sensors

MDPI

*Article*

# Deep Reinforcement Learning-Based Resource Management in Maritime Communication Systems

**Xi Yao [1], Yingdong Hu [1], Yicheng Xu [1] and Ruifeng Gao [2,*]**

[1] School of Information Science and Technology, Nantong University, Nantong 226019, China; 2110310028@stmail.ntu.edu.cn (X.Y.); huyd@ntu.edu.cn (Y.H.); yc.x@ntu.edu.cn (Y.X.)
[2] School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China
* Correspondence: grf@ntu.edu.cn

**Abstract:** With the growing maritime economy, ensuring the quality of communication for maritime users has become imperative. The maritime communication system based on nearshore base stations enhances the communication rate of maritime users through dynamic resource allocation. A virtual queue-based deep reinforcement learning beam allocation scheme is proposed in this paper, aiming to maximize the communication rate. More particularly, to reduce the complexity of resource management, we employ a grid-based method to discretize the maritime environment. For the combinatorial optimization problem of grid and beam allocation under unknown channel state information, we model it as a sequential decision process of resource allocation. The nearshore base station is modeled as a learning agent, continuously interacting with the environment to optimize beam allocation schemes using deep reinforcement learning techniques. Furthermore, we guarantee that grids with poor channel state information can be serviced through the virtual queue method. Finally, the simulation results provided show that our proposed beam allocation scheme is beneficial in terms of increasing the communication rate.

**Keywords:** deep reinforcement learning; beam allocation scheme; deep Q-network



**Citation:** Yao, X.; Hu, Y.; Xu, Y.; Gao, R. Deep Reinforcement Learning-Based Resource Management in Maritime Communication Systems. *Sensors* **2024**, *24*, 2247. https://doi.org/10.3390/s24072247

Academic Editor: Jingjing Wang

Received: 12 March 2024
Revised: 24 March 2024
Accepted: 28 March 2024
Published: 31 March 2024



## 1. Introduction

The sixth-generation (6G) wireless communication aims to expand network coverage and improve network performance [1,2]. Maritime communication, as an important component of wireless communication, has received increasing attention with the growing maritime economy.

In general, maritime communication is composed of satellite communication and communication based on nearshore base stations (BSs). Satellite communication systems, such as the Global Maritime Distress and Safety System (GMDSS), Iridium system, and International Maritime Satellite System (Inmarsat) [3–5] can cover the large maritime environment, meeting the communication requirements of maritime users. However, the high cost and latency of satellite communication are the main challenges faced in maritime satellite communication [6]. Communication systems based on nearshore BSs can be integrated with terrestrial communication systems, which effectively reduces cost and latency. Nevertheless, compared to terrestrial communication, maritime communication is subject to multiple factors. The refractive index fluctuation caused by the uneven atmospheric pressure and temperature, namely turbulence, reduces the performance of the communication system [7], and, due to the lack of scatterers in the vast maritime environment, the scattering of electromagnetic waves affects communication performance [8]. Furthermore, maritime communication users exhibit locally dense, and overall sparse, distribution characteristics. These factors make it inappropriate to directly address the communication requirements of maritime users through traditional terrestrial communication.

Currently, there are some nearshore communication systems and networks designed for maritime communication. The Long Term Evolution (LTE)–Maritime project aims to

meet the communication requirements of maritime users using ground infrastructure [9]; it can support a high data rate while providing coverage around 100 km from BSs. The mesh TRITON network, based on IEEE 802.16, focuses on dense wireless mesh networks in maritime nearshore areas for maritime users [10]. The nearshore communication systems, such as the navigation telex (NAVTEX) system and the automatic identification system (AIS) provide services for information broadcasting, voice, and ship identification [11]. However, the above communication systems and networks are merely direct applications of terrestrial communication systems in maritime environments. At present, there is still a lack of communication schemes based on the characteristics of the maritime environment and maritime users. To solve the above problems, beamforming technology can provide directional coverage, ensuring communication quality for users within the coverage area [12,13].

In maritime communication systems, beamforming technology can be used to solve the issue of communication distance and improve user communication quality [14,15]; it can further enhance communication system performance through its combined application with non-orthogonal multiple access (NOMA) technologies [16]. However, further research is needed on how to manage beam resources, and when and where to make beam management decisions. On the one hand, maritime communication faces challenges such as long distances between BSs and users, significant transmission delays, and high interference [17]. It is difficult to obtain the CSI in the maritime environment [18], and the traditional communication resource allocation schemes based on CSI are difficult to employ in the maritime environment [19]. On the other hand, there is still a lack of relevant research on the distribution characteristics of communication users in maritime environments. So, it is necessary to design an efficient beam management scheme to cater to the characteristics of the maritime environment and users.

The essence of a beam management scheme is essentially a combinatorial optimization problem; its computational complexity increases continuously with the growth of the dimension of combinatorial optimization states. Traditional methods are not applicable when the state space is large. Reinforcement learning (RL) can solve this problem by structuring combinatorial optimization as a sequential decision process. It continuously interacts with the environment and updates iterations based on real-time data, optimizing the choices made. RL has been widely applied in the scenario of resource allocation in maritime environments [20,21]. In a rapidly changing vehicular environment, Liang et al. [22] effectively improved the transmission rate in the end-to-end communication link. It has been shown that multi-agent RL can achieve significant performance gains by allocating appropriate resources in the face of uncertain environments [23].

Nevertheless, Shi et al. [24] also point out that, as the number of serviced users increases, which corresponds to the growth in the state space of RL, there is a higher requirement for online devices, which will reduce the efficiency of service devices. Meanwhile, due to the dynamic nature of the maritime environment, it is challenging to fully describe the state information of all communication users using a Q-network in RL. Hence, we consider using the method of deep reinforcement learning (DRL) to address the resource allocation problem. DRL as an effective method has achieved great performance in resource management [25,26]. In [27], Hu et al. propose a DRL framework model to address the decision problem of dynamic resource allocation in satellite communication. In [28], Qian et al. effectively reduced the total energy consumption of the entire maritime Internet of Things network by utilizing relay devices for resource allocation through a DRL framework. All of these papers demonstrate the effectiveness of using a DRL method for resource allocation in communication networks. Furthermore, in order to further reduce the complexity of communication systems, an optimized algorithm using a grid-based method to tackle the resource allocation management problem is proposed in [29]. Similarly, we can divide the sea area covered by the BS's transmitted signal into multiple grids.

In this paper, a virtual queue-based deep reinforcement learning (VQDRL) beam allocation scheme for maritime communication systems is proposed, wherein the CSI is unknown, aiming to maximize the communication rate of the maritime communication

system. Due to the sparse features of maritime users, an optimized algorithm using a grid-based method is employed to tackle the resource allocation management problem. After modeling the maritime communication system and VQDRL framework, we employ neural networks to allocate beam resources. Furthermore, the neural network is trained based on the communication rate of maritime users to obtain the optimal beam allocation scheme. This paper has three main contributions which are as follows:

1. Due to the complexity of the maritime environment, a grid-based method is adopted to discretize the coverage area, reducing the complexity of resource management.
2. A VQDRL resources allocation scheme is employed in grids with unknown CSI. By continuously training the neural network to optimize its output, we obtain the most effective beam allocation scheme.
3. A virtual queue method is employed in maritime communication, it can ensure that grids with poor channel states can be served.

The remainder of this paper is organized as follows. Section 2 introduces the system model. A VQDRL resources allocation scheme is proposed in Section 3, and the simulation results are given in Section 4. Finally, we conclude this paper in Section 5.

