemes. For the convenience of observation, we smoothed the curves. It can be seen that the proposed TTO-PCAR scheme could be convergent after 1000 episodes, while the compared TTO-OMA scheme needs 3000 episodes to be convergent. Moreover, the compared DTO-PCAR scheme cannot be convergent after 6000 episodes. Therefore, the performance of our proposed scheme is significantly better than the other two schemes. Figure 5 shows the UAV trajectory comparison with TTO-PCAR scheme under different SIC thresholds. The SIC thresholds are 10 dB, 12 dB, and 15 dB, respectively. We find that the average total mission completion time of UAV is basically the same, which is 33 s, 36 s, and 37 s, respectively. However, the UAV trajectory is closer to the farther buoy with the increase of SIC threshold. This is because when the UAV

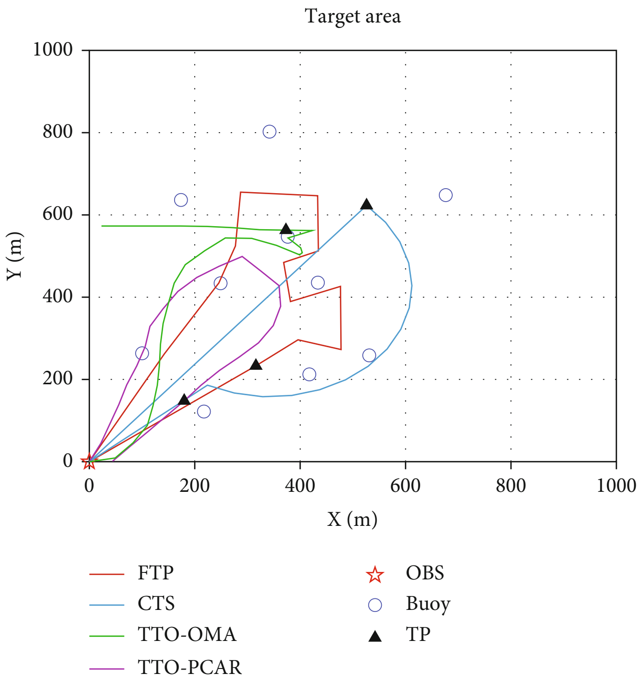

Figure 7: UAV trajectory of different schemes.

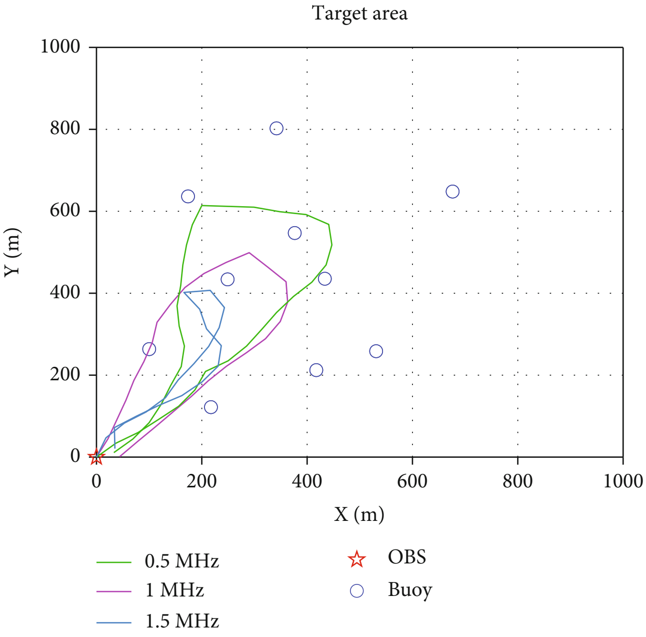

Figure 8: UAV trajectory with different spectrums.

uses NOMA technology for data collection, the channel gain of the farther buoy is poor. Therefore, in order to meet the SIC constraint, the UAV will gradually fly to the farther buoys whose data has not yet been collected.

Figure 6 shows the comparison of the total mission completion time of under different spectrums. Figure 7 shows the UAV trajectory obtained by our proposed scheme with $M = 10$ buoys and compares it with the other three schemes. It can be seen that the total mission completion time of

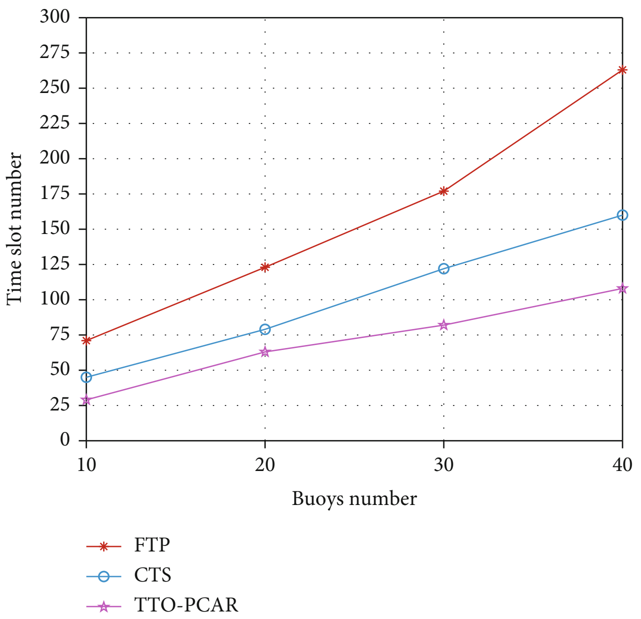

Figure 9: Total mission completion time with different buoy numbers.

