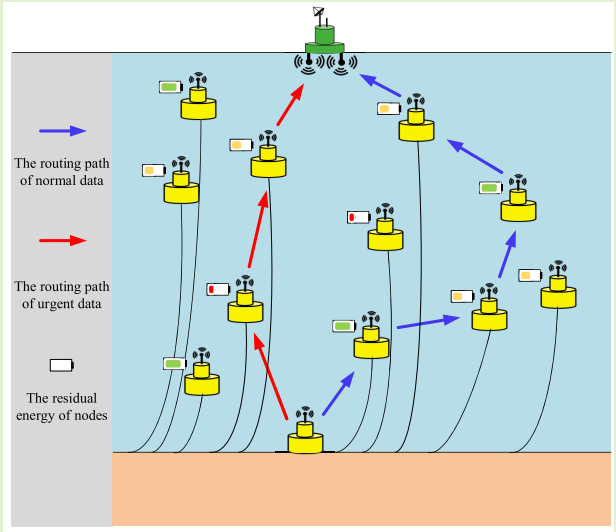
Multi-Agent Reinforcement Learning-Based Routing Protocol for Underwater Wireless Sensor Networks With Value of Information

Chao Wang, Xiaohong Shen, Haiyan Wang, Weiliang Xie, Hongwei Zhang, and Haodi Mei

***Abstract*—Efficient data transmission plays a crucial role in the applications of underwater wireless sensor networks (UWSNs). In this article, by considering the differences in transmission requirements for data of varying importance degrees in UWSNs, a multi-agent reinforcement learning-based routing protocol with value of information (MARV) is proposed. First, to distinguish the difference of transmission requirements, we introduce the value of information (VoI) to characterize the importance degree of data to reflect the requirement for the real-time characteristic. Moreover, to ensure the efficient routing for different importance degree of data, we establish a multi-agent reinforcement learning (MARL)-based framework by enabling nodes to learn from the environment and interact with neighbors and elaborately design a reward function by considering the timeliness and energy efficiency of transmission. In addition, to improve the transmission efficiency, we design a packet holding mechanism by designing a priority list and variable holding interval according to transmission requirements. The simulation results show that the proposed protocol performs well for the transmission of different data.**



***Index Terms*— Multi-agent reinforcement learning (MARL), routing protocol, transmission requirements, underwater wireless sensor networks (UWSNs), value of information (VoI).**

I. Introduction

RECENTLY, the Internet of Things (IoT) has attracted wide attention due to the convenience and efficiency of its services [1], [2], [3]. Underwater wireless sensor networks (UWSNs) as an important branch of the IoT in underwater environment have been widely used in multiple applications, such as disaster prevention, ocean monitoring, and underwater target detection [4], [5], [6]. To ensure efficient collection, data transmission is of great importance for UWSNs [7], [8].

With the increase of application requirements, nodes in UWSNs are required to transmit data of different importance degrees to the sink node [9], [10]. Specifically, urgent data have a high importance degree and should be forwarded promptly and reliably; other data can be forwarded by considering delay and residual energy. Obviously, it is not appropriate for nodes to forward different importance degree data in the same manner, especially under the constraints of battery capacity and delay requirements. UWSNs have to transmit different data according to corresponding transmission requirements. Therefore, this article focuses on designing a routing protocol that meets different transmission requirements.

Due to the harsh underwater environment, UWSNs exhibit unique characteristics and face more challenges compared with terrestrial wireless sensor networks (TWSNs) [6], [11], [12]. First, given the significant attenuation of radio and optical waves in underwater environment, acoustic waves are the most common communication method in underwater environment currently. However, acoustic communication is bandwidth-limited, and the propagation speed is five orders of magnitude

Manuscript received 23 September 2023; revised 17 December 2023; accepted 19 December 2023. Date of publication 29 December 2023; date of current version 29 February 2024. This work was supported by the National Natural Science Foundation of China under Grant 62031021. The associate editor coordinating the review of this article and approving it for publication was Dr. Vinay Chakravarthi Gogineni. *(Corresponding author: Xiaohong Shen.)*

Chao Wang, Xiaohong Shen, Weiliang Xie, Hongwei Zhang, and Haodi Mei are with the School of Marine Science and Technology and the Key Laboratory of Ocean Acoustics and Sensing, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: whc_wdc@sina.cn; xhshen@nwpu.edu.cn; weiliang@mail.nwpu.edu.cn; zhanghongwei@mail.nwpu.edu.cn; hdmei@mail.nwpu.edu.cn).

Haiyan Wang is with the School of Marine Science and Technology and the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710072, China (e-mail: hywang@sust.edu.cn).

Digital Object Identifier 10.1109/JSEN.2023.3345947

slower than other waves. Second, due to water currents, nodes in UWSNs move passively and irregularly. In addition, considering the influence of time-varying channels, the topology of UWSNs dynamically changes in both space and time. Third, considering that the battery of nodes is difficult to recharge or replace in underwater environment, the energy supply of UWSNs is also limited. Therefore, TWSN routing protocols cannot be directly applied to UWSNs.

In addition, considering that the collection of global information incurs a significant cost due to the limited communication range, high transmission power, and challenges in the underwater environment, centralized routing decisions are not practical in UWSNs [13]. Although distributed routing can make decisions based on local information, traditional heuristic algorithms are difficult to make the global optimal routing decisions due to their reliance on fixed control rules. By taking long-term rewards into account, reinforcement learning (RL) can learn from the environment to maximize the reward of policies from a global perspective. Therefore, RL-based frameworks have been widely applied to distributed routing decisions to obtain the global optimal transmission path [14], [15]. Moreover, multi-agent RL (MARL) as an extension of single-agent RL, allows multiple agents to collaborate through local information processing and mutual information exchange to achieve network goals [16], [17]. Although MARL requires more communication overhead and energy consumption than single-agent RL, the information exchange enables networks to rapidly adapt to dynamic topology and effectively accelerates the discovery of the global optimal transmission path, which saves significant energy consumption.

In recent years, research on underwater routing has mainly focused on the design of transmission paths to optimize energy efficiency [18], end-to-end delay [12], and other factors. Although some papers have considered the differences in data [19], these studies do not sufficiently consider the precise transmission requirements of different importance degrees. In this article, an MARL-based routing protocol with value of information (MARV) is proposed to improve the timeliness and energy efficiency of transmission in UWSNs according to different transmission requirements. MARV is an opportunistic-based routing that exploits the broadcasting characteristics of underwater acoustics. Considering that abnormal/normal data have different importance degrees for the sink, we define the value of information (VoI) to reflect the transmission requirements of data. Meanwhile, an MARL-based framework is introduced into the protocol, which allows UWSNs to make the global optimal routing decision by iteration. To ensure efficient transmission of sensed data, a reward function that considers the VoI of data to meet different transmission requirements is designed under the MARL-based framework. Moreover, a packet holding mechanism is designed to schedule an appropriate waiting time by considering the VoI of data.

The main contributions of this article are summarized as follows.

1) The VoI is introduced to indicate the transmission requirements of different data. According to the actual situation in different scenarios, the importance degree of data is related to the setting of corresponding threshold ranges. As the transmission requirements vary for data with different importance degrees, we map the importance degree into VoI to distinguish these differences, which reflects the requirement for the real-time characteristic of data. The VoI is fundamental for the MARL-based framework and the protocol.
2) An MARL-based framework is proposed for efficient routing. The routing decision process is conducted under an MARL-based framework, and the reward function of MARL is elaborately designed. By considering the factors of depth, residual energy, congestion, and the VoI of different data, the reward function enables the protocol to forward data with high VoI quickly and reliably, thereby ensuring timely processing in the sink. Conversely, data with low VoI are forwarded energy efficient to prolong the lifetime of UWSNs. The framework enables UWSNs to forward packets via appropriate transmission paths according to VoI by learning from the environment and interacting with neighbors.
3) A packet holding mechanism is designed to improve transmission efficiency. By correlating the holding time with the sequence number of nodes in the priority list, the holding mechanism properly schedules the holding time of different nodes, which enables the optimal relay node to forward packets preferentially. Moreover, by considering different transmission requirements, we design a variable holding interval to reduce transmission delay for data with high VoI and improve energy efficiency for data with low VoI.

