Multi-Agent Deep Reinforcement Learning Framework Strategized by Unmanned Aerial Vehicles for Multi-Vessel Full Communication Connection

Jiabao Cao [1], **Jinfeng Dou** [2,*] , **Jilong Liu** [1], **Xuanning Wei** [2] **and Zhongwen Guo** [2]

[1] School of Science, Qingdao University of Technology, Qingdao 266520, China; caojiabao@qut.edu.cn (J.C.)
[2] College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China; guozhw@ouc.du.cn (Z.G.)
* Correspondence: jinfengdou@ouc.edu.cn

Abstract: In the Internet of Vessels (IoV), it is difficult for any unmanned surface vessel (USV) to work as a coordinator to establish full communication connections (FCCs) among USVs due to the lack of communication connections and the complex natural environment of the sea surface. The existing solutions do not include the employment of some infrastructure to establish USVs' intragroup FCC while relaying data. To address this issue, considering the high-dimension continuous action space and state space of USVs, we propose a multi-agent deep reinforcement learning framework strategized by unmanned aerial vehicles (UAVs). UAVs can evaluate and navigate the multi-USV cooperation and position adjustment to establish a FCC. When ensuring FCCs, we aim to improve the IoV's performance by maximizing the USV's communication range and movement fairness while minimizing their energy consumption, which cannot be explicitly expressed in a closed-form equation. We transform this problem into a partially observable Markov game and design a separate actor–critic structure, in which USVs act as actors and UAVs act as critics to evaluate the actions of USVs and make decisions on their movement. An information transition in UAVs facilitates effective information collection and interaction among USVs. Simulation results demonstrate the superiority of our framework in terms of communication coverage, movement fairness, and average energy consumption, and that it can increase communication efficiency by at least 10% compared to DDPG, with the highest exceeding 120% compared to other baselines.

Keywords: communication coverage; full communication connection; internet of vessels; multi-agent deep reinforcement learning; unmanned aerial vehicles; unmanned surface vessels

1. Introduction

With the development of marine technology and the potential to greatly improve the operational efficiency and fuel economy of various marine engineering projects, the marine Internet of Vessels (IoV) is widely used in marine fisheries, marine pollution treatment, marine disaster monitoring, disaster rescuing, etc., providing them with massive data collection, data exchange, and data processing services [1]. It uses the effective communication of intelligent devices, such as ocean ships and unmanned surface vessels (USVs), to monitor and share data. Hence, the full communication connection (FCC) of USVs and high communication coverage of the target sea area are the primary tasks for the reliable and effective implementation of offshore missions in modern eceanographic observations [2,3].

The unstable resource consumption and the uncontrollable communication connection between USVs often lead to an ineffective FCC and poor data collection quality [4]. For example, most communication connections among the randomly deployed USVs are uncertain in the initial stage of IoV deployment, and some vessels often sail out of communication range or exhaust energy prematurely when the marine task is executed. In the

IoV, coastal base stations, satellites, and unmanned aerial vehicles (UAVs) often act as relay nodes to assist and provide communication connections with USVs. Nevertheless, USVs in extreme weather conditions or some offshore zones cannot maintain communication connections with coastal base stations. Furthermore, considering the real economic situation, it is also not ideal to equip all USVs with high-gain antennas to request communication services from satellites because of the relatively high costs and low data rate [5]. Due to their agile flexibility and mobility, relatively low cost, and high data rate, UAVs attract lots of attention [6–10]. At the same time, UAVs have some resource limitations and do not work all day. Compared to participating in long-term relay communication, UAVs have a better advantage in on-demand communication [11]. Therefore, in addition to the relay communication provided by the above infrastructure, the intragroup FCC of USVs is imperative for offshore tasks.

First, the current research lacks the effective infrastructure to establish a multi-USV FCC. Second, it is difficult to rely solely on USVs to achieve the establishment of an FCC among the USVs because the existence of sea clutter makes the marine environment have a long transmission distance and an irregularly deteriorated channel due to its blocking [12]. USVs also cannot work as coordinators to deal with multi-USV FCC tasks due to incomplete communication among them before establishing the FCC. Third, the existing technologies support heterogeneous communication between UAVs and USVs [13]. Fourth, the movement of USVs will impact economic costs and operation efficiency, such as fuel energy consumption and communication coverage [8]. Motivated by the above cases, this paper proposes that UAVs be deployed over the target sea area to navigate multiple USVs and adjust their positions when multi-USV FCC demand is driven. This would establish a multi-USV FCC and maximize the communication coverage of the target sea area in a fair and energy-saving manner. We elaborate on the challenges of designing such mechanisms as well as our approaches to addressing them.

The first challenge comes from the fact that the process of achieving an FCC by multi-USV position adjustment is a continuous control task. The state space and action space of the agents are continuous and interrelated in this environment, and it is difficult to discretize actions into separate and effective action vectors, especially for large-scale action vectors. Therefore, the single-agent reinforcement learning algorithm is not suitable for USVs. And the movement decisions of USVs need to be made in each time slot by the position selection and feedback of each USV, whereas USVs cannot effectively communicate with each other before establishing an FCC. Additionally, the IoV needs to both maintain an FCC among USVs and optimize their performance, including maximizing communication coverage, reducing resource waste among USVs, and relatively improving the fairness of resource utility. We need to explicitly design the quantitative indexes and balance the multi-object optimization problem. Hence, we propose a multi-agent deep reinforcement learning (MADRL) scenario strategized by UAVs for a multi-USV FCC, called "UST-MADRL," to solve the high-dimensional, continuous action, and state space problem. We transform our problem into a partial observable Markov game (POMG) and propose a novel multi-agent actor–critic structure (ACS), referred to as UST-ACS, based on the multi-agent deep deterministic policy gradient algorithm (MADDPG). In this setup, the UAVs act as critics to evaluate the actions of USVs, optimize parameters, and help USVs centrally train. Meanwhile, the USVs with distributive execution act as actors and simulated samples.

Furthermore, to enable effective decision making on USV movement through centralized training, the critics of UAVs need to obtain global information from USVs. The state–action information of each USV changes with the policy decision, which will change the effective messages passing between UAVs and USVs. We need to encode the local messages of a single USV and fuse multi-USV messages to form a global message. Therefore, we design an information transition module that obtains the state–action information of USVs in each time slot, fuses it into global information, and sends it to the critics in the UAV.

With the increase in the number of USVs, the number of agents in the environment and the dimension explosion of the state space and action space will intensify, making

it difficult for the solution algorithms to converge. In this paper, based on the strategy of signal-weighted Voronoi cell division [9], we divide the target sea area into several sub-areas, which are called sea service areas (SSAs). Therefore, our solution algorithm can be executed in each SSA, and the contact among the UAVs ensures the cooperation of all UAVs and USVs. Based on this design, we can effectively reduce the computational complexity of the algorithm.

Overall, we integrate the above components into an integrated UST-MADRL framework and jointly solve the challenges mentioned above. This work includes the first mechanism to study concept that the IoV realizes the intragroup FCC of USVs by reinforcement learning (RL) based on the evaluation of UAVs. The main contributions of this paper are summarized as follows:

- We use a deep neural network (DNN) to model each USV and dispatch UAVs to evaluate the movement policy of USVs. The USVs not only interact with the environment, but also interact with each other. In the distributed scenes, multiple USV agents learn to make decisions based on their local observations and the global cooperation by UAVs for the same system targets.
- Considering the marine communication feature and the joint behavior of multiple USVs, we define the optimization indexes, including communication coverage, energy consumption, USVs movement fairness, and communication efficiency. Balancing the optimization among them is a coupled and non-convex problem. Therefore, we transform it into a POMG. Then, to achieve an FCC with the improved optimization indexes, we design a UST-ACS, separating the the agents in the USVs and UAVs, i.e., USVs act as actors and UAVs act as critics. The UAVs can efficiently communicate with the USVs, evaluate the movement policy of USVs in the critic networks of UAVs, and feed the evaluation back to the USVs.
- The information transition module is responsible for collecting the action–state information of USVs and for message fusion. Meanwhile, we divide the target sea area into several SSAs and execute the UST-ACS in each SSA in order to effectively reduce the computational complexity and facilitate the convergence of solution algorithms.
- We perform extensive simulations and compare them to the deep deterministic policy gradient (DDPG) and three baselines. A large number of experiment results show that our proposed framework has better communication coverage performance, higher communication efficiency, and fairer movement decision making.

The rest of this paper is summarized as follows: In Section 2, the related work is reviewed. In Section 3, the system model and problem definition are introduced. The problem formulation is presented in Section 4. In Section 5, the solution is proposed. In Section 6, simulation results are presented. Section 7 concludes the paper.

2. Related Work

2.1. Full Communication Connections among USVs

Some maritime communication research has been dedicated to maintaining the communication connections within the group of USVs and ensuring the stable communication link between ship users. The authors of [14] studied how the USVs applied a long-range (LR) Wi-Fi wireless communication system to support the transmission of important information with fishing vessels or other vessels users. The vessel cooperative waterway intersection scheduling [15] was proposed to allocate a desired arrival time for the vessels, which helps avoid collisions and improves the communication connectivities in the urban waterway networks. The deployment of multiple USVs from the perspective of game theory was studied in [16], quantifying the transmission information of USVs through a uniform quantizer and ensuring the communication connections of multiple USVs. The study developed a marine broadband communication network to improve the transmission quality of the video packets between vessels and proposed three offline algorithms to improve the weighted throughput of the video packet transmission [17]. In [18], the communication topology is built to optimize the intra-group and inter-group communication

of USVs with lower communication cost. The existence of sea clutter makes the marine environment have a long transmission distance and an irregularly deteriorated channel. It is difficult for USVs to maintain the FCC in an efficient manner while only depending on their own communication and coordination abilities.