TTO-PCAR is significantly lower than that of other schemes. In particular, the data collection time of our proposed scheme is 20 s with $B = 1$ MHz and that of TTO-OMA is 33 s; thus, NOMA is more efficient in data collection than OMA. This is because the designed reward (shown in Equation (16)) is related to the total transmission rate in each time slot. Second, the trajectory of UAV data collection process is fixed with the FTP and CTS scheme, resulting in the UAV trajectory of data offloading process longer. TTO-PCAR scheme takes the coupling of two stages into the consideration of UAV trajectory optimization; thus, the time of data offloading process is less. The total flight distance based on TTO-PCAR is also significantly lower than that of the other two schemes.

Figure 8 shows the UAV trajectory based on TTO-PCAR with different spectrums. It can be seen from Figure 8 that the flight distance of UAV decreases with the increase of spectrum. This is because the transmission rate of buoys is reduced with the reduction of spectrum. If the data of the buoy far from the OBS has not been collected, the agent chooses to make the UAV closer to the buoy in order to increase the transmission rate and obtain greater reward according to (16).

Figure 9 shows the total mission completion time of different schemes with different buoy numbers. FTP scheme is to find the hover points to collect data and classify the problem as a travelling salesman problem, so as to traverse the hover points. Hence, FTP takes a lot of time on UAV flight. Although CTS scheme can collect data in each time slot, it does not consider the data collection requirements of different buoys, because the UAV just flies based on circle. The proposed scheme TTO-PCAR dynamically adjusts the UAV trajectory according to the data collection requirements of different buoys. Therefore, the total mission completion time of TTO-PCAR is significantly lower than that of FTP and CTS.

5. Conclusion

This paper has investigated the joint optimization problem of the buoy-UAV association relationship, transmit powers, and the UAV trajectory for NOMA-enabled UAV data collection and offloading in MIoT. First, we propose a TD3-based UAV trajectory optimization algorithm to solve the UAV trajectory subproblem. Second, we design a heuristic algorithm to solve the subproblem of power control and buoy-UAV association relationship. Finally, we propose a joint TD3-based trajectory optimization, power control, and buoy-UAV association relationship scheme. The proposed scheme can effectively solve the mixed-integer nonconvex problem. Simulation results show that the proposed scheme significantly shortens the total mission completion time of UAV. In future work, we will investigate the problem of UAV trajectory optimization based on NOMA to shorten the time for UAV to perform mission in the MIoT.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant 2019YFE0111600, in part by the National Natural Science Foundation of China under Grant Nos. 62101089, 62002042, 51939001, and 61971083, in part by the China Postdoctoral Science Foundation under Grant Nos. 2021M700655 and 2021M690022, in part by the Liaoning Revitalization Talents Program under Grant No. XLYC2002078, and in part by the Fundamental Research Funds for the Central Universities under Grant No. 3132022231.

References

[1] T. Wei, W. Feng, Y. Chen, C.-X. Wang, N. Ge, and J. Lu, "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, 2021.

[2] Y. Li, Y. Zhang, W. Li, and T. Jiang, "Marine wireless big data: efficient transmission, related applications, and challenges," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 19–25, 2018.

[3] A. Laun and E. Pittman, "Development of a small, low-cost, networked buoy for persistent ocean monitoring and data acquisition," in *OCEANS 2018 MTS/IEEE*, pp. 1–6, Charleston, 2018.

[4] Y. Huo, X. Dong, and S. Beatty, "Cellular communications in ocean waves for maritime Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9965–9979, 2020.

[5] X. Li, W. Feng, Y. Chen, C.-X. Wang, and N. Ge, "Maritime coverage enhancement using UAVs coordinated with hybrid satellite-terrestrial networks," *IEEE Transactions on Communications*, vol. 68, no. 4, pp. 2355–2369, 2020.

[6] Y. Wang, W. Feng, J. Wanga, and T. Q. Quek, "Hybrid satellite-UAV-terrestrial networks for 6g ubiquitous coverage: a maritime communications perspective," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 11, pp. 3475–3490, 2021.

[7] H. Shen, Q. Ye, W. Zhuang, W. Shi, G. Bai, and G. Yang, "Drone-small-cell-assisted resource slicing for 5G uplink radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 7, pp. 7071–7086, 2021.

[8] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: a connectivity-constrained trajectory optimization perspective," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2580–2604, 2019.

[9] C. Zhan and Y. Zeng, "Aerial–ground cost tradeoff for multi-UAV-enabled data collection in wireless sensor networks," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1937–1950, 2020.

