Letter

Secure Underwater Distributed Antenna Systems: A Multi-Agent Reinforcement Learning Approach

Chaofeng Wang, Zhicheng Bi, and Yaping Wan

Dear Editor,

Underwater distributed antenna systems (DAS) are stationary infrastructures consisting of multiple geographically distributed antenna elements (DAEs) which are interconnected through high-rate backbone networks [1]. Compared to centralized systems, the DAS could provide a larger coverage area and higher throughput for underwater acoustic (UWA) transmissions. In this work, exploiting the low sound speed in water, a multi-agent reinforcement learning (MARL)-based approach is proposed to secure underwater DAS against eavesdropping at the physical layer. Specifically, the theoretical secrecy rate is firstly derived for time-slotted UWA networks (UWANs) considering the large propagation delays. Furthermore, we investigate the long-term sum secrecy rate optimization problem under the MARL framework, where each DAE learns its optimal transmission strategy online. Simulation results show that the proposed method achieves higher secrecy performance compared to competing benchmark methods.

Typical physical-layer security approaches against eavesdropping in UWANs include: secret key generation based on the randomness of UWA channels [2]; cooperative jamming where the transmission strategies of friendly jammers are optimized to blind the eavesdropper (EVE) [3]; and secure coordinated multipoint (CoMP) transmissions to enforce time-domain self-interference at the EVE [4], [5]. Particularly about the secure CoMP transmissions, by coordinated transmission scheduling of multiple DAEs with low sound speed in water, the decoding performance of the EVE can be significantly suppressed by collisions of useful signals while the signals received by the legitimate user (LU) are collision-free. However, this type of security mechanism is effective with specific transmission protocols and cannot be applied directly to general UWANs. In addition, the CoMP transmissions heavily rely on efficient coordination of all the transmitters while sharing information with UWA transmissions results in great overhead and latency. In this work, we study the security enhancement against eavesdropping for time-slotted UWANs with CoMP transmissions by taking advantage of the low coordination cost of underwater DAS.

Reinforcement learning (RL) has been leveraged to secure terrestrial radio networks against eavesdropping at the physical layer [6]–[8]. The basic idea is to let the system learn optimal transmission strategies, e.g., transmitting nodes, transmission power, or beamforming vectors, to maximize the secrecy performance through dynamically interacting with environments. However, due to the non-negligible transmission latency of UWA transmissions, those methods cannot be directly applied to UWANs. Although RL has been introduced to secure UWANs with privacy-preserving localization [9] and anti-jamming relay design [10], to the best of our knowledge, there is no work that exploits RL to secure UWA transmissions against eavesdropping at the physical layer. Hence, considering the large propagation delays, we propose an MARL-based framework to secure the underwater DAS, where all the DAEs coordinately learn their transmission strategies online to improve the network secrecy performance.

System model and secrecy rate: We first consider an underwater system where N DAEs coordinate with each other to transmit signal blocks to a LU while an EVE collects transmitted signals from the DAEs. In this study, we consider that the EVE's location information is known *a priori* to the DAS, which has also been considered in many existing works, e.g., [5], [6] and [9]. The underwater system operates in a slotted-based manner. Specifically, in each time slot, each DAE decides its transmission strategy including whether to transmit one signal block to the LU and the transmission power of each block. Denote $d_\mu(\ell) \in \{0,1\}$ as the transmission schedule in the ℓth time slot. $d_\mu(\ell) = 1$ indicates that the μth DAE is active while $d_\mu(\ell) = 0$ implies that the μth DAE keeps silent. Denote $p_\mu(\ell)$ as the transmission power of the signal block sent by the μth DAE in the ℓth time slot. We further assume that if $d_\mu(\ell) = 0$, the transmission power $p_\mu(\ell) = 0$. Denote g_μ and $g_\mu^{(\mathrm{e})}$ as the transmission losses from μth DAE to the LU and the EVE, respectively. The transmission losses depend on the transmission center frequency of the acoustic signals and the propagation distances of transmission links.