## 2. System Model and Problem Formulation

### *2.1. System Model*

The maritime communication system model is shown in Figure 1. A uniform antenna array with a total of $N_t = N_t^h \times N_t^v$ antennas is deployed on the base station (BS). There are $N_t^h$ antennas in the horizontal direction and $N_t^v$ antennas in the vertical direction. Based on the characteristics of the antennas, a total of $N$ beams are generated. After dividing the maritime environment using the grid-based method, we can get $M$ grids. Now, $\mathbf{N} = \{1, \dots n, \dots N\}$ and $\mathbf{M} = \{1, \dots m, \dots M\}$ represent the set of beams and grids, respectively. Here, $n$ and $m$ denote the index of beams and grids. When we obtain the distance $d_m$ and horizontal angle $\theta_m$ from grid $m$ to the BS, we can describe the grid information by using the set $s_m = \{d_m, \theta_m\}$. And the received signal at the $m$-th grid with the $n$-th beam can be expressed as

$$y_{m,n} = \sqrt{\frac{P_t G_m}{d_m^\alpha}} \mathbf{h}_m \mathbf{w}_n x_m + \sqrt{\frac{P_t G_m}{d_m^\alpha}} \sum_{\substack{b \in \mathbf{N} \\ b \neq n}} \sum_{\substack{a \in \mathbf{M} \\ a \neq m}} \mathbf{h}_m \mathbf{w}_b x_a + n_0 \tag{1}$$

where $P_t$ represents the transmit beam power, $G_m$ denotes the antenna gain for grid $m$, $\alpha$ is the path loss coefficient, and $\mathbf{h}_m$ and $\mathbf{w}_n$ denote the channel from grid $m$ to the BS and the unit-norm precoding vectors of beam $n$, respectively, which satisfy $||\mathbf{w}_m|| = ||\mathbf{w}_b|| = 1$. Let $x_m$ and $x_a$ denote the transmission signal from the BS to the grid $m$ and $a$; moreover, the $n_0 \sim \mathcal{CN}(0, 1)$ is the additive complex white Gaussian noise.

### *2.2. Problem Formulation*

From Equation (1), when we transmit a signal from the BS to the grid $m$ with beam $n$, we can get the received SINR of the users at time slot $t$ using the following equation:

$$\mathrm{SINR}_{m,n}(t) = \frac{P_t G_m d_m^{-\alpha} |\mathbf{h}_m(t) \mathbf{w}_n(t)|^2}{1 + P_t G_m d_m^{-\alpha} \sum_{b \neq n}^{b \in \mathbf{N}} |\mathbf{h}_m(t) \mathbf{w}_b(t)|^2} \tag{2}$$

The communication rate can be expressed as

$$R_{m,n}(t) = \log_2(1 + \mathrm{SINR}_{m,n}(t)) \tag{3}$$

Consequently, the achievable communication sum-rate of the maritime communication system model is given by

$$R_{\text{sum}}(t) = \sum_{m \in \mathbf{M}} \sum_{n \in \mathbf{N}} R_{m,n}(t) \tag{4}$$

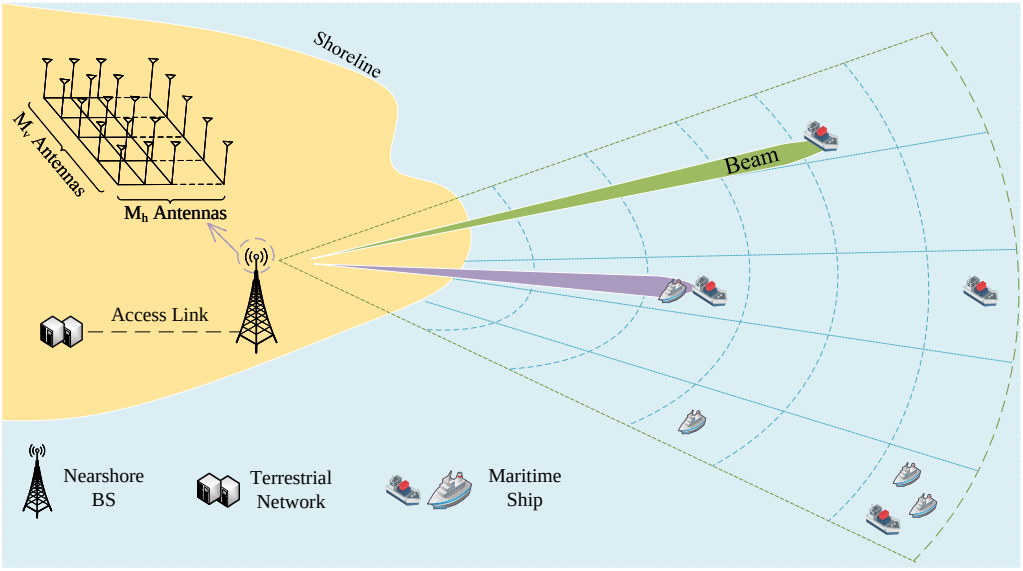


**Figure 1.** Maritimecommunication system model.

To maximize the accumulated communication throughput in a period set $\mathbf{T} = \{1, 2, \ldots, T\}$, the optimization problem can be expressed as

$$\max \sum_{t \in \mathbf{T}} \sum_{m \in \mathbf{M}} \sum_{n \in \mathbf{N}} R_{m,n}(t) K_{m,n}(t) \tag{5}$$

$$s.t. \quad K_{m,n}(t)(1 - K_{m,n}(t)) = 0, \forall m \in \mathbf{M}, n \in \mathbf{N}, t \in \mathbf{T}, \tag{5a}$$

$$\sum_{m \in \mathbf{M}} K_{m,n}(t) \leq P, \forall m \in \mathbf{M}, n \in \mathbf{N}, t \in \mathbf{T}, \tag{5b}$$

$$\sum_{n \in \mathbf{N}} K_{m,n}(t) \leq P, \forall m \in \mathbf{M}, n \in \mathbf{N}, t \in \mathbf{T}, \tag{5c}$$

$$R_{m,n}(t) K_{m,n}(t) \geq R_{th}(t) K_{m,n}(t), \forall m \in \mathbf{M}, n \in \mathbf{N}, t \in \mathbf{T}. \tag{5d}$$

where $R_{th}(t)$ denotes the minimum achievable communication rate for an $m$ user in time slot $t$, and $K_{m,n}(t)$ represents whether the communication of beam $n$ for grid $m$ in the current time slot $t$ was successful.

Constraints (5a) ensure that the value of $K_{m,n}(t)$ can be either 0 or 1. Constraints (5b) and (5c) denote that a maximum of $P$ beams can be used to serve $P$ grids at any given time slot. Constraint (5d) means that when the current communication rate is greater than the minimum rate or equal to the minimum rate, the value of $K_{m,n}(t)$ is 1, otherwise, the value of $K_{m,n}(t)$ is 0.

### *2.3. Deep Reinforcement Learning Model*

For the optimization equation described above, it is impossible to obtain real-time channel state information $\mathbf{h}_m(t)$ to adjust beam allocation schemes. The beam allocation problem is constructed as an RL system model and the RL algorithm can optimize the action-choosing behavior through massive interactions between the agent and environment. However, when the dimensions of the state space and action space are too large, traditional tabular-based RL algorithms face issues such as being time-consuming. The DRL addresses the limitations of traditional RL algorithms by using deep neural networks to select actions;

it updates the weight parameters by minimizing the loss function to optimize the action-choosing behavior. The structure of the DRL model is shown in Figure 2.