[10] X. Li, W. Feng, Y. Chen, C.-X. Wang, and N. Ge, "UAV-enabled accompanying coverage for hybrid satellite-UAV-terrestrial maritime communications," in *2019 28th Wireless and Optical Communications Conference (WOCC)*, pp. 1–5, Beijing, China, 2019.

[11] D. S. Lakew, A. Masood, and S. Cho, "3D UAV placement and trajectory optimization in UAV assisted wireless networks," in *2020 International Conference on Information Networking (ICOIN)*, pp. 80–82, Barcelona, Spain, 2020.

[12] Y. Dai, J. Liu, M. Sheng, N. Cheng, and X. Shen, "Joint optimization of BS clustering and power control for NOMA-enabled comp transmission in dense cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1924–1937, 2021.

[13] J. Zhao, Y. Wang, Z. Fei, X. Wang, and Z. Miao, "Noma-aided UAV data collection system: trajectory optimization and communication design," *IEEE Access*, vol. 8, pp. 155843–155858, 2020.

[14] D. Hu, Q. Zhang, Q. Li, and J. Qin, "Joint position, decoding order, and power allocation optimization in UAV-based NOMA downlink communications," *IEEE Systems Journal*, vol. 14, no. 2, pp. 2949–2960, 2020.

[15] N. Senadhira, S. Durrani, X. Zhou, N. Yang, and M. Ding, "Uplink NOMA for cellular-connected UAV: impact of UAV trajectories and altitude," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5242–5258, 2020.

[16] W. Chen, S. Zhao, R. Zhang, Y. Chen, and L. Yang, "UAV-assisted data collection with nonorthogonal multiple access," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 501–511, 2021.

[17] R. Tang, W. Feng, Y. Chen, and N. Ge, "NOMA-based UAV communications for maritime coverage enhancement," *China Communications*, vol. 18, no. 4, pp. 230–243, 2021.

[18] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4576–4589, 2019.

[19] X. Liu, M. Chen, Y. Liu, Y. Chen, S. Cui, and L. Hanzo, "Artificial intelligence aided next-generation networks relying on UAVs," *IEEE Wireless Communications*, vol. 28, no. 1, pp. 120–127, 2021.

[20] X. Shen, J. Gao, W. Wu et al., "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 45–66, 2020.

[21] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2021.

[22] R. Zhong, X. Liu, Y. Liu, and Y. Chen, "Multi-agent reinforcement learning in noma-aided uav networks for cellular offloading," 2021, https://arxiv.org/abs/2010.09094.

[23] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and A. Nallanathan, "Deep reinforcement learning based dynamic trajectory control for UAV assisted mobile edge computing," 2021, https://arxiv.org/abs/1911.03887.

[24] R. Zhang, M. Wang, L. X. Cai, and X. Shen, "Learning to be proactive: self-regulation of UAV based networks with UAV and user dynamics," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4406–4419, 2021.

[25] Y. Wang, W. Fang, Y. Ding, and N. Xiong, "Computation offloading optimization for UAV-assisted mobile edge computing: a deep deterministic policy gradient approach," *Wireless Networks*, vol. 27, no. 4, pp. 2991–3006, 2021.

[26] Q. Ye, W. Shi, K. Qu, H. He, W. Zhuang, and X. Shen, "Joint ran slicing and computation offloading for autonomous vehicular networks: a learning-assisted hierarchical approach," *IEEE Open Journal of Vehicular Technology*, vol. 2, pp. 272–288, 2021.

[27] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," 2018, http://arxiv.org/abs/1802.09477.

[28] M. Sun, X. Xu, X. Qin, and P. Zhang, "Aoi-energy-aware UAV-assisted data collection for iot networks: a deep reinforcement learning method," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17275–17289, 2021.

[29] J. Zhang, F. Liang, B. Li, Z. Yang, Y. Wu, and H. Zhu, "Placement optimization of caching UAV-assisted mobile relay maritime communication," *China Communications*, vol. 17, no. 8, pp. 209–219, 2020.

[30] Y. Dai, M. Sheng, J. Liu, N. Cheng, X. Shen, and Q. Yang, "Joint mode selection and resource allocation for d2d-enabled NOMA cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6721–6733, 2019.

[31] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7244–7257, 2016.

[32] Y. Wang, Z. Gao, J. Zhang et al., "Trajectory design for UAV-based Internet-of-Things data collection: a deep reinforcement learning approach," 2021, https://arxiv.org/abs/2107.11015.

[33] R. Duan, J. Wang, C. Jiang, H. Yao, Y. Ren, and Y. Qian, "Resource allocation for multi-UAV aided IoT NOMA uplink transmission systems," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 7025–7037, 2019.

[34] R. Ding, F. Gao, and X. S. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: a deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7796–7809, 2020.

[35] L. Lyu, Z. Chu, B. Lin, Y. Dai, and N. Cheng, "Fast trajectory planning for UAV-enabled maritime IoT systems: a Fermat-point based approach," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 328–332, 2022.