Denote D_μ and $D_\mu^{(\mathrm{e})}$ as the propagation delays measured in time slots from the μth DAE to the LU and the EVE, respectively. Denote $\mathcal{I}(\ell)$ and $\mathcal{I}^{(\mathrm{e})}(\ell)$ as interfering DAE sets indicating the DAEs whose transmitted blocks are interfered with each other at the LU and the EVE in the ℓth time slot, respectively. An illustration of the sets $\mathcal{I}(\ell)$ and $\mathcal{I}^{(\mathrm{e})}(\ell)$ is shown in Fig. 1. It is worth noting that the two sets $\mathcal{I}(\ell)$ and $\mathcal{I}^{(\mathrm{e})}(\ell)$ are *unknown* until the end of ℓth time slot.

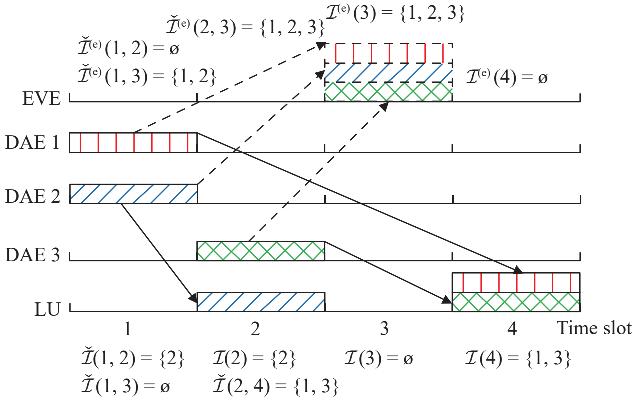
$\mathcal{I}^{(\mathrm{e})}(3) = \{1, 2, 3\}$
$\check{\mathcal{I}}^{(\mathrm{e})}(2, 3) = \{1, 2, 3\}$
$\check{\mathcal{I}}^{(\mathrm{e})}(1, 2) = \varnothing$
$\check{\mathcal{I}}^{(\mathrm{e})}(1, 3) = \{1, 2\}$
$\mathcal{I}^{(\mathrm{e})}(4) = \varnothing$
EVE
DAE 1
DAE 2
DAE 3
LU
1
2
3
4
Time slot
$\check{\mathcal{I}}(1, 2) = \{2\}$
$\check{\mathcal{I}}(1, 3) = \varnothing$
$\mathcal{I}(2) = \{2\}$
$\check{\mathcal{I}}(2, 4) = \{1, 3\}$
$\mathcal{I}(3) = \varnothing$
$\mathcal{I}(4) = \{1, 3\}$

Fig. 1. An illustration of secure CoMP transmissions against eavesdropping in underwater DAS with large propagation delays and examples of interfering DAE sets. $\mathcal{I}^{(\mathrm{e})}(3) = \{1,2,3\}$ indicates that, in the 3rd time slot, the signal blocks from DAEs 1−3 interfere with each other at the EVE while $\mathcal{I}(4) = \{1,3\}$ indicates that, in the 4th time slot, the signal blocks from DAEs 1 and 3 are interfered with each other at the LU. $\check{\mathcal{I}}^{(\mathrm{e})}(1,3) = \{1,2\}$ shows that, when viewed from the end of the 1st time slot, the DAEs 1 and 2 will be interfered at the EVE in the coming 3rd time slot while $\check{\mathcal{I}}(1,2) = \{2\}$ shows that, when viewed from the end of the 1st time slot, the LU will only receive a signal block from DAE 2 in the 2nd time slot.