2.2. UAV-Assisted USV Networks

Some works noted that UAVs assisted USVs to relay the data. In maritime tasks, UAVs can provide the communication connections and the data processing services for USVs. The authors of [7] deployed a UAV-relay to carry out cooperative communication with USVs. The deep reinforcement learning (DRL) algorithm is used to obtain the optimal position of the UAV. A UAV-assisted mobile relay communication system in a downlink maritime communication is proposed to make the average reachability between the UAV and the offshore users meet the communication requirements. The UAV is dispatched to accompany the vessels users to sail, so as to ensure the communication connections among the vessels, shore base stations and satellites in [8]. The study proposed a network system structure where UAVs assist USVs to collect the scientific data offshore, in which UAVs are deployed as relay nodes to assist the backbone network to complete the communication [19]. A solution combining the RL strategy with the whale optimization algorithm was introduced to improve the data transmission rate and the delay between UAVs and USVs [20]. Another work [21] proposed a downlink maritime communication to make the average reachability between the UAV and offshore users meet the communication requirements by optimizing the positions of the UAV. In 5G and beyond networks, the authors established an on-demand style and a ubiquitous trust evaluation framework to eliminate malicious mobile data collectors and create a clean data collection and communication environment by dispatching UAVs [22]. In [23], the authors proposed a multi-path long-term evolution protocol by deploying a UAV, which effectively promotes data exchange among USVs and transmits them to the base station. The on-demand communication for the ship users [24] was provided by deploying UAVs to ensure the communication connections among the ships, shore base stations and satellites. In [5], the authors improved the coverage range of the maritime communication network by deploying UAVs and mobile vessels as mobile base stations to provide the communication services for other vessels. The above research employed the UAVs as relays or mobile base stations to participate in the communication with USVs in the target areas and supported the hardware heterogeneity between UAVs and USVs. They did not consider deploying UAVs to help USVs realize and assure their intragroup FCC.

2.3. DRL for UAV-Assisted Networks

DRL has recently attracted much attention from various industries and in academia. UAVs and USVs were integrated into a cognitive mobile computing network [25] to carry out the search and rescue path planning based on a distributed DRL algorithm. In [9], a DRL framework adopted a DDPG algorithm to optimize a group of UAV trajectories to improve wireless communication coverage, maximize the number of vehicles covered by the minimum number of UAVs and reduce energy consumption. A distributed DRL framework was proposed to use UAVs as air mobile base stations to provide long-term communication coverage for ground users [26]. MADRL considers learning through multiple agents in RL, which has been developed for distributed scenes, in which multiple agents learn to make decisions based on their local observations and communication cooperation for the same system targets [27]. MADRL also necessitates the exploration of environment dynamics and the joint action space between agents. This is a difficult problem due to non-stationarity caused by concurrently learning agents [28]. In order to achieve better coordination, the authors of [29] considered a multi-agent partially observable Markov decision process and proposed a MADRL framework to help agents convey message reliably in a noisy channel. UAVs were deployed as mobile edge computing servers, and MADDPG was used to make joint decisions to allocate computing resources for vehicles.

Compared to DDPG, MADDPG is more suitable for dynamic and complex environments. It employs centralized training and distributed execution, adding the information of other agents during the training process to increase convergence efficiency, and using agents' own information during the testing process. MADDPG has been empirically proved to outperform some DRL algorithms including DDPG in cooperative and competitive multi-agent environments and to be also suitable for the continuous control task [30]. In [31], the authors investigated the spectrum-sharing problem in vehicular networks based on multi-agent reinforcement learning. A fingerprint-based multi-agent Q-network method was proposed to achieve centralized resource management for the base station agent. However, in practical engineering, the deep Q-network (DQN) will encounter many difficulties, such as low sample utilization and unstable training value. The authors investigated a double deep Q-network-based resource allocation framework that maximizes energy efficiency and total network throughput in UAV-assisted terrestrial networks [32] and studied nonlinear energy-harvesting for UAV-assisted device-to-device networks using multi-agent DQN (MADQN) [33]. The training and testing of MADQN are the same network, their input information must be consistent, and both the training phase and the testing phase need the information from other agents, whereas the communication among USVs is incomplete before establishing the FCC. Therefore, MADDPG is suitable for our scenario, centralized training, and distributed execution. A graph-embedded value-decomposition actor–critic algorithm was proposed to embed the interaction information of agents and learn a locally optimal solution through a distributed policy [34]. It studied the non-convex, strongly coupled, and highly complex mixed integer nonlinear programming problem.

The existing technologies verify that UAVs have adequate communication, caching, and computing abilities and can be enhanced with artificial intelligence. MADRL can make navigation decisions through capturing the IoV dynamics and the collaboration of USVs. However, it is difficult for the USVs to obtain each other's state action information over the vast sea surface. Therefore, the existing algorithms cannot be directly applied to the establishment of the FCC, as illustrated in Table 1. Thus, in this paper, we consider the USVs marine communication characteristics and design a MADRL framework strategized by a UAV for the intragroup FCC of USVs based on the advantages of UAVs, which have a larger communication range and on-demand communication. We separate the critics from USVs and propose UST-ACS. USVs act as actors and simulate samples. UAVs act as critics to obtain the global state–action information, evaluate the actions of USVs, optimize parameters, help USVs make effective decisions on movement, and adjust the position of USVs in each timeslot.

Table 1. Related work summary.

Literature	UAV-Assisted USVs	UAV-Assisted Terrestrial Networks	UAV Role	FCC among USVs	DRL
[14–18]	×	×	×	Partial	×
[5,7,8,19–24]	√	×	Relay/participant	×	DRL, RL, etc.
[25]	√	×	Relay/participant	×	A distributed DRL framework
[26]	×	√	Relay/participant	×	A distributed DRL framework
[9]	×	√	Relay/participant	×	DDPG
[29]	×	√	Relay/participant	×	MADDPG
[30]	×	√	Relay/participant	×	MADDPG
[31]	×	√	Relay/participant	×	DQN
[32]	×	√	Relay/participant	×	DDQN
[33]	×	√	Relay/participant	×	MADQN
Our framework	√	×	On-demand/transient	√	UST-MADRL

3. System Model

In this section, we introduce the network model, including the definition of the target sea area, the set of USVs and UAVs, and the motion model of USVs and UAVs. Then, the communication model among USVs, the communication model between USVs and UAVs, and the SSA division strategy are proposed. Finally, we model the problem in this paper.

3.1. Network Model

We consider a three-dimensional sea scenario consisting of a number of USVs and several UAVs. USVs are randomly distributed in the target sea area G, monitoring the information or collecting the information from other USVs, as shown in Figure 1. During the demand for establishing the FCC, UAVs are used to assist the USVs to optimize their positions to guarantee the FCC. After the FCC is established, the UAVs will return or execute another on-demand task. The target sea area G can be divided into $k = \{1, 2, \ldots, |k|\}$ grids, and their intersection can be used as the positions where USVs can stay. Let $N \triangleq \{N_i | i = 1, 2, \ldots, |N|\}$ be a set of USVs with a limited communication range. The communication range of each USV is defined as R_N, which is a connecting constraint between USVs. If the distance is $d(N_i, N_j) \leq R_N$, we consider that USV N_i and USV N_j are interconnected. Otherwise, they cannot communicate with each other. In addition, we let UAVs set be $M \triangleq \{M_u | u = 1, 2, \ldots, |M|\}$. We define the communication range of UAVs as R_M. Generally, the communication range of UAVs is much larger than that of USVs, i.e. $R_M \gg R_N$.
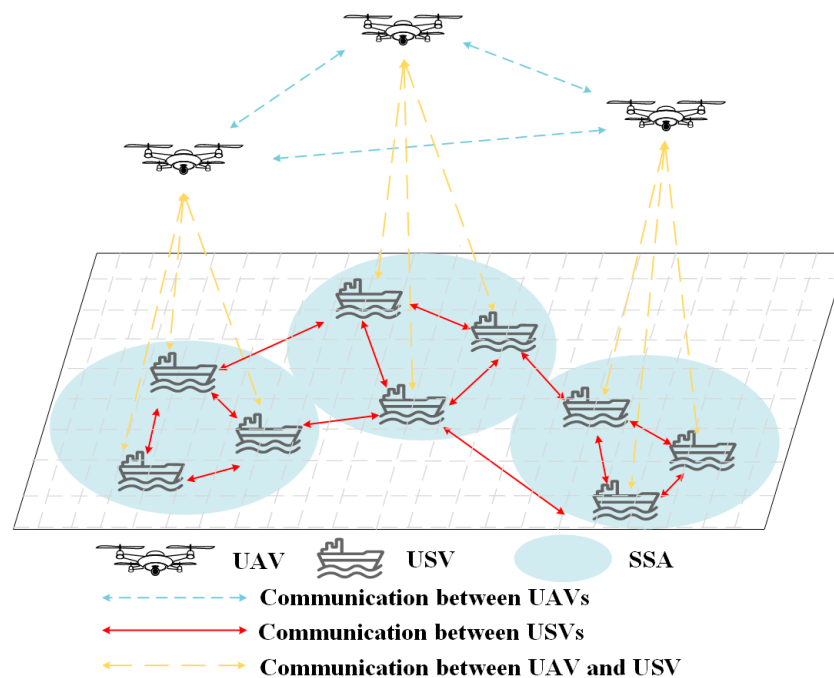
Figure 1. Network model, in which UAVs are guides to evaluate the USVs' movement to establish the FCC on demand. After the FCC is established, UAVs will return or execute another on-demand task.

USVs need to achieve the FCC in the target sea area in a fair and energy-saving manner via a task. We assume that the task period is divided into T timeslots, denoted as $T = \{1, 2, \ldots, T\}$, and that each timeslot is equal, $t \in [1, T]$. At the initial time, all USVs are randomly and statically deployed on the target sea area. Without loss of generality, the position of N_i in any timeslot t can be modeled as

$$P_t(N_i) = [x_t(N_i), y_t(N_i), \varphi_t(N_i), \omega_t(N_i)] \tag{1}$$

where $[x_t(N_i),\ y_t(N_i)]$ denotes the instantaneous coordinates of USV N_i at the grid intersection in the timeslot t, and $[\varphi_t(N_i),\ \omega_t(N_i)]$ denotes the action of selecting a rotatable move to next state for N_i in timeslot t. The direction is $\omega_t(N_i) \in [0, 2\pi]$, and the distance is $\varphi_t(N_i) \in (0, \varphi_{Max})$.

In the beginning, the position of USV N_i is $P_0(N_i) = [x_0(N_i), y_0(N_i), \varphi_0(N_i), \omega_0(N_i)]$, and the position of USV N_i is $P_T(N_i) = [x_T(N_i), y_T(N_i), 0, 0]$ in the end. When the length of timeslot t is small enough, the position of USV N_i in each timeslot is considered to be fixed. Therefore, we can express the continuous position of each USV within the finite time horizon T as

$$P(N_i) = \{P_0(N_i), \ldots, P_t(N_i), \ldots, P_T(N_i)\}. \tag{2}$$

For simplicity, we consider that UAV M_u can move freely in the plane of height H_u, $H_u \le R_M$. Then, the position of UAV M_u in timeslot t is defined as

$$P_t(M_u) = [x_t(M_u), y_t(M_u), H_u]. \tag{3}$$

Therefore, in the task period T, the continuous position of UAV M_u can be derived as

$$P(M_u) = \{P_0(M_u), \ldots, P_t(M_u), \ldots, P_T(M_u)\}. \tag{4}$$

3.2. Communication Channel Models and UAV SSA Partition

3.2.1. Communication Channel Model of USVs

The communication of USVs on the sea surface usually works in the complex and changeable channel environment [35]. We assume that all USVs are equipped with omnidirectional antennas because the channel of the USVs is impacted by the earth radian, wave and ship shielding, as well as the fading and multi-path effects. Generally, the free space propagation loss [14] is

$$L_p = 32.45 + 20lgf + 20lgd(N_i, N_j) \tag{5}$$

where f is the operating frequency in MHz, and $d(N_i, N_j)$ is the distance between two USVs, in which $d(N_i, N_j) = \sqrt{(y_t(N_j) - y_t(N_i))^2 + (x_t(N_j) - x_t(N_i))^2}$. The channel between USV N_i and USV N_j on the sea surface in timeslot t can be denoted as the antenna gain (G_t and G_r) minus all losses occurring in a link, i.e.,

$$L_t(N_i, N_j) = G_t + G_r - L_p - L_w \tag{6}$$

where G_t is the transmitting antenna gain, and G_r is the receiving antenna gain. L_w represents other possible losses, such as sea surface reflection and atmospheric absorption loss [36].