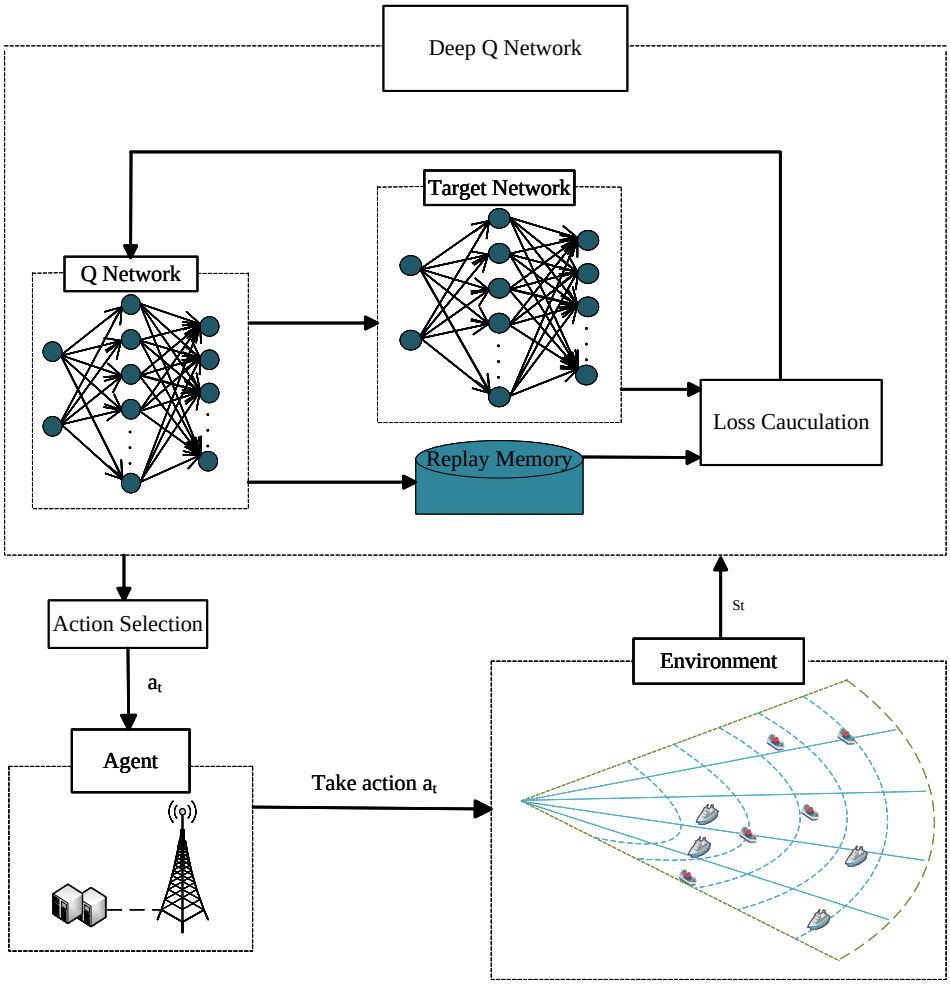


**Figure 2.** Deep reinforcement learning model.

#### 2.3.1. State Definition

In our model, we can obtain the state information of the grid that needs communication, namely as $s_m = \{d_m, \theta_m\}$. The set of state information can be denoted as $\mathbf{S} = \{s_1, ...s_m, ...s_M\}$. Moreover, the BS can handle communication requirements from $P$ grids at any given communication time slot $t$, and the system state $s(t) \in \mathbb{R}^{P\times 1} \subseteq \mathbf{S}$.

#### 2.3.2. Action Definition

We define $\mathcal{A}(s(t))$ as the set of available actions under state $s(t)$. For any available action $a^*$, we have $a^* \in \mathcal{A}(s(t))$. Additionally, we use $a(t)$ to denote the beam chosen scheme at time slot $t$.

#### 2.3.3. Reward Definition

We use reward $r(t)$ to describe the degree of goodness or badness of taking an action $a(t)$ in the current state $s(t)$. In our model, we use $r(t)$ to denote the condition of the communication rate. Typically, the reward value should be normalized in range $r(t) \in [0, 1]$.

#### 2.3.4. Action Selection

To avoid the local optimum of DRL, the resource allocation decision is made by adopting the $\epsilon$-greedy strategy, where the action is randomly selected with a probability of $\epsilon$, while the action with largest action-values is selected with a probability of $1 - \epsilon$. We use

$Q(s(t), a(t); \theta)$ to denote the chosen action through the Q-network. Action selection with the $\epsilon$-greedy strategy can be expressed as:

$$a(t) = \begin{cases} \text{random}, & \text{with probability } \epsilon \\ \arg\max_{a^*} Q(s(t), a^*; \theta), & \text{otherwise} \end{cases} \quad (6)$$

2.3.5. Replay Memory

To alleviate the problems of related data and non-stationary distributions in our system model, a replay memory technique, which randomly samples previous transitions, and, thereby, smooths the training distribution over many past behaviors, is adopted. The experience item is stored in the form of quad-tuples $(s(t), a(t), r(t), s(t+1))$.

2.3.6. Loss Calculation

In order to improve the deep neural network performance, we use $Q(\theta^-)$ to denote the target network. During Q-network training, we update the weight parameters of the Q-network $\theta$ by minimizing the loss function, as given in Equation (3):

$$L(\theta) = \mathbb{E}[(Q_{target} - Q(s(t), a(t); \theta))^2] \quad (7)$$

where the $Q_{target}$ is the target value defined below:

$$Q_{target} = r(t) + \gamma \max_{a^*} Q(s(t+1), a^*; \theta^-) \quad (8)$$

$\gamma$ is the discount factor.

## 3. VQDRL Resources Allocation Scheme

In this section, a VQDRL resources allocation scheme is proposed to address the beam resource management problem in the maritime communication system. First, a virtual queue method is introduced to choose the grids of communication [30]. Then, the BS allocates the beam to the selected communication grids by using the proposed VQDRL resources allocation scheme, which can maximize the communication rate.

### *3.1. Virtual Queue Method*

In communication, we can use DRL to select beam and grid pairs to maximize the communication rate. However, each grid has a different CSI, and directly using DRL to allocate beams may result in some grids with poorer channel conditions being unable to communicate. To address this issue, the method of virtual queue is introduced. By recording communication requirements and queue lengths in different grids, this method can ensure that communication is implemented in grids with poor channel conditions.

When we transmit signals to maritime users with communication requirements, we also need to allocate beam resources as frequently as possible to ensure the communication quality of those users with higher requirements. Meanwhile, we aim to send signals fairly to each maritime area with communication requirements. By recording the requirements of different grids, the virtual queue method can more frequently select grids with higher communication requirements for communication; it also ensures fairness in the communication among grids.

Let $\mathbf{r} = [r_1, \ldots, r_m, \ldots, r_M]$ represents the vector of communication requirements of maritime users within each grid, and, for any grid $m$, we can calculate the proportion of a user's requirement to the total grid requirements as follows:

$$p_m = \frac{r_m}{\sum_{i=1}^{M} r_i} \quad (9)$$

Now, for each grid $m$, we create a virtual queue $V_m$ to denote the queue length at the beginning of the communication; it can be seen as the accumulated length of communication requirements up to the current communication time slot. The virtual queue length $V_m(t)$ is denoted according to the following dynamics:

$$V_m(t) = \left[V_m(t-1) + p_m - d_m(t-1)\right]^+ \tag{10}$$

where $\left[x\right]^+ \triangleq \max\{x, 0\}$, $d_m(t-1)$ denotes whether the communication rate in the current grid at time slot $t-1$ meets the communication rate requirement; it can be denoted as $d_m(t-1) = \eta \sum_{n=1}^{N} K_{m,n}(t-1)$, and $\eta$ is a coefficient to avoid excessively queue length.