If the μth DAE decides to transmit in the time slot ℓ, its transmitted signal block will be received by the LU and the EVE in the $(\ell{+}D_\mu)$th and the $(\ell{+}D_\mu^{(\mathrm{e})})$th time slots, respectively. Hence, the set of DAEs interfered with the μth DAE at the LU and the EVE can be described by $\mathcal{I}(\ell{+}D_\mu)$ and $\mathcal{I}^{(\mathrm{e})}(\ell{+}D_\mu^{(\mathrm{e})})$, respectively. If $\nu \in \mathcal{I}(\ell{+}D_\mu)|_{\nu\neq\mu}$, there must be one signal block transmitted by the νth DAE in the $(\ell{+}D_\mu{-}D_\nu)$th time slot. Based on [3], considering single-block pro-

Corresponding author: Chaofeng Wang.

Citation: C. F. Wang, Z. C. Bi, and Y. P. Wan, "Secure underwater distributed antenna systems: A multi-agent reinforcement learning approach," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 7, pp. 1622–1624, Jul. 2023.

The authors are with the School of Computer Science, University of South China, Hengyang 421001, China (e-mail: cwang@usc.edu.cn; zcbi@stu.usc.edu.cn; ypwan@usc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123366

cessing, the signal-to-interference-and-noise ratio (SINR) of the received signal block at the LU sent by the μth DAE in the ℓth time slot can be formulated as

$$\lambda_\mu(\ell;\mathcal{I}(\ell+D_\mu)) := \frac{p_\mu(\ell)g_\mu}{\sigma_n^2+\sum_{\nu\in\mathcal{I}(\ell+D_\mu),\nu\neq\mu}p_\nu(\ell+D_\mu-D_\nu)g_\nu} \tag{1}$$

where $\lambda_\mu(\ell;\mathcal{I}(\ell+D_\mu))$ is conditioned on the set $\mathcal{I}(\ell+D_\mu)$, σ_n is the power of background noise, and $p_\nu(\ell+D_\mu-D_\nu)$ corresponds to the transmission power performed by the νth DAE which interferes with the μth DAE in the $(\ell+D_\mu)$th time slot. Similarly, the SINR of the same signal block received by the EVE can be obtained by

$$\lambda_\mu^{(\mathrm{e})}(\ell;\mathcal{I}^{(\mathrm{e})}(\ell+D_\mu^{(\mathrm{e})})) := \frac{p_\mu(\ell)g_\mu^{(\mathrm{e})}}{\sigma_n^2+\sum_{\nu\in\mathcal{I}^{(\mathrm{e})}(\ell+D_\mu^{(\mathrm{e})}),\nu\neq\mu}p_\nu(\ell+D_\mu^{(\mathrm{e})}-D_\nu^{(\mathrm{e})})g_\nu^{(\mathrm{e})}} \tag{2}$$

where $\lambda_\mu^{(\mathrm{e})}(\ell;\mathcal{I}^{(\mathrm{e})}(\ell+D_\mu^{(\mathrm{e})}))$ is conditioned on the set $\mathcal{I}^{(\mathrm{e})}(\ell+D_\mu^{(\mathrm{e})})$.

To derive the theoretical secrecy rate (SR), we first assume that the global information of transmission strategy is available. For instance, $\{d_\mu(\ell)\}_{\ell=0}^{\infty}$ is assumed to be known in order to obtain the sets $\mathcal{I}(\ell)$ and $\mathcal{I}^{(\mathrm{e})}(\ell)$. Based on the SINRs defined in (1) and (2), the SR of the transmitted signal block from the μth DAE in the ℓth time slot can be cast as

$$\begin{aligned} C_\mu(\ell) = \frac{1}{2}\Big[&\log(1+\lambda_\mu(\ell;\mathcal{I}(\ell+D_\mu))) \\ &-\log(1+\lambda_\mu^{(\mathrm{e})}(\ell;\mathcal{I}^{(\mathrm{e})}(\ell+D_\mu^{(\mathrm{e})})))\Big]^+ \end{aligned} \tag{3}$$