3.2.2. Communication Channel Model between USVs and UAVs

We assume that the UAVs in the air and the USVs on the sea surface are equipped with a single antenna and communicate through the line-of-sight [8,37]. At timeslot t, the signal transmitted from UAV M_u is denoted as $b_t(M_u)$, and the received signal of USV N_i by UAV M_u at the position $P_t(M_u)$ can be expressed as

$$H_t(N_i, M_u) = TP_t(M_u)G_{M_u}G_{N_i}C_t(N_i, M_u)b_t(M_u) + \sigma_t(N_i, M_u) \tag{7}$$

where $TP_t(M_u)$ denotes the transmission power of UAV M_u, G_{Mu} denotes the antenna gain of UAVs, G_{Ni} denotes the antenna gain of USVs served by UAVs, and $\sigma_t(N_i, M_u)$ denotes the white Gaussian noise. $C_t(N_i, M_u)$ denotes the channel between UAV M_u and USV N_i that can be denoted as $C_t(N_i, M_u) = (h_t(N_i, M_u))^{-1/2} C^-_t(N_i, M_u)$, where $C^-_t(N_i, M_u)$ is Rician fading during transmission, as defined in [8], and $h_t(N_i, M_u)$ is the path loss,

$$h_t(N_i, M_u) = \varepsilon + 10\tau log10\left(\frac{d(N_i, M_u)}{d_0}\right) + Z_t(N_i, M_u). \tag{8}$$

Here $d(N_i, M_u)$ is the Euclidean distance between USV N_i and UAV M_u, d_0 denotes the reference distance, ε denotes the path loss at d_0, τ denotes the path-loss exponent, and $Z_t(N_i, M_u)$ is a zero-mean Gaussian random variable with standard deviation.

3.2.3. SSA Partition of UAV

The communication range of UAV R_M is much larger than that of USV R_N, so the number of UAVs set M can be estimated according to the scale of the target sea area and the communication range of UAV on the sea. When the number of UAVs is determined, we adopt the deployment strategy of UAVs to achieve the full coverage of the target sea area by minimizing the total deployment delay proposed in [38]. When USV N_i is within the communication range R_M of UAV M_u, USV N_i can receive the service of UAV M_u. Due to different signal/noise ratio (SNR), we express the SNR of UAV M_u to USV N_i in timeslot t as

$$SNR_t(N_i, M_u) = \delta(d(N_i, M_u))H_t(N_i, M_u). \tag{9}$$

We define $\delta(d(N_i, M_u))$ as a binary variable. If $d(N_i, M_u) \leq R_M$, we define $\delta(d(N_i, M_u)) = 1$; otherwise, $\delta(d(N_i, M_u)) = 0$. We next use the signal-weighted Voronoi cell partition strategy [39] to divide the target sea area into M sub-areas, referred to as UAV SSAs. The SNR between a UAV and any USV in its SSA is always higher than that in other UAV SSAs. Therefore, the UAV only cooperates with the USVs in its own SSA. The SSA of UAV M_u is defined as a signal-weighted Voronoi cell $V_t(M_u)$,

$$V_t(M_u) = \{N_i | \forall M_v \in M \setminus \{M_u\} : SNR_t(N_i, M_u) > SNR_t(N_i, M_v)\}. \tag{10}$$

From (10), we can construct an associated signal-weighted Delaunay graph by choosing the set of vertices as M and the set of edges as pairs of UAVs whose signal-weighted Voronoi cells are adjacent [40]. We demonstrate that each UAV provides communication services for its connecting USVs in its SSA with the strategy of distributed cooperation. Therefore, UAVs can be used as the guide to evaluate and assist the position adjustment of USVs in its own SSA $V_t(M_u)$ by observing the overall situation and the communication among UAVs. The contact among UAVs ensures the cooperation of all UAVs and USVs.

3.3. Problem Definition

In order to realize the FCC and optimize the performance of the IoV, we define and model the optimization aims, including optimizing the communication coverage and USV movement fairness, saving energy, and increasing communication efficiency. We regard the intersection of each grid in the target sea area as the sampling point k, and the SNR between USVs N_i and sampling point k can be expressed as

$$SNR_t(N_i, k) = \frac{TP_t(N_i) + L_t(N_i, k)}{\sigma^2} \tag{11}$$

where σ^2 is the variance of additive white gaussian noise, and the sum of the transmit power $TP_t(N_i)$ and $L_t(N_i, k)$ denotes the signal power [14] from USVs N_i to sampling point k. $L_t(N_i, k)$ can be calculated by (6).

The communication coverage of USVs N_i in the target sea area is determined by SNR and is defined as

$$c_t(N_i, k) = \begin{cases} 1, & if\ SNR_t(N_i, k) \geq \lambda_{SNR} \\ 0, & else \end{cases} \tag{12}$$

where λ_{SNR} is the SNR threshold [39]. When the SNR between USV N_i and sampling point k is greater than the given threshold, the position will be covered by USV N_i. Hence, we approximately define the communication coverage score of the target sea area covered by

USVs set at timeslot t as the ratio of the number of sampling points that are covered by all USV communication in the target area to the total number of sampling points,

$$\mathbb{S}_t \approx \frac{\sum_i^N \sum_j^k c_t(N_i, k)}{k}. \tag{13}$$

The movement of USVs is accompanied by energy consumption. If the energy of the USVs are exhausted, the communication quality of service will not meet the IoV. Therefore, this paper aims to reduce the average moving distance of USVs as much as possible and improve the average residual energy of USVs. The initial energy of all USVs at the initial time is equal and denoted as e_0. The energy consumed by USV N_i to remain stationary in timeslot t is defined as $e_t(N_i) = \alpha$, where t is the stationary timeslot. And the energy consumption of USV N_i in the process of moving on the sea is defined as $e_t(N_i) = le_z$, where e_z is the movement distance of USV N_i, and e_z is the normalized energy consumed (NEC) of the USV moving unit distance. Thus, the energy consumption of USV N_i in a specific task period T is calculated as

$$e_T(N_i) = \sum_{t=0}^{T} e_t(N_i). \tag{14}$$

Then, the average energy consumption of all USVs is expressed as

$$e_T(avg) = \frac{\sum_{i=1}^{N} e_T(N_i)}{N}. \tag{15}$$

Moreover, when only some of the USVs in the IoV are in constant motion, they will unfairly consume more energy. If these vessels run out of energy, there will be more communication coverage holes in the target area and the IoV will not work any more. In order to balance the energy consumption and the residual energy of all USVs, we use Jain's fairness index [41] to define the movement fairness index as

$$F_t = \frac{\left(\sum_{i=1}^{|N|} e_t(N_i)\right)^2}{\left|N\right|\sum_{i=1}^{|N|} e_t(N_i)} \tag{16}$$

where $F_t \in [\frac{1}{|N|}, 1]$. When all $e_t(N_i)$ are equal, $F_t = 1$. Then, the final achieved fairness index in the whole mission cycle is $F_T = F_t|_{t=T}$, which helps all USVs to move fairly and reasonably throughout the task period. The communication connection performance of the IoV is positively correlated with the movement fairness index.

4. Problem Formulation

This paper aims to realize FCC and optimize the IoV performance, including optimizing the communication coverage, energy consumption and movement fairness index. Since the optimizations among USVs are coupled and non-convex, we cannot integrate these three indexes into a mixed-integer linear programming optimization problem. The performance of these three indexes depends on the position decision-making process of multiple USVs, and the performance of the next timeslot is related to the decision of multiple USVs in the current timeslot, which has the Markov property [42]. Therefore, we transform our problem into a POMG in each SSA, which is defined as a tuple $(N, S, A_t, \mathcal{F}, and\ R_t)$. The details are clarified as follows.

4.1. Multi-Agent Set N

In this paper, the USV set N is defined as the $|N|$ agents in the multi-agent environment. Each agent partially observes the environment and obtains local observation information.

4.2. Observation Space O_t and State S

The set of O_t is defined as the observation space of USVs at timeslot t. At each timeslot t, each USV actor N_i can partially observe the environment and obtain partial observation $o_t(N_i)$, including the coordinates $[x_t(N_i), y_t(N_i)]$, the movement distance $\varphi_t(N_i) \in (0, \varphi_{Max})$, the directions $\omega_t(N_i) \in [0, 2\pi]$, and the energy consumption $e_t(N_i)$. Hence, it is expressed as $o_t(N_i) = \{x_t(N_i), y_t(N_i), \varphi_t(N_i), \omega_t(N_i), e_t(N_i)\}$, and the observation space of all USVs is $O_t \triangleq \{o_t(N_i) | i \in N, t = 1, 2, \ldots, T\}$.

In the multi-agent environment, the set S represents the state, which obeys the markov property, i.e., each state only depends on the previous state and the action taken by the agents. The set S includes the observation space O_t and the communication coverage score $\mathbb{S}_t$ of USVs in timeslot t, denoted as $S \triangleq \{s_t\} = O_t \cup \{\mathbb{S}_t\}$.

4.3. Action Space A_t

At each timeslot t, each USV N_i chooses an action $a_t(N_i)$ from its action space according to the current policy π_{N_i} and the corresponding observation. The actions of USVs can be described as $A_t \triangleq \{a_t(N_i) | N_i \in N, t = 1, 2, \ldots, T\}$.

4.4. State Transition Function $\mathcal{F}$

The environment state s_t will transit into next state s_{t+1} after the USVs perform their actions. The state transition function $\mathcal{F}$ is defined as $\mathcal{F} : s_t \times A_t \rightarrow s_{t+1}$.