At any given time slot $t$, the BS selects the grid to be served by beam resources and obtains the grid information $s_m$ using the following equation:

$$s_m \in \arg\max_{i \in \mathbf{M}}(V_i(t)) \tag{11}$$

### *3.2. VQDRL Resources Allocation Scheme*

After obtaining the communication grids for the current time slot using the virtual queue method, we aim to maximize the communication rate by applying the DRL algorithm for beam allocation in these grids. However, the obtained grid information cannot be directly used as inputs for the neural network. In this case, the VQDRL resources allocation scheme is proposed to effectively utilize the obtained grid information and employ it as input to the neural network to obtain beam allocation schemes that maximize the communication rate.

According to Equation (11), when we obtain the transmitted grid information $s_m$ in the current time slot $t$, the VQDRL algorithm uses $s(t) = s_m = \{d_m, \theta_m\}$ as an input to the neural network to obtain the output, which is the index of the allocated beam.

We can get different $Q(s(t), a^*; \theta)$ values through the Q-network after inputting the state information $s(t)$. In order to avoid the local optimum of VQDRL, we choose the action $a(t)$ taken by the agent in the current state to allocate beam based on the probability of $\epsilon$. When a randomly generated value is more than $\epsilon$, we randomly select an action $a$, otherwise, we choose the action $\arg\max_{a^*} Q(s(t), a^*; \theta)$.

After selecting the action $a(t)$ for the current time slot, when the BS sends beams to the grids, the maritime users will get their communication rates $R_{m,n}(t)$. The reward $r(t)$ is used to describe the condition of the communication rate, and we can obtain the next state $s(t+1)$. It is obvious that when the constraint condition (5d) is satisfied under condition $\sum_{m,n} R_{m,n}(t)K_{m,n}(t) = \sum_{m,n} R_{th}(t)K_{m,n}(t) = 0$, the reward $r(t) = 0$. Otherwise, we define the function of reward $r(t)$ as follows:

$$r(t) = \begin{cases} 1 & \text{if } \sum\limits_{m,n} R_{m,n}(t)K_{m,n}(t) \geq \sum\limits_{m,n} R_{th}(t)K_{m,n}(t) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

To further train our neural network while avoiding the issue of local optimization, we often use the replay memory D to store training data. Let us use the buffer size to describe the maximum amount of data stored in D and store the data set $(s(t), a(t), r(t), s(t+1))$ from the communication process into D. When the amount of data set stored is larger than the pre-defined batch size, we can randomly select a batch size of data from D to update the neural network. Otherwise, we continue the aforementioned communication process until the amount of data stored in D exceeds the batch size.

After selecting the data to be used for training, we update the weight parameters of the Q-network through the following process. We typically use the Bellman equation to update the Q-values like Equation (13) in RL:

$$Q(s(t),a(t)) \leftarrow Q(s(t),a(t)) + \beta[r(t) + \gamma \max_{a^* \in \mathcal{A}(s(t))} Q(s(t+1),a^*) - Q(s(t),a(t))] \tag{13}$$

where $\beta$ is the learning rate. Similar to RL, in DRL, based on the Bellman equation, we can obtain the weight update through gradient descent using the loss function. According to Equations (7) and (8), the gradient of loss function is calculated by calculating parameters as follows:

$$\frac{dL(\theta)}{d\theta} = \mathbb{E}[Q_{target} - Q(s(t),a(t);\theta)\frac{dQ(s(t),a(t);\theta)}{d\theta}]. \tag{14}$$

In our model, we used the Adaptive Moment Estimation (Adam) optimizer to solve the gradient descent problem. It can adaptively adjust the learning rate to more effectively update the model's weights. As a hyperparameter, the learning rate has a significant impact on weight updates for different values. In this model, we used a learning rate of $10^{-3}$ to train the model at the beginning of the training period and hyperparameter $\gamma$ represents the discount factor.

After updating the weights of the Q-network, we determine whether it is time to update the weights of the target network for the current time slot. Typically, the target network parameter weights update every $N_t$ step. The parameter weights of the Q-network are assigned to the target network's parameters, completing the update of the neural network.

## 4. Simulation Result

In this section, we simulated a maritime environment and deployed an antenna array at the BS, constructing beam resources firstly, then we employed the VQDRL algorithm to obtain the beam allocation schemes and observed the communication rate within the beam coverage area through simulation results, further demonstrating the effectiveness of our algorithm.

### *4.1. Simulation Enviroment Configuration*

We deployed Uniform Planar Array (UPA) antenna arrays at the BS, and considered taking the Kronecker product of the Discrete Fourier Transform (DFT) codebook in the horizontal direction and vertical direction by using the Kronecker-product method. The 3D Kronecker-product-based codebook can select the appropriate beam to enhance the channel gain for the grid in both the horizontal and vertical direction [31]. It is generated as

$$\begin{aligned} \mathbf{C}_v &= [1, e^{\frac{j2\pi\mathbf{m}}{\xi N_v}}, \ldots, e^{\frac{j2\pi(N_t^v-1)\mathbf{m}}{\xi N_v}}]^T \\ \mathbf{C}_h &= [1, e^{\frac{j2\pi\mathbf{n}}{N_h}}, \ldots, e^{\frac{j2\pi(N_t^h-1)\mathbf{n}}{N_h}}]^T \\ \mathbf{C} &= \mathbf{C}_v \otimes \mathbf{C}_h \end{aligned} \tag{15}$$

where $\mathbf{m} = 0, 1, \ldots, N_v - 1$, $\mathbf{n} = 0, 1, \ldots, N_h - 1$, $N_v$, and $N_h$ are the number of codewords in the vertical direction and horizontal direction, respectively, $\otimes$ denotes the Kronecker product, and $\xi$ is a parameter to adjust the proportion which is determined by the maximum downtilt.

According to Equations (3) and (12), we need to simulate the maritime communication channel model to obtain the communication rate and the reward value to train the neural network. The maritime communication channel is assumed to follow the Rician distribution, expressed as follows:

$$\mathbf{h}_m = \sqrt{\frac{K}{K+1}}\overline{\mathbf{h}}_m + \sqrt{\frac{1}{K+1}}\hat{\mathbf{h}}_m \tag{16}$$

where $K$ is the rice factor, $\hat{\mathbf{h}_m}$ is the complex Gaussian random variables with zero mean and unit variance, which belong to a set of $\mathbb{C}^{1\times N_t}$, and $\overline{\mathbf{h}}_m$ is the channel mean vector. When we consider the antenna array arranged in a UPA, the channel mean vector of the $m$-th grid vector is expressed by [32,33]:

$$\overline{\mathbf{h}}_m = [1, \ldots, e^{j\frac{2\pi}{\lambda}d(\mathbf{n}_t^h \sin\theta_m \sin\phi_m + \mathbf{n}_t^v \cos\phi_m)}, \ldots, e^{j\frac{2\pi}{\lambda}d(N_t^h \sin\theta_m \sin\phi_m + N_t^v \cos\phi_m)}] \tag{17}$$

where $\lambda$ is the wavelength, $d$ is the inter-antenna spacing, and $\theta_m$ is the horizontal angle of the BS to the $m$-th grid, while $\phi_m$ is the vertical angle of the BS to the $m$-th grid, which can be calculated based on the distance from the grid to the BS under the condition of determining the height of the BS and the antenna, $\mathbf{n}_t^h = 0, 1, \ldots, N_t^h$, $\mathbf{n}_t^v = 0, 1, \ldots, N_t^v$.