where $[\cdot]^+=\max\{0,\cdot\}$. The SR depicts how much secure information can be delivered to the LU, taking into account the information leakage to the EVE. To prevent the underwater DAS against eavesdropping, by properly determining the transmission strategy including the transmission schedule $d_\mu(\ell)$ and transmission power $p_\mu(\ell)$ of each DAE, an optimization problem to maximize the long-term sum SR of all the DAEs can be cast as

$$\max_{\{d_\mu(\ell),p_\mu(\ell):1\leq\mu\leq N\}_{\ell=0}^{\infty}} \sum_{\ell=0}^{\infty}\sum_{\mu=1}^{N}C_\mu(\ell) \tag{4a}$$

$$\text{s.t.} \quad \lambda_\mu(\ell;\mathcal{I}(\ell+D_\mu))\geq\Gamma_{\mathrm{th}},\ \text{if}\ d_\mu=1 \tag{4b}$$

where Γ_{th} is a predetermined decoding threshold. Inequality (4b) ensures that the received block at the LU can be successfully decoded. The optimization problem (4) is a mixed-integer nonlinear programming problem combined with sequential decision-making, which in general is difficult to solve. Specifically, without the global information $\{\mathcal{I}(\ell),\mathcal{I}^{(\mathrm{e})}(\ell)\}_{\ell=0}^{\infty}$, its decision space is notably large, which results in the curse of dimensionality (CoD) and suboptimal solutions. In this work, we adopt RL to solve (4) in a tractable way by learning the transmission strategies of all the DAEs online.

RL reformulation: To reformulate (4) in the RL paradigm, a Markov decision process (MDP) $\langle\mathcal{A},\mathcal{S},T(\mathbf{S},\mathbf{A},\mathbf{S}'),R(\mathbf{S},\mathbf{A})\rangle$ must be defined where $\mathcal{A}$ is the action space, $\mathcal{S}$ is the system state space, $T(\mathbf{S},\mathbf{A},\mathbf{S}')$ is the state transition function indicating how the system state evolves from the states $\mathbf{S}$ to $\mathbf{S}'$ by performing the action $\mathbf{A}$, where $\mathbf{S},\mathbf{S}'\in\mathcal{S}$ and $\mathbf{A}\in\mathcal{A}$, and $R(\mathbf{S},\mathbf{A})$ is the reward obtained by performing the action $\mathbf{A}$ under the system state $\mathbf{S}$. Considering that each DAE acts as an RL agent, the action, state, and reward function are defined as follows.

1) Action: Denote $\mathbf{a}_\mu(\ell)$ as the transmission action for the μth DAE in the ℓth time slot. It consists of the transmission schedule and transmission power, i.e., $\mathbf{a}_\mu(\ell):=\{d_\mu(\ell),p_\mu(\ell)\}$. Denote $\mathbf{A}(\ell)$ as the action of the DAS in the ℓth time slot which includes actions from all the DAEs, i.e., $\mathbf{A}(\ell)=\{\mathbf{a}_1(\ell),\mathbf{a}_2(\ell),\ldots,\mathbf{a}_N(\ell)\}$. $\{\mathbf{A}(\ell)\}_{\ell=0}^{\infty}$ can now fully describe the optimization variables in (4).

2) State: As the SR $C_\mu(\ell)$ with $d_\mu(\ell)=1$ can only be calculated in the future time slot, to describe the system status in the current ℓth time slot, we define the intermediate SR observation as

$$\begin{aligned} \check{C}_\mu(\ell) := \frac{1}{2}\Big[&\log(1+\lambda_\mu(\ell;\check{\mathcal{I}}(\ell,\ell+D_\mu))) \\ &-\log(1+\lambda_\mu^{(\mathrm{e})}(\ell;\check{\mathcal{I}}^{(\mathrm{e})}(\ell,\ell+D_\mu^{(\mathrm{e})})))\Big]^+ \end{aligned} \tag{5}$$