4.5. Reward–Penalty Mechanism for USVs R_t

The reward function $r_t(N_i)$ of USV N_i represents the immediate reward after agent N_i executes its action $a_t(N_i)$ at each timeslot, which measures the effect of the action taken by a USV at a given state. In order to obtain the effective reward, we define the reward-penalty mechanism in the following normalized quantities, including communication efficiency, penalty constrains, and corresponding penalty values:

(1) Communication Efficiency X_t: It integrates three optimization indexes, namely the communication coverage score of USVs $\mathbb{S}_t$, the average energy consumption of USVs $e_T(avg)$, and the movement fairness index of USVs F_t, defined as

$$X_t = \frac{\mathbb{S}_t F_t}{e_T(avg)} \tag{17}$$

(2) Non-connectivity Penalty $p_t(N_i)_1$: If the action selected by USV N_i were to change its position from inside of the communication range to outside of the communication range, i.e., $\exists d_t(N_i, N_j) > R_N$, where N_j is any neighbor of N_i, this kind of action should receive a penalty, and the penalty value is $p_t(N_i)_1$. The reason is that it will cause disconnections among USVs and finally undermine the FCC.

(3) Redundancy Penalty $p_t(N_i)_2$: If the action selected by USV N_i causes the Euclidean distance between USV N_i and any neighbor USV N_j to be less than the distance threshold D, i.e., $\exists d_t(N_i, N_j) < D$, this kind of action should receive a redundancy penalty, and the value is $p_t(N_i)_2$. This is because it will increase the communication coverage redundancy and reduce the communication coverage score.

(4) Cross Border Penalty $p_t(N_i)_3$: When USV N_i moves beyond the target sea area, we will impose a penalty for this kind of action, and the penalty value is $p_t(N_i)_3$. This operation ensures that USVs learn how to move continually on the given target sea area.

The reward function $r_t(N_i)$ of USV N_i is defined as

$$r_t(N_i) = X_t(N_i) - p_t(N_i)_1 - p_t(N_i)_2 - p_t(N_i)_3 \tag{18}$$

And we construct the holistic joint reward function as $R_t \triangleq \{r_t(N_i) | i \in N\}$, representing the joint reward of all agents in the environment. Here, the communication efficiency

ensures the reward for effective communication coverage, average energy consumption, and movement fairness. Additionally, if any USV N_i executes one of the above penalty actions, it will obtain the corresponding penalty value, which will reduce the reward and reduce the selection probability of this action. Note that this penalty value is global and not limited to a single SSA, which makes our algorithm effective in the global environment. The reward–penalty mechanism can lead each agent to its optimal policy, and the policy directly determines the optimal trajectory and position of the USV. The reward function is designed based on the objectives of the original formulated problems. The optimization problem can be formulated as

$$\max_{\{N,M,O_t,A_t\}} \left\{ R_t \triangleq \{r_t(N_i)|i \in N\} \right\} \tag{19}$$

$$\text{s. t.} \quad (10), (12), (13), (15), (16), (17), (18); \tag{19a}$$

$$p_t(N_i)_1, \ p_t(N_i)_2, \ p_t(N_i)_3; \tag{19b}$$

$$D \leq d_t(N_i, \ N_j) \leq R_N; \tag{19c}$$

$$N_i \text{ is within } G; \tag{19d}$$

where (19a) denotes FCC and USVs' performance constraints, and (19b)–(19d) describe the movement constraints of USVs.

5. Proposed Solution

The above multi-object optimization is an infinite control task; USVs can carry out continuous actions because of the infinite sailing angle and distance, and the reward function depends on the joint action of all USVs in the global environment. It cannot be solved using the conventional dynamic programming method [10], which is a model-based approach. We use DRL to find suboptimal solutions. However, since the USV reward is affected by the actions of many other USVs in our scenario, it needs central UAV training and distributed USV execution as a multi-agent environment. Traditional policy-gradient-based methods, such as DDPG, require that the reward only depends on a USV's own action, which cannot be directly applied to this problem. We propose a UST-ACS with the centralized training and decentralized execution. We separate the critic networks from USV agents and put them into UAVs. UAVs are responsible for multi-USV information collection, sharing, and evaluation. The UST-ACS includes actor networks of USVs, critic networks of UAVs, and information transitions in UAVs. At a given state, the USV actor uses the observation as the input for its independent parameterized policy, generates the movement action, and executes it. The execution of each action will be rewarded accordingly. The transition samples including state, action, and reward, and are collected and stored in the experience replay buffer. The critic in the UAV corresponding to the actor, which can be represented as a value function, evaluates the action generated by the actor and training the value function with the time difference (TD) error by sampling the mini-batch from the experience replay buffer. After training, the UAV's critic networks return the optimized policy parameters to the USVs. The actor will be guided to update the policy, and then produce the action with higher communication performance. The above procedure is repeated until convergence, and finally obtain an optimal strategy to realize the FCC as well as the optimal IoV network performance. SSA partition strategy can solve the problem of space dimension explosion caused by the large number of agents, and reduce the complexity of algorithm. Figure 2 shows its framework in one SSA.
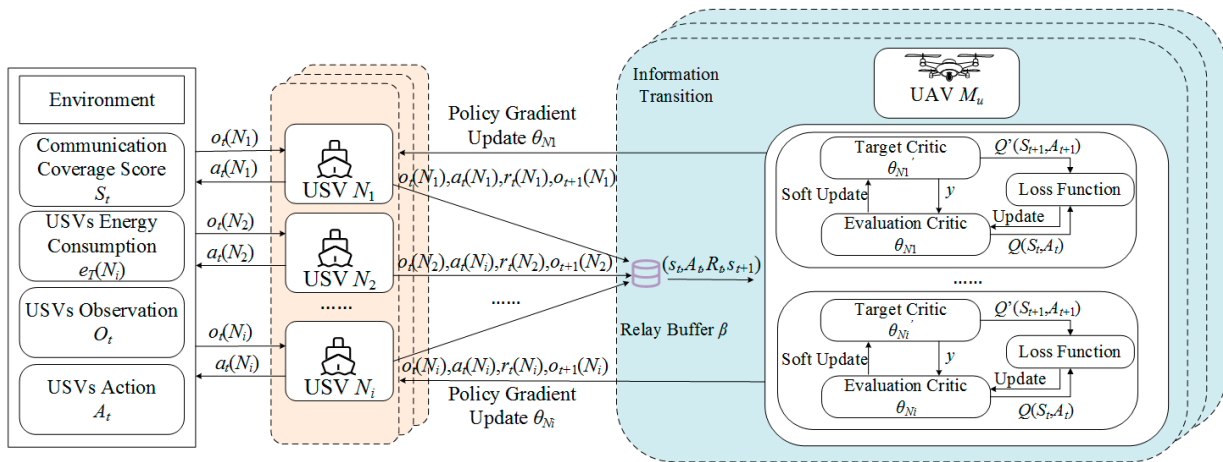
Figure 2. The framework overview of UST-MADRL, in which USV agents act as actors, and UAVs act as critics and deal with information transition.

5.1. Actor Network Design for USVs

In our framework, each USV processes a local actor network that generates an action according to its own policy and observation information, modeled as a DNN with the weights matrix θ and policy π. The actor network aims to optimize the action probability sets for each state by the policy π so as to achieve the FCC and the optimization of the indexes. Thus, we use a deterministic policy rather than a random policy. In any SSA $V_t(M_u)$, each USV agent N_i corresponds to a policy $\pi_{N_i}(o_t(N_i), \theta_{N_i})$, where θ_{N_i} is policy parameter, and this function maps an observation definitively to an action. USV N_i determines its action $a_t(N_i)$ by policy function, described as

$$a_t(N_i) = \pi_{N_i}\big(o_t(N_i), \theta_{N_i}\big) \tag{20}$$

Each USV N_i generates and executes the determined action $a_t(N_i)$ according to its own independent actor network, and accordingly obtains reward $r_t(N_i)$. After the actions of all actors are executed, the environment state will transition from s_t to s_{t+1} by state transition function $\mathcal{F}$ and obtain a global reward R_t. Each USV's local information, including observation $o_t(N_i)$, action $a_t(N_i)$, reward $r_t(N_i)$, and next observation $o_{t+1}(N_i)$, will be stored in the experience replay buffer β with capacity size B, and the transition samples of all USVs are stored as tuples in each UAV.

As for the update procedure of actor network, we use $\pi = \left[\pi_{N_1}, \ldots, \pi_{N_{|N|}}\right]$ to denote the $|N|$ continuous policy selected by the USVs in timeslot t. Thus, the actor network of USV N_i is updated by the gradient of the expected return in the critic of UAV as

$$\nabla\theta_{N_i}J\big(\theta_{N_i}\big) = \mathbb{E}_{O_t,a\sim\beta}\Big[\nabla\theta_{N_i}\pi_{Ni}(a_t(N_i)|o_t(N_i))\nabla_{ai}Q^{\pi}_{N_i}(S_t, a_t(N_1), \ldots, a_t(N_N)) \\ \big|a_t(N_i) = \pi_{Ni}\big(o_t(N_i), \theta_{N_i}\big)\Big] \tag{21}$$

where $Q^{\pi}_{N_i}(S_t, a_t(N_1), \ldots, a_t(N_N))$ is a centralized action-value function that takes the actions of all USVs as input in addition to S_t, and then outputs the Q-value of the joint action for USV N_i. S_t includes O_t and the available additional state information (environmental state information). $O_t = [o_t(N_1), \ldots, o_t(N_N)]$ denotes the joint observation of USVs in timeslot t. Our algorithm is model-free, directly learning from the experiences of USVs. During training, at each timeslot, actors can generate the transition samples of position state–action information. We leverage the experience replay buffer to store the transition samples of all USVs as tuples $\Big(O_t, a_t(N_1), \ldots, a_t\big(N_{|N|}\big), r_t(N_1) \ldots, r_t\big(N_{|N|}\big)\Big)$. The information transition fuses the transition samples into the global information. The pseudocode of "actor network" is shown in Algorithm 1.