The rest of this article is organized as follows. In Section II, we provide an overview of recent work on routing protocols for UWSNs. In Section III, we illustrate the network scenario, importance degree model, and MARL model in UWSNs. In Section IV, we describe the proposed MARV protocol in detail. In Section V, we present the simulation results and analysis. Finally, we draw the conclusion in Section VI.

II. Related Work

The improvement of transmission efficiency in UWSNs has been extensively studied in many papers.

In [20], a geographic and cooperative opportunistic routing protocol is proposed. By exploiting geographic information and opportunistic routing, a relay forwarding set can be determined and the best relay node can be selected using a weight calculation scheme. The protocol has a good performance in end-to-end delay and network lifetime. In [21], a distance vector-based opportunistic routing is proposed. The protocol establishes distance vectors for each node and coordinates packet forwarding based on the distance vectors thereby, which has the ability to avoid void regions and long detour to improve packet delivery ratio (PDR) and end-to-end delay. In [22], a depth-based routing for UWSNs is proposed. The protocol introduces depth information to guide the forward direction and exploits opportunistic routing to improve transmission reliability, which achieves high PDRs. In [23], an energy-efficient depth-based routing protocol is proposed. The protocol considers the depth and residual

energy of each node comprehensively for relay selection, which reduces delay and improves network lifetime efficiently. In [24], an energy-efficient probabilistic depth-based routing is proposed. The protocol exploits vertical depth, residual energy, and the number of neighbor nodes to calculate forwarding probability and extends the operation within two-hop neighborhood, which improves PDR and decreases energy consumption. In [25], an energy-efficient guiding-network-based routing is proposed. The protocol introduces a query mechanism into a guiding network and a concurrent working mechanism to forward packets, which reduces network delay while ensuring reliable transmission. In [26], a diagonal and vertical routing protocol is proposed. The protocol relies on the flooding-based technique to increase reliability and adopts the triangular inequality theorem to avoid unnecessary horizontal communication, which realizes scalability and efficiency of end-to-end delay and energy consumption. However, these methods above only transmit data in the same manner and cannot meet the efficient transmission requirements according to the type of data.

There are also some papers that consider the differences between data. Cheng and Li [27] integrated two data gathering mechanisms using AUVs and multihop routing to reduce the problem of unbalanced energy consumption and long delay time, which decreases delay time and prolongs the network lifetime. In [28], a hybrid data collection scheme is proposed. The scheme clusters all nodes into groups based on their location and takes both real-time data collection and energy efficiency issues into consideration for normal and urgent data, which effectively increases timeliness while significantly improving energy efficiency. In [29], a value-based hierarchical information collection system is proposed. The scheme designs a value-energy balanced sink node selection algorithm, which exploits AUVs to collect network information and focuses on path planning. The scheme can prolong the network's lifetime and improve the timeliness. In [30], an efficient opportunistic routing is proposed. By introducing two heuristics for candidate set selection to jointly select the most suitable acoustic modem for data transmission, the protocol aims at both improving data delivery and reducing energy consumption in multimodal UWSNs. The protocol can reduce energy consumption while maintaining data delivery ratio efficiently. In [31], to enable urgent data to be rapidly transmitted compared with other data, the authors design a minimum cost flow model considering the transmission latency, energy balance, and transfer load. Furthermore, a distributed multilevel transmission strategy is proposed. The method improves lifetime and reduces latency efficiently. However, due to the lack of a global perspective, these methods may not be able to find the appropriate transmission paths efficiently.

Moreover, RL as a promising technique has been widely applied to UWSNs routing protocols [14], [32]. In [33], an adaptive, energy-efficient and lifetime-aware routing protocol based on RL is proposed. The protocol applies the Q learning technique to balance the residual energy of each node and learn from the environment effectively, which makes the energy distribution of UWSNs more even and prolongs the lifetime of networks thereby. In [34], a Q learning-based localization-free anypath routing protocol is proposed. The protocol comprehensively considers the energy and depth information of nodes throughout the forwarding process and designs a holding time mechanism to schedule packet transmission, which reduces end-to-end delay and extends network lifetime efficiently. In [35], an RL-based opportunistic routing protocol is proposed. The protocol exploits nodes' peripheral status to select appropriate relay nodes and employs a recovery mechanism to enable packets to bypass void regions efficiently, which has a good performance in end-to-end delay, reliability, and energy efficiency. What is more, RL-based underwater routing protocols can also be used to transmit data with different transmission requirements. In [36], a channel-aware RL-based multipath adaptive routing is proposed. The protocol can switch between single-path and multipath routing guided by a distributed RL framework and jointly optimizes route-long energy consumption and PDR. In [37], an RL-based multimodal communications framework is proposed. By constructing an RL-based framework and defining two classes of soft quality of service, UWSNs can smartly learn the selection of relay nodes to realize reliable and low-latency underwater data delivery. In [38], an energy-efficient multilevel routing strategy based on RL in multimodal UWSNs is proposed. The protocol exploits model knowledge collection to preliminarily learn the network and channel environment before operating and introduces RL to select the appropriate modems and relay nodes for packet transmission. The protocol prolongs the network lifetime and increases the data delivery quantity. However, these methods are difficult to meet the precise transmission requirements according to importance degrees due to the rough categorization of data, and the use of multimodal communication in UWSNs increases the network cost. Meanwhile, these methods neglect the broadcasting characteristics of underwater acoustic signals. The oversight leads nodes to fail to utilize information from neighbor nodes in RL methods and results in a severe waste of information resources.

III. System Model

In this section, we first introduce the network scenario of UWSNs. Then, we give the importance degree model of data and construct the underwater routing process as an MARL model.

A. Network Scenario

In this article, we consider a multihop UWSN for underwater monitoring. The network consists of N sensor nodes and a sink node. Sensor nodes are randomly deployed in a 3-D area to sense and relay data to the sink node. In addition, the sink is deployed on the water surface to gather data and transmit it to a terrestrial data center. The network architecture is shown in Fig. 1.

The underwater communication of nodes uses acoustic channels, while maritime communication uses radio channels. Each node can directly interact with other nodes within its communication range. Obviously, the depth of the sink is 0 and that of each sensor node is larger than 0. The sink on the water surface is easy to maintain and can be regarded as

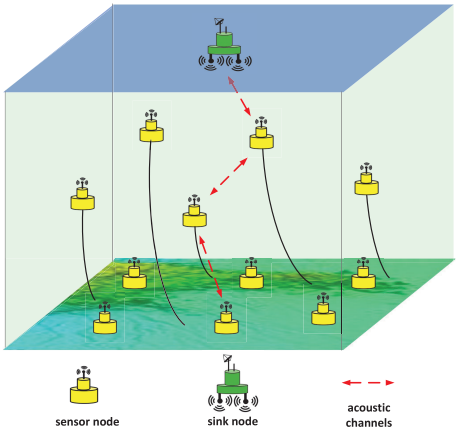


Fig. 1. Network architecture of UWSNs.

energy-sufficient. However, because it is difficult to charge or replace the battery, the energy of underwater nodes is limited. Furthermore, packets will be stored in the buffer of sensor nodes to prepare for transmission in sequence, and considering the irregular mobility of nodes caused by water currents, the UWSN has a dynamic topology.

B. Importance Degree Model

In underwater monitoring scenarios, sensed data have different importance degrees, and the timeliness of data is positively related to the degree. For example, in underwater volcano temperature monitoring, normal temperatures fluctuate around the average temperature and are widely recorded in the historical statistical information of sensor nodes. When a volcano erupts, the temperature of the sensor node near the eruption rises and becomes significantly above the average temperature. Moreover, the larger the difference between the sensed abnormal data and the average data, the higher the probability of a volcano eruption [39]. To deal with disasters effectively, abnormal data are important and time-sensitive, which needs to be transmitted to the sink as soon as possible. However, the transmission of normal data is less important and can take a tradeoff between delay and residual energy into consideration.

