IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 March 2024, accepted 14 April 2024, date of publication 19 April 2024, date of current version 26 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391386

RESEARCH ARTICLE

SQMCR: Stackelberg Q-Learning-Based Multi-Hop Cooperative Routing Algorithm for Underwater Wireless Sensor Networks

WANG BIN[1], BEN KERONG[1], HAO YIXUE[2], (Member, IEEE), AND ZUO MINGJIU[1]

[1]College of Electronic Engineering, Naval University of Engineering, Wuhan 430033, China
[2]College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Wang Bin (wbnavy@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 52071153.

ABSTRACT The underwater wireless sensor network (UWSNs) is an important communication facility supporting underwater monitoring applications. However, the transmission channel has the characteristics of high bit error rate, strong multipath effect, and many interference factors, and the network node has the characteristics of high energy consumption, difficult energy supply, and the node position vulnerable to change, which makes it extremely difficult for UWSNs to realize the reliable and efficient packet forwarding. To address the problem, we propose the Stackelberg Q-learning based multi-hop cooperative routing algorithm (SQMCR). The SQMCR builds the transmission routes based on the Q-learning algorithm, considering factors such as the delay, the remaining energy, and the network topology, which improves the rationality and adaptability of selecting the next-hop node. By balancing the packet forwarding benefits and the energy consumption costs based on the Stackelberg Q-learning algorithm, the SQMCR establishes the cooperative communication policy to ensure both the reliability and efficiency of underwater communications. It also adopts initializing Q-values and dynamic exploration probabilities optimization methods to further improve the performance of routing algorithms. Experimental results show that the SQMCR can help UWSNs increase the packet forwarding reliability and prolong the network lifetime by 17%. It has a better environment and application adaptability and is more suitable for underwater high-reliability applications.

INDEX TERMS Underwater wireless sensor networks (UWSNs), routing algorithm, cooperative communication, Q-learning, Stackelberg game.

I. INTRODUCTION

Underwater wireless sensor networks are an important part of the construction of the marine Internet of Things [1] and an important part of the underwater direction of the future 6G network [2]. They are widely used in many fields, such as disaster early warning, pollutant monitoring, hydrological data monitoring, marine resource exploration, auxiliary navigation, and as an important infrastructure for studying, building, and developing the ocean [3]. Underwater wireless sensor networks are composed of sensor nodes, communication nodes, and sink nodes [4]. At present, the long distance underwater wireless transmission of data mainly depends on the acoustic channel [5]. The underwater acoustic channel has many problems, such as large transmission delay, limited transmission bandwidth, many interference factors, and serious multipath phenomena [6], [7]. Underwater communication nodes are affected by water flow, and their positions and node relationships change dynamically, their communication energy consumption is high, and the energy supplies for the nodes are difficult [6], [7], [8]. All the unfavorable factors make reliable underwater communication extremely difficult. But the reliable communication is the base of various applications in underwater networks [9]. The reliable communication in underwater wireless sensor networks is reflected not only in the reliability of single packet

The associate editor coordinating the review of this manuscript and approving it for publication was Chien-Fu Cheng.

forwarding but also the persistence of packet forwarding services [10]. So, it is very important to design the routing algorithms which can not only overcome the influence of time varying underwater transmission environment and achieve reliable packet forwarding but also ensure the stability of packet forwarding services.

The routing algorithms are used to find the optimal paths from the sensor node to the sink node through the communication nodes for packet forwarding [4]. A good transmission route can not only reduce the packet transmission delay and improve the packet delivery rate but also balance the energy consumption during the underwater nodes and prolong the lifetime of underwater networks [11], [12]. Due to the limitation of underwater acoustic channels and the dynamic change of underwater network topology, it is difficult for the data source node to get the network topology in real time and optimize the route globally [13]. It can only select the next-hop node according to the state information between the local and neighbor nodes to obtain the approximation of the globally optimal path [14]. A good next-hop node should be closer to the target node, which can decrease the number of transmission hops, and reduce energy consumption [14]. A good next-hop node should have sufficient remaining energy for continuous forwarding, and the energy consumption should be close to that of the neighbor node to ensure the channel durability [15]. A good next-hop node must have a good channel state with the source node to make the signal-to-noise of the receiving node meet the receiving requirements [16], [17]. A good next-hop node should have sufficient storage capacity to avoid congestion and reduce the transmission delay [18]. A good next-hop node should also have stable, good, and sufficient neighbor relationships so that it can further forward the packet [19]. In underwater wireless sensor networks, the traditional routing protocols generally select routes based on the current status between the data source node and the next-hop node or node cluster, such as location, energy, delay, number of neighbors, received signal strength, and so on. Based on the routing protocol implemented by reinforcement learning, the communication node in the underwater network is modeled as an agent, the reward function is established based on domain knowledge, and the policy of forwarding packets is established through iterative learning [20], [21]. The introduction of reinforcement learning enables routing algorithms to optimize routes based on the current state and long-term forwarding experience, improving the rationality of route selection and the reliability of packet forwarding [22], [23].

Cooperative communication is also a typical way to improve the reliability of packet forwarding [24]. A cooperative communication system includes the sending node, the receiving node, and the cooperative nodes. The sending node sends the packet to the receiving node by broadcasting, and the selected cooperative node receives and forwards the packet. The signals come from different directions, which can make the receiving node obtain the required signal-to-noise ratio for the correct reception. Because if there is some interference or occlusion on the transmission channel, the signals from the sending node and the cooperative nodes can be properly superimposed on the receiving node. The virtual multiple input multiple outputs (MIMO) systems composed of the sending node, the receiving node, and the cooperative nodes enable single-hop transmission in the network to obtain additional spatial diversity benefits and can improve the ability of single-hop transmission to combat channel fading [25], [26]. Compared to the single-hop networks, the multi-hop networks can transmit longer distances, provide greater bandwidth, and consume lower communication energy per node [27]. Therefore, a multi-hop cooperative system can help underwater wireless sensor networks improve the reliability of packet forwarding. However, unnecessary cooperative communication will also cause excessive energy consumption [28]. Therefore, the cooperative communication system also needs to solve the problems of ‘‘whether to cooperate’’ and ‘‘who will cooperate ’’. Cooperative nodes need to balance the benefits of packet forwarding and the cost of energy consumption. The Stackelberg game is a two-stage complete information dynamic game investigated for multi-agent systems [29]. In this game mode, the player who makes the decision first is called the leader, and the remaining players are called the followers. The global goal of the game is to maximize the benefit of the leader. Each of the agents can sense the environment, learn the policies and assist the leader to achieve the global goal while making an effort to maximize their own benefits [29]. The sending node sends the data to the receiving node by broadcasting, and the cooperative node selects whether to cooperate and who to cooperate according to the status of itself, the sending node, the receiving node and the other neighbor nodes. Due to the full consideration of the status, the cooperative communication system makes it easier to balance the data receiving benefits representing the short-term interests and the energy consumption costs affecting the long-term interests, which improves the packet delivery rate, prolongs the network life, and realizes the overall improvement of network reliability.

II. RELATED WORKS

To overcome the adverse transmission conditions of underwater wireless sensor networks, researchers have studied the routing protocols from different perspectives.

The routing algorithm based on the deterministic rules generally determines the best relay node according to the current position, depth, and remaining energy of the underwater nodes. For example, Xie et al. proposed VBF [30] to establish a virtual pipe on the vector between the source node and the destination node. VBF limits the set of the candidate forwarding nodes by controlling the radius of the virtual pipeline, and selects the best next-hop node according to the distance between the node and the vector. Yan et al.

proposed DBR [31] to select the best relay node according to the depth of the candidate forwarding nodes. The node close to the surface will give priority to the forwarding packet. The routing algorithm based on the deterministic rules, which fully considers the current state of neighbor nodes, makes the routing algorithm simple, efficient, and adaptive. However, due to the lack of a long-term iterative learning process, it is unable to obtain prior knowledge from the forwarding experience, which makes the packet forwarding partially sighted to a certain extent.

The routing algorithm based on reinforcement learning abstracts the underwater network data forwarding process as a Markov process. Historical forwarding can be used as a priori knowledge, which affects the selection of subsequent forwarding nodes and improves the accuracy of the next-hop node selection [32]. Hu and Fei proposed QELAR [33] to select the next-hop node based on Q-learning. QELAR's reward function considers not only the packet delivery rate of the transmission channel but also the remaining energy and energy consumption balance of the nodes. It makes the routing policy not only select the shortest transmission path but also comprehensively consider the remaining energy of the receiving nodes, to avoid the energy depletion of the node in the optimal path. Jin et al. proposed RCAR [18] to take account of the congestion avoidance method of the relay nodes in the case of heavy traffic and to optimize the transmission delay and energy consumption distribution in underwater data communication by the reinforcement learning algorithm. Wang et al. proposed EP-ADTA [34] to select the relay nodes and data transmission accuracy based on the Q-learning. EP-ADTA can optimize the transmission path and dynamically adjust the transmission data accuracy according to the change in the transmission environment. The routing algorithm based on reinforcement learning can dynamically select routes according to the experience accumulated by each agent in data forwarding and the current state of neighbor nodes, thereby improving the reliability of the transmission routing.

In underwater data transmission, due to the influence of path loss, shadow fading, and multipath fading, the receiving node cannot receive the packet whose signal-to-noise ratio does not meet the receiving requirements. The cooperative node is used to provide relay assistance to enhance the signal gain of receiving nodes and achieve reliable communication. Zhang et al. proposed SA-FRL [35] to realize the cooperative communication of underwater networks based on Q-learning. SA-FRL selects cooperative nodes with good link quality and low access delay to improve the efficiency of cooperative communication. However, SA-FRL only provides the solutions for single-hop network scenarios and does not involve multi-hop networks. Chen et al. proposed QMCR [32] to form the routes in the multi-hop networks and achieves the gain of underwater communication performance through cooperative communication based on Q-learning. However, QMCR only selects the cooperative mode according to the deterministic rule, without considering the remaining energy of cooperative nodes and the necessity of cooperative transmission, which may result in the waste of communication energy consumption. Therefore, it is necessary to further optimize the cooperative node selection by introducing the reinforcement learning algorithm and the game theory.

In order to further compare the merits and demerits of several typical routing algorithms, their strategy and characteristics are listed in Table 1.

III. UNDERWATER WIRELESS SENSOR NETWORKS

A. UNDERWATER ACOUSTIC TRANSMISSION CHANNEL

In underwater wireless sensor networks, the underwater acoustic channel is complex, unreliable and time-varying.

1) SIGNAL ATTENUATION

The signal attenuation of underwater acoustic channel depends not only on the distance between the sending node and the receiving node, but also on the signal frequency. According to [36], for the signal with frequency f, the attenuation generated by the underwater acoustic channel at distance l is:

$$A(l,f) = A_0 l^k a(f)^l. \tag{1}$$

In *dB* form, it can be expressed as:

$$10\log\frac{A\left(l,f\right)}{A_0} = k \times 10\log l + l \times 10\log a(f). \tag{2}$$

where the first item is expansion loss, and the second item is absorption loss, A_0 is the normalization constant, $a(f)$ is the absorption constant, and k is the expansion factor that describes the transmission geometry. According to [37], when the spherical expansion occurs in deep water, k is usually set to 2; when the cylindrical expansion occurs in the shallow water, k is usually set to 1; and in the actual expansion, k is set to 1.5.