Algorithm 1. Actor network of USVs

1: **Input:** USVs observation $o_t(N_i)$, policy π_{Ni} for each USV N_i with parameter θ_{N_i}, discount factor γ, learning rate for actor network of USVs η;
2: **Output:** $a_t(N_1),\ldots,a_t(N_N)$.
3: **for** USV N_i: = 1,…, |N| **do**
4:      Initialize a random process for action exploration with target parameter θ_{N_i};
5:      Receive initial state s_t;
6: **end for**
7: **for** episode: = 1,…, Episode Length **do**
8:      **for** Timeslot t: = 1,…, T **do**
9:          **for** USV N_i = 1,…, |N| **do**
10:              USV i obtain $a_t(N_i) = \pi_{Ni}\left(o_t(N_i),\ \theta_{N_i}\right)$ according to the local policy and the observation;
11:          **end for**
12:      Execute all actions of USVs $a_t(N_1),\ldots, a_t\left(N_{|N|}\right)$ and get reward $r_t(N_1)\ldots,\ r_t\left(N_{|N|}\right)$;
13:      Send $o_t(N_i)$, $a_t(N_i)$, $r_t(N_i)$ and $o_{t+1}(N_i)$ to Information Transition;
14:          $s_t \leftarrow s_{t+1}$;
15:      **end for**
16: **end for**

5.2. Information Transition

The information transition in the UAV is used to better establish the message transmission between UAVs and USVs. In an SSA, the observation and action of each agent will be encoded into a message and sent to the UAV managing the SSA. Then, the UAV decodes each message and integrates it into a global information, which includes the historical observations and behaviors of all USVs. Thus, the critic network will obtain the global environmental state and make a global evaluation, as shown in Figure 3. In timeslot t, USV N_i encodes its observation $o_t(N_i)$, action $a_t(N_i)$, and reward $r_t(N_i)$, adds it to its own message vector $h_{t-1}(N_i)$, and updates it to a new local environment message $h_t(N_i)$. Next, the local messages of all USVs will be summarized into a global message into $\{h_t(N_1),\ h_t(N_2),\ldots,\ h_t\left(N_{|N|}\right)\} \cup \mathbb{S}_t$ in the information transition, where the message $h_t(N_i)$ contains all previous observations and behaviors. Finally, the global state information is decoded in the UAV's critic network and the critic network evaluates the action-value functions of all USVs in each SSA. The pseudocode is shown in Algorithm 2.

Algorithm 2. Information transition of UAVs

1: **Input:** USVs observation $o_t(N_i)$, USVs action $a_t(N_i)$, USVs reward $r_t(N_i)$, USVs observation in next state $o_{t+1}(N_i)$;
2: **Output:** Global environment information $\{s_t, A_t, R_t\}$.
3: Initialize the capacity of replay buffer β to B;
4: **for** UAV M_u: = 1, ..., |M| **do**
5:      Receive the USV local information in its SSA;
6:      Generate information sequence $h_t(N_i)$;
7:      Integrate $\{h_t(N_1), h_t(N_2),\ldots,h_t\left(N_{|N|}\right)\} \cup \mathbb{S}_t$;
8: Store s_t, A_t, R_t, s_{t+1}, in replay buffer β;
9: **end for**
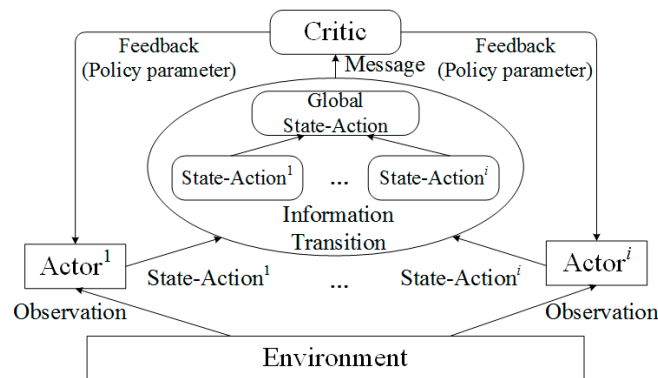
Figure 3. Information transition.

5.3. Critic Network Design for UAVs

We design the critic network in each UAV and access the global state s_t and joint actions $a_t(N_1),\ldots, a_t\left(N_{|N|}\right)$ in the training process. The action value function

$Q^{\pi}_{N_i}(s_t,\,a_t(N_1),\dots,\,a_t(N_{|N|})) = \mathbb{E}_{\pi}\left[\sum_{i=t}^{T} r(N_i)\big|s_t, a_t(N_i)\right]$ is approximated under the current policy π_{N_i}, where $\pi = [\pi_{N_1},\dots,\pi_{N_{|N|}}]$ is the joint policy with parameter set $\theta = [\theta_{N_1},\dots,\theta_{N_{|N|}}]$. From a global perspective, all USVs are designed to achieve a common goal. When the critic network of a UAV updates its parameters, the state–action information, $(s_t, a_t(N_i), r_t(N_i), s_{t+1}, a_{t+1}(N_i)|a_{t+1}(N_i) = \pi_{N_i}(o_{t+1}(N_i), \theta_i))$, is necessary, where $a_{t+1}(N_i)$ comes from its target policy. Thus, a UAV critic can randomly sample mini-batches of multi-USV experiences from the experience replay buffer and utilize the TD error method to update the parameter of the critic network by minimizing the loss function

$$L(\theta_{N_i}) = \mathbb{E}_{O_t,\,a_t(N_1),\dots,\,a_t(N_{|N|}),\,r_t(N_1)\dots,\,r_t(N_{|N|}),O_{t+1}}\left[\left(Q^{\pi}_{N_i}(O_t,\,a_t(N_1),\dots,\,a_t(N_{|N|})) - y\right)^2\right] \tag{22}$$

where y is the target value generated by the target network of critic and is calculated by

$$y = r_t(N_i) + \gamma Q^{\theta'}_{N_i}\left(O_{t+1},\,a_{t+1}(N_1),\dots,\,a_{t+1}\left(N_{|N|}\right)\Big|a_{t+1}(N_j) = \theta'_{N_j}(O_{t+1}(N_j))\right) \tag{23}$$

where the target policy parameter set is $\theta' = [\theta'_{N_1},\dots,\theta'_{N_{|N|}}]$.

After training, each UAV critic will feed the optimized policy parameters back to each USV actor. A UAV critic can use joint observations and actions in the training process as the guide to master the overall environment, and evaluate the action-value functions of all USVs in each SSA. Therefore, the critic can obtain the global environmental state and make the global evaluation. Considering the common objective of the formulated optimization problems, $|N|$ agents should cooperatively maximize the communication coverage score and movement fairness index and minimize the average energy consumption of the USVs. The pseudocode of "critic network" is shown in Algorithm 3.

Algorithm 3. Critic network of UAVs

1: **Input:** s_t, A_t, R_t, s_{t+1}, action selection by policy $\pi^{\theta}{}_t(N_i)$ for each USV;
2: **Output:** Updated policy parameter set θ.
3: **for** each episode: = 1,..., Episode Length **do**
4:     **for** Timeslot t: = 1,..., T **do**
5:         **for** UAV M_u: = 1,..., $|M|$ **do**
6:             Sample a random mini-batch of k samples, s_t, A_t, R_t, s_{t+1} from β;
7:             Set target value y by (23);
8:             Update critic network by minimizing the loss $L(\theta_i)$ by (22);
9:             Update actor network using the sampled policy gradient by (21);
10:         **end for**
11:         Update two target network parameters for each USV N_i;
12:         $\theta_{N_i}' \leftarrow \varsigma\theta_{N_i} + (1 - \varsigma)\,\theta_{N_i}'$;
13:     **end for**
14: **end for**

5.4. Training Process

According to [29,30], offline training is available, and the training dataset and loss function can be designed to guarantee the generalization capability. Our architecture is a centralized training and distributed implementation framework. In the centralized training, actors can generate transition samples of observation, action, state, and reward information at each timeslot, and store them in the experience replay buffer of the corresponding UAV. A UAV's critic network masters the global information, evaluates USVs policies and feeds evaluation results back to the USVs. The actor and the critic update the parameters according to the input mini-batch of transitions. By this process, the policies of USVs are gradually optimized until they become as optimal as possible.

5.5. Testing Process

In the testing process, since our scenario is a distributed execution, each USV can make a parameterization policy according to its observation $o_t(N_i)$ and policy parameter θ_{N_i}. Therefore, the action $a_t(N_i)$ of each USV is generated through its actor network without the direct communication of other USVs' information.

5.6. Complexity Analysis

Based on [43], we analyzed the complexity of algorithm. In the distributed execution procedure, each USV obtained its action from its actor networks through the state, including the 2D coordinate, directions, moving distance, energy consumption, and communication coverage score. Hence, the input size was 6, and the output size was 8. In the centralized training process, UAVs collected the global information from all USVs. The input and output sizes in each critic network in a UAV were $8N$ and 1, respectively. According to [44], given the fully-connected neural network with fixed numbers of hidden layers and neurons, the computational complexity of the back-propagation algorithm was proportional to the product of the input size and the output size. The centralized training complexity was $O(N^2)$ in the critic network, while the decentralized execution procedure complexity was $O(N)$ in the actor network. Therefore, the overall complexity was $O(N^2)$.

6. Simulation Results

To evaluate the performance of our solution, we conducted a series of experiments step by step. We compared our solution with DDPG and the three baselines employed in the multi-agent cooperation and competition search. The results show that our framework was superior to the other solutions in communication coverage, movement energy consumption, movement fairness index, and communication efficiency.

6.1. Setup and Evaluation Metrics

We implemented the simulation results in Windows 10, TensorFlow 1.14, and python 3.7. We simulated a sea surface target area with the scale of 20×20 grids, in which the grid points were the sampling points. Our parameters were set according to [26,29]. The communication range of the USVs was defined as R_N = 5 units, and each USV was able to choose any grid point to move to. The communication range of the UAVs was defined as about 35 units. We gave a non-connectivity penalty $p_t(N_i)_1 = 3$, a redundancy penalty $p_t(N_i)_2 = 2$, and a cross border penalty $p_t(N_i)_3 = 1$, respectively. Every time the action selected by a USV met a non-connection penalty, redundancy penalty, or cross-border penalty, it obtained the corresponding penalty value and reduced the reward. We trained the proposed model into $4K$ episodes, each episode had T = 500 timeslots, with every 100 sets of modeling producing 40 models. During the test period, we tested each model 100 times, took the average value, and chose the best one among the 40 models. The main parameters of simulation are listed in Table 2.

Table 2. Main simulation parameters.

Parameters Configuration	Quantity
Noise power	-120 dBm
Channel power gain	-50 dB
Size of replay buffer	10,000
Size of mini-batch	100
Activation functions	ReLU
Discount factor γ	0.96
Actor's learning rate η	Decaying from 0.0002 to 0.0000001
Critic's learning rate ζ	Decaying from 0.002 to 0.000001
Reward discount factor γ	Augmenting from 0.8 to 0.99

Our simulation adopted four layers of DNN, including an input layer, two hidden layers, and an output layer of 80 neurons. Each USV maintained the actor network, and each UAV maintained the critic network, which were activated by the rectified linear unit (ReLU) function. The hyperbolic tangent activation function was used in the outermost layer. The actor network of a USV received the state value of USV N_i at the input layer, output an action at the output layer, and then sent it to the associated UAV. The critic network of a UAV received the current environment status and the actions of all USVs in

its SSA in the input layer. In the output layer, it produced a Q value. The training of the DNN was carried out by using the generated samples, and the best moving route of each USV in the SSA was realized by using Tensor Processing Unit (TPU).