where $\check{\mathcal{I}}(\ell_1,\ell_2)$ and $\check{\mathcal{I}}^{(\mathrm{e})}(\ell_1,\ell_2)$ are interfering DAE sets indicating that when viewed from the end of the ℓ_1th time slot, which DAEs will be interfered at the LU and the EVE in the ℓ_2th time slot, respectively. Please note that $\check{\mathcal{I}}(\ell_1,\ell_2)\subseteq\mathcal{I}(\ell_2)$ and $\check{\mathcal{I}}^{(\mathrm{e})}(\ell_1,\ell_2)\subseteq\mathcal{I}^{(\mathrm{e})}(\ell_2)$, as shown in Fig. 1. Compared to the SR defined in (3), the intermediate SR describes the secrecy level induced by the previous and current transmission actions and does not take future actions into account.

We construct a tuple for the μth DAE as $\mathbf{o}_\mu(\ell)=\{\mathbf{a}_\mu(\ell),\check{C}_\mu(\ell)\}$ containing the action-observation pair in the ℓth time slot. The system state for the μth DAE can now be defined as $\mathbf{s}_\mu(\ell):=\{\mathbf{o}_\mu(\ell-1),\mathbf{o}_\mu(\ell-2),\ldots,\mathbf{o}_\mu(\ell-K)\}$ where K historical action-observation pairs are included. Denote $\mathbf{S}(\ell)$ as the system state in the ℓth time slot containing the states of all the DAEs, i.e., $\mathbf{S}(\ell)=\{\mathbf{s}_1(\ell),\mathbf{s}_2(\ell),\ldots,\mathbf{s}_N(\ell)\}$. The action $\mathbf{A}(\ell)$ is determined based on the current state $\mathbf{S}(\ell)$ and then the system state evolves to $\mathbf{S}(\ell+1)$ after observing the intermediate SRs $\check{C}_\mu(\ell)|_{1\leq\mu\leq N}$.

3) Reward: Denote $r_\mu(\ell)$ as the reward function for the μth DAE in the ℓth time slot after performing the action $\mathbf{a}_\mu(\ell)$ under the state $\mathbf{s}_\mu(\ell)$. The reward function is crucial as it guides each agent to learn the mapping from the state to the optimal action while achieving the desired goal. In this work, to pursue the optimization objective (4a) and satisfy the constraint (4b), we consider the reward function as

$$r_\mu(\ell)=C_\mu(\ell)-\beta[\Gamma_{\mathrm{th}}-\lambda_\mu(\ell;\mathcal{I}(\ell+D_\mu))]^+d_\mu(\ell) \tag{6}$$

where $\beta>0$ is a penalty factor for violation of the constraint (4b). Please note that the reward function (6) contains the SR $C_\mu(\ell)$ rather than $\check{C}_\mu(\ell)$. Hence, it cannot be collected immediately by the end of the ℓth time slot and is only available to the system in the future. In other words, due to the large propagation delays, the system receives delayed rewards. The sum reward received by the whole system in the ℓth time slot can be calculated by $R(\ell)=\sum_{\mu=1}^{N}r_\mu(\ell)$ where $R(\ell)$ is obviously a function of the action $\mathbf{A}(\ell)$ and state $\mathbf{S}(\ell)$.

Based on the MDP elements defined above, the optimization problem (4) can be reformulated to an RL problem with the aim of maximizing the long-term expected sum reward as

$$\max_{\{\mathbf{A}(\ell),\mathbf{S}(\ell)\}_{\ell=0}^{\infty}} \mathbb{E}\left\{\sum_{\ell=0}^{\infty}\gamma^\ell R(\ell)\right\} \tag{7}$$

where $\gamma\in(0,1]$ is a discounted factor and $\mathbb{E}(\cdot)$ is the expectation w.r.t. the joint distribution of state visitation and policy. In this work, an MARL method, i.e., multi-agent policy gradient (MADDPG) algorithm [11], is exploited to solve the optimization problem (7). In the MADDPG algorithm, each agent has an actor and a critic that learn to generate better actions and evaluate the generated actions of all the agents, respectively, where the two can be mathematically described by deep neural networks.