According to [36], when the unit of $a(f)$ is dB/km (when the frequency unit is kHz) and the frequency ranges from 100Hz to 1MHz, the commonly used empirical formula for estimating the absorption coefficient is the Thorps formula, as follows:

$$\begin{aligned}10\log a(f) = 0.11\frac{f^2}{1+f^2} + 44\frac{f^2}{4100+f^2}\\ + 2.75 \times 10^{-4}f^2 + 0.003.\end{aligned} \tag{3}$$

2) MULTIPATH INTERFERENCE

Multipath propagation is caused by the reflection of acoustic signals on the sea surface, seabed and other objects. The signal sent by the sending node arrives at the receiving node through different paths. Assuming that the propagation speed of the p acoustic signal in the multipath signal is c and the propagation distance is l_p. Routing delay is $\tau_p = l_p/c$, the cumulative reflection coefficient after multiple reflections is Γ_p. The propagation loss is $A\left(l_p, f\right)$, and according to [37],

TABLE 1. Several typical routing algorithms and their characteristics.

Algorithm	Strategy		Characteristics				Year
	Routing	Cooperative communication	Network	Reliability	Efficiency	Lifetime	
VBF [30]	Deterministic rule	-	Multi-hop	Low	Low	Short	2006
DBR [31]	Deterministic rule	-	Multi-hop	Low	Low	Short	2008
QELAR [33]	Q-learning	-	Multi-hop	Middle	High	Middle	2010
RCAR [18]	Q-learning	-	Multi-hop	Middle	High	Middle	2019
EP-ADTA [34]	Q-learning	-	Multi-hop	Middle	High	Middle	2022
SA-FRL [35]	-	Q-learning	Single-hop	High	-	-	2022
QMCR [32]	Q-learning	Deterministic rule	Multi-hop	High	Middle	Middle	2021
SQMCR	Q-learning	Q-learning	Multi-hop	High	High	Long	-

the frequency response of the p path is:

$$H_p(f) = \frac{\Gamma_p}{\sqrt{A\left(l_p, f\right)}}. \tag{4}$$

Then, the overall response of the acoustic signal after the superposition of multiple paths is:

$$h(t) = \sum_p h_p\left(t - \tau_p\right). \tag{5}$$

where $h_p(t)$ is the inverse Fourier transform of $H_p(f)$.

3) NOISE INTERFERENCE

Another factor that affects the quality of underwater acoustic channels is the presence of a large amount of noise in the underwater environment, typically including the environmental noise, the vehicles noise, and the marine organism noise. According to [37], there are four types of noise sources. The noise power empirical formula (in dB) as a function of frequency (in kHz) is:

$$10\log N_t(f) = 17 - 30\log f. \tag{6}$$

$$10\log N_s(f) = 40 + 20\left(s - 0.5\right) + 26\log f - 60\log\left(f + 0.03\right). \tag{7}$$

$$10\log N_w(f) = 50 + 7.5\omega^{\frac{1}{2}} + 20\log f - 40\log\left(f + 0.4\right). \tag{8}$$

$$10\log N_{th}(f) = -15 + 20\log f. \tag{9}$$

where ω is the wind speed, in m/s; $N_t(f)$ is the turbulent noise that only affects a very low frequency range of less than 10Hz; $N_s(f)$ is the ship noise, dominant between 10 and 100Hz; $N_w(f)$ is the noise caused by waves and other surface movements caused by wind and rain in the range of 100Hz to 100kHz; $N_{th}(f)$ is the dominant thermal noise with frequencies exceeding 100kHz.

Then, the overall ambient noise is:

$$N(f) = N_t(f) + N_s(f) + N_w(f) + N_{th}(f). \tag{10}$$

4) OCCLUSION INTERFERENCE

The interference caused by the object occlusion often has a significant impact on the transmission of acoustic signals, leading to a significant decrease in the amplitude response $h_p(t)$ of the acoustic signal on the transmission path where the occluded object is located, and even leading to signal interruption. The frequency of occlusion interference occurring within a fixed time $\lambda_{block-out}$ and the transmission distance l_p between communication nodes is in the direct proportion, namely:

$$\lambda_{block-out} \propto l_p. \tag{11}$$

5) TRANSMISSION CAPACITY

Assuming that the communication node sends the sound signals with a fixed power (P) and bandwidth (B), and the signal-to-noise ratio (SNR) of the receiving node is [38]:

$$SNR = \frac{|h_{sd}(t)|^2 \times P}{N(f) \times B}. \tag{12}$$

where, $h_{sd}(t)$ is the amplitude value of the channel response from the sending node to the receiving node, $h_{sd}(t)$ is determined by transmission loss, multipath interference, and occlusion interference, and is inversely proportional to the transmission distance.

Assuming that the noise follows Gaussian distribution, according to the Shannon-Hartley theorem, the maximum transmission capacity of the transmission channel between the sending and receiving nodes is:

$$C_{sd}(t) = B\log_2\left(1 + SNR\right). \tag{13}$$

When the data transmission rate $R(t)$ satisfies $C_{sd}(t) \geq R(t)$, the data can be accurately received at the receiving node.

Joining cooperative communication nodes in data forwarding can effectively improve the signal power and SNR at the receiving node due to shorter signal transmission distances or signal superposition. Therefore, the cooperative communication can increase the maximum transmission capacity of the underwater transmission channels and further improve the reliability of data reception.

6) OUTAGE PROBABILITY

The outage probability refers to the ratio of the number of interruption events to the total number of transmission times during data transmission. According to the Information Theory, the outage probability can be expressed as:

$$P_{out-sd}(t) = P\left[C_{sd}(t) < R(t)\right]. \tag{14}$$

The transmission rate is determined by the business application. When the transmission rate is relatively fixed, the main factor affecting the outage probability is the varying transmission capacity. Increasing the transmission capacity can effectively reduce the outage probability. According to the Shannon-Hartley theorem, the transmission capacity depends on the SNR, which is related to signal attenuation, multipath interference, noise interference, occlusion interference. It is also related to the health status of the communication node and whether there are enough neighboring nodes that can participate in cooperative communication. Therefore, methods to reduce the probability of interruption include controlling the communication distance between the transmitting and receiving nodes, and selecting nodes with sufficient remaining energy and many neighboring nodes as the next-hop.

B. UNDERWATER WIRELESS SENSOR NETWORKS BASED ON MULTI-HOP COOPERATIVE COMMUNICATION

Underwater wireless sensor networks provide services for underwater monitoring applications. The underwater wireless sensor network (see Figure 1), includes an underwater sensor node, several underwater communication nodes, and a surface sink node. The obtained data is transmitted to the surface sink node hop by hop through the communication nodes. After the sink node obtains the data, it sends the data to the shore-based or ship-based data center by radio. In each hop of the transmission route, the cooperative communication system is composed of the sending node, the receiving node, and the candidate cooperative nodes. The transmission path is the direct packet forwarding path from the sending node to the receiving node, and the cooperative communication path is the packet forwarding path through the cooperative nodes. The packet forwarding channel is divided into a broadcast channel and a multiaccess channel. The sending ends (including the sending node and the cooperative nodes) send packets through multiple independent transmission channels, and the receiving end (including the receiving node) appropriately combines multiple copies of signals that carry the same data but are statistically independent of each other to combat channel fading. Modeling from the perspective of minimizing the symbol error rate and using the optimal single cooperative node, we can get a lower symbol error rate than multi-node participation in cooperative communication [28]. Therefore, in this paper, only the case of a single cooperative node is considered, and the receiving node only realizes the signal synthesis from two directions.

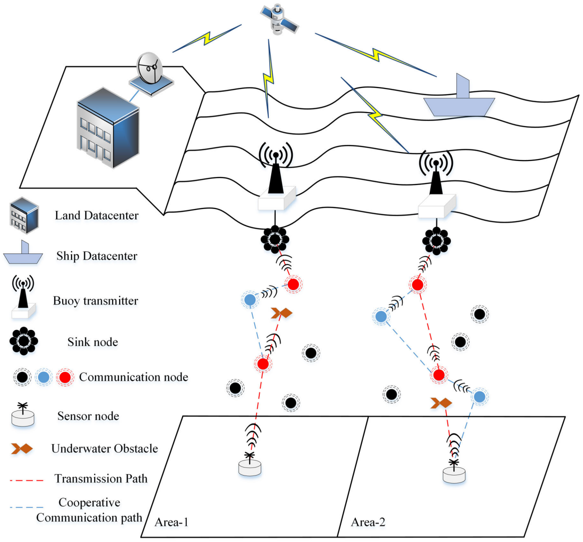

FIGURE 1. Underwater wireless sensor network based on multi-hop cooperative communication.

IV. SQMCR ALGORITHM

A. FRAMEWORK OF SQMCR ALGORITHM

Stackelberg game is a tool of dealing with the situation where the players take actions sequentially [29]. The main agent is the leader and the subagents are the followers in the modal. The leader takes actions firstly, considering the policy of the followers, and the followers make the best response based on the leader's action [29], [39]. The global goal of the multi-agent system is to maximize the benefit of the leader. The goal of the followers is to maximize the leader's and their own benefits by sensing the environment [29], [40].

In each hop of packet forwarding in underwater wireless sensor network based on multi-hop cooperative communication, it is the cooperative communication system consisting of the sending nodes, the receiving nodes, and the candidate cooperative nodes. The global goal of the cooperative communication system is to improve the success rate of forwarding packets to the receiving nodes. For the sending node, blindly forwarding packets without paying attention to the status of receiving node and candidate cooperative nodes can reduce the reliability of packet forwarding. For the candidate cooperative nodes, blindly participating in cooperative communication without paying attention to the packet forwarding policy of sending node and the status of other candidate cooperative nodes can not only reduce the reliability of packet forwarding, but also causes unnecessary communication energy consumption. Therefore, it is necessary to set certain rules to coordinate and control the relationship between the sending node and the candidate cooperative nodes, to achieve maximum the global benefits. The rules in the cooperative communication system can be defined based on the Stackelberg game and the nodes in the underwater wireless sensor network can be defined as agents, as shown in Fig 2. The sending node is the leader agent, and the candidate cooperative nodes are the follower agents. The benefit of the leader is to realize the reliable packet reception of receiving node, reduce the outage probability, and improve the packet forwarding efficiency as much as possible. The benefits of the followers is to reduce the communication energy consumption and prolong the network life based on achieving reliable packet reception of the receiving node.

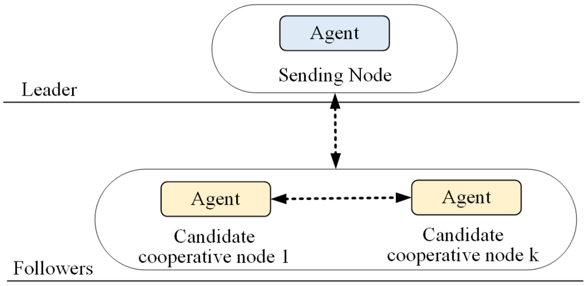

FIGURE 2. Model of Stackelberg game in cooperative communication system.