Therefore, we introduce the concept of VoI to characterize the importance degrees of data, which can be exploited to indicate the transmission requirements on timeliness or energy efficiency. By setting different threshold ranges for importance degrees according to the actual situation in different scenarios, data can be divided into different VoIs based on numerical values. In fact, the classification method for importance degrees is widely used in many underwater scenarios such as earthquake warnings and pollution monitoring. We assume that there are n importance degrees and the degree of data x is $\phi(x)$. VoI is directly proportional to the importance degree of data, and the VoI can be set as follows:

$$\text{VoI}(x) = \frac{k * \phi(x)}{\phi_n}, \quad \text{s.t. VoI} \in (0, 1] \tag{1}$$

where k is a weighted factor and ϕ_n is the highest importance degree. We consider the data with low VoI as normal data, and the source node tends to select relay nodes by considering energy efficiency. Moreover, a higher VoI indicates that the data are more likely to be important, which needs to be forwarded to the sink promptly. Then, the data with different importance degrees can be forwarded to the sink with different transmission paths based on VoI.

C. MARL Model

In this article, we introduce an MARL method, i.e., distributed value function-distributed RL (DVF-DRL), into the protocol to forward packets. DVF-DRL is a distributed Q learning method for multiple agents that considers the local action and reward of an agent, and the Q value of adjacent agents, rather than merely exploiting the single state–action pair information in Q learning [40], [41]. For underwater acoustic communication, nodes can overhear the packets from neighbor nodes due to the broadcasting characteristics, which allow nodes to obtain the routing information required by DVF-DRL. Moreover, the high transmission cost due to long propagation delay, low bandwidth, and limited energy in UWSNs requires nodes to make decisions without a central decision-maker as efficiently as possible. Therefore, we construct the UWSNs routing protocol under an MARL-based framework.

The underwater routing process can be modeled as a Markov decision process (MDP). An MDP is usually composed of a quadruple $\langle S, A, P, R\rangle$, and the definition of each component is given as follows.

1) *State:* S denotes a finite state set. We define all nodes in the networks as a state set, and node i holding a packet is expressed as state s_i. When node i forwards a packet to node j successfully, the state transfers from s_i to s_j.
2) *Action:* A denotes a finite action set. In our framework, we define the neighbor node set of node i as action set A_i. When node j is selected as a relay node to forward packets, the action is expressed as a_j.
3) *Transition Probability:* P denotes a state transition probability. We define $P_{s_i s_j}^{a_j}$ as the probability that node i takes action a_j and transfers from state s_i to state s_j successfully. Moreover, $P_{s_i s_i}^{a_j} = 1 - P_{s_i s_j}^{a_j}(s_j \neq s_i)$ is the failure probability.
   Here, we exploit the number of packets during a period to estimate the state transition probability between nodes. Specifically, let the number of packets sent from node i to node j be n_{ij} and the number of corresponding packets successfully received by node j be m_{ij}; then, the state transition probability can be calculated as follows:
$$P_{s_i s_j}^{a_j} = \frac{m_{ij}}{n_{ij}}. \tag{2}$$
4) *Reward:* R denotes the immediate reward and is modeled as a reward function. We define $R_{s_i s_j}^{a_j}$ as the reward that node i takes action a_j and transfers from state s_i to state s_j. The reward function is designed to reflect the specific task goal and needs to take the network requirement into consideration comprehensively. The detailed design is shown in Section IV-B.

In addition, the direct reward r_t when an agent takes an action from state s^t at time t is defined as follows:

$$r_t = \sum_{s^{t+1} \in S} P^{a^t}_{s^t s^{t+1}} R^{a^t}_{s^t s^{t+1}}. \tag{3}$$

The expected cumulative reward $V(s^t)$ of a state under a policy can be calculated as follows:

$$V\left(s^t\right) = \mathbb{E}\left[r_t + \sum_{u=1}^{\infty} \gamma^u r_{t+u}\right] \tag{4}$$

where $\gamma \in [0, 1]$ is the discount factor used to adjust the influence of the current reward and the future reward. Obviously, the expected cumulative reward considers long-term reward more than the direct reward, which endows a global perspective for policy decisions. The optimal value $V^*(s^t)$ of a state can be obtained as follows:

$$\begin{aligned} V^*\left(s^t\right) &= \max V\left(s^t\right) \\ &= \max \mathbb{E}\left[r_t + \sum_{u=1}^{\infty} \gamma^u r_{t+u}\right] \\ &= \max_a \left[r_t + \gamma \sum_{s^{t+1} \in S} P^a_{s^t s^{t+1}} V^*\left(s^{t+1}\right)\right]. \end{aligned} \tag{5}$$

According to Bellman's principle of optimality, the optimal solution in each state constitutes the optimal policy collectively, i.e., the optimal policy can be obtained by selecting $V^*(s^t)$ [42], [43].

However, by considering the scale and computing power, it is difficult to obtain full information in a large and distributed system. As a temporal difference RL technique, Q learning enables systems to approximate the optimal policy recursively. The state–action pairs $Q(s^t, a^t)$ in Q learning are defined as the value when taking action a^t in state s^t under a policy and can be expressed as follows:

$$\begin{aligned} Q\left(s^t, a^t\right) &= \mathbb{E}\left[r_t + \sum_{u=1}^{\infty} \gamma^u r_{t+u}\right] \\ &= r_t + \gamma \sum_{s^{t+1} \in S} P^{a^t}_{s^t s^{t+1}} V\left(s^{t+1}\right) \end{aligned} \tag{6}$$

and the optimal value $Q^*(s^t, a^t)$ can be expressed as follows:

$$Q^*\left(s^t, a^t\right) = r_t + \gamma \sum_{s^{t+1} \in S} P^{a^t}_{s^t s^{t+1}} V^*\left(s^{t+1}\right). \tag{7}$$

Moreover, the Q value in DVF-DRL can be approximated recursively as follows:

$$\begin{aligned} &Q\left(s^t, a^t\right) \\ &\leftarrow (1-\alpha) * Q\left(s^t, a^t\right) + \alpha \\ &\quad * \left(r_t + \gamma * \lambda_1 * V\left(s^{t+1}\right) + \gamma * \lambda_2 * \sum_{s' \in \Omega(s^t)} V\left(s'\right)\right) \end{aligned} \tag{8}$$

where $\alpha \in (0, 1]$ is the learning rate used to adjust the rate of updating Q values, $\Omega(s^t)$ is the state transition set of the current state s^t, and λ_1 and λ_2 are the weight factors of the future reward for the next state s^{t+1} and the other state s' in $\Omega(s^t)$, respectively. Also, the optimal policy can be derived through iterative Q learning to select the optimal state.

In the scenario of routing decisions, transmission paths can be regarded as policies, and the optimal transmission path can be achieved by selecting the optimal relay node for transmission. It is worth noting that routing decisions are related to the design of the reward function, and different functions may lead to different optimal transmission paths, i.e., the paths are determined by the reward function designed for specific task goals. In addition, the environment also affects routing decisions by changing the agent's actions. With the change of environment, the optimal transmission path under a specific reward function will also vary. However, although transmission paths may vary with different task goals and environment changes, they are all explored and exploited based on the same RL principles and make optimal routing decisions accordingly.

Therefore, by constructing the underwater routing process as an MARL-based framework, the global optimal transmission path in accordance with network requirements can be theoretically achieved through learning from the environment and collaborating with neighbor nodes.

IV. MARV Protocol

In this section, we first give an overview of the proposed protocol and design the reward function of MARL by considering network requirements. Then, the detailed protocol design is proposed.

A. Overview

In this article, the MARV protocol is proposed to transmit packets from source nodes to the sink via appropriate transmission paths. By designing the protocol under an MARL-based framework, MARV learns from the environment continuously and allows fast adaptation to underwater dynamic topology. In the packet forwarding process, the VoI of data is determined at source nodes according to the importance degree model first. Moreover, by embedding the information of VoI into packets, nodes can select appropriate relay nodes under the MARL-based framework to forward packets via different transmission paths efficiently. Specifically, the protocol enables UWSNs to forward data with high VoI via quick and reliable transmission paths to ensure the transmission requirement for time sensitivity. The protocol is also able to forward data with low VoI and considers delay and residual energy comprehensively in the selection of transmission paths. The schematic of transmission paths under two conditions with different types of VoI data is shown in Fig. 2.