We used four metrics specified in (13) and (15)–(17) to measure performance evaluation, including the communication coverage score $\mathbb{S}_t$, average energy consumption $e_T(avg)$, movement fairness index F_t, and communication efficiency X_t.

6.2. Baselines

We compared our proposed UST-MADRL framework with the frameworks using the following solutions:

- DDPG [45]: A policy-gradient-based approach for continuous control tasks, which uses one actor network and one critic network to output control decisions for all UAVs. The state and reward functions in DDPG are consistent with those in our framework.
- Genetic algorithm (GA) [46]: A stochastic global search optimization method that simulates the replication, crossover, and mutation phenomena occurring in natural selection and genetics, starting from any initial population by the genetic operation to produce a group of individuals better suited for the environment. Here, the USV's position is adjusted according to the fitness that is calculated from the reward of our paper. In each timeslot t, each USV chooses an action with high fitness value. Combined with the gene crossover and mutation, we use the roulette to eliminate the fittest.
- Particle swarm optimization (PSO) [47]: A bionic optimization algorithm based on multiple agents, which is derived from the study of bird predation behavior. The velocity of the particle is updated according to its own previous best position and the previous best position of its companions. The particles fly with the updated velocities.
- Virtual force algorithm (VFA) [48]: VFA constructs a virtual force field, which is composed of the attractive force field of the target orientation and the repulsive force field around the other agents. It searches in the descending direction of the potential function to find an optimal path and makes the agent move along the direction of the resultant force of virtual attractive force and virtual repulsive force. The attractive force and virtual repulsive force are mainly reflected by the distance among the agents. For a given number of USVs, our optimization indicators can be provided as inputs to the VFA, thereby ensuring flexibility.

According to the above references, we set parameters of these four baselines as follows: In GA, the number of iterations was set to 500 times, the crossover probability was 0.4, the mutation probability was 0.005, and the initial population was generated randomly. In PSO, the initial position and initial speed were randomly generated, the initial inertia weight was set to 0.9, the inertia weight when iterating to the maximum evolutionary algebra was set to 0.4, and the speed interval was $[0.1, \sqrt{2}]$ unit per timeslot. The VFA needed to consider the threshold factor of the attractive force and repulsive force function that USVs can establish at the current position. We set the threshold distance to USV's communication range considering the coverage rate. The maximum step size of USV movement was set to 0.5 unit. We tested these four methods 100 times and took the average value.

6.3. Evaluation Results

In this section, we show the trends of the communication coverage score; movement fairness index; average energy consumption and communication efficiency, which varied with various USV's communication range; the number of USVs in each SSA; and the NEC by USVs moving a unit distance.

Figure 4 shows the impact of the communication range on the above four metrics. We can clearly see that the average energy consumption of all solutions decreased with the increase in a USV's communication range because a higher communication range was able to reduce the moving distance of USVs when establishing the FCC. At the same time, the movement fairness and the coverage range correspondingly increased.
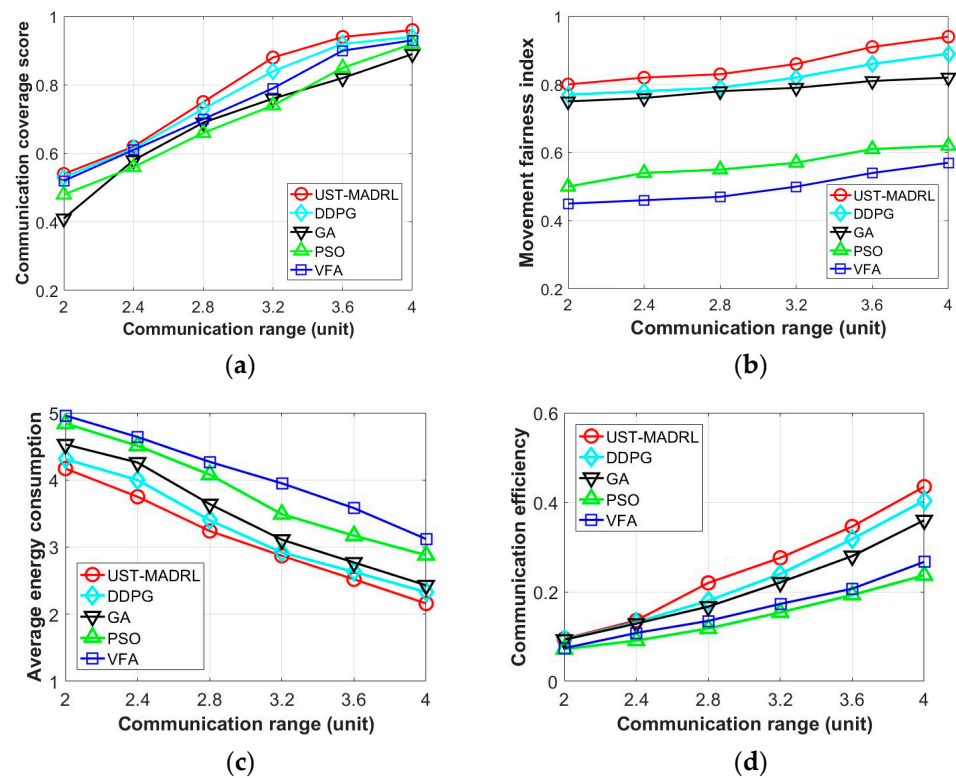
Figure 4. Impact of USV's communication range on (**a**) communication coverage score, (**b**) movement fairness index, (**c**) average energy consumption, (**d**) communication efficiency.

In Figure 4a, it can be seen that our proposed framework was better in the communication coverage score because in our framework, each USV needed to consider that the selected moving position was able to cover more sampling points in order to obtain more rewards in the training process. DDPG focused on the local information of a single USV in training and lacked comprehensive consideration, while our approach used the global information of all USVs and leveraged the cooperation and competition among USVs, hence DDPG was worse than UST-MADRL. In the VFA, the attractive force and virtual repulsive force were mainly represented by the distance between the agents; therefore, it showed better communication coverage than GA and PSO, whereas the movement fairness and average energy consumption in the VFA were relatively worse than GA and PSO in Figure 4b,c.

Figure 4b shows that UST-MADRL was always optimal in terms of USVs' movement fairness when the communication range of USVs varied because we quanitifed the energy consumption by Jain's fairness index and considered energy consumption fairness by multi-USV cooperative movement in the target sea area. Compared to this, DDPG, with one policy, only relayed on one agent, which might have led to the potential unbalances and conflicts among USVs. We can see that its movement fairness index were about 6% higher than DDPG, 15% higher than GA, 65% higher than VFA, and 52% higher than PSO when communication range *R* = 4.

In Figure 4c, it can be seen tha the average energy consumption of USVs was the lowest under the control of UST-MADRL. DDPG was inferior to UST-MADRL. The average energy consumption of GA was lower than PSO and the VFA because the energy consumption of USVs was fully considered in the fitness function of GA as an important standard of the chromosome evolution in the iterative process. However, the velocity and position of the particles played a more important role than the fitness information in PSO. In the meantime, the VFA paid more attention to the distance. We input some optimization indicators into the VFA to try to enhance it, so the average energy consumption of PSO was the highest.

Figure 4d shows that the communication efficiency of our framework was always better than that of the others, with an increase in USV's communication range. When the UAVs in the target sea area were trained by UST-ACS, there was a good solution for the cooperation and competition among the UAVs; that is, the policy parameters of the UAVs in the strategy selection was able to help USVs to improve the overall communication efficiency. Based on the above trends, DDPG is inferior to UST-MADRL. UST-MADRL exceeded GA by an average of 10%. Compared to GA, the communication efficiency of UST-MADRL exceeded that of GA by 2.2%, 5.4%, 31.7%, 24.9%, 23.6%, and 20.8%, respectively. This was because even though GA can find the optimal solution according to the evolution characteristics in multi-agent environment and the fitness was the same as our reward, GA had a preference for local optimization.

Figure 5 shows the influence of the number of USVs in each SSA on the four indicators when we changed the number of USVs in each SSA from 2 to 7. With the increase in the number of USVs, the communication coverage score and movement fairness index showed an upward trend, while the average energy consumption decreased. Accordingly, the communication efficiency also kept increasing.
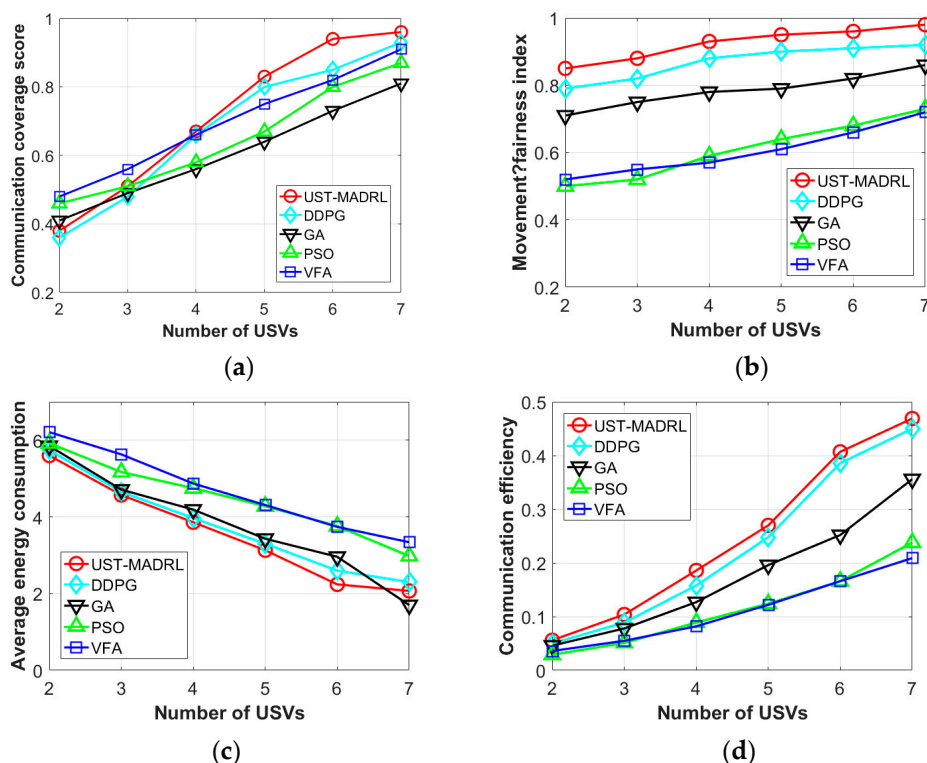
Figure 5. Impact of USVs' number in each SSA on (**a**) communication coverage score, (**b**) movement fairness index, (**c**) average energy consumption, (**d**) communication efficiency.