B. TRANSMISSION ROUTING SUB-ALGORITHM

The routing process of underwater wireless sensor networks can be defined as Markov decision processes (MDPs). The selection of each next hop is not only based on the current state but also related to the historical packet forwarding status. Reinforcement learning is often used to solve the problem of MDPs. It is defined by five tuples (S, A, R, P, γ). S represents the environment state, A represents the action set, R represents the reward, P represents the transition probability, and γ represents the discount rate. According to the state and the cumulative reward, the agent updates the transition probability, action, and new state. The nodes of the underwater wireless sensor network are defined as agents based on Q-learning which is a kind of widely used reinforcement learning algorithms. Through the iterative training, the Q-learning agents form the optimal transmission route policy for packet forwarding.

Suppose that the underwater wireless sensor network consists of m nodes, which can be expressed as:

$$N = \{n_1, n_2, \cdots, n_i, \cdots, n_m\}. \tag{15}$$

where N represents the nodes set, n_i represents the current node. Then, the candidate next-hop nodes set $N_{relay}(i)$ of the current node n_i can be expressed as:

$$N_{relay}(i) = \{dep(n_j) - dep(n_i) \leq 0\} \cap neighbors(n_i). \tag{16}$$

where $\left\{dep\left(n_j\right) - dep\left(n_i\right) \leq 0\right\}$ represents the node set with a shallower depth than the current node. The set of the candidate next-hop nodes is in the upper hemisphere near the water surface covered by the current node's transmitted signal.

Definition 1: At time slot t, if the packet is located at the node n_i, the state S can be defined as:

$$S = \{n_i\} \cup N_{relay}(i). \tag{17}$$

At time slot t, the action A can be defined as:

$$A = \left\{a_j \middle| n_j \in S\right\}. \tag{18}$$

where n_j is the next-hop node to which the action is forwarding the packet. At time slot t, if the packet is at the node n_i and the node n_j is the next-hop, the reward function is:

$$R^{a_j}_{n_i n_j} = -C_0 - [\varphi_e \times co(e) + \varphi_t \times co(t) + \varphi_n \times co(n)]. \tag{19}$$

(1) C_0 is the fixed cost.

(2) co(e) is the energy cost, which can be expressed as:

$$co(e) = 1 - \frac{e^j_{res}}{\sum\limits_{k \in N_{relay}(i)} e^k_{res}}. \tag{20}$$

where e^j_{res} is the remaining energy of the next-hop node, $\sum\limits_{k \in N_{relay}(i)} e^k_{res}$ is the total remaining energy of the candidate next-hop nodes set, φ_e is the sensitivity coefficient of energy cost.

(3) co(t) is the delay cost, which can be expressed as:

$$co(t) = 1 - \frac{1}{t^{i \to j} + 1}. \tag{21}$$

where $\overline{t^{i \to j}}$ is the average transmission delay from node n_i to node n_j, φ_t is the sensitivity coefficient of delay cost.

(4) co(n) is the robustness cost, which can be expressed as:

$$co(n) = \frac{1}{2} \times (2 - \beta_1 \frac{num^j_{input}}{\sum\limits_{k \in N_{relay}(i)} num^k_{input}} - \beta_2 \frac{num^j_{output}}{\sum\limits_{k \in N_{relay}(i)} num^k_{output}}). \tag{22}$$

where, num^j_{input} and num^j_{output}, represent the number of in-degree and out-degree neighbor nodes of the node n_j respectively, β_1 and β_2 are the adjustment weight, and φ_n is the sensitivity coefficient of the robustness cost.

The reward function determines the optimization direction of the transmission routing policy. To achieve the benefits of the leader, the reward function is designed in Definition 1.

According to the Q-learning algorithm, the Q-value iteration as follows.

$$Q^{t+1}_i\left(s^t_i, a^t_i\right) = (1-\alpha)\, Q^t_i\left(s^t_i, a^t_i\right) + \alpha \left\{R^{a_j}_{n_i n_j} + \gamma * \omega_1 * V^t_i\left(s^{t+1}_j\right)\right\}. \tag{23}$$

where α is the learning rate, $V^t_i\left(s^{t+1}_j\right)$ is the historical forwarding experience from the node n_i to the node n_j, and ω_1 is the adjustment weight.

In the Stackelberg game modal, the global goal of the multi-agents system is to maximize the benefits of the leader, [29], [40], which is consistent with the optimization direction of the transmission routing sub-algorithm based on reinforcement learning. So the optimal policy of the transmission routing sub-algorithm implemented based on reinforcement learning, that is, the maximum Q value, is consistent with the Q value selected by Stackelberg game.

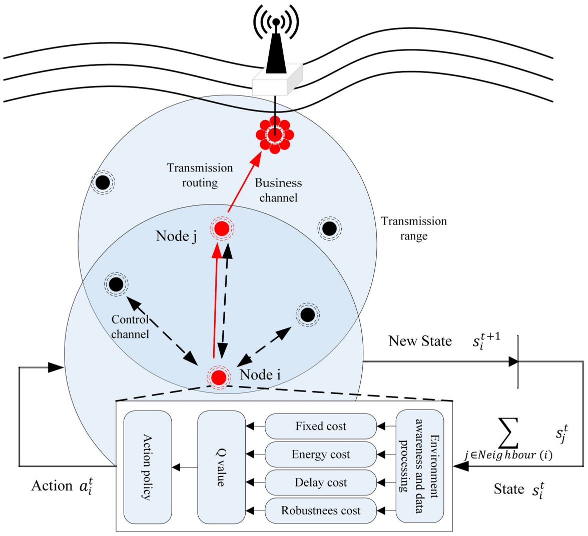

FIGURE 3. Transmission routing sub-algorithm.

And according to the Bellman equation, the value function of the sending node is:

$$\begin{aligned} V^t(s) &= Stackelberg_{leader}(Q(s,a)) \\ &= \max_a Q^*(s,a). \end{aligned} \tag{24}$$

At the beginning of reinforcement learning, the V-value is given an initial value. With the iteration of learning, the Q-value and V-value will be continuously updated and converged by (23) and (24), and a good transmission routing policy will finally be highlighted. Figure 3 shows how the policy is generated based on the transmission routing sub-algorithm. The sending node n_i learns the routing policy to select the next-hop node n_j and forwards the packet to it. The node n_j continues to select the optimal next-hop node based on the routing policy until the packet is forwarded to the sink node. Due to the comprehensive consideration of the location, remaining energy, transmission delay, and structural robustness of neighbor nodes in the routing policy, it can improve the reliability of packet forwarding.

C. COOPERATIVE COMMUNICATION SUB-ALGORITHM

In the cooperative communication system, whether candidate cooperative nodes participate in cooperative communication and who participates in cooperative communication is determined by the Stackelberg game. The decision-making process of candidate cooperative nodes participating in cooperative communication can be described by the partially observable Markov process (POMDP) which can be solved based on the Stackelberg Q-learning [29], [40].

Suppose that the cooperative communication system contains p candidate cooperative nodes, which can be expressed as:

$$R = \left\{ n_{r_1}, \cdots, n_{r_p} \right\} \tag{25}$$

At time slot t, the candidate cooperative node n_{r_x} has three states: state *i*) the packet is forwarded by the cooperative node but fails to be received by the receiving node; state *ii*) the packet is not forwarded; state *iii*) the packet is forwarded by the cooperative node and received successfully by the receiving node.

Definition 2: The status of the candidate cooperative nodes n_{r_x} can be defined as:

$$\begin{aligned} S_{r_x} = \{&relay\ and\ without\ ack, \\ &\neg relay, relay\ and\ with\ ack\}. \end{aligned} \tag{26}$$

The action A_{r_x} of the candidate cooperative node n_{r_x} can be defined as:

$$A_{r_x} = \{relay, \neg relay\}. \tag{27}$$

where, relay and ¬relay, represent the actions of to forward the packet or not to forward the packet respectively, 'ack' is the feedback representing the packet has been received successfully. At time slot t, the reward of the candidate cooperative node n_{r_x} is:

$$R_{r_x}^{a_{r_x}} = \begin{cases} -\varphi_{co} \times co\left(\rho_{relay}\right) \times co\left(e_{res}^{r_x}\right), \\ \qquad when\ relay\ and\ without\ ack. \\ 0, \\ \qquad when\ \neg relay. \\ R_0 + [\varphi_{cop} \times re\left(d_{n_{r_x}-n_j}\right) + \\ \quad \varphi_{com} \times re\left(d_{n_i-n_j}\right)] \\ \quad -\varphi_{co} \times co\left(e_{res}^{r_x}\right) \times co\left(\rho_{relay}\right), \\ \qquad when\ relay\ and\ with\ ack. \end{cases} \tag{28}$$

(1) R_0 is the fixed benefit.

(2) $re\left(d_{n_{r_x}-n_j}\right)$ is the cooperative distance benefit, which can be expressed as:

$$re(d_{n_{r_x}-n_j}) = \frac{2}{\pi}\tan^{-1}(\overline{d_{n_{r_k}-n_j}} - d_{n_{r_x}-n_j}),\ n_{r_k} \in R. \tag{29}$$

where $d_{n_{r_x}-n_j}$ is the distance between the selected candidate cooperative node and the receiving node, $\overline{d_{n_{r_k}-n_j}}$ is the average distance between the candidate cooperative nodes set and the receiving node, and φ_{cop} is the adjustment coefficient.

(3) $co\left(e_{res}^{r_x}\right)$ is the communication distance benefit, which can be expressed as:

$$re\left(d_{n_i-n_j}\right) = \frac{d_{n_i-n_j}}{d_{com}}. \tag{30}$$

where $d_{n_i-n_j}$ is the distance between the sending node n_i and the receiving node n_j, d_{com} is the communication distance, φ_{com} is the adjustment coefficient.

(4) $co\left(e_{res}^{r_x}\right)$ is the remaining energy cost, which can be expressed as:

$$\begin{aligned} co\left(e_{res}^{r_x}\right) &= \beta_1 \times \left(1 - \frac{e_{res}^{r_x}}{e_{ini}^{r_x}}\right) \\ &\quad + \beta_2 \times \frac{2}{\pi}\tan^{-1}\left(\overline{e_{res}^{r_k}} - e_{res}^{r_x}\right),\ n_{r_k} \in R. \end{aligned} \tag{31}$$

where e_{res}^{rx} and e_{ini}^{rx} are the remaining energy and initial energy of the candidate cooperative node n_{r_x}, $\overline{e_{res}^{rk}}$ is the average remaining energy of all candidate cooperative nodes, $\left(1-\frac{e_{res}^{rx}}{e_{ini}^{rx}}\right)$ reflects the proportion of the remaining energy of the candidate cooperative node, β_1 and β_2 are the adjustment coefficients.

(5) $co\left(\rho_{relay}\right)$ is the density cost, which can be expressed as:

$$co\left(\rho_{relay}\right)=1-\frac{1}{p}. \tag{32}$$

where p is the number of nodes in the candidate cooperative node set, which represents the node density of the candidate cooperative node set, and φ_{co} is the adjustment coefficient.