In this article, the relay nodes participating in the routing process are selected by senders. In addition, to reduce the unnecessary transmission paths caused by the broadcasting characteristics of underwater acoustics, only the neighbor nodes whose depth is smaller than the sender's are qualified to forward packets. To select the optimal relay node, the sender calculates the Q value of all its qualified neighbors before packet forwarding. Obviously, the node with a larger Q value is more suitable for relaying packets. By considering

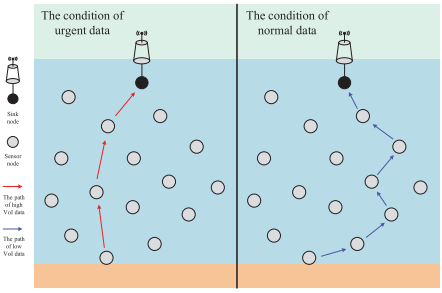


Fig. 2. Schematic of transmission paths under two conditions with different types of VoI data.

the transmission requirement for data with high VoI, depth and congestion information of nodes are considered in Q value. Moreover, by considering data with low VoI and limited energy of UWSNs, residual energy of nodes should also be considered.

When the sender calculates the Q value of the qualified neighbor nodes, it sorts these nodes in a priority list according to the descending order of Q values. Then, the ID of each neighbor node is included in the priority list and embedded into the packet header to forward. According to the proposed packet holding mechanism, the node with a higher priority has a shorter holding time, which ensures the forwarding priority of the relay node with a high Q value. Moreover, the holding interval is correlated with the VoI of data, which adjusts the transmission delay and energy efficiency of different types of VoI data. In addition, the proposed transmission suppression mechanism reduces the redundant packets and unnecessary energy consumption during transmission.

B. Reward Function

Considering that the immediate reward obtained from the reward function will directly affect the calculation of Q value and that the node with a higher Q value should forward packets earlier, the design of the reward function is very important for the performance of routing protocols.

To ensure that the proposed routing can forward packets with different types of VoI efficiently, the depth, residual energy, and congestion of nodes are considered in the reward function. Then, the reward function for a packet transmitted from node i to node j can be defined as follows:

$$R_{s_i s_j}^{a_j} = -g_c - (1-\omega-\mathrm{VoI}) * \left[E(i)+E(j)\right] - (\omega+\mathrm{VoI}) * \left[D(i,j)+\delta * C(j)\right] \quad (9)$$

where g_c is the constant cost for packet transmission, ω is the weight coefficient for balancing, $E(i)$ and $E(j)$ are residual energy-related cost functions for sending node i and receiving node j, respectively, $D(j,i)$ is a depth-related cost function for receiving node j and sending node i, $C(j)$ is a congestion-related cost function for receiving node j, and δ is the congestion parameter to adjust the impact of congestion on relay selection.

Due to the existence of constant cost g_c, nodes prefer to choose the shortest transmission path to the sink to minimize relay times and delay. However, the path may not be the best path. This is because the residual energy and network congestion as well as the transmission requirement of a packet will affect the selection. Therefore, three cost functions are introduced as follows.

The residual energy-related cost function $E(x)$ is defined as the residual energy ratio of node x and can be expressed as follows:

$$E(x) = 1 - \frac{e_{\mathrm{res}}(x)}{e_{\mathrm{init}}(x)} \quad (10)$$

where $e_{\mathrm{res}}(\cdot)$ is the residual energy of a node and $e_{\mathrm{init}}(\cdot)$ is the initial energy. The design of the function encourages sending nodes to select the neighbor node with a large residual energy to forward.

The depth-related cost function $D(y,x)$ is defined as the depth difference between receiving node y and sending node x and can be expressed as follows:

$$D(y,x) = 1 - \frac{d(y)-d(x)}{R} \quad (11)$$

where $d(\cdot)$ is the depth of a node and R is the maximum communication range of a node. The design of the function makes sending nodes tend to select the neighbor node with a small depth to forward.

Due to the fact that a packet cannot be forwarded immediately if there are other earlier packets in the buffer waiting for transmission, the selection of relay nodes needs to take the congestion condition into consideration. The congestion-related cost function $C(y)$ is defined as the congestion ratio of neighbor node y of node x and can be expressed as follows:

$$C(y) = \frac{\mathrm{cg}(y)}{\sum_{z\in\Omega(x)} \mathrm{cg}(z)+1} \quad (12)$$

where $\mathrm{cg}(\cdot)$ is the number of packets in the buffer of a node and $\Omega(\cdot)$ is the set of neighbor nodes. It is worth noting that the coefficient 1 is used to prevent the denominator from being 0. By introducing the function into the reward function, sending nodes are more likely to choose the neighbor node without congestion to forward.

Moreover, considering the transmission requirement for different data, VoI is exploited to adjust the importance of delay and network lifetime. For data with high VoI, the weight of depth and congestion increases, while the weight of residual energy decreases, which brings the node with smaller depth and less congestion a large Q value and compels the system to select the node to forward packets preferentially thereby. Meanwhile, for data with low VoI, there is an opposite tendency for the weight of depth, congestion, and residual energy, and the node with large residual energy is more likely to be selected to forward packets.

It is worth noting that the communication link quality has been considered in the transition probability of the MARL-based framework, which gives the node with high link quality a large reward according to (3) and ensures the reliable transmission for packets. Then, by using the reward function and the transition probability, the corresponding Q value can be calculated and used to make routing decisions.

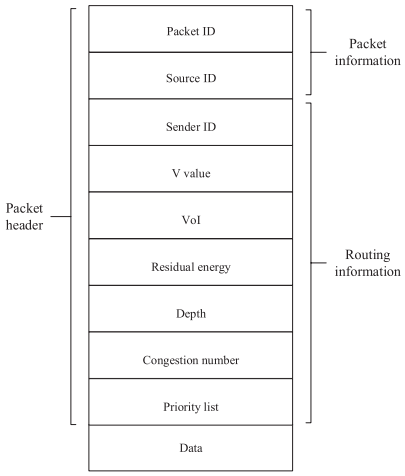


Fig. 3. Structure of packet.

C. Protocol Design

In this section, we introduce the proposed protocol from four parts, including the packet structure, the information exchange method, the packet holding mechanism, and the transmission suppression mechanism.

1) Packet Structure: The goal of a routing protocol is to enable data forward to the sink successfully. The packet structure used in the MARV protocol is shown in Fig. 3. The packet is composed of a header and a payload, and the definition of each field in the header is as follows.

1) *Packet ID:* The unique ID of a packet.
2) *Source ID:* The unique ID of the source node.
3) *Sender ID:* The unique ID of the current node.
4) *V Value:* The V value of the current node.
5) *VoI:* The VoI of data.
6) *Residual Energy:* The residual energy of the current node.
7) *Depth:* The depth information of the current node.
8) *Congestion Number:* The number of the packet waiting for transmission in the buffer of the current node.
9) *Priority List:* The ordering of node ID based on forwarding priority.

After receiving a packet, each node extracts the corresponding value and stores it into local information. Then, if the node ID is included in the priority list, the node updates the corresponding value in the packet and waits for forwarding; otherwise, the node drops the packet directly. The detailed description of the process is illustrated in the information exchange method, packet holding mechanism, and transmission suppression mechanism as follows.

2) Information Exchange Method: In the MARL-based framework, the information of neighbors is required for a node to make routing decisions. Therefore, each node needs to broadcast its local information to neighbors. Meanwhile, nodes that receive this information will update the corresponding data recorded by themselves, regardless of whether they are qualified to forward.

There are two methods to realize the process of information exchange. On the one hand, each node attaches its local information to the packet header before transmitting. Due to the broadcasting characteristics of UWSNs, all neighbors of a sender can obtain this information from the process of packet transmission. On the other hand, each node broadcasts a control packet, including its local information periodically. This method ensures that all nodes can obtain the information of their neighbors even if they do not have packets to forward and can be regarded as a supplement of information exchange in packet transmission.

3) Packet Holding Mechanism: In this article, the neighbor node with the largest Q value is defined as the optimal relay node. To ensure the priority forwarding of the optimal relay node and the reliable transmission of packets simultaneously, we design a packet holding mechanism and enable the node included in the priority list of a packet to forward the packet at different holding times. In addition, the mechanism ensures efficient forwarding for different types of VoI data.

Specifically, by taking the sequence number in the priority list into consideration, the mechanism ensures that the node with a smaller sequence number can forward earlier. Moreover, VoI is also considered in the holding time to ensure that data with high VoI can be forwarded to the sink earlier. Assuming that the sequence number of node i in the priority list of a packet is n, the holding time can be defined as follows:

$$T_i(n) = \frac{k * T_{\max}}{h * \mathrm{VoI}}(n-1) \tag{13}$$

where h is a global parameter for holding time calculation. $T_{\max}$ is the predefined maximum holding interval between two adjacent nodes in the priority list and can be defined as follows:

$$T_{\max} = \frac{2R}{v} \tag{14}$$

where v is an underwater acoustic velocity. $T_{\max}$ is defined as twice the maximal propagation delay R/v, which is long enough to enable all lower priority neighbors of node i in the priority list to overhear the packet before forwarding.