Performance evaluation: To evaluate the proposed method, we consider an underwater network consisting of 4 DAEs and assume that the DAEs, LU, and EVE are within a disk area of a radius of 4 km. We consider that the sound speed in water is 1500 m/s, the center frequency of the transmitted signal block is 13 KHz, and the duration of each block is 0.1 s. Similar to the actor proposed in [12], the actor network in this study passes the state into a two-layer perceptron (TLP) followed by two ResNet blocks to generate an action. For the critic network, the state and action are firstly inputted into a CNN encoder and a TLP, respectively. Next, the concatenation of the outputs from the two network components is fed into three ResNet blocks to obtain an estimation of the expected reward for the inputting state-action pair. The Adam optimizer is used to update the two networks.

We compare the proposed method to three methods: 1) Nearest: Select the DAE closest to the LU to transmit signal blocks all the time with the maximal transmission power; 2) SA: A modified version of the signal alignment method with the known location of the

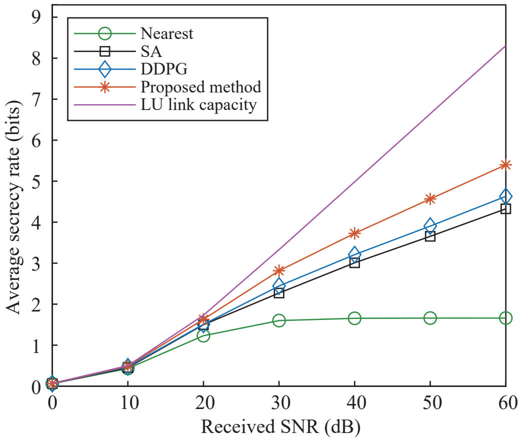

Fig. 2. Comparison of average SRs achieved by different methods.

Table 1. Comparison of SRs With Different Total Numbers of DAEs

Methods	Numbers of DAEs		
	4	6	8
Nearest	1.65	2.12	2.51
SA	3.01	3.23	5.06
DDPG	3.21	3.35	5.14
Proposed method	**3.73**	**3.99**	**5.75**

EVE [5] for time-slotted systems; 3) DDPG: Determine the transmission strategies of all the DAEs by a single agent with the DDPG algorithm.

The average SRs with different methods as well as the link capacity of the LU in Nearest are shown in Fig. 2. The received SNR corresponds to the ratio of the received signal power to the noise power. One can see that the average SR achieved by Nearest eventually converges as the received SNR increases while the rates achieved by SA, DDPG, and the proposed method grow monotonically with the received SNR. The efficacy of the proposed method is demonstrated as it outperforms all the other methods. It also shows that, compared to DDPG where only one agent learns the actions for all the DAEs, the proposed method enables higher secrecy performance, thanks to cooperative learning with multiple agents. Moreover, the average SRs gained by different methods under different total numbers of DAEs are presented in Table 1. It shows that the SRs of all the methods increase with the total number of DAEs as more DAEs could offer higher degrees of freedom to optimize the secrecy performance. Nevertheless, the proposed method still achieves the greatest SRs, which further validates its effectiveness.

Fig. 3 shows the learning performance of the proposed method in the case that an EVE stays in one location until the 15000th time slot and then moves to another location. It shows the moving average of the rewards over 100 time slots. One can see that the system collects negative rewards at the beginning due to violation of the constraint (4b) while the transmission policy gets improved through the learning process and the average reward eventually converges to the optimum. Immediately after the movement of the EVE, the rewards drop significantly since the current transmission policy is outdated w.r.t. the new location of the EVE. However, the system could adjust its policy to the new environment and finally reach its equilibrium, which exhibits the adaptivity of the proposed method.