For the convenience of distinction, the above model parameters and channel transmission parameters are represented in Table 1.

**Table 1.** Summary of Key Notations.

| Notations | Meaning |
|---|---|
| $\mathbf{M}; \mathbf{N}$ | Set of grids and beam resources |
| $\theta_m; \phi_m$ | Horizontal and vertical angle from grid $m$ to BS |
| $y_{m,n}$ | Receive signal from beam $n$ to grid $m$ |
| $x_m$ | Transmit signal to grid $m$ |
| $P_t; \alpha$ | Transmit power and path loss coefficient |
| $G_m$ | Antenna gain for grid $m$ |
| $n_0$ | Additive complex white Gaussian noise |
| $\mathbf{h}$ | Channel transfer matrix |
| $\mathbf{w}$ | Unit-form precoding vector |
| $\mathbf{S}; \mathcal{A}$ | Set of state information and available action |
| $r(t)$ | Reward function |
| $R_{m,n}$ | Communication rate of grid $m$ with beam $n$ |
| $Q(\theta); Q(\theta^-)$ | Q-network and target network |
| $r_m$ | Requirements of maritime users within grid $m$ |
| $p_m$ | Communication requirement of grid $m$ |
| $V_m; K$ | Virtual queue length of grid $m$ and rice factor |

### *4.2. Average Communication Rate Analysis*

The effectiveness of the VQDRL algorithm under this model will be proven by analyzing the average communication rate with the entire grid and the average communication rate with different grids. All simulations were conducted on a desktop equipped with an Intel Core i7-10700 2.9 GHz CPU(Intel, Santa Clara, CA, USA), with each iteration of $1 \times 10^5$ time slots taking approximately 12 min in our model.

We divided the coverage area of the nearshore BS communication into grids of size $10 \times 10$. The parameters of each grid and model value can be represented as follows in Tables 2 and 3.

**Table 2.** Grid parameters.

| | | | | |
|---|---|---|---|---|
| $s_0\{5\text{ km}, \frac{1}{6}\pi\}$ | ... | $s_4\{5\text{ km}, 0\}$ | ... | $s_9\{5\text{ km}, -\frac{1}{6}\pi\}$ |
| ... | ... | ... | ... | ... |
| $s_{50}\{2.5\text{ km}, \frac{1}{6}\pi\}$ | ... | $s_{54}\{2.5\text{ km}, 0\}$ | ... | $s_{59}\{2.5\text{ km}, -\frac{1}{6}\pi\}$ |
| ... | ... | ... | ... | ... |
| $s_{90}\{0.5\text{ km}, \frac{1}{6}\pi\}$ | ... | $s_{94}\{0.5\text{ km}, 0\}$ | ... | $s_{99}\{0.5\text{ km}, -\frac{1}{6}\pi\}$ |

**Table 3.** Simulation parameters and values.

| Parameters | Values |
|---|---|
| $N_t^h; N_t^h$ | $N_t^h = 8; N_t^h = 8$ |
| $N_h; N_v$ | $N_h = 4; N_v = 4$ |
| $M; N$ | $M = 100; N = 16$ |
| $P_t; G_m$ | $P_t = 1000; G_m = 1$ |
| $\alpha; \beta$ | $\alpha = 1; \beta = 0.001$ |
| $\epsilon; \gamma$ | $\epsilon = 0.9; \gamma = 0.99$ |
| $\xi; R_{th}$ | $\xi = 4; R_{th} = 1$ |
| buffer size; batch size | buffer size = 512; batch size = 64 |
| learning rate; $K$ | learning rate = $1 \times 10^{-4}$; $K = 9$ |
| $N_t; \eta$ | $N_t = 20; \eta = 10$ |

To observe how the communication rate of the entire maritime environment changes with the communication time slots, we used the average communication rate to show the variation in Figure 3. Here, we employed random beam allocation and round robin beam allocation schemes as comparative simulation results. The scheme that randomly allocates a beam to the current communication grid is called the random beam allocation scheme. Meanwhile, the scheme that allocates a beam to the current grid in ascending order by beam index, and restarts the cycle when the maximum beam index is reached, is called the round robin beam allocation scheme.

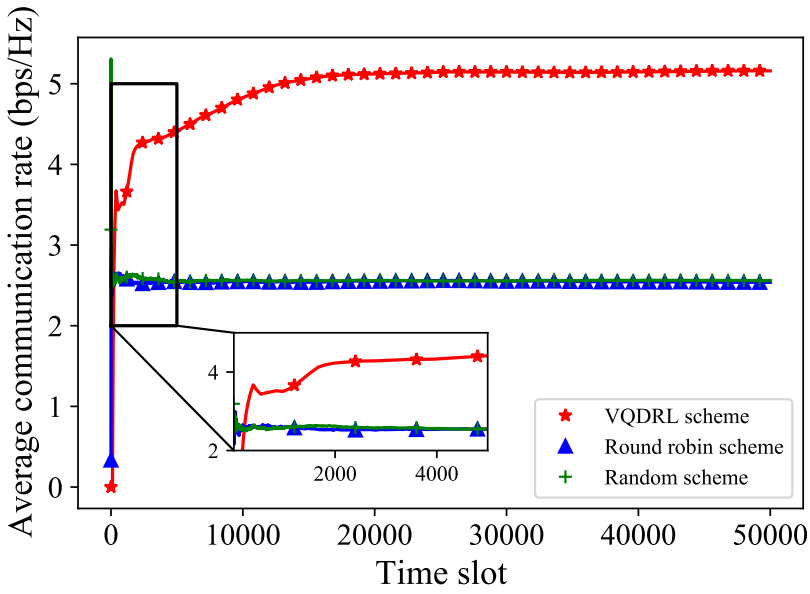


**Figure 3.** The average communication rate variation of different beam allocation schemes with entire grids.

It is shown that, as the time slots increase, the difference in average communication rates between the three beam allocation schemes becomes larger. In particular, the round robin beam allocation scheme and random beam allocation scheme do not show a significant difference in the average communication rate for entire maritime grids. Furthermore, with the increase of time slots, the average communication rate for the entire maritime environment under the two schemes essentially remains consistent. Compared to the other two beam allocation schemes, the proposed VQDRL resources allocation scheme significantly improves the average communication rate for the entire maritime environment, and it gradually converges with the increase of the time slot.

To further analyze the average communication rate of the maritime environment, we sampled some grids to observe their average communication rate. In Figure 4, we take two grids, $\text{Grid}_{55}$ and $\text{Grid}_{95}$, as our observation points to observe the average communication

rate. The left subplot shows the variation of the average communication rate over time for $\text{Grid}_{55}$, while the right subplot shows the variation of the average communication rate over time for $\text{Grid}_{95}$. It is shown that, as the time slots increase, the average communication rate of the proposed VQDRL algorithm gradually increases until it stabilizes and converges. Meanwhile, the average communication rates of the two beam schemes in the comparative simulation gradually converge and eventually become consistent as the number of time slots increases. The performance of the random and round robin beam allocation schemes is worse than the performance of the proposed VQDRL algorithm. This trend aligns with the observations in Figure 3.