In (13), by setting $n-1$, the node with the highest priority, i.e., the first node in the priority list, can forward a packet without waiting and avoid unnecessary long delay. In addition, by introducing k/VoI, the holding interval decreases with the data's VoI increase, which ensures transmission requirements for different data efficiently.

Specifically, on the one hand, the packet with low VoI data has a long holding interval, which enables the neighbor nodes of a relay node to have a higher probability to overhear the packet before forwarding. Although a long holding interval inevitably increases transmission delay, energy consumption can be effectively reduced according to the transmission suppression mechanism given next, which improves the energy efficiency of UWSNs. On the other hand, the packet with high VoI data has a short holding interval, which enables the node in the priority list to forward the packet as early as possible. The short holding interval can decrease the transmission delay but increase the energy consumption inevitably. However, considering the transmission requirement, a short holding interval is necessary.

By exploiting the packet holding mechanism, each node included in the priority list of a packet can calculate a suitable

Algorithm 1 Procedure of Packet Forwarding in MARV

$Ne(i)$: The neighbors of node i
$U(i)$: The neighbor nodes qualified to forward packets
Input: The residual energy, depth and V value of node i, the priority list for packet forwarding and the packet ID
Output: The holding time of node i and updated Q value of state-action pair $Q(s_i, a_j)$

```
Update the local information of node i
Check the priority list
if The ID of node i is included in the list then
    Check the Packet's ID
    if Packet's ID has not been recorded then
        for each node j ∈ Ne(i) do
            if d_j < d_i then
                U(i) = U(i) ∪ j
            end if
        end for
        for each node j ∈ U(i) do
            Calculate and update Q(s_i, a_j) using Eq. (8)
        end for
        Set the priority list in the packet according to their Q value
        Calculate the holding time of node i using Eq. (13), and wait for transmitting the packet
        if Holding time is not expired then
            Overhear packets
            if The packet has been forwarded then
                Discard the packet
            end if
        else
            Forward the packet
        end if
    else
        Discard the packet
    end if
else
    Discard the packet
end if
```

holding time. After the holding time expires, nodes will forward the packet immediately.

4) Transmission Suppression Mechanism: In MARV, the node included in the priority list of a packet will forward the packet in sequence with different holding times. However, the redundant transmission for the same packet leads to unnecessary energy consumption and decreases the energy efficiency of UWSNs. Therefore, a mechanism should be designed to suppress redundant transmission.

Based on the broadcasting characteristics of UWSNs, the packet can be overheard by all neighbor nodes of a sender. When overhearing a packet, a node will check whether it is included in the priority list or not first. If not, the node discards the packet directly; otherwise, the node checks the packet ID. In that case, if the packet has been forwarded before, the node discards the packet; otherwise, the node calculates holding time based on the packet holding mechanism and waits for forwarding. Moreover, if the node overhears the forwarding of the same packet during the holding time, the node discards the packet; otherwise, the node will forward the packet when the holding time expires.

By summarizing the various components of the protocol designed above, we present the pseudocode of packet forwarding, as shown in Algorithm 1.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed protocol MARV. First, we give the simulation configurations and the investigated metrics. Then, we present a parameter analysis to show the influence of different parameters on the proposed protocol. Finally, we compare MARV with three other opportunistic routing protocols DBR [22], EEDBR [23], and QLFR [34] to evaluate the performance.

A. Simulation Configurations

In our simulation experiment, a 4000 × 4000 × 4000 m 3-D underwater area is deployed and nodes are randomly distributed. All nodes are equipped with the same acoustic modem and have the same features, such as transmitting/receiving power, transmission rate, and communication range. All types of VoI data are forwarded to the sink using relay nodes via different transmission paths.

Due to the expensive cost of underwater nodes [44], we assume that there is one sink deployed in the UWSN, and the number of nodes changes from 40 to 90. The transmitting and the receiving power are set to be 3 and 0.8 W, respectively; the communication range is 700 m; and the transmission rate is 1500 b/s. The underwater acoustic velocity is set to be 1500 m/s.

We employ a normal distribution to describe the distribution of the historical statistical information of sensor nodes [27]. Let $\mu = 0$ and $\sigma = 1$ denote the mean and standard deviation of a normal distribution, respectively. We define four importance degrees and set $k = 0.4$. Then, we calculate the VoI of data x as follows:

$$\mathrm{VoI}(x) = \begin{cases} 0.1, & 0 < |x-\mu| < \sigma \\ 0.1, & \sigma < |x-\mu| < 2\sigma \\ 0.3, & 2\sigma < |x-\mu| < 3\sigma \\ 0.4, & \text{otherwise.} \end{cases}$$

Moreover, in the parameter configuration of the MARL model, we set the learning rate to 0.9, the discount rate to 0.5, the weight coefficient to 0.2, the congestion parameter to 1, and the constant cost to 1. The detailed parameters settings required for the simulations are shown in Table I.

B. Investigated Metrics

In this article, we use average end-to-end delay, PDR, and energy consumption per packet as the metrics to assess routing protocol performance.

1) Average End-to-End Delay: It is defined as the average time that a source node successfully transmits a packet to the sink and can be expressed as follows:

$$\hat{t} = \frac{\sum_{l=1}^{m} t_l}{m} \tag{15}$$

TABLE I
SIMULATION CONFIGURATIONS

Parameters	Values
Deployment area	4km × 4km × 4km
Number of deployed nodes	40 ∼ 90
Number of sink node	1
Transmitting power	3 W
Receiving power	0.8 W
Communication range	700 m
Packet size	20 Bytes
Transmission rate	1500 bps
Signal frequency	10 KHz
Processing delay	0.05 s
underwater acoustic velocity v	1500m/s
Learning rate α	0.9
Discount factor γ	0.5
Weight coefficient ω	0.2
Congestion parameter δ	1
Constant cost g_c	1
VoI weight factor k	0.4

where m denotes the number of successfully received nonduplicate packets at the sink and t_l denotes the end-to-end delay of the lth packet. Specifically, it is comprised of the following four parts: *transmission delay*, which is defined as the time a sender spends transmitting a packet to the channel; *propagation delay*, which is defined as the time a packet spends forwarding from senders to receivers; *queuing delay*, which is defined as the time a packet spends waiting in transmission queues due to congestion; and *retention delay*, which is defined as the time a sender spends keeping a packet before the holding time expires.

2) Packet Delivery Ratio: It is defined as the ratio of nonduplicate packets successfully received by the sink to packets transmitted by source nodes and can be expressed as follows:

$$\text{PDR} = \frac{m}{n} \tag{16}$$

where n denotes the number of transmitted packets at source nodes.

3) Energy Consumption per Packet: It is defined as the network energy consumption required for source nodes to transmit a packet to the sink successfully and can be expressed as follows:

$$\hat{e} = \frac{\sum_{l=1}^{m} e_l}{m} \tag{17}$$

where e_l denotes the network energy consumption for successfully transmitting the lth packet.

C. Parameter Analysis

1) Date VoI: In the practical process of underwater data collection, the VoI changes depending on the importance degree of the data at source nodes. Therefore, to assess the performance of MARV under different types of VoI, we model the value of sensing data as a standard normal distribution in Section V-A, and the VoI is divided into four levels $\{0.1, 0.2, 0.3, 0.4\}$. Different types of VoI data will be generated at source nodes and transmitted to the sink during the data collection period.

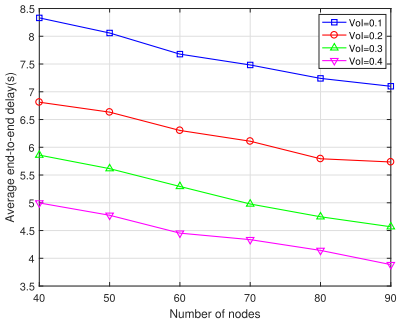


Fig. 4. Comparisons of average end-to-end delay with different types of VoI.