Figure 5a shows that the communication coverage score of our framework kept growing. This growth slowed when the number of USVs was 6 or 7. At that time, the communication coverage gradually tended toward saturation. The curve of DDPG was still below our framework. The VFA kept better coverage than PSO and GA, which was the same as Figure 4, but worse average energy consumption and movement fairness, as shown in Figure 5b,c. When the number of USVs in each SSA was 2 or 3, the communication coverage score of our method was even lower than that of the VFA. This was because the communication coverage of a small number of USVs was not enough to fully cover the target sea area. In order to obtain greater communication efficiency, our method will prefer to improve the movement fairness and reduce the energy consumption.

Moreover, in Figure 5b,c, it can be seen that the other baselines were inferior to our proposed method. This is because the benefits from our comprehensive reward as well as

the global policy optimization process, which ensures the overall return of the policy set in the multi-agent environment. Thus, UST-MADRL was still able to achieve the lowest average energy consumption and the highest movement fairness. When more USVs were deployed, our framework still performed better. For instance, in Figure 5c, when the number of USVs was 6 in each SSA, our method showed a 24% improvement compared to GA and a 40.4% improvement compared to PSO, while VFA performed the worst. UST-MADRL performed worse than GA when the number of USVs was 7 in each SSA. This was because when the number of USVs increased, UST-MADRL needed to consider both the higher movement fairness and average energy consumption of USVs. Contrarily, GA only focused on the energy consumption of USV while ignoring the unfair movement that could have made partial USVs run out of the energy and could have caused the void network communication. As for the VFA, the increase in the USVs number in each SSA will make the balance of the attraction and repulsion among USVs difficult.

In Figure 5d, it can be seen that UST-MADRL was always superior to the others in terms of communication efficiency. For example, when the USVs in each SSA was 4, our framework was 17.7% higher than DDPG, 46.5% higher than GA, 109% higher than PSO, and 127% higher than the VFA. When the USVs in each SSA was 7, our algorithm was 4.2% higher than DDPG, 31.7% higher than GA, 97% higher than PSO, and 124% higher than VFA. Our framework showed an average 10.9% improvement compared to DDPG. The communication efficiency of VFA was the worst because it was not able to comprehensively consider the movement energy consumption and movement fairness of USVs while ensuring the maximum communication coverage score. With the increase in the number of USVs, the difference between UST-MADRL and DDPG reduced because UST-MADRL employed central training and decentral execution based on the improvement of MADDPG, which showed poor convergence with the larger number of agents.

Figure 6 verifies the influence of the NEC of a unit movement distance on the four indicators when it changed from 0.4 to 1.4. Our UST-MADRL outperformed the others. It is obvious that the four indicators become worse with the increase in the NEC value.
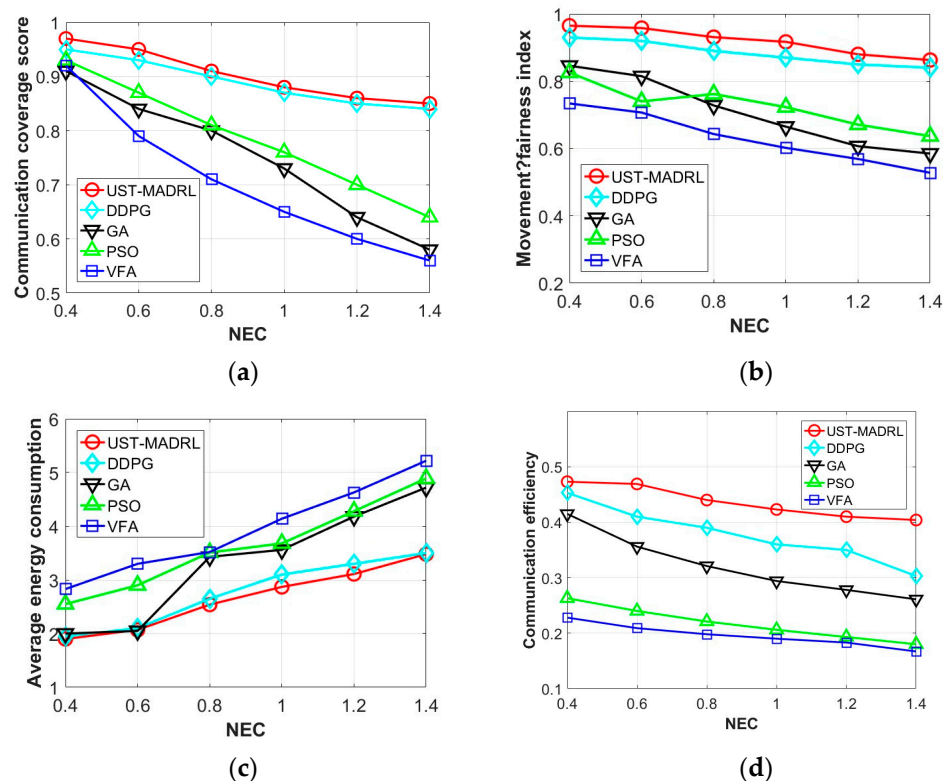
Figure 6. Impact of NEC on (**a**) communication coverage score, (**b**) movement fairness index, (**c**) average energy consumption, (**d**) communication efficiency.

Although UST-MADRL tended to decrease in terms of the communication coverage score when the NEC increased, UST-MADRL was still better than the other baselines, as shown in Figure 6a. GA always selected an action with the smallest energy cost, thus the more the NEC impacted its results, and it obtained a lower communication coverage score. PSO received less impact from the NEC because it focused on velocity and position rather than other indicators. As the NEC became greater and greater, the VFA, considered the distance, lost the advantage and its communication coverage score decreased. UST-MADRL both considered the average energy consumption of USVs and focused on the movement fairness index, which made USVs obtain a higher communication coverage score. DDPG considered the same reward as our framework, therefore, it outperforms the other three baselines.

In Figure 6b, we have the similar observations that UST-MADRL outperformed all baselines in terms of the movement fairness index. For example, when the NEC was 1.0, the movement fairness index of UST-MADRL was 52% higher than that of VFA. In addition, Figure 6c shows that UST-MADRL had the absolute advantage in the aspect of the average energy consumption. When the NEC was 1.4, the average energy consumption of UST-MADRL was 3.48, which was the lowest in all methods. This was because UST-MADRL calculated the global reward value by weighing the average energy consumption, which reduced the overall movement energy consumption of USVs as much as possible.

Finally, we can observe that the communication efficiency of all methods decreased with the increase in the NEC, as shown in in Figure 6d. However, the UST-MADRL achieved the best performance. For example, when the NEC was 1.4, the communication efficiency of UST-MADRL was 0.404, the communication efficiency of DDPG was 0.303, the communication efficiency of GA was 0.261, the communication efficiency of PSO was 0.18, and the communication efficiency of the VFA was 0.167. UST-MADRL gave an average increase of 16.6% compared to DDPG. It was proved that our framework was able to obtain a stable communication coverage score, movement fairness index, and average energy consumption.

7. Conclusions

In this paper, we proposed a UST-MADRL framework that enables UAVs to efficiently navigate the movement of USVs to establish a multi-USV FCC based on MADRL. The optimization of both FCC and IoV performance, including the communication coverage of USVs, USVs' movement fairness, and energy consumption, is multi-objective, mutually coupled, and non-convex. Accordingly, we designed quantitative indexes and transformed the optimization problem into a POMG. Unlike existing research, we put the critic networks into UAVs, taking the communication features of both USVs and UAVs into consideration. We then proposed a UST-ACS, in which USV agents operated in the actor network to execute the mobility strategy. In the meantime, the UAVs acted as critic to evaluate the action of USV agents. Moreover, the information transition efficiently collected the local information from multiple USVs and provided the global experience information to the critic networks in the UAV for better assessment. The definition of SSA further facilitated the convergence of the algorithms. Finally, the numerical results and theoretical analyses demonstrated that our UST-MADRL framework was able to effectively establish the FCC and improve the IoV's communication coverage, the USVs' movement fairness, energy consumption, and communication efficiency. Because our model is based on the improvement of MADDPG, it may be faced with the poor convergence performance associate with a larger scale of USVs. In the future, we will study this issue.

Author Contributions: J.C. and J.D. conceived and designed the whole procedure of this paper. J.C. contributed to the introduction, system model sections, and manuscript writing. J.L. and X.W. performed and analyzed the computer simulation results and drew partial figures. J.D. and Z.G. reviewed and amended writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Department of Science & Technology of Shandong Province through the Natural Science Foundation of Shandong Province under Grant ZR2022MF299.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAVs	Unmanned Aerial Vehicles
USVs	Unmanned Surface Vessels
IoV	Internet of Vessels
FCC	Full Communication Connection
MADRL	Multi-Agent Deep Reinforcement Learning
POMG	Partial Observable Markov Game
MADDPG	Multi-agent Deep Deterministic Policy Gradient
UST-MADRL	MADRL Scenario Strategized by UAVs for Multi-USV FCC
ACS	Multi-agent Actor-Critic Structure
SSA	Sea Service Area
DNN	Deep Neural Network
LR	Long Range
DRL	Deep Reinforcement Learning
RL	Reinforcement Learning
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-network
MADQN	Multi-Agent Deep Q-network
SNR	Signal/Noise Ratio
TD	Time Difference
ReLU	Rectified Linear Unit
TPU	Tensor Processing Unit
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
VFA	Virtual Force Algorithm
NEC	Normalized Energy Consumed