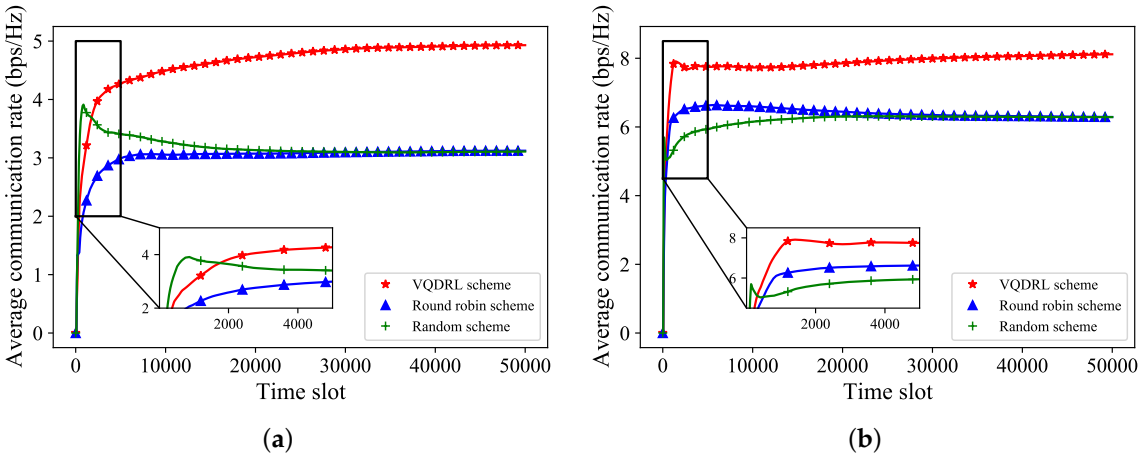


**Figure 4.** The average communication rate of different beam allocation schemes with different grids. (**a**) $\text{Grid}_{55}$. (**b**) $\text{Grid}_{95}$.

### *4.3. Awaiting Time per Transmission and Average Virtual Queue Length*

In this subsection, we discuss the waiting time per transmission and the average virtual queue length of different grids. Let us set an upper bound on the length of the virtual queue, denoted as $V_{max}$, for all grids. When the virtual queue length of any grid exceeds this value, we consider the grid to be in a communication waiting state. Here, we set $V_{max} = 1$, and Figure 5 shows the average waiting time per transmission of different grids over $5 \times 10^4$ time slots.

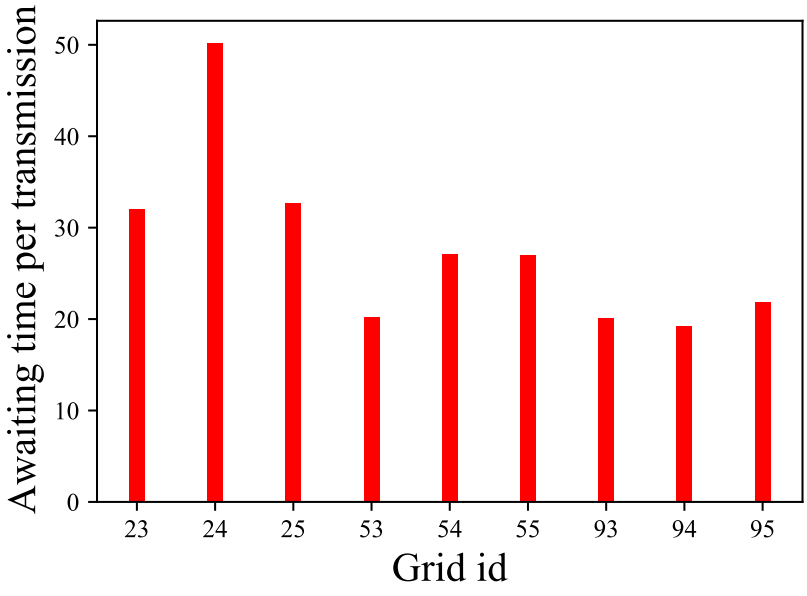


**Figure 5.** The average waiting time slot of different grids.

Meanwhile, we sampled some grids to obtain the average virtual queue length and the confidence interval for the average queue length in Figure 6. The average virtual queue length denotes the growth trend of virtual queues for different grids. It can be calculated by $\frac{\sum_{k=1}^{t} V_m(k)}{t}$. The average virtual queue lengths and their respective confidence intervals under multiple simulations for $Grid_{55}$, $Grid_{53}$, and $Grid_{95}$ are displayed in red, blue, and green, respectively. The purple line represents the upper bound of the virtual queue, and the grids whose virtual queue length exceeds the line at any time slot are considered to be in a waiting state until they are successfully communicated, reducing the virtual queue length to below that line. We observed that the average virtual queue length under these three different grids increases with the growth of time slots and eventually converges to around 0.6, which is often associated with the ability to communicate in each time slot.

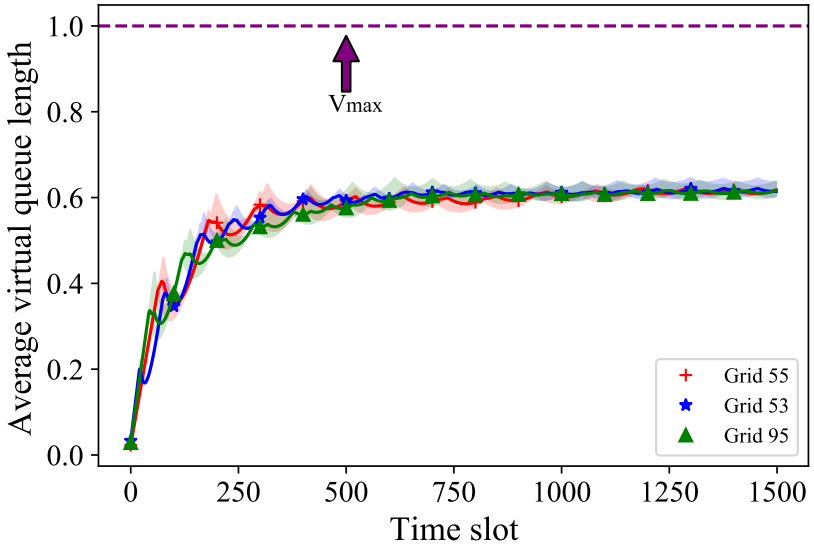


**Figure 6.** The average virtual queue length over communication time slots.

### *4.4. Hyperparameters Analysis*

In this subsection, we adjust buffer size, learning rate, and batch size to observe the variation of the average communication rate.

We adjust the buffer size in Figure 7. At the beginning of the simulation, the average communication rate with a buffer size of 256 is significantly higher than that of the other two batch sizes. As the number of time slots increases, the average communication rate of the buffer size 512 gradually increases and converges near the average communication rate of the buffer size 256. The average communication rate of the largest buffer size remained much lower than the aforementioned two buffer sizes throughout the simulation. This is because when the buffer size is too large, selecting data randomly from replay memory may include some outdated data, which affects the effectiveness of the simulation. In the simulation, we can adjust different buffer sizes to improve the average communication rate in the maritime environment.

As shown in Figure 8, when we set the learning rate to $1 \times 10^{-3}$ and $1 \times 10^{-5}$ , we can see that the convergence curves of the average communication rate are relatively less for both grids. However, when the learning rate is set to $1 \times 10^{-4}$, the average communication rate shows the best performance. This is because a learning rate that is too large or too small can cause the beam selection to get stuck in local optima, resulting in poorer performance. We can set a moderate learning rate value to obtain the best performance.