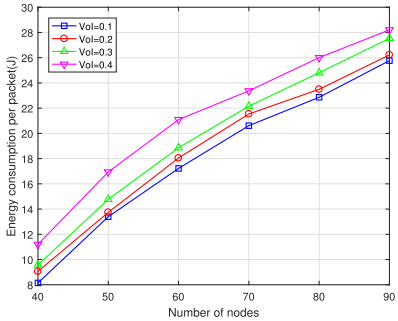


Fig. 5. Comparisons of energy consumption per packet with different types of VoI.

Fig. 4 shows that the end-to-end delay of increasing VoI decreases gradually. This is because, as VoI increases, the holding interval of relay nodes decreases because of the designed packet holding mechanism. The shortened interval reduces the holding time of each relay node and decreases the end-to-end delay thereby. Moreover, the end-to-end delay also decreases gradually with the node number increasing. This is because, as the network deployment changes from sparse to dense, more appropriate relay node for packet forwarding appears and end-to-end delay can be further reduced.

Fig. 5 shows that the energy consumption per packet of increasing VoI increases gradually. This is because, as VoI increases, the shortened holding interval cannot effectively suppress the transmission for redundant packets due to the designed transmission suppression mechanism, thereby increasing network energy consumption. Moreover, as the number of nodes increases, the energy consumption per packet increases correspondingly. This is because more nodes can overhear packets, which leads to more energy consumption during packet sending and receiving.

Fig. 6 shows that the PDR is similar under different types of VoI. On the one hand, the characteristics of opportunistic routing enable multiple neighbors to receive and cooperate in forwarding packets, thereby improving the reliability of transmission. On the other hand, the MARL-based framework enables each node to learn from the environment according to VoI and select appropriate transmission paths. Moreover, the PDR also increases as the number of deployed nodes increases. This is because more qualified nodes can participate

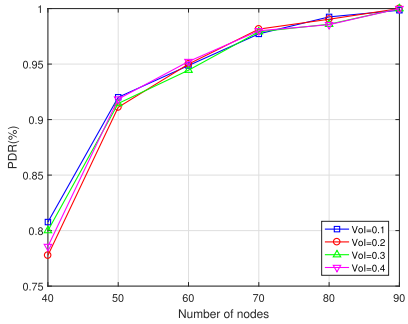


Fig. 6. Comparisons of PDR with different types of Vol.

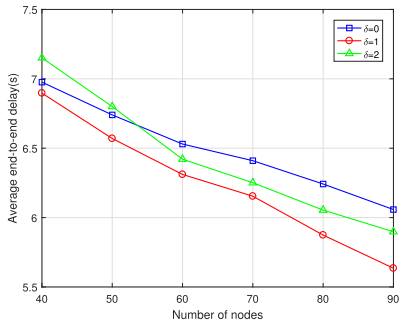


Fig. 7. Average end-to-end delay with different congestion parameters.

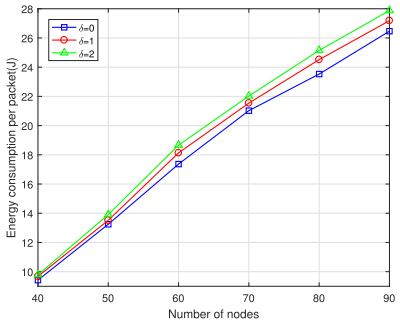


Fig. 8. Energy consumption per packet with different congestion parameters.

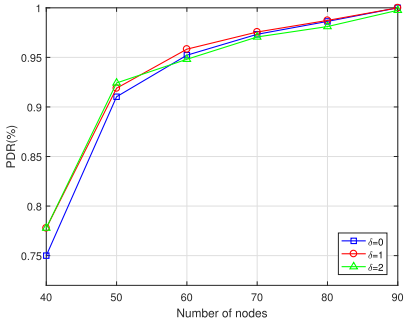


Fig. 9. PDR with different congestion parameters.

in forwarding packets and the number of void regions reduced, and a packet loss will be decreased.

2) Congestion Parameter: To assess the influence of congestion parameter δ on UWSNs, we compare the performance of MARV under $\delta = \{0, 1, 2\}$. Moreover, to effectively reveal the impact of parameter changes on network performance, we set the VoI of all data to 0.2 during the packet collection period.

Fig. 7 shows that UWSNs have the lowest end-to-end delay when $\delta = 1$; the delay of $\delta = 0$ is lower than $\delta = 2$ when the node number is small and gradually surpasses with the increase of the node number. This is because when the node number is small, an oversized congestion parameter, i.e., $\delta = 2$, will lead the network to excessively select nodes with less congestion to forward, which requires a long detour due to the lack of appropriate relays and results in a large delay. Meanwhile, ignoring the influence of congestion, i.e., $\delta = 0$, will lead the network to continuously select the same node to forward, which requires queuing for transmission and results in a large delay. Therefore, there exists a tradeoff between the delay due to queuing and the delay due to detour. Moreover, more appropriate relay nodes appear with the increase of the node number, which reduces the delay due to queuing and is lower than the delay due to detour. At this time, the delay of $\delta = 2$ is lower than $\delta = 0$.

Fig. 8 shows that the energy consumption per packet of different congestion parameters increases with δ increasing. This is because when $\delta = 0$, nodes will select the direction with the shortest path for transmission without considering congestion. Due to not bypassing congested nodes, fewer hops are required for packet forwarding and lower energy is consumed. However, when $\delta = 1$, nodes tend to select other paths for transmission if congestion occurs, which causes a detour to some extent and additional energy consumption. Similarly, when δ is oversized, such as $\delta = 2$, nodes will be oversensitive to congestion, which leads to unnecessary detours and energy consumption.

Fig. 9 shows that the PDR of different congestion parameters is similar with δ increasing. This is because of the broadcasting characteristics of opportunistic routing, which make the change of forwarding path caused by congestion parameter not affect the reliability of data transmission.

Moreover, end-to-end delay, energy consumption per packet, and PDR all improve with the node number increasing due to the appearance of appropriate paths and the disappearance of void regions.

3) Weighted Factor: To assess the influence of weighted factor λ_1 for the next state and λ_2 for other states in (8) on network performances, we give the simulation under different (λ_1, λ_2) at $(0.2, 0.01)$, $(0.5, 0.01)$, and $(0.2, 0.1)$. For the same reason, we set the VoI of all data to 0.2 during the packet collection period in the simulation.

As can be seen from Figs. 10 to 12, the selection of different values for λ_1 and λ_2 will affect the transmission paths directly. Figs. 10 and 11 show that the end-to-end delay and energy consumption per packet have better performance than others when $\lambda_1 = 0.2$ and $\lambda_2 = 0.01$, respectively. This can be explained by the fact that a larger λ_1 will overshadow the influence of λ_2 when $\lambda_1 = 0.5$ and $\lambda_2 = 0.01$, which results in nodes ignoring the information of neighbor nodes and being unable to select the better transmission path. Similarly, a larger λ_2 will unduly influence on λ_1 when $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$, which results in nodes overvaluing the information of neighbor nodes and selecting inappropriate relay nodes. Fig. 12 shows that PDR is similar with under different (λ_1, λ_2). This can be explained by the multipath transmission in opportunistic routing.

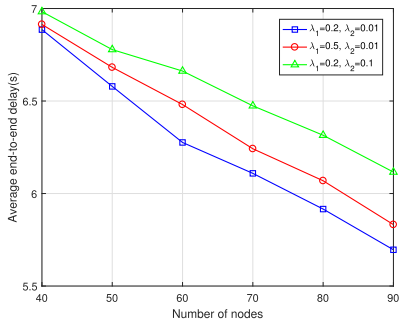


Fig. 10. Average end-to-end delay with different λ_1 and λ_2 pairs.

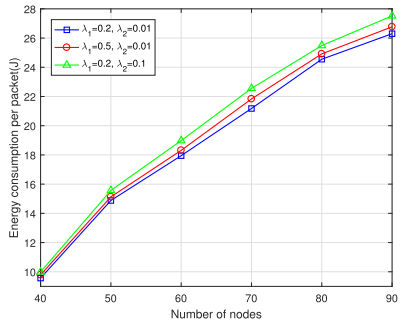


Fig. 11. Energy consumption per packet with different λ_1 and λ_2 pairs.