The reward function is designed based on the benefits of the followers, which not only encourages cooperative nodes to participate in communication to improve the packet forwarding reliability and generate the communication benefits but also inhibits the unnecessary communication energy consumption caused by cooperative nodes' partially sighted and frequent participation in cooperative communication.

The Stackelberg game model in the cooperative communication systems belongs to the complete information game model. Candidate cooperative nodes obtain each other's status by overhearing the packets in the underwater network. According to the multi-agent Q-learning algorithm, the Q-value iteration formula is:

$$\begin{aligned}
Q_{r_x}^{t+1}(s_{r_x}^t,a_{r_x}^t)&=(1-\alpha_{r_x})Q_{r_x}^t(s_{r_x}^t,a_{r_x}^t)\\
&\quad+\alpha_{r_x}\left\{R_{r_x}^{a_{r_x}}+\gamma_{r_x}\times\omega_2\times V_{r_x}^t(s_{r_x}^{t+1})\right.\\
&\quad+\gamma_{r_x}\times\omega_3\times V_i^t(s_j^{t+1})\\
&\quad\left.-\gamma_{r_x}\times\omega_4\times\sum_{i'\in I,i'\neq n_{r_x}}V_{i'}^t\left(s_{i'}^{t+1}\right)\right\},\\
&\quad I=R\cap neighbors(n_{r_x}).
\end{aligned} \tag{33}$$

where α_{r_x} is the learning rate, $V_{r_x}^t\left(s_{r_x}^{t+1}\right)$ represents the current cooperative node's historical cooperative communication experience, $V_i^t\left(s_j^{t+1}\right)$ represents the historical forwarding experience from the node n_i to the node n_j, $\sum_{i'\in I,i'\neq n_{r_x}}V_{i'}^t\left(s_{i'}^{t+1}\right)$ represent the historical cooperative communication experience of the neighbor candidate cooperative nodes, ω_2, ω_3, and ω_4 are the adjustment coefficients.

In the Stackelberg game modal, the relationship between the followers is competitive, and the followers' goal is to maximize the leader and their own benefits [29], [40]. Due to the fact that the cooperative communication sub-algorithm implemented based on multi-agent reinforcement learning comprehensively considers the states of the current node, the sending node, the receiving node and other candidate cooperative nodes, the policy of the candidate cooperative node with the maximum Q value represents the optimal policy for the candidate cooperative nodes set, that is, the policy of the followers. Therefore, the candidate cooperative nodes with the maximum Q value will be selected, and its optimal policy will be used as the policy in the Stackelberg game. The optimal policy of the followers can be defined as:

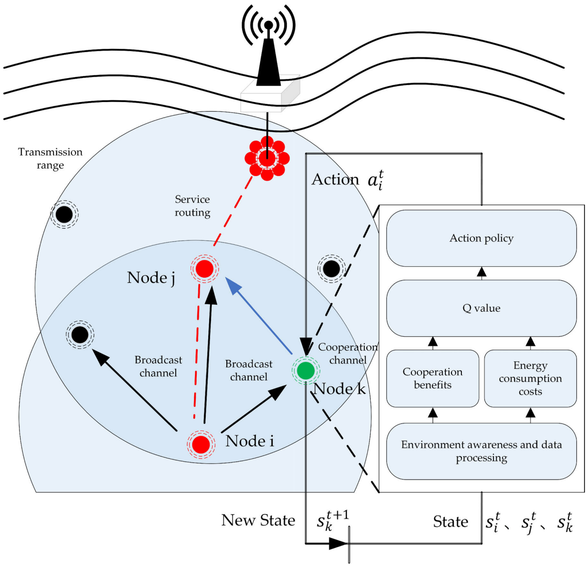

FIGURE 4. Cooperative communication sub-algorithm.

$$\begin{aligned}
&Stackelberg_{follower}(Q_{r_1}^*,\cdots,Q_{r_x}^*,\cdots Q_{r_p}^*)\\
&\quad=max(Q_{r_1}^*,\cdots,Q_{r_x}^*,\cdots Q_{r_p}^*)=Q_{r_x}^*.
\end{aligned} \tag{34}$$

So according to the Bellman equation, the V-value of the selected cooperative node n_{r_x} is:

$$V_{r_x}^t(s)=\max_a Q_{r_x}^*(s,a). \tag{35}$$

The V-value of the not-selected candidate cooperative node nodes is not updated. With the iteration of training, the V-value will be continuously updated according to (34) and (35) and gradually converge, and a good cooperative communication policy will finally be highlighted. Figure 4 shows how the policy is generated based on the cooperative communication sub-algorithm. According to the transmission route formed by the sending node n_i and the receiving node n_j, the candidate cooperative node n_k is determined to participate in the cooperative communication based on the policy, it forwards the packet by the cooperation channel. Due to the comprehensive consideration of the cooperative benefits and the energy consumption costs, on the premise of maximizing the packet delivery rate, the communication energy consumption can be reduced and balanced as much as possible.

D. OPTIMIZATION METHODS

1) Q-VALUE INITIALIZATION

The Q table of the sending node in the transmission routing sub-algorithm can be initialized according to the nodes' location in the underwater network.

At each status update time, according to the neighbor relations of the nodes, the Q-value of the sending node n_i to its non-out-degree node $n_{\bar{j}}$ is initially set as:

$$Q^{ini}_{n_i \to n_{\bar{j}}} = -100,\ n_i,\ n_{\bar{j}} \in N,\ n_{\bar{j}} \notin N_{relay}(i). \tag{36}$$

At the initial time, a hemisphere with a layered structure is established with the sink node as the center and the integer multiple of the communication distance d_{com} as the radius. When the distance d_{i-sink} between the node n_i and the sink node n_{sink} meets:

$$(T-1) \times d_{com} < d_{i-sink} \leq T \times d_{com},\ T \in \{1, 2, \cdots\}. \tag{37}$$

The initial Q-value of the sending node n_i to the out-degree node n_j is set as:

$$Q^{ini}_{n_i \to n_j} = -100 \times \frac{T}{T_{max}},\ n_i,\ n_j \in N,\ n_j \in N_{relay}(i). \tag{38}$$

$$T_{max} = \frac{d_{sensor-sink}}{d_{com}}. \tag{39}$$

where $d_{sensor-sink}$ is the distance between the sensor node and the sink node.

Initializing the Q table according to the topology relationship can shorten the iteration rounds of the Q-learning algorithm.

2) DYNAMIC EXPLORATION PROBABILITY

The reinforcement learning algorithm adjusts the degree of “exploration” and “utilization” by exploring probability ε to ensure that it always converges to the optimal result. Because of the time-varying transmission channel, it is necessary to dynamically adjust the exploration probability ε in the transmission routing sub-algorithm, according to the convergence degree of the algorithm.

At time slot t, the current state value of the current node is $V(t)$ and the new state action value is $Q(t+1)$, then $\varepsilon(t)$ should meet:

$$\varepsilon(t) = \begin{cases} \omega_\varepsilon \times e^{-1 \times |V(t) - Q(t+1)|}, & \\ \qquad 0 < V(t) - Q(t+1) < \varepsilon_{thres}. & \\ \varepsilon_{ini}, \quad otherwise. & \end{cases} \tag{40}$$

where ω_ε is the adjustment coefficient for the dynamic exploration probability, ε_{thres} is the threshold of the variation range of the V-value.

E. PROCESS OF SQMCR

The SQMCR algorithm process is based on the Stackelberg game and is divided into two stages. The first is the leader stage and the second is the follower stage.

1) LEADER STAGE

In this stage, the goal of SQMCR is to select the optimal next-hop node and improve the reliability of packet forwarding, (see Algorithm 1).

where $times_{send}$ is the number of retransmissions and $times_{max}$ is the retransmission threshold.

Algorithm 1 SQMCR-Leader

```
Initialize the $Q_{leader}$ by (36) - (39)
While(true)
  If (the new packet to send)
    While ($times_{send} \leq times_{max}$)
      Update the $S_{leader}$ and $A_{leader}$ by (17) – (18)
      Calculate the $R_{leader}$ by (19) – (22)
      Calculate the $Q_{leader}$ by (23)
      Calculate the $\varepsilon_{leader}$ by (40)
      Choose the next-hop node
      Form the packet and forward it
     If (the packet has been received)
       Update the $V_{leader}$ using (24)
       Break
     Else
       Update the $V_{leader}$ using (24)
       $times_{send}++$
     End If
     Update the $\varepsilon_{leader}$ using (40)
    End While
  End If
  Update the $Q_{leader}$ by (36)
End While
```

2) FOLLOWER STAGE

In this stage, the goal of SQMCR is to determine whether to participate in the communication according to the route selected by the leader, (see Algorithm 2).

Algorithm 2 SQMCR-Follower

```
Initialize the $Q_{follower}$ as 0
While(true)
  If (the new packet has been sent from the leader)
    If (the destination is the neighbor node)
      Update the $S_{follower}$ and $A_{follower}$ by (26) – (27)
      Calculate the $R_{follower}$ by (28) – (32)
      Calculate the $Q_{follower}$ by (33)
      Determine whether to $relay$ or $\neg relay$
      If (the selected action is $relay$)
        Update the packet, and forward it
        Update the $V_{follower}$ using (35)
      End If
    End If
  End If
End While
```

F. PACKET OF SQMCR

The SQMCR algorithm defines two types of packets: business packets and control packets. The business packets are forwarded in the business channel, carrying the monitoring data. The transmission routes for the business packets are from the underwater sensor node to the surface sink node through the communication nodes. The control packets forwarded between neighboring nodes are transmitted in the

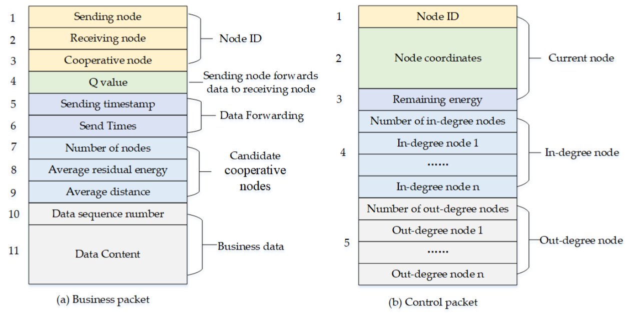

FIGURE 5. Packet of SQMCR algorithm.

TABLE 2. Hyper-parameters of SQMCR.

Parameter	Value
C_0	1
$\varphi_e,\ \varphi_t,\ \varphi_n$	$0.8, 0.2, 0.2$
$\beta_1,\ \beta_2$	$0.5, 0.5$
$\alpha,\ \gamma$	$0.8, 0.8$
ω_1	1
R_0	0
$\varphi_{cop},\ \varphi_{com},\ \varphi_{co}$	$0.6, 0.4, 1$
$\beta_3,\ \beta_4$	$0.5, 0.5$
$\alpha_{r_x},\ \gamma_{r_x}$	$0.8, 0.8$
$\omega_2,\ \omega_3, \omega_4$	$1, 0.2, 0.2$
$\omega_\varepsilon, \varepsilon_{ini}, \varepsilon_{thres}$	$1, 0.8, 0.1$
$times_{max}$	2

control channel, which is used to synchronize the status of neighbor nodes. Each node updates the status of the current node, neighbor nodes and the transmission channel in the database by listening to the business or control packets.