References

1. Lin, B.; Wang, X.; Yuan, W.; Wu, N. A Novel OFDM Autoencoder Featuring CNN-Based Channel Estimation for Internet of Vessels. *IEEE Internet Things J.* **2020**, *7*, 7601–7611. [CrossRef]
2. Singh, R.; Bhushan, B. Condition Monitoring Based Control Using Wavelets and Machine Learning for Unmanned Surface Vehicles. *IEEE Trans. Ind. Electron.* **2021**, *68*, 7464–7473. [CrossRef]
3. Yuan, S.; Li, Y.; Bao, F.; Xu, H.; Yang, Y.; Yan, Q.; Zhong, S.; Yin, H.; Xu, J.; Huang, Z.; et al. Marine environmental monitoring with unmanned vehicle platforms: Present applications and future prospects. *Sci. Total. Environ.* **2023**, *858*, 159741. [CrossRef] [PubMed]
4. Gaugue, M.A.; Menard, M.; Migot, E.; Bourcier, P.; Gaschet, C. Development of an Aquatic USV with High Communication Capability for Environmental Surveillance. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019; pp. 1–8.
5. Wang, Y.; Feng, W.; Wang, J.; Quek, T.Q.S. Hybrid Satellite-UAV-Terrestrial Networks for 6G Ubiquitous Coverage: A Maritime Communications Perspective. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 3475–3490. [CrossRef]
6. Do-Duy, T.; Nguyen, L.D.; Duong, T.Q.; Khosravirad, S.R.; Claussen, H. Joint Optimization of Real-Time Deployment and Resource Allocation for UAV-Aided Disaster Emergency Communications. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 3411–3424. [CrossRef]
7. Zeng, J.; Sun, J.; Wu, B.; Su, X. Mobile edge communications, computing, and caching (MEC3) technology in the maritime communication network. *China Commun.* **2020**, *17*, 223–234. [CrossRef]
8. Li, X.; Feng, W.; Chen, Y.; Wang, C.; Ge, N. Maritime Coverage Enhancement Using UAVs Coordinated with Hybrid Satellite-Terrestrial Networks. *IEEE Trans. Commun.* **2020**, *68*, 2355–2369. [CrossRef]
9. Ao, T.; Zhang, K.; Shi, H.; Jin, Z.; Zhou, Y.; Liu, F. Energy-Efficient Multi-UAVs Cooperative Trajectory Optimization for Communication Coverage: An MADRL Approach. *Remote Sens.* **2023**, *15*, 429. [CrossRef]
10. Liu, C.H.; Dai, Z.; Zhao, Y.; Crowcroft, J.; Wu, D.O.; Leung, K. Distributed and Energy-Efficient Mobile Crowdsensing with Charging Stations by Deep Reinforcement Learning. *IEEE Trans. Mob. Comput.* **2019**, *20*, 130–146. [CrossRef]

11. Samir, M.; Ebrahimi, D.; Assi, C.; Sharafeddine, S.; Ghrayeb, A. Leveraging UAVs for Coverage in Cell-Free Vehicular Networks: A Deep Reinforcement Learning Approach. *IEEE Trans. Mob. Comput.* **2021**, *20*, 2835–2847. [CrossRef]

12. Zhang, Y.; Jiang, L.; Ewe, H.T. A Novel Data-Driven Modeling Method for the Spatial–Temporal Correlated Complex Sea Clutter. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 5104211. [CrossRef]

13. Liu, H.; Weng, P.; Tian, X.; Mai, Q. Distributed adaptive fixed-time formation control for UAV-USV heterogeneous multi-agent systems. *Ocean Eng.* **2023**, *267*, 113240. [CrossRef]

14. Zainuddin, Z.; Wardi; Nantan, Y. Applying Maritime Wireless Communication to Support Vessel Monitoring. In Proceedings of the International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 18–19 October 2017; pp. 158–161. [CrossRef]

15. Chen, L.; Huang, Y.; Zheng, H.; Hopman, H.; Negenborn, R. Cooperative Multi-Vessel Systems in Urban Waterway Networks. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 3294–3307. [CrossRef]

16. Fang, X.; Zhou, J.; Wen, G. Location Game of Multiple Unmanned Surface Vessels with Quantized Communications. *IEEE Trans. Circuits Syst. II Express Briefs* **2022**, *69*, 1322–1326. [CrossRef]

17. Yang, T.; Liang, H.; Cheng, N.; Deng, R.; Shen, X. Efficient Scheduling for Video Transmissions in Maritime Wireless Communication Networks. *IEEE Trans. Veh. Technol.* **2015**, *64*, 4215–4229. [CrossRef]

18. Huang, Z.; Xue, K.; Wang, P.; Xu, Z. A nested-ring exact algorithm for simple basic group communication topology optimization in Multi-USV systems. *Ocean Eng.* **2022**, *266*, 113239. [CrossRef]

19. Zolich, A.; Sœgrov, A.; Vågsholm, E.; Hovstein, V.; Johansen, T.A. Coordinated Maritime Missions of Unmanned Vehicles—Network Architecture and Performance Analysis. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–7. [CrossRef]

20. Cao, H.; Yang, T.; Yin, Z.; Sun, X.; Li, D. Topological Optimization Algorithm for HAP Assisted Multi-unmanned Ships Communication. In Proceedings of the 2020 IEEE 92nd Vehicular Technology Conference, Victoria, BC, Canada, 18 November–16 December 2020; pp. 1–5. [CrossRef]

21. Zhang, J.; Liang, F.; Li, B.; Yang, Z.; Wu, Y.; Zhu, H. Placement optimization of caching UAV-assisted mobile relay maritime communication. *China Commun.* **2020**, *17*, 209–219. [CrossRef]

22. Huang, M.; Liu, A.; Xiong, N.N.; Wu, J. A UAV-Assisted Ubiquitous Trust Communication System in 5G and Beyond Networks. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 3444–3458. [CrossRef]

23. Güldenring, J.; Koring, L.; Gorczak, P.; Wietfeld, C. Heterogeneous Multilink Aggregation for Reliable UAV Communication in Maritime Search and Rescue Missions. In Proceedings of the 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob'19), Barcelona, Spain, 21–23 October 2019; pp. 215–220. [CrossRef]

24. Li, X.; Feng, W.; Wang, J.; Chen, Y.; Ge, N.; Wang, C.X. Enabling 5G on the Ocean: A Hybrid Satellite-UAV-Terrestrial Network Solution. *IEEE Wirel. Commun.* **2020**, *27*, 116–121. [CrossRef]

25. Yang, T.; Jiang, Z.; Sun, R.; Cheng, N.; Feng, H. Maritime Search and Rescue Based on Group Mobile Computing for Unmanned Aerial Vehicles and Unmanned Surface Vehicles. *IEEE Trans. Ind. Inform.* **2020**, *16*, 7700–7708. [CrossRef]

26. Liu, C.; Ma, X.; Gao, X.; Tang, J. Distributed Energy-Efficient Multi-UAV Navigation for Long-Term Communication Coverage by Deep Reinforcement Learning. *IEEE Trans. Mob. Comput.* **2020**, *19*, 1274–1285. [CrossRef]

27. Bae, H.J.; Koumoutsakos, P. Scientific multi-agent reinforcement learning for wall-models of turbulent flows. *Nat. Commun.* **2022**, *13*, 1443. [CrossRef] [PubMed]

28. Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; Basar, T. Fully Decentralized Multi-agent Reinforcement Learning with Networked Agents. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5872–5881.

29. Peng, H.; Shen, X. Multi-Agent Reinforcement Learning Based Resource Management in MEC- and UAV-Assisted Vehicular Networks. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 131–141. [CrossRef]

30. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, O.P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6379–6390.

31. Liang, L.; Ye, H.; Li, G.Y. Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2282–2292. [CrossRef]

32. Ouamri, M.A.; Barb, G.; Singh, D.; Adam, A.B.M.; Muthanna, M.S.A.; Li, X. Nonlinear Energy-Harvesting for D2D Networks Underlaying UAV with SWIPT Using MADQN. *IEEE Commun. Lett.* **2023**, *27*, 1804–1808. [CrossRef]

33. Ouamri, M.A.; Alkanhel, R.; Singh, D.; El-Kenaway, E.-S.M.; Ghoneim, S.S.M. Double Deep Q-Network Method for Energy Efficiency and Throughput in a UAV-Assisted Terrestrial Network. *Comput. Syst. Sci. Eng.* **2023**, *46*, 73–92. [CrossRef]

34. Xu, X.; Chen, Q.; Mu, X.; Liu, Y.; Jiang, H. Graph-Embedded Multi-Agent Learning for Smart Reconfigurable THz MIMO-NOMA Networks. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 259–275. [CrossRef]

35. Liu, Y.; Wang, C.-X.; Chang, H.; He, Y.; Bian, J. A Novel Non-Stationary 6G UAV Channel Model for Maritime Communications. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2992–3005. [CrossRef]

36. Khawaja, W.; Guvenc, I.; Matolak, D.W.; Fiebig, U.; Schneckenburger, N. A Survey of Air-to-ground Propagation Channel Modeling for Unmanned Aerial Vehicles. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2361–2389. [CrossRef]

37. Tran, T.A.; Sesay, A.B. A Generalized Linear Quasi-ML Decoder of OSTBCs for Wireless Communications over Time-Selective Fading Channels. *IEEE Trans. Wirel. Commun.* **2004**, *3*, 855–864. [CrossRef]

38. Zhang, X.; Duan, L. Fast Deployment of UAV Networks for Optimal Wireless Coverage. *IEEE Trans. Mob. Comput.* **2019**, *18*, 588–601. [CrossRef]

39. Kimura, T.; Ogura, M. Distributed Collaborative 3D-Deployment of UAV Base Stations for On-Demand Coverage. In Proceedings of the IEEE INFOCOM 2020—IEEE Conference on Computer Communications, Toronto, ON, Canada, 6–9 July 2020; pp. 1748–1757. [CrossRef]

40. Eledlebi, K.; Ruta, D.; Hildmann, H.; Saffre, F.; Alhammadi, Y.; Isakovic, A.F. Coverage and Energy Analysis of Mobile Sensor Nodes in Obstructed Noisy Indoor Environment: A Voronoi-Approach. *IEEE Trans. Mob. Comput.* **2022**, *21*, 2745–2760. [CrossRef]

41. Jain, R.K.; Chiu, D.M.; Hawe, W.R. A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems. In Proceedings of the DEC Research Report TR-301, Hudson, MA, USA, 26 September 1984; 38p.

42. Littman, M.L. Markov Games as A Framework for Multi-Agent Reinforcement Learning. In Proceedings of the 11th International Conference Machine Learning, San Francisco, CA, USA, 10–13 July 1994; pp. 157–163.

43. Sipper, M. A serial complexity measure of neural networks. In Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, USA, 28 March–1 April 1993; pp. 962–966.

44. Zhao, N.; Ye, Z.; Pei, Y.; Liang, Y.-C.; Niyato, D. Multi-Agent Deep Reinforcement Learning for Task Offloading in UAV-Assisted Mobile Edge Computing. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 6949–6960. [CrossRef]

45. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the International Conference Learn Representations, San Juan, Puerto Rico, 2–4 May 2016.

46. Sivanandam, S.; Deepa, S. *Introduction to Genetic Algorithms*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 15–37.

47. Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In Proceedings of the ICNN'95—International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995.

48. Zou, Y.; Chakrabarty, K. Sensor deployment and target localization based on virtual forces. In Proceedings of the the IEEE INFOCOM 2003, Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428), San Francisco, CA, USA, 30 March–3 April 2003; Volume 2, pp. 1293–1303. [CrossRef]