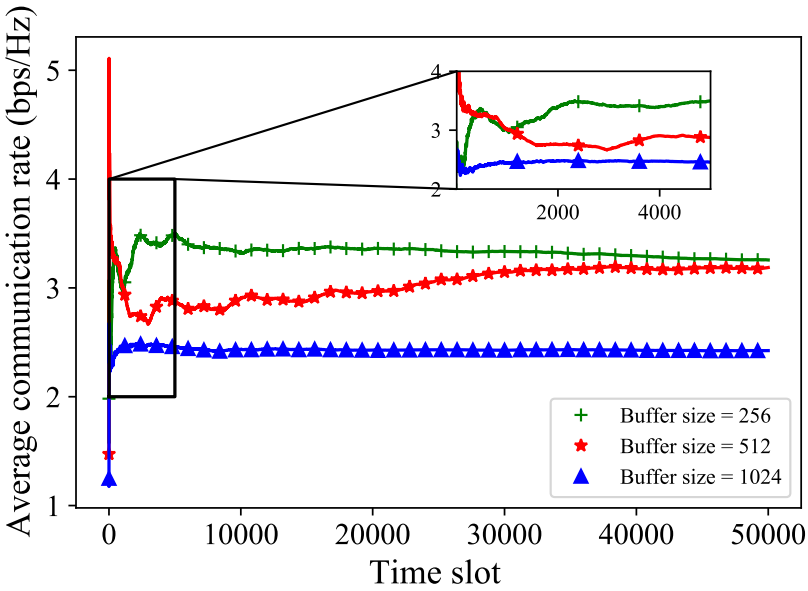


**Figure 7.** The average communication rate with different buffer sizes.

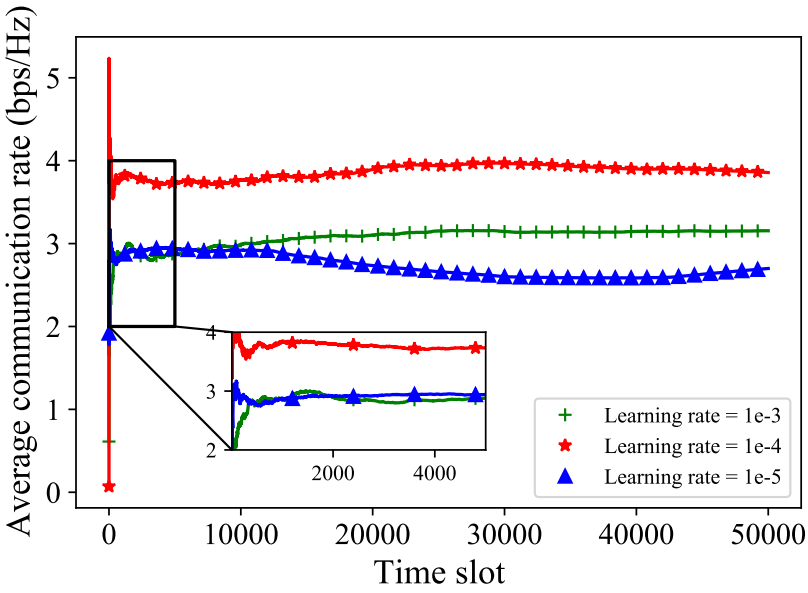


**Figure 8.** The average communication rate with different learning rates.

In Figure 9, we adjust the batch size. We can see that when the batch size is 64, the average communication rate performs better compared to the case with a batch size of 32 and 256, as the number of communications increased. This is because, when the batch size is 32, the training data is not sufficient to adequately represent the data features stored in the replay memory, leading to poor performance. And when the batch size is 256, a larger batch size will rely on training data that are too "old" and degrade the convergence performance. Furthermore, a large batch size will consume more time when training neural networks. Therefore, an appropriate batch size plays a crucial role in training neural networks.

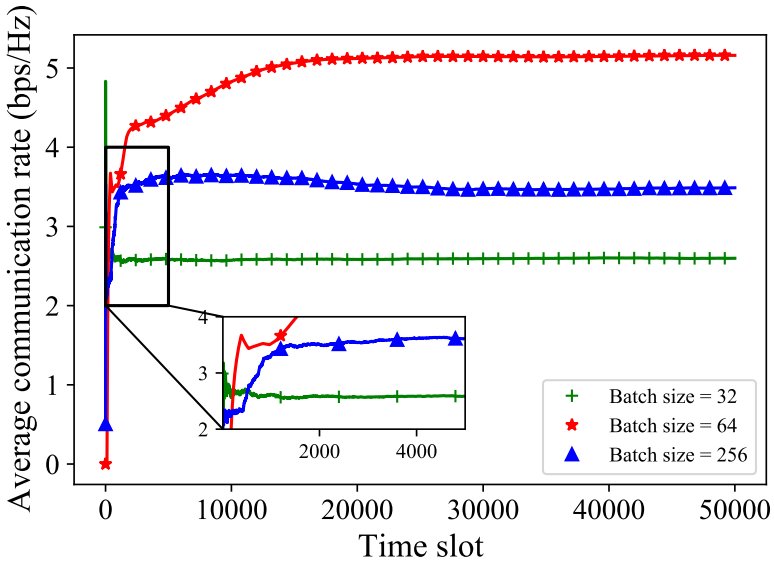


**Figure 9.** The average communication rate with different batch sizes.

## 5. Conclusions

In this paper, a VQDRL resources allocation scheme was proposed and investigated in the maritime environment, and we discussed the average communication rate performance. Firstly, a maritime communication model with grid-based partitioning was employed, and we utilized VQDQN for allocating beam resources. Secondly, the average communication rate of all grids and the average communication rates of different grids were simulated and the simulation results demonstrate the effectiveness of our beam allocation scheme. Additionally, we discussed the performance of average waiting per transmission, average virtual queue length, and confidence intervals of different grids. Finally, we adjusted hyperparameters to obtain a better performance on the average communication rate. The simulation results demonstrated that we can construct suitable beam allocation schemes to maximize the communication rate in the maritime environment using our algorithm. Furthermore, further work can be conducted by adopting the objective function of our proposed scheme to analyze various network performance metrics, including energy efficiency and spectrum utilization. Moreover, improving the generalization performance of our proposed scheme requires further investigation.