The structure of business packets is shown in Figure 5-a. Fields 1 to 3 are the ID of the sending node, the receiving node, and the cooperative node. Field 4 is the Q-value forwarded by the current node to the receiving node. Fields 5 and 6 are the timestamp of the packet forwarded and the number of times the packet has been repeatedly forwarded. Fields 7 to 9 are the number, the average remaining energy of the candidate cooperative nodes, the average distance from the receiving node to the candidate cooperative nodes. Fields 10 to 11 are the sequence number and data content of the forwarded packet. The business packet is forwarded by the sending node and the cooperative node. When the sending node sends out, the third field is empty. When the cooperative node forwards, the current node ID is filled in. When the receiving node feeds back the receiving result, set the field 4 to 255, indicating that the data has been received successfully, and update the sending timestamp of the receiving node.

The structure of the control packet is shown in Figure 5-b. Fields 1 to 3 are the current node's ID, coordinates, and remaining energy. Field 4 contains the number and IDs of the in-degree nodes. Field 5 contains the number and IDs of the out-degree nodes.

V. PERFORMANCE EVALUATION

A. EXPERIMENTAL ENVIRONMENT AND METHODS

The performance evaluation mainly focuses on the transmission hops, packet delivery rate, transmission delay, network lifetime and remaining energy provided by the SQMCR under different node densities, traffic flow, outage probability, and node location dynamic ranges. The baseline algorithms include VBF, QELAR, and QMCR. The simulation environment is based on the application of underwater temperature monitoring, and the data are from the KEO station in the NOAA database. The three dimensional underwater area is set to 500m × 500m × 500m. The MAC layer protocol is implemented by S-FAMA. The simulation environment is built based on Python. Table 2 shows the settings of the hyper-parameters of SQMCR.

The experiment is divided into 9 scenarios. The environment parameters of scenario (a) (100-Nodes) are set as follows: the number of communication nodes $N_c = 100$, the maximum dynamic range of node position change $P_d = 0$ meters/minute, the outage probability $O_p = 0$ times/packet, and the business flow $B_r = 6$ packets/minute. The environmental parameters of scenario (b) (150-Nodes) is $(N_c = 150, P_d = 0, O_p = 0, B_r = 6)$. The environmental parameters of scenario (c) (200-Nodes) is $(N_c = 200, P_d = 0, O_p = 0, B_r = 6)$.The environmental parameters of scenario (d) (DR-5) is $(N_c = 100, P_d = 5, O_p = 0, B_r = 6)$. The environmental parameters of scenario (e) (IC-0.01) is $(N_c = 100, P_d = 0, O_p = 0.01, B_r = 6)$.The environmental parameters of scenario (f) (TF-2) is $(N_c = 100, P_d = 0, O_p = 0, B_r = 2)$. The environmental parameters of scenario (g) (DR-10) is $(N_c = 100, P_d = 10, O_p = 0, B_r = 6)$.The environmental parameters of scenario (h) (IC-0.1) is $(N_c = 100, P_d = 0, O_p = 0.1, B_r = 6)$. The environmental parameters of scenario (i) (TF-1) is $(N_c = 100, P_d = 0, O_p = 0, B_r = 1)$.

B. TRANSMISSION HOPS COMPARISON

The number of transmission hops refers to the number of nodes experienced by the packets sent from the underwater sensor nodes to the surface sink nodes.

From the overall view of Figure 6, the shortest path algorithm corresponds to the least number of transmission hops and always keeps at 5. However, after less than 50 packets are forwarded, the nodes cannot continue to forward due to the depletion of the remaining energy of the nodes on the transmission route. The number of transmission hops of VBF generally changes dynamically between 5 and 7, but after less than 70 packets are forwarded, it is also because the remaining energy of nodes in the transmission route is exhausted and the node cannot continue to forward. The number of transmission hops of QELAR, QMCR, and SQMCR varies dynamically from 5 to 15. With the increase of the forwarded packets, the number of transmission hops shows an increasing trend. Among them, the number of transmission hops required by SQMCR is less than that of

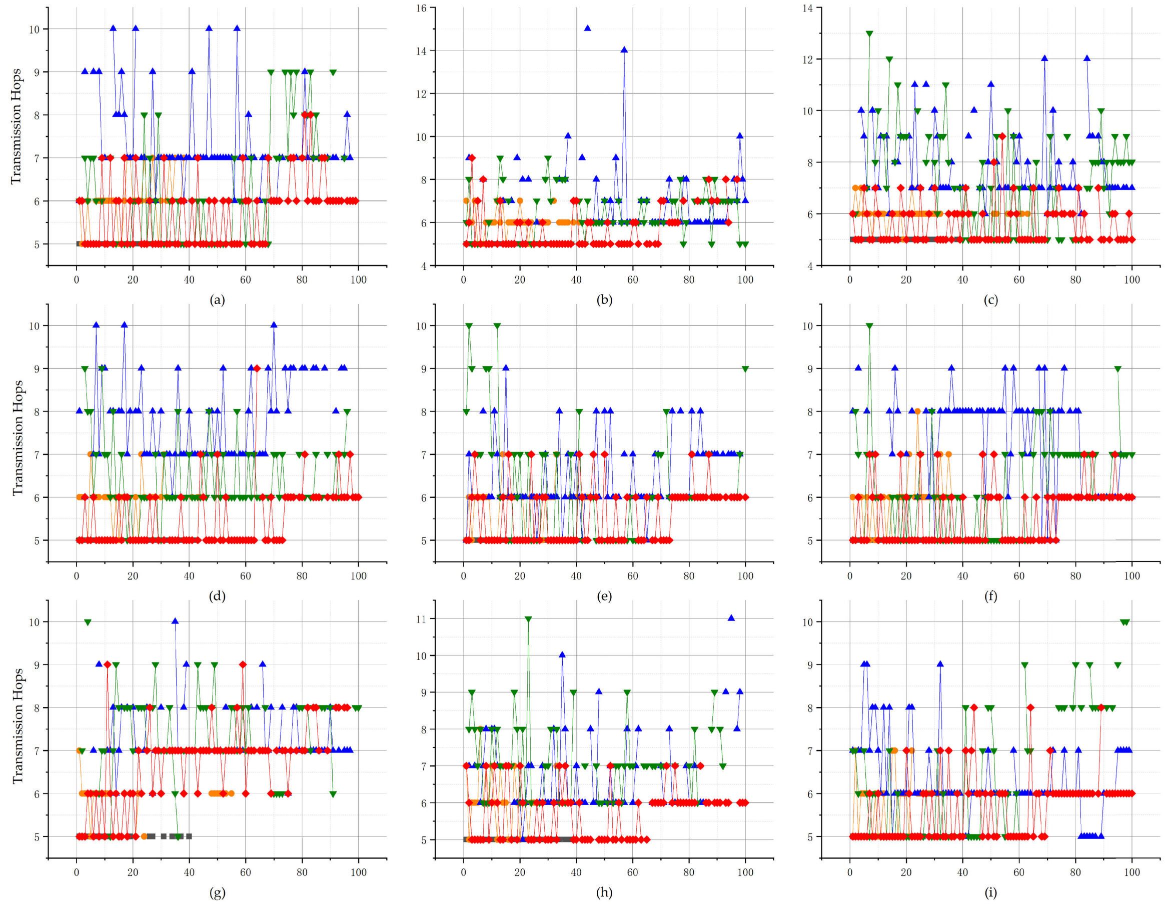

FIGURE 6. Changes in the number of transmission hops for 100 packets sent. Figure 6-a reflects the change in the number of transmission hops for 100 packets sent in the scenario (a). The horizontal axis is the sequence number of packets, and the vertical axis is the number of transmission hops. Figure 6-b reflects the scenario (b), Figure 6-c reflects the scenario (c), Figure 6-d reflects the scenario (d), Figure 6-e reflects the scenario (e), Figure 6-f reflects the scenario (f), Figure 6-g reflects the scenario (g), Figure 6-h reflects the scenario (h), and Figure 6-i reflects the scenario (i). The black (□) line represents the shortest path algorithm. The orange (○) line represents the VBF. The blue (△) line represents the QELAR. The green (▽) line represents the QMCR. The red (◇) line represents the SQMCR.

QELAR and QMCR. The shortest path and VBF based on the determined rules have the advantages of smaller transmission hops and smaller dynamic range, but the disadvantage is that the route is relatively fixed and the distribution of communication energy consumption is uneven. The routing algorithm based on reinforcement learning can adaptively construct and optimize the routes according to the number of hops, remaining energy, delay, and other factors, and can provide packet forwarding service for a long time.

Comparing Figures 6-a, 6-b, and 6-c, with the increase of node density and the change of network topology, the minimum number of the transmission hops required does not change, which is 5. The shortest path and VBF have no obvious packet loss, showing the same change as the whole. For QELAR, QMCR, and SQMCR, with the increase in node density, the number of optional transmission paths increases, and the increased trend in transmission hops slows down. Comparing Figures 6-a, 6-d, and 6-g, as the dynamic range of the node location increases, the shortest path, and VBF have a certain number of packet losses, the dynamic range of transmission hops of QELAR, QMCR, and SQMCR increases, and the overall transmission hops show an increasing trend. The QMCR and SQMCR can reduce the impact of increasing the dynamic range of the node position and maintain a relatively stable change in the number of transmission hops due to the use of cooperative communication. Comparing Figures 6-a, 6-e, and 6-h, with the increase of the outage probability, the shortest path and VBF have a certain number of packet losses, and the dynamic range of transmission hops of QELAR, QMCR, and SQMCR increases. Because the small movements of the node positions will not cause the change in node communication, the influence of the increasing dynamic range of node position has a certain cumulative effect. However, the impact of the increase in the outage probability is rapid, there have been significant changes in the number of transmission

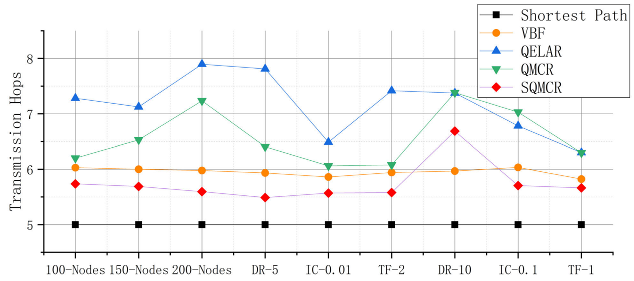

FIGURE 7. Change in the average transmission hops number for 100 packets sent.

hops for QELAR, QMCR, and SQMCR since the initial stage. SQMCR relies on the cooperative communication, the retransmission mechanism and the optimization methods, so the number of transmission hops of SQMCR is still the minimum compared with QELAR and QMCR. Comparing Figures 6-a, 6-f, and 6-i, with the change of packet traffic demand, the number of transmission hops of the shortest path, VBF, QELAR, QMCR, and SQMCR is consistent with the overall trend.