Conclusion: This letter explored an MARL-based method to secure transmissions in underwater DAS against eavesdropping. Considering the large propagation delays of UWA transmissions, the secrecy rate was first derived for practical time-slotted UWANs. Then, the long-term sum secrecy rate maximization problem was studied in the RL paradigm, where each DAE learned its transmission schedule and transmission power online based on the MADDPG algorithm. The simulation results showed the efficacy of the proposed method compared to benchmark methods. In future, we will extend the proposed method for underwater DAS with multiple

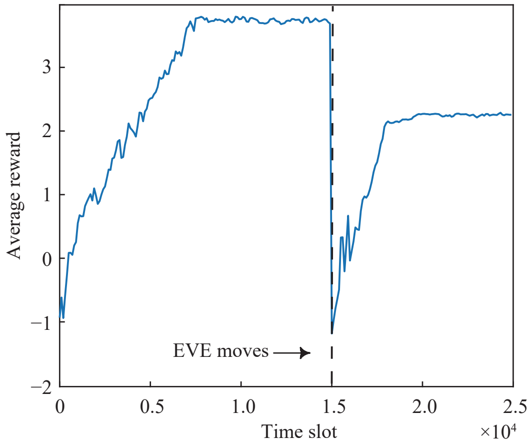

Fig. 3. Average reward with the proposed method in the case that an EVE moves to another location in the 15000th time slot.

LUs and multiple EVEs, where the interference relation is more challenging. In addition, how to adapt the proposed framework to underwater systems where their nodes, e.g., LUs, EVEs, or even DAEs, can move over time, is also an important research direction for our future work.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (62201248), and the Startup Foundation of the University of South China (200XQD056).

References

[1] Z. Wang, S. Zhou, and Z. Wang, "Underwater distributed antenna systems: Design opportunities and challenges," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 178–185, 2018.

[2] M. Xu, Y. Fan, and L. Liu, "Multi-party secret key generation over underwater acoustic channels," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 1075–1079, 2020.

[3] Y. Huang, P. Xiao, S. Zhou, and Z. Shi, "A half-duplex self-protection jamming approach for improving secrecy of block transmissions in underwater acoustic channels," *IEEE Sensors J.*, vol. 16, no. 11, pp. 4100–4109, 2015.

[4] Z. Peng, X. Han, and Y. Ye, "Enhancing underwater sensor network security with coordinated communications," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[5] C. Wang and Z. Wang, "Signal alignment for secure underwater coordinated multipoint transmissions," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6360–6374, 2016.

[6] Y. Yang, B. Li, S. Zhang, W. Zhao, and H. Zhang, "Cooperative proactive eavesdropping based on deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1857–1861, 2021.

[7] Y. Zhang, Z. Mou, F. Gao, J. Jiang, R. Ding, and Z. Han, "UAV-enabled secure communications by multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11599–11611, 2020.

[8] H. Sharma, I. Budhiraja, N. Kumar, and R. K. Tekchandani, "Secrecy rate maximization for THz-enabled femto edge users using deep reinforcement learning in 6G," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2022, pp. 1–6.

[9] J. Yan, Y. Meng, X. Yang, X. Luo, and X. Guan, "Privacy-preserving localization for underwater sensor networks via deep reinforcement learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1880–1895, 2021.

[10] L. Xiao, D. Jiang, Y. Chen, W. Su, and Y. Tang, "Reinforcement learning-based relay mobility and power allocation for underwater sensor networks against jamming," *IEEE J. Ocean. Eng.*, vol. 45, no. 3, pp. 1148–1156, 2020.

[11] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6382–6393.

[12] X. Ye, Y. Yu, and L. Fu, "Deep reinforcement learning based MAC protocol for underwater acoustic networks," *IEEE Trans. Mobile Comput.*, vol. 21. no. 5, pp. 1625–1638. 2022.