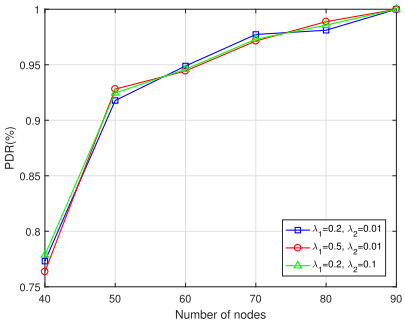


Fig. 12. PDR with different λ_1 and λ_2 pairs.

Moreover, the network performances all have improved due to the appearance of better relay nodes with the increase of the node number.

D. Performance Comparison

In this section, we compare the performance of DBR, EEDBR, QLFR, and MARV. The VoI of data is set to 0.2 during the packet collection period.

Fig. 13 shows that the average end-to-end delay of the four protocols decreases gradually. This is because the neighbors of a node increase with the number of deployed nodes increasing, and more relay nodes can be used to forward packets thereby. Among the four protocols, MARV has the lowest end-to-end delay and DBR has the highest compared with the others. This is because DBR and EEDBR exploit a greedy strategy to forward packets without a global perspective, which leads to a detour and high delay. Furthermore, the value-based holding mechanism in both protocols also results in an unnecessary waiting time before transmitting packets. However, both MARV and QLFR are RL-based routing protocols. With the help of RL-based frameworks, MARV and QLFR are able to

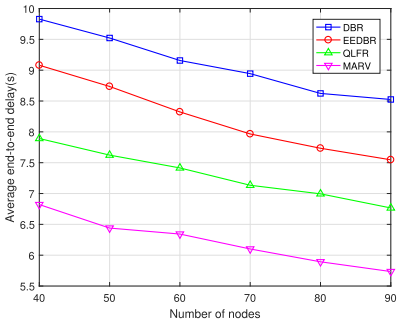


Fig. 13. Comparisons of average end-to-end delay among different protocols.

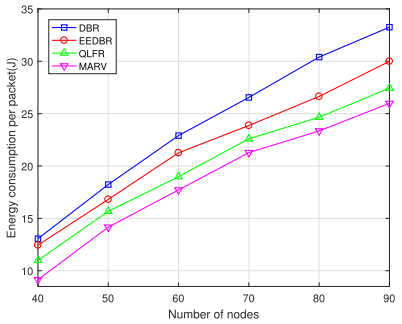


Fig. 14. Comparisons of energy consumption per packet among different protocols.

learn transmission paths by interacting with the environment. The priority-based holding mechanism in both protocols can effectively reduce waiting time and conflicts compared to the value-based mechanism. It is worth noting that MARV further considers the VoI of data and congestion of each node, which reduces end-to-end delay efficiently.

Fig. 14 shows the energy consumption per packet of the four protocols increases gradually. This is because with the node number varying from sparse to dense, the qualified nodes that can forward packets also increase, which consumes more energy while sending and receiving packets. Among the four protocols, MARV has the lowest energy consumption and DBR has the highest compared with the others. This is because the greedy strategy exploited by both DBR and EEDBR consumes additional energy due to the lack of path information for transmission. However, the RL-based framework enables both MARV and QLFR to adaptively learn and select appropriate paths, which reduces the hop number and energy consumption thereby. The priority-based holding mechanism also effectively suppresses the transmission of redundant packets and further reduces unnecessary energy consumption. Moreover, due to the MARL method in MARV, the proposed protocol is able to obtain environmental information quickly and comprehensively, which helps the network make routing decisions and consume less energy than QLFR.

Fig. 15 shows that the PDR of the four protocols increases gradually. This is because with the node number increasing, more neighbor nodes are able to participate in forwarding packets, and the decreased void regions reduce packet loss. Among the four protocols, MARV has the highest PDR and DBR has the lowest compared with the others when the network is sparse. This is because there exist many void

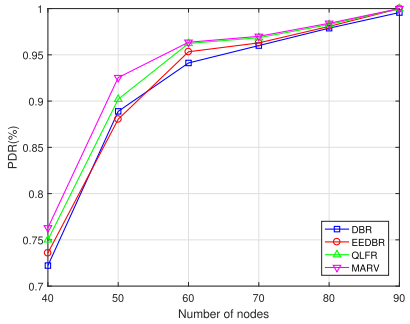


Fig. 15. Comparisons of PDR among different protocols.

regions in this case, which leads to packet loss inevitably. Considering that DBR and EEDBR transmit packets greedily and cannot bypass void regions, both protocols have a relatively lower PDR. In comparison, due to the learning mechanism in MARV and QLFR, both protocols can bypass void regions in subsequent transmission, which improves PDR to some extent. Specifically, MARV has a higher PDR compared with QLFR, and this is because the MARL-based framework can further improve the learning ability of a system than the RL-based one. Moreover, due to the reduction of void regions, the PDR of the four protocols is improved and tends to be similar with the node number increasing.

VI. Conclusion

In this article, an MARL-based routing protocol with VoI is proposed. The protocol can realize different transmission requirements in UWSNs efficiently. In the proposed protocol, by exploiting the threshold range to set the importance degree of data, VoI is introduced to distinguish the difference of data and indicate transmission requirements. On this basis, an MARL-based framework is constructed. The framework enables UWSNs to select appropriate transmission paths by allowing multiple nodes to learn and collaborate. Also, the reward function in the framework takes the factors of depth, residual energy, congestion, and the VoI of different data into consideration, which ensures the timeliness and energy efficiency for the transmission of different types of VoI data. Moreover, to properly schedule the holding time of relay nodes, a packet holding mechanism is designed. The mechanism can adaptively change the holding interval based on transmission requirements. The simulation results show that the protocol can improve the timeliness, PDR, and energy efficiency of different types of VoI data accordingly. The limitation of the work is that we have not discussed the void recovery problem when packets drop in void regions and the packet priority transmission problem when congestion happens in nodes, and this should be addressed in future works.

References

[1] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[2] T. Qiu, Z. Zhao, T. Zhang, C. Chen, and C. L. P. Chen, "Underwater Internet of Things in smart ocean: System architecture and open issues," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4297–4307, Jul. 2020.

[3] C.-C. Kao, Y.-S. Lin, G.-D. Wu, and C.-J. Huang, "A comprehensive study on the Internet of Underwater Things: Applications, challenges, and channel models," *Sensors*, vol. 17, no. 7, p. 1477, Jun. 2017.

[4] E. Felemban, F. K. Shaikh, U. M. Qureshi, A. A. Sheikh, and S. B. Qaisar, "Underwater sensor network applications: A comprehensive survey," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 11, Nov. 2015, Art. no. 896832.

[5] J. Luo, Y. Yang, Z. Wang, and Y. Chen, "Localization algorithm for underwater sensor network: A review," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13126–13144, Sep. 2021.

[6] K. M. Awan, P. A. Shah, K. Iqbal, S. Gillani, W. Ahmad, and Y. Nam, "Underwater wireless sensor networks: A review of recent issues and challenges," *Wireless Commun. Mobile Comput.*, vol. 2019, pp. 1–20, Jan. 2019.

[7] M. Ahmed, M. Salleh, and M. I. Channa, "Routing protocols based on node mobility for underwater wireless sensor network (UWSN): A survey," *J. Netw. Comput. Appl.*, vol. 78, pp. 242–252, Jan. 2017.

[8] N. Li, J.-F. Martínez, J. M. Chaus, and M. Eckert, "A survey on underwater acoustic sensor network routing protocols," *Sensors*, vol. 16, no. 3, p. 414, Mar. 2016.

[9] G. Han, Z. Tang, Y. He, J. Jiang, and J. A. Ansere, "District partition-based data collection algorithm with event dynamic competition in underwater acoustic sensor networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 10, pp. 5755–5764, Oct. 2019.

[10] K. F. Haque, K. H. Kabir, and A. Abdelgawad, "Advancement of routing protocols and applications of underwater wireless sensor network (UWSN)—A survey," *J. Sensor Actuator Netw.*, vol. 9, no. 2, p. 19, Apr. 2020.

[11] K. Chen, M. Ma, E. Cheng, F. Yuan, and W. Su, "A survey on MAC protocols for underwater wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1433–1447, 3rd Quart., 2014.