From the overall view of Figure 7, when 100 packets sent, the average transmission hops number of SQMCR is smaller than that of VBF, QELAR, and QMCR in most scenarios, and is close to that of the shortest path algorithm. As different scenarios switched, the change in the average transmission hop number of SQMCR is relatively moderate, which indicates that SQMCR has better robustness. When the dynamic range of node location changes further increases, the average transmission hop number of SQMCR increases slightly, mainly due to the increase in dynamic range of node position change, which leads to the changes of the neighbor relationships and thus affects the transmission hop number.

C. PACKET DELIVERY RATE COMPARISON

Packet delivery rate refers to the ratio of the number of packets received by the sink node to the number of packets sent by the sensor node.

From the overall view of Figure 8, regardless of changes in node density, network topology, node position dynamic range, outage probability and data transmission flow, the SQMCR maintains the highest packet delivery rate, followed by the QMCR, and the QELAR has the lowest packet delivery rate. This shows that the SQMCR can provide higher reliability for packet forwarding in underwater wireless sensor networks.

Comparing the histogram of “100-Nodes”, “150-Nodes” and “200-Nodes” data sub-groups, the packet delivery rates of QELAR, QMCR, and SQMCR all change with the overall trend as the node density and network topology change. Although in the “150-Nodes” and “200-Nodes” data subgroups, the optional transmission routes increase with the increase of node density, due to the change of network topology, the packet delivery rates of QELAR, QMCR, and SQMCR do not increase. The packet delivery rates of QELAR, QMCR, and SQMCR in the “150-Nodes” data subgroup decrease compared with the other two scenarios. Compared with the histogram of “100-Nodes”, “DR-5” and “DR-10” data subgroups, the packet delivery rate of QELAR, QMCR, and SQMCR showed a downward trend as the dynamic range of the node location increased. Compared with QELAR and QMCR, SQMCR has the smallest drop and can always achieve reliable packet forwarding in the case of large changes in network topology. Comparing the histogram of “100-Nodes”, “IC-0.01” and “IC-0.1” data subgroups, with the increase of the outage probability, the packet delivery rate of QELAR, QMCR, and SQMCR showed a downward trend. Compared with QELAR and QMCR, SQMCR has the smallest decrease. Comparing the histogram of “100-Nodes”, “TF-2” and “TF-1” data subgroups, with the reduction of packet traffic, the packet loss caused by congestion is further avoided, and the packet delivery rate of QELAR, QMCR, and SQMCR shows an overall upward trend.

D. TRANSMISSION DELAY COMPARISON

Transmission delay refers to the time from the packet left from the sensor node to the arrival at the sink node.

From the overall view of Figure 9, in most cases, the SQMCR maintains the minimum transmission delay, the QMCR takes the second place, and the QELAR has the maximum transmission delay. Only in the “IC-0.01” and “DR-10” data subgroups, the transmission delay of SQMCR is slightly higher than that of QMCR. SQMCR uses the optimization method based on Q-value initialization, which shortens the training time and reduces the transmission delay. In addition, the consideration of the transmission delay in the SQMCR’s reward function further avoids the possibility of congestion. It shows that SQMCR has high transmission efficiency.

Comparing the histogram of “100-Nodes”, “150-Nodes” and “200-Nodes” data sub-groups, with the increase of node density, the transmission delay of QELAR and SQMCR decreased, while the transmission delay of QMCR fluctuated slightly. As the node density decreases, more and better transmission routes will be generated, which is conducive to the reduction of transmission delay. Comparing the histogram of “100-Nodes”, “DR-5” and “DR-10” data subgroups, with the increase of the dynamic range of node location, the transmission delay of QMCR and SQMCR shows an upward trend, while the transmission delay of QELAR fluctuates slightly. As the dynamic range of node position increases and the transmission route changes, the number of transmission hops increases, the packet loss is serious, the probability of packet retransmission increases, and the transmission delay increases. Comparing the histogram of “100-Nodes”, “IC-0.01” and “IC-0.1” data subgroups, with the increase of the outage probability, the transmission delay of QMCR and SQMCR shows an upward trend, while the transmission delay of QELAR fluctuates slightly. As the outage probability increases, the packet loss is serious, which increases the probability of packet retransmission, and the transmission

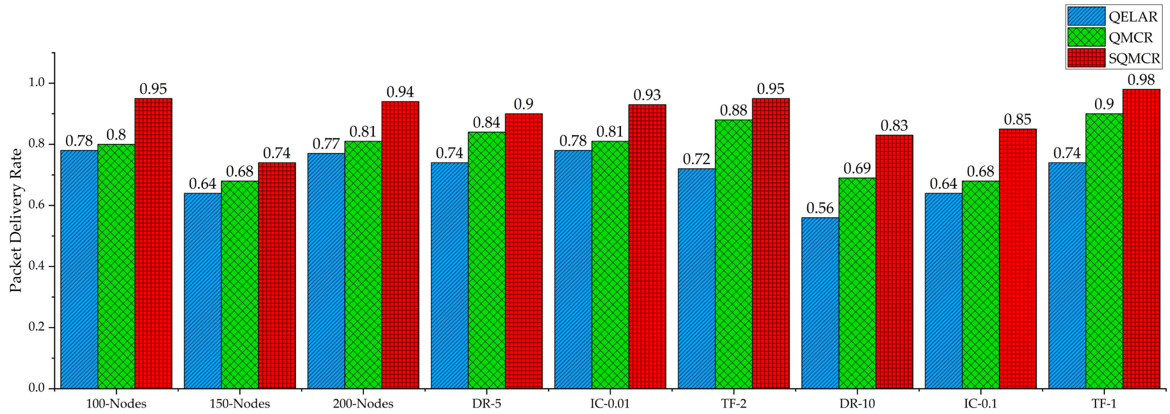

FIGURE 8. Change in the packet delivery rate of 100 packets sent.

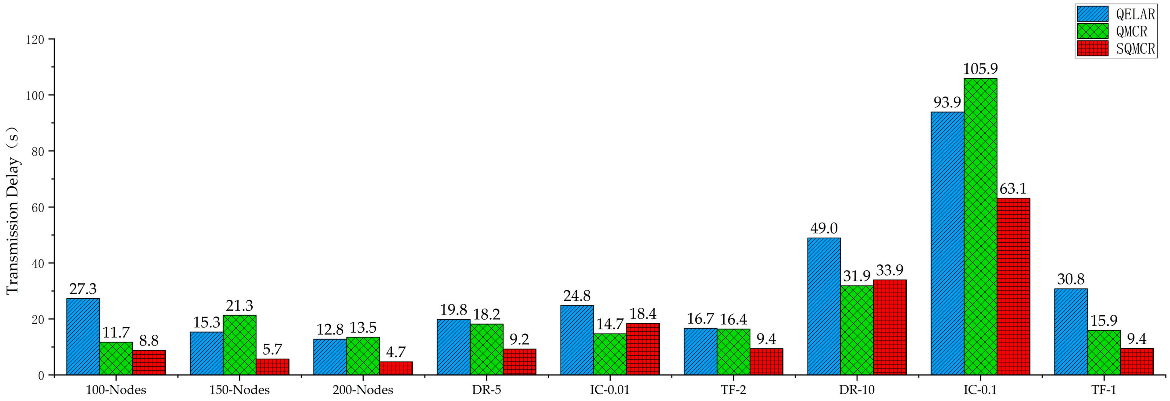

FIGURE 9. Change in the transmission delay of 100 packets sent.

delay increases. Among them, the outage probability of 0.1 times/packet has the greatest impact on the transmission delay. Compared with the histogram of "100-Nodes", "TF-2" and "TF-1" data subgroups, the overall transmission delay of QELAR, QMCR, and SQMCR changes little with the reduction of packet traffic.

E. NETWORK LIFETIME COMPARISON

Network lifetime refers to the time from the first packet sent by the sensor node to the last packet received by the sink node. The network lifetime is determined by the maximum sequence number of the packets that can be received by the sink node. The maximum packet forwarding capacity of the network refers to the number of packets forwarded by the underwater network within the network lifetime, which represents the actual ability of the underwater network to provide packet forwarding services.

In Figure 10, compared with QELAR and QMCR, overall, SQMCR has the largest number of received packets, the highest packet delivery rate, and the highest maximum sequence number of received packets. It shows that SQMCR maximizes the efficiency of the communication energy utilization among the nodes and can provide reliable packet transmission services over a longer time range. Only in some cases, the highest maximum sequence number of SQMCR is slightly inferior to that of QELAR. Because QELAR does not adopt cooperative communication, it can save a certain of communication energy and prolong the network life, but it cannot achieve more reliable packets forwarding. QMCR uses the determined rules to implement cooperative communication. Although it will improve the packet delivery rate, it wastes the remaining energy of nodes and reduces the network lifetime when cooperative communication is not required. SQMCR controls cooperative communication based on Q-learning, which not only achieves better packet forwarding benefits but also avoids unnecessary communication energy consumption and prolongs the network lifetime.

Comparing Figures 10-a, 10-b, and 10-c, with the increase of node density, the maximum sequence number of received packets and the number of received packets in the underwater networks using QELAR, QMCR, and SQMCR increases, but the packet delivery rate decreases. This is because the node density increases, the available transmission routes increase, and the network life is extended. With the easy-to-use transmission routes gradually consumed, the packet delivery rate is reduced. Compared with Figures 10-a, 10-d, and 10-g, as the dynamic range of node location increases, the maximum sequence number of received packets in underwater networks using QELAR, QMCR, and SQMCR is not significantly affected, but the number of received packets