**Author Contributions:** Conceptualization, X.Y. and R.G.; methodology, X.Y., Y.H. and R.G.; investigation, X.Y. and Y.X.; writing, X.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the 22KJB510039 Jiangsu Provincial Higher Education Institutions General Program of Basic Science Research, the Fujian Province Special Fund Project for Promoting High-Quality Development of Marine and Fisheries Industries, FJHYF-ZH-2023-0, and the Natural Science Foundation of Nantong (JC2023074).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. You, L.; Chen, X.; Song, X.; Jiang, F.; Wang, W.; Gao, X.; Fettweis, G. Network Massive MIMO Transmission over Millimeter-Wave and Terahertz Bands: Mobility Enhancement and Blockage Mitigation. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 2946–2960. [CrossRef]
2. 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]
3. You, L.; Li, K.; Wang, J.; Gao, X.; Xia, X.; Ottersten, B. Massive MIMO Transmission for LEO Satellite Communications. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 1851–1865. [CrossRef]
4. You, L.; Qiang, X.; Li, K.; Tsinos, C.; Wang, W.; Gao, X.; Ottersten, B. Hybrid Analog/Digital Precoding for Downlink Massive MIMO LEO Satellite Communications. *IEEE Trans. Wireless Commun.* **2022**, *21*, 5962–5976. [CrossRef]
5. Boiardt, H.; Rodriguez, C. Low Earth Orbit nanosatellite communications using Iridium's network. *IEEE Aerosp. Electron. Syst. Mag.* **2010**, *25*, 35–39. [CrossRef]
6. Xu, Y. Quality of Service Provisions for Maritime Communications Based on Cellular Networks. *IEEE Access* **2017**, *5*, 23881–23890. [CrossRef]
7. Zuo, Y.; Xiao, H.; Wu, J.; Hong, X.; Lin, J. Effect of atmospheric turbulence on non-line-of-sight ultraviolet communications. In Proceedings of the 2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications—(PIMRC), Sydney, Australia, 9–12 September 2012; pp. 1682–1686.
8. Wang, J.; Zhou, H.; Li, Y.; Sun, Q.; Wu, Y.; Jin, S.; Quek, T.Q.S.; Xu, C. Wireless Channel Models for Maritime Communications. *IEEE Access* **2018**, *6*, 68070–68088. [CrossRef]
9. Jo, S.; Shim, W. LTE-Maritime: High-Speed Maritime Wireless Communication Based on LTE Technology. *IEEE Access* **2019**, *7*, 53172–53181. [CrossRef]
10. Zhou, M.; Hoang, V.; Harada, H.; Pathmasuntharam, J.; Wang, H.; Kong, P.; Ang, C.; Ge, Y.; Wen, S. TRITON: High-speed maritime wireless mesh network. *IEEE Wireles Commun.* **2013**, *20*, 134–142. [CrossRef]
11. Wei, T.; Feng, W.; Chen, Y.; Wang, C.; Ge, N.; Lu, J. Hybrid Satellite-Terrestrial Communication Networks for the Maritime Internet of Things: Key Technologies, Opportunities, and Challenges. *IEEE Internet Things J.* **2021**, *8*, 8910–8934. [CrossRef]
12. Koppenborg, J.; Halbauer, H.; Saur, S.; Hoek, C. 3D beamforming trials with an active antenna array. In Proceedings of the 2012 International ITG Workshop on Smart Antennas (WSA), Dresden, Germany, 7–8 March 2012; pp. 110–114.
13. He, S.; Huang, Y.; Jin, S.; Yang, L. Coordinated Beamforming for Energy Efficient Transmission in Multicell Multiuser Systems. *IEEE Trans. Commun.* **2013**, *61*, 4961–4971. [CrossRef]
14. Yue, D.; Zhang, Y.; Jia, Y. Beamforming Based on Specular Component for Massive MIMO Systems in Ricean Fading. *IEEE Wireles Commun.* **2015**, *4*, 197–200. [CrossRef]
15. You, L.; Xu, J.; Alexandropoulos, G.; Wang, J.; Wang, W.; Gao, X. Energy Efficiency Maximization of Massive MIMO Communications with Dynamic Metasurface Antennas. *IEEE Trans. Wireless Commun.* **2023**, *22*, 393–407. [CrossRef]
16. Nguyen, T.; Choi, J.; Park, S.; Kim, B.; Shim, W. NOMA with Cache-Aided Maritime D2D Communication Networks. *IEEE Access* **2023**, *11*, 123784–123797.
17. Du, W.; Zhengxin, M.; Bai, Y.; Shen, C.; Chen, B.; Zhou, Y. Integrated Wireless Networking Architecture for Maritime Communications. In Proceedings of the 2010 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, London, UK, 9–11 June 2023; Volume 22, pp. 393–407. [CrossRef]
18. Danklmayer, A.; Förster, J.; Fabbro, V.; Biegel, G.; Brehm, T.; Colditz, P.; Castanet, L.; Hurtaud, Y. Radar Propagation Experiment in the North Sea: The Sylt Campaign. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 835–846.
19. Fang, X.; Feng, W.; Wang, Y.; Chen, Y.; Ge, N.; Ding, Z.; Zhu, H. NOMA-Based Hybrid Satellite-UAV-Terrestrial Networks for 6G Maritime Coverage. *IEEE Trans. Wireless Commun.* **2023**, *22*, 138–152. [CrossRef]
20. Xu, Y.; Yao, J.; Jacobsen, H.; Guan, H. Cost-efficient negotiation over multiple resources with reinforcement learning. In Proceedings of the 2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS), Vilanova i la Geltru, Spain, 14–16 June 2017; pp. 1–6. [CrossRef]
21. You, L.; Qiang, X.; Tsinos, C.; Liu, F.; Wang, W.; Gao, X.; Ottersten, B. Beam Squint-Aware Integrated Sensing and Communications for Hybrid Massive MIMO LEO Satellite Systems. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 2994–3009. [CrossRef]
22. Liang, L.; Ye, H.; Li, G. Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2282–2292.
23. Cui, J.; Liu, Y.; Nallanathan, A. Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks. *IEEE Trans. Wireless Commun.* **2020**, *19*, 729–743. [CrossRef]
24. Shi, Y.; Sagduyu, Y.; Erpek, T. Reinforcement Learning for Dynamic Resource Optimization in 5G Radio Access Network Slicing. In Proceedings of the 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Pisa, Italy, 14–16 September 2020; pp. 1–6.
25. Xu, F.; Yang, F.; Zhao, C.; Wu, S. Deep reinforcement learning based joint edge resource management in maritime network. *China Commun.* **2020**, *17*, 211–222. [CrossRef]
26. Ye, H.; Li, G. Deep Reinforcement Learning for Resource Allocation in V2V Communications. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6. [CrossRef]
27. Hu, X.; Liu, S.; Chen, R.; Wang, W.; Wang, C. A Deep Reinforcement Learning-Based Framework for Dynamic Resource Allocation in Multibeam Satellite Systems. *IEEE Commun. Lett.* **2018**, *22*, 1612–1615.

28. Qian, L.; Zhang, H.; Wang, Q.; Wu, Y.; Lin, B. Joint Multi-Domain Resource Allocation and Trajectory Optimization in UAV-Assisted Maritime IoT Networks. *IEEE Internet Things J.* **2023**, *10*, 539–552. [CrossRef]
29. Guo, W.; Liu, F.; Chen, Z.; Li, Y. Grid Resource Allocation and Management Algorithm Based on Optimized Multi-task Target Decision. In Proceedings of the 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, 12–13 January 2019; pp. 560–564.
30. Li, F.; Liu, J.; Ji, B. Combinatorial Sleeping Bandits with Fairness Constraints. *IEEE Trans. Netw. Sci. Eng.* **2020**, *7*, 1799–1813. [CrossRef]
31. Xie, Y.; Jin, S.; Wang, J.; Zhu, Y.; Gao, X.; Huang, Y. A limited feedback scheme for 3D multiuser MIMO based on Kronecker product codebook. In Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, UK, 8–11 September 2013; pp. 1130–1135. [CrossRef]
32. McKay, M.; Collings, I. General capacity bounds for spatially correlated Rician MIMO channels. *IEEE Trans. Inf. Theory* **2005**, *51*, 3121–3145.
33. Jin, S.; Gao, X.; You, X. On the Ergodic Capacity of Rank-1 Ricean-Fading MIMO Channels. *IEEE Trans. Inf. Theory* **2007**, *53*, 502–517.