[12] H. Khan, S. A. Hassan, and H. Jung, "On underwater wireless sensor networks routing protocols: A review," *IEEE Sensors J.*, vol. 20, no. 18, pp. 10371–10386, Sep. 2020.

[13] T. Wang, D. Zhao, S. Cai, W. Jia, and A. Liu, "Bidirectional prediction-based underwater data collection protocol for end-edge-cloud orchestrated system," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4791–4799, Jul. 2020.

[14] R. T. Rodoshi, Y. Song, and W. Choi, "Reinforcement learning-based routing protocol for underwater wireless sensor networks: A comparative survey," *IEEE Access*, vol. 9, pp. 154578–154599, 2021.

[15] R. A. Nazib and S. Moh, "Reinforcement learning-based routing protocols for vehicular ad hoc networks: A comparative survey," *IEEE Access*, vol. 9, pp. 27552–27587, 2021.

[16] L. Canese et al., "Multi-agent reinforcement learning: A review of challenges and applications," *Appl. Sci.*, vol. 11, no. 11, p. 4948, May 2021.

[17] T. Li et al., "Applications of multi-agent reinforcement learning in future Internet: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1240–1279, 2nd Quart., 2022.

[18] S. Khisa and S. Moh, "Survey on recent advancements in energy-efficient routing protocols for underwater wireless sensor networks," *IEEE Access*, vol. 9, pp. 55045–55062, 2021.

[19] F. Alawad and F. A. Kraemer, "Value of information in wireless sensor network applications and the IoT: A review," *IEEE Sensors J.*, vol. 22, no. 10, pp. 9228–9245, May 2022.

[20] S. Karim et al., "GCORP: Geographic and cooperative opportunistic routing protocol for underwater sensor networks," *IEEE Access*, vol. 9, pp. 27650–27667, 2021.

[21] Q. Guan, F. Ji, Y. Liu, H. Yu, and W. Chen, "Distance-vector-based opportunistic routing for underwater acoustic sensor networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3831–3839, Apr. 2019.

[22] H. Yan, Z. J. Shi, and J. H. Cui, "DBR: Depth-based routing for underwater sensor networks," in *NETWORKING 2008 Ad Hoc and Sensor Networks, Wireless Networks, Next Generation Internet*. Berlin, Germany: Springer, 2008.

[23] A. Wahid, S. Lee, H. J. Jeong, and D. Kim, "EEDBR: Energy-efficient depth-based routing protocol for underwater wireless sensor networks," in *Advanced Computer Science and Information Technology*. Berlin, Germany: Springer, 2011.

[24] M. Zhang and W. Cai, "Energy-efficient depth based probabilistic routing within 2-hop neighborhood for underwater sensor networks," *IEEE Sensors Lett.*, vol. 4, no. 6, pp. 1–4, Jun. 2020.

[25] Z. Liu, X. Jin, Y. Yang, K. Ma, and X. Guan, "Energy-efficient Guiding-Network-Based routing for underwater wireless sensor networks," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21702–21711, Nov. 2022.

[26] T. Ali, L. T. Jung, and I. Faye, "Diagonal and vertical routing protocol for underwater wireless sensor network," *Proc.-Social Behav. Sci.*, vol. 129, pp. 372–379, May 2014.
[27] C.-F. Cheng and L.-H. Li, "Data gathering problem with the data importance consideration in underwater wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 78, pp. 300–312, Jan. 2017.
[28] Z. Liu, X. Meng, Y. Liu, Y. Yang, and Y. Wang, "AUV-aided hybrid data collection scheme based on value of information for Internet of Underwater Things," *IEEE Internet Things J.*, vol. 9, no. 9, pp. 6944–6955, May 2022.
[29] R. Duan, J. Du, C. Jiang, and Y. Ren, "Value-based hierarchical information collection for AUV-enabled Internet of Underwater Things," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9870–9883, Oct. 2020.
[30] R. W. L. Coutinho and A. Boukerche, "OMUS: Efficient opportunistic routing in multi-modal underwater sensor networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5642–5655, Sep. 2021.
[31] Z. Zhao, C. Liu, W. Qu, and T. Yu, "An energy efficiency multi-level transmission strategy based on underwater multimodal communication in UWSNs," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Jul. 2020, pp. 1579–1587.
[32] J. Luo, Y. Chen, M. Wu, and Y. Yang, "A survey of routing protocols for underwater wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 137–160, 1st Quart., 2021.
[33] T. Hu and Y. Fei, "QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 6, pp. 796–809, Jun. 2010.
[34] Y. Zhou, T. Cao, and W. Xiang, "Anypath routing protocol design via Q-learning for underwater sensor networks," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8173–8190, May 2021.
[35] Y. Zhang, Z. Zhang, L. Chen, and X. Wang, "Reinforcement learning-based opportunistic routing protocol for underwater acoustic sensor networks," *IEEE Trans. Veh. Technol.*, vol. 70, pp. 2756–2770, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233178288
[36] V. Di Valerio, F. Lo Presti, C. Petrioli, L. Picari, D. Spaccini, and S. Basagni, "CARMA: Channel-aware reinforcement learning-based multi-path adaptive routing for underwater wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2634–2647, Nov. 2019. Accessed: Aug. 17, 2022.
[37] S. Basagni, V. D. Valerio, P. Gjanci, and C. Petrioli, "MARLIN-Q: Multi-modal communications for reliable and low-latency underwater data delivery," *Ad Hoc Netw.*, vol. 82, pp. 134–145, Jan. 2019.
[38] Z. Zhao, C. Liu, X. Guang, and K. Li, "MLRS-RL: An energy efficient multi-level routing strategy based on reinforcement learning in multi-modal UWSNs," *IEEE Internet Things J.*, to be published.
[39] *Earthquake Hazards Program*. [Online]. Available: https://www.usgs.gov/programs/earthquake-hazards
[40] X. Liang, I. Balasingham, and S.-S. Byun, "A multi-agent reinforcement learning based routing protocol for wireless sensor networks," in *Proc. IEEE Int. Symp. Wireless Commun. Syst.*, Oct. 2008, pp. 552–557.
[41] X. Li, X. Hu, R. Zhang, and L. Yang, "Routing protocol design for underwater optical wireless sensor networks: A multiagent reinforcement learning approach," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9805–9818, Oct. 2020.
[42] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to reinforcement learning," in *Deep Reinforcement Learning*. 2020, pp. 47–123.
[43] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[44] H. Yu, N. Yao, and J. Liu, "An adaptive routing protocol in underwater sparse acoustic sensor networks," *Ad Hoc Netw.*, vol. 34, pp. 121–143, Nov. 2015.

Chao Wang is currently pursuing the Ph.D. degree in information and communication engineering with the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, China.

His research interests include underwater acoustic sensor networks.

Xiaohong Shen received the B.S., M.S., and Ph.D. degrees in electrical engineering from Northwestern Polytechnical University (NPU), Xi'an, China, in 1987, 1998, and 2008, respectively.

She is currently a Full Professor with the School of Marine Science and Technology, NPU. Her research interests cover signal processing and underwater acoustic communication and networking. Recently, her research has focused on underwater acoustic high data rate communication and underwater acoustic sensor networks.

Haiyan Wang received the B.S., M.S., and Ph.D. degrees in electrical engineering from the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, China, in 1987, 1990, and 2004, respectively.

He has been a Faculty Member with NPU since 1990 and a Professor since 2004. He has also been with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, since 2018. He teaches and conducts research in the areas of signal and information processing, electronic engineering, and tracking and locating of maneuvering targets. His general research interests include modern signal processing, array signal processing, underwater acoustic communications, tracking and locating of maneuvering targets, data mining techniques, and its application.

Weiliang Xie received the M.S. degree from the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, China, where he is currently pursuing the Ph.D. degree.

His research interests include protocol design and implementation in underwater acoustic sensor networks, and cross-layer design.

Hongwei Zhang received the B.S. degree from the School of Physical Engineering, Zhengzhou University, Zhengzhou, China, in 2017, and the M.S. degree from the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, China, in 2020, where he is pursuing the Ph.D. degree.

His research direction is weak signal detection.

Haodi Mei received the M.S. degree in signal and information processing from the School of Marine Science and Technology, Northwestern Polytechnical University (NPU), Xi'an, China, in 2012, where he is currently pursuing the Ph.D. degree.

His research interest is underwater acoustic networks.