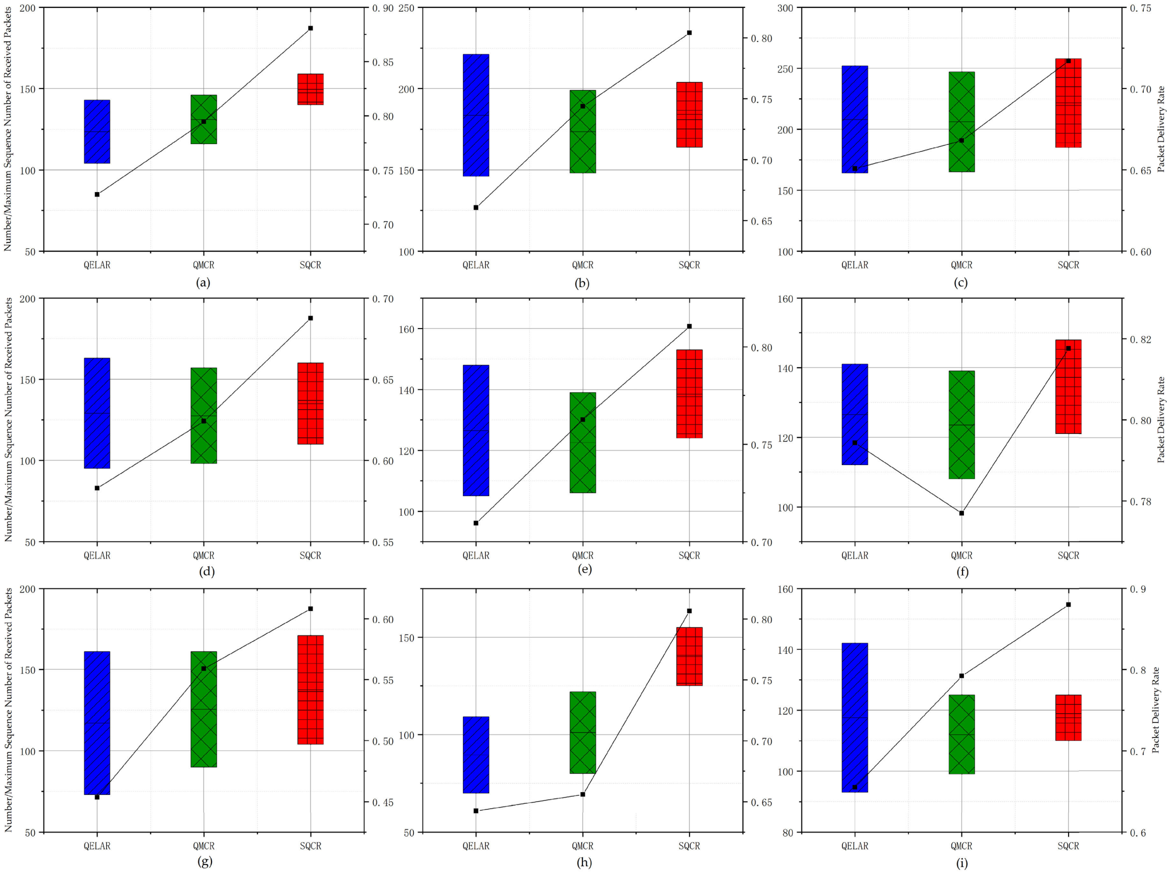

FIGURE 10. Changes in the number of received packets, maximum sequence number of received packets, and packet delivery rate within the network lifetime. Figure 10-a shows the number of received packets, the maximum sequence number of received packets, and the change of packet delivery rate, within the network lifetime in the scenario (a). The lower edge of the box is the number of received packets, and the upper edge of the box is the maximum sequence number of received packets, identified by the left vertical axis. The black dotted line reflects the change in packet delivery rate for different routing algorithms and is identified by the right vertical axis. The Figure 10-b reflects the scenario (b), Figure 10-c reflects the scenario (c), Figure 10-d reflects the scenario (d), Figure 10-e reflects the scenario (e), Figure 10-f reflects f, Figure 10-g reflects the scenario (g), Figure 10-h reflects the scenario (h), and Figure 10-i reflects the scenario (i). The blue box reflects QELAR, the green box reflects QMCR, the red box reflects SQMCR and the three identification points in the black line reflect QELAR, QMCR, and SQMCR respectively.

and the packet delivery rate are reduced. Compared with Figures 10-a, 10-e, and 10-h, with the increase of outage probability, the maximum sequence number of received packets, the number of received packets, and the packet delivery rate are all reduced in underwater networks using QELAR, QMCR, and SQMCR. Comparing Figures 10-a, 10-f, and 10-i, with the change of packet traffic demand, the maximum sequence number of received packets, the number of received packets, and the packet delivery rate in the underwater network using QELAR, QMCR, and SQMCR are not significantly affected.

F. REMAINING ENERGY COMPARISON

When the remaining energy of a node exceeds the energy required for transmission, the node can participate in packet forwarding and is called an alive node. When the remaining energy of a node is lower than the energy required for transmission, the node cannot participate in packet forwarding and is called a dead node. When there are dead nodes in each transmission route of the network, the network lifetime ends. Energy tax refers to the average of energy required for each packet forwarded over the network lifetime relative to all nodes.

In Figure 11, compared with QELAR and QMCR, overall, SQMCR has the lowest energy tax. It shows that SQMCR requires less energy consumption for forwarding each packet and has higher forwarding efficiency. Effective cooperative communication control not only reduces the energy consumption generated by unnecessary cooperative communication, but also reduces the energy consumption caused by packet retransmission. The energy consumption generated by unnecessary cooperative communication makes the energy tax of QMCR the highest, greatly reducing the forwarding efficiency of QMCR. At lifetime end of the networks configured by QELAR, QMCR, and SQMCR separately, the number of alive and dead nodes is similar. It indicates that the three routing algorithms based on reinforcement learning use the similar key-nodes in the transmission routes. As the number of packets forwarding increases, some nodes on the transmission route become

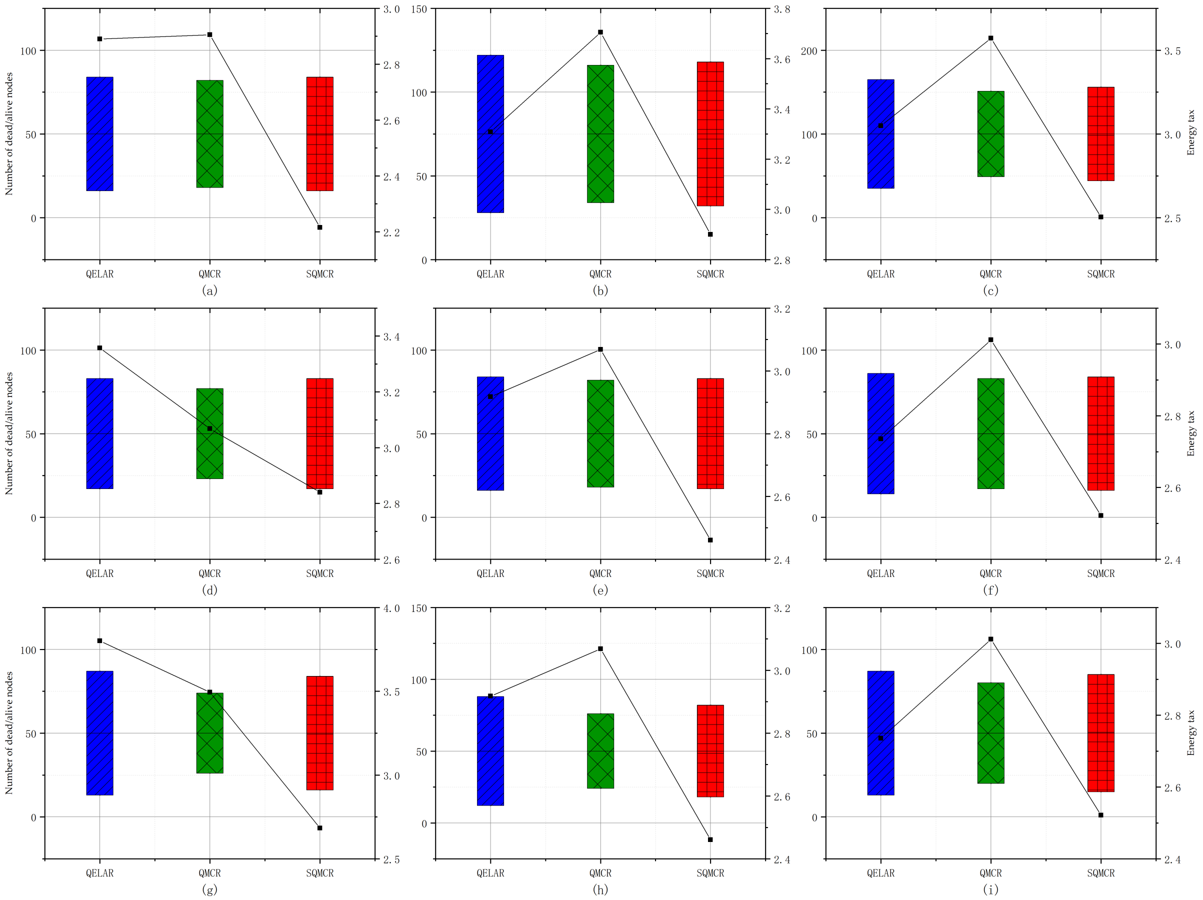

FIGURE 11. Changes in the number of dead/alive nodes and the energy tax. Figure 11-a shows the number of dead/alive nodes and the energy tax in the scenario (a). The lower edge of the box is the number of dead nodes, and the upper edge of the box is the number of alive nodes, identified by the left vertical axis. The black dotted line reflects the change in energy tax for different routing algorithms and is identified by the right vertical axis. The Figure 11-b reflects the scenario (b), Figure 11-c reflects the scenario (c), Figure 11-d reflects the scenario (d), Figure 11-e reflects the scenario (e), Figure 11-f reflects the scenario (f), Figure 11-g reflects the scenario (g), Figure 11-h reflects the scenario (h), and Figure 11-i reflects the scenario (i). The blue box reflects QELAR, the green box reflects QMCR, the red box reflects SQMCR and the three identification points in the black line reflect QELAR, QMCR, and SQMCR respectively.

dead nodes. The number of alive nodes for the QMCR is relatively lower, while the number of dead nodes is relatively higher, which is because some nodes frequently participate in cooperative communication, leading to the energy depletion.

Comparing Figures 11-a, 11-b, and 11-c, with the increase of node density, the number of dead and alive nodes in the underwater networks using QELAR, QMCR, and SQMCR increase, but the energy tax dynamically change with the changes in network topology. It is because with the node density increases, more nodes participate in packet forwarding, resulting in the increase of the number of alive and dead nodes. As the node density and the packet delivery rate increases, the energy consumption caused by the unnecessary cooperative communication of QMCR becomes more significant which make the energy tax for QMCR the highest. Compared with Figures 11-a, 11-d, and 11-g, as the dynamic range of node location increases, the number of dead and alive nodes in underwater networks using QELAR, QMCR, and SQMCR are not significantly affected, but the overall of energy tax has been improved to a certain extent. Retransmission results in the significant energy consumption, especially for QELAR. Compared with Figures 11-a, 11-e, and 11-h, in the underwater networks using SQMCR, with the increase of outage probability, the number of dead nodes has slightly increased, while the number of alive nodes has slightly decreased, indicating that more nodes are participating in the packet forwarding. And with the increase of outage probability, the energy tax for SQMCR is still the lowest compared to QELAR and QMCR, and the energy tax for QELAR has significantly increased, due to the increase in retransmission times. Comparing Figures 11-a, 11-f, and 11-i, with the change of packet traffic demand, the number of dead and alive nodes in the underwater network using QELAR, QMCR, and SQMCR vary very little, the energy tax for QELAR, QMCR, and SQMCR have slightly increased as the interval of packet forwarding increases.

VI. DISCUSSION

SQMCR uses multi-hop cooperative communication to improve the reliability of underwater packet forwarding. Compared with the single-hop communication, the multi-hop communication can provide a larger transmission bandwidth and longer transmission distance. Compared with point-to-point communication, cooperative communication can avoid packet forwarding outages caused by fading, multipath, occlusion, and the Doppler effect. The simulation results show that it is not difficult to find that the underwater wireless sensor network configured with SQMCR can provide a higher packet delivery rate, regardless of the conditions of different node densities, different maximum dynamic ranges of node positions, different outage probability and different packet traffic. Compared with the baseline algorithms, the packet delivery rate provided by SQMCR is more than 17% higher.

SQMCR uses the Stackelberg game to coordinate the relationship among the sending node, the receiving node, and the candidate cooperative nodes. The goal is to achieve reliable packet forwarding, reduce and balance the communication energy consumption, and extend the lifetime of the underwater networks. When selecting the transmission route, the sending node will comprehensively consider the location, remaining energy, transmission delay, historical forwarding experiences of the candidate receiving node, and whether the candidate receiving node has enough suitable neighbor nodes to act as the cooperative nodes, to improve the reliability of packet forwarding. When selecting whether to participate in cooperative forwarding, the candidate cooperative nodes will comprehensively consider the distance to the receiving node, the historical forwarding status between the sending node and the receiving node, the number of candidate cooperative nodes, as well as its remaining energy and its advantages in the candidate cooperative nodes set, to ensure reliable packet forwarding and minimize communication energy consumption. As the leader, the sending node considers the followers before making the transmission routing decisions. The candidate cooperative nodes, as the followers, choose whether to participate in the cooperative communication according to the leader's decision and its conditions. Therefore, compared with the cooperative communication method based on the determined rule, the cooperative communication decision making method based on the Stackelberg Q-learning can further improve the rationality of the cooperative routing selection, and further approach the Nash equilibrium point of the communication benefits and energy consumption costs. The simulation results show that it is also not difficult to find that the underwater wireless sensor network with SQMCR can provide more durable and reliable packet forwarding services compared with the baseline algorithms under different conditions. The SQMCR can reduce the energy tax by 23% and increase the number of received packets by 17%, during the network lifetime.

SQMCR also adopts the method of Q-value initialization and dynamic exploration probability, which further improves the convergence speed and stability of the routing algorithm.

VII. CONCLUSION

The underwater wireless sensor network based on SQMCR can provide more reliable packet forwarding services as much as possible based on ensuring efficient packet forwarding services. The SQMCR realizes the long-distance and broadband packet forwarding in the underwater network based on multi-hop cooperative communication. By coordinating the relationships between the sending node, the receiving node, and the candidate cooperative nodes based on the Stackelberg game, the SQMCR learns the best routing policy and cooperative communication policy with the optimized Q-learning. The simulation results show that the SQMCR can help the underwater wireless sensor network increase the packet delivery rate and the maximum packet forwarding capacity of the network by 17%, with better environment and application adaptability. Therefore, SQMCR is more suitable for underwater high-reliability applications.

REFERENCES

[1] S. Jiang, "Networking in oceans: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–33, Jan. 2022.

[2] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, Jan. 2020.

[3] 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.

[4] 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.

[5] Y. Ma, Q. Zhang, and H. Wang, "6G: Ubiquitously extending to the vast underwater world of the oceans," *Engineering*, vol. 8, pp. 12–17, Jan. 2022.

[6] 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.

[7] A. Boukerche and P. Sun, "Design of algorithms and protocols for underwater acoustic wireless sensor networks," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–34, Nov. 2021.

[8] S. Li, W. Qu, C. Liu, T. Qiu, and Z. Zhao, "Survey on high reliability wireless communication for underwater sensor networks," *J. Netw. Comput. Appl.*, vol. 148, Dec. 2019, Art. no. 102446.

[9] M. Pundir and J. K. Sandhu, "A systematic review of quality of service in wireless sensor networks using machine learning: Recent trend and future vision," *J. Netw. Comput. Appl.*, vol. 188, Aug. 2021, Art. no. 103084.

[10] 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.

[11] 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.

[12] Y. Yuan, M. Liu, X. Zhuo, Y. Wei, X. Tu, and F. Qu, "A Q-learning-based hierarchical routing protocol with unequal clustering for underwater acoustic sensor networks," *IEEE Sensors J.*, vol. 23, no. 6, pp. 6312–6325, Mar. 2023.

[13] 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.

[14] C. S. Nandyala, H.-W. Kim, and H.-S. Cho, "QTAR: A Q-learning-based topology-aware routing protocol for underwater wireless sensor networks," *Comput. Netw.*, vol. 222, Feb. 2023, Art. no. 109562.

[15] H. Chang, J. Feng, and C. Duan, "Reinforcement learning-based data forwarding in underwater wireless sensor networks with passive mobility," *Sensors*, vol. 19, no. 2, p. 256, Jan. 2019.

[16] 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.

[17] 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, no. 3, pp. 2756–2770, Mar. 2021.

[18] Z. Jin, Q. Zhao, and Y. Su, "RCAR: A reinforcement-learning-based routing protocol for congestion-avoided underwater acoustic sensor networks," *IEEE Sensors J.*, vol. 19, no. 22, pp. 10881–10891, Nov. 2019.

[19] C. Wang, X. Shen, H. Wang, H. Zhang, and H. Mei, "Reinforcement learning-based opportunistic routing protocol using depth information for energy-efficient underwater wireless sensor networks," *IEEE Sensors J.*, vol. 23, no. 15, pp. 17771–17783, Aug. 2023, doi: 10.1109/JSEN.2023.3285751.

[20] Y. Chen, J. Zhu, L. Wan, S. Huang, X. Zhang, and X. Xu, "ACOA-AFSA fusion dynamic coded cooperation routing for different scale multi-hop underwater acoustic sensor networks," *IEEE Access*, vol. 8, pp. 186773–186788, 2020.

[21] J. Zhu, Y. Chen, X. Sun, J. Wu, Z. Liu, and X. Xu, "ECRKQ: Machine learning-based energy-efficient clustering and cooperative routing for mobile underwater acoustic sensor networks," *IEEE Access*, vol. 9, pp. 70843–70855, 2021.

[22] W. Guo and W. Zhang, "A survey on intelligent routing protocols in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 38, pp. 185–201, Feb. 2014.

[23] X. Geng and B. Zhang, "Deep Q-network-based intelligent routing protocol for underwater acoustic sensor network," *IEEE Sensors J.*, vol. 23, no. 4, pp. 3936–3943, Feb. 2023.

[24] X. Wei, H. Guo, X. Wang, X. Wang, and M. Qiu, "Reliable data collection techniques in underwater wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 404–431, 1st Quart., 2022.

[25] Z. Shen, H. Yin, L. Jing, Y. Liang, and J. Wang, "A cooperative routing protocol based on Q-learning for underwater optical-acoustic hybrid wireless sensor networks," *IEEE Sensors J.*, vol. 22, no. 1, pp. 1041–1050, Jan. 2022.

[26] Y. Su, M. Liwang, Z. Gao, L. Huang, X. Du, and M. Guizani, "Optimal cooperative relaying and power control for IoUT networks with reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 791–801, Jan. 2021.

[27] M. Ismail, M. Islam, I. Ahmad, F. A. Khan, A. B. Qazi, Z. H. Khan, Z. Wadud, and M. Al-Rakhami, "Reliable path selection and opportunistic routing protocol for underwater wireless sensor networks," *IEEE Access*, vol. 8, pp. 100346–100364, 2020.

[28] Y. Zhao, R. Adve, and T. J. Lim, "Symbol error rate of selection amplify-and-forward relay systems," *IEEE Commun. Lett.*, vol. 10, no. 11, pp. 757–759, Nov. 2006.

[29] C. Cheng, Z. Zhu, B. Xin, and C. Chen, "A multi-agent reinforcement learning algorithm based on Stackelberg game," in *Proc. 6th Data Driven Control Learn. Syst. (DDCLS)*, M. Sun and H. Gao, Eds., May 2017, pp. 727–732.

[30] P. Xie, J.-H. Cui, and L. Lao, "VBF: Vector-based forwarding protocol for underwater sensor networks," in *Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, vol. 3976, F. Boavida, T. Plagemann, B. Stiller, C. Westphal, and E. Monteiro, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 1216–1221.

[31] 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* (Lecture Notes in Computer Science), vol. 4982. Berlin, Germany: Springer-Verlag, 2008, pp. 72–86.

[32] Y. Chen, K. Zheng, X. Fang, L. Wan, and X. Xu, "QMCR: A Q-learning-based multi-hop cooperative routing protocol for underwater acoustic sensor networks," *China Commun.*, vol. 18, no. 8, pp. 224–236, Aug. 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] B. Wang, K. Ben, H. Lin, M. Zuo, and F. Zhang, "EP-ADTA: Edge prediction-based adaptive data transfer algorithm for underwater wireless sensor networks (UWSNs)," *Sensors*, vol. 22, no. 15, p. 5490, Jul. 2022.

[35] Y. Zhang, Y. Su, X. Shen, A. Wang, B. Wang, Y. Liu, and W. Bai, "Reinforcement learning based relay selection for underwater acoustic cooperative networks," *Remote Sens.*, vol. 14, no. 6, p. 1417, Mar. 2022.

[36] S. Jiang, "Wireless networking principles: From terrestrial to underwater acoustic," in *Wireless Networking Principles: From Terrestrial To Underwater Acoustic*. Cham, Switzerland: Springer, 2018, pp. 233–243, doi: 10.1007/978-981-10-7775-3.

[37] M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 84–89, Jan. 2009.

[38] D. D. Tan, T. T. Le, and D.-S. Kim, "Distributed cooperative transmission for underwater acoustic sensor networks," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Apr. 2013, pp. 205–210.

[39] D. Liu, Y. Chen, T. Zhang, K. K. Chai, J. Loo, and A. Vinel, "Stackelberg game based cooperative user relay assisted load balancing in cellular networks," *IEEE Commun. Lett.*, vol. 17, no. 2, pp. 424–427, Feb. 2013.

[40] C. Zhu, W. Yu, and H. Wang, "A multi-agent Q-learning with value function approximation based on single-leader multi-followers stackelberg game," in *Proc. IEEE 13th Int. Conf. CYBER Technol. Autom., Control, Intell. Syst. (CYBER)*, Jul. 2023, pp. 34–1229.

WANG BIN was born in Yuci, Shanxi, China, in 1978. He received the M.S. degree in communication and information systems from the Naval University of Engineering, Wuhan, China, in 2004, where he is currently pursuing the Ph.D. degree. He is a Lecturer with the Naval University of Engineering. He has published two monographs, written 12 articles, and one patent. Since 2019, he has been mainly engaged in the research of underwater wireless sensor networks and intelligent networks.

BEN KERONG was born in 1963. He is currently a Ph.D. Supervisor with the Naval University of Engineering. His current research interests include software quality assurance and artificial intelligence. He is a Senior Member of China Computer Federation.

HAO YIXUE (Member, IEEE) received the Ph.D. degree in computer science from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2017. He is currently an Associate Professor with the School of Computer Science and Technology, HUST. His current research interests include 5G networks, the Internet of Things, edge computing, edge caching, and cognitive computing.

ZUO MINGJIU was born in Huangshi, Hubei, China, in 1979. He received the B.S. degree in automation from Central South University, Changsha, China, in 2002, the M.S. degree in control theory and control engineering from Huazhong University of Science and Technology, Wuhan, China, in 2008, and the Ph.D. degree in marine engineering from the School of Shipping and Oceanography, Huazhong University of Science and Technology, in 2016. Since 2008, he has been engaged in teaching and scientific research with the College of Electronic Engineering, Naval University of Engineering, Wuhan. He has translated and published two monographs, written 15 articles, and three patents. His research interests include submarine networks, unmanned underwater systems, and underwater operation technology.