Establishment of Line-of-Sight Optical Links Between Autonomous Underwater Vehicles: Field Experiment and Performance Validation

Yang Weng[a,*], Takumi Matsuda[b], Yuki Sekimori[a], Joni Pajarinen[c], Jan Peters[d] and Toshihiro Maki[a]

[a]Institute of Industrial Science, The University of Tokyo, 153-8505 Tokyo, Japan
[b]School of Science and Technology, Meiji University, 214-8571 Kanagawa, Japan
[c]Department of Electrical Engineering and Automation, Aalto University, 02150 Espoo, Finland
[d]Intelligent Autonomous Systems Laboratory, Technische Universität Darmstadt, 64289 Darmstadt, Germany

ARTICLE INFO

Keywords:
Reinforcement learning
Autonomous underwater vehicle (AUV)
Underwater wireless optical communication(UWOC)
Line-of-sight (LOS) link

ABSTRACT

Establishing a line-of-sight link between autonomous underwater vehicles (AUVs) is an unavoidable challenge for realizing high data rate optical communication in ocean exploration. We propose a method for link establishment by maintaining the relative position and orientation between AUVs. Using a reinforcement learning algorithm, we search for the policy that can suppress external disturbances and optimize the link establishment efficiency. To evaluate the performance of the proposed method, we prepared a hovering AUV to conduct the link establishment experiments. The reinforcement learning policy trained in a simulation environment was deployed on the AUV in real environments. In field experiments, our approach successfully performed the link establishment from the hovering AUV to an autonomous surface vehicle. Based on the experimental results, we evaluate the performance of the AUV in executing the link establishment policy. Comparisons with existing optical search-based link establishment methods are presented.

1. Introduction

The emerging Underwater Internet of Things (UIoT) concept is developed for connecting underwater platforms, which is an extension of the Internet of Things to underwater scenarios (Jahanbakht et al., 2021). The UIoT technology can enhance the real-time or near-real-time data sharing between multiple platforms in underwater monitoring and survey missions, such as oil plume detection (Wang et al., 2022b), seafloor mapping (Matsuda et al., 2020), and environment survey Zhang et al. (2020).

In the past few decades, the fixed underwater observatories have been connected via fiber optic cables for high-speed data sharing (Heesemann et al., 2014). Mobile platforms, like autonomous underwater vehicles (AUVs), remotely operated vehicles, and underwater gliders, play a prominent role in ocean investigation, but are limited by underwater communications (Kong et al., 2022). The state-of-the-art underwater acoustic communication systems only support the data transmission of about 1 - 100 Kilobit per second (Kbps) (Huang et al., 2018). Meanwhile, the optical communication systems with LEDs can achieve an underwater communication link in the order of Gigabits per second (Gbps) (Wang et al., 2018). The underwater optical communication (UWOC) technology can provide a high data rate solution for underwater mobile platforms to exchange information. Monitoring and exploration of the underwater environment will be more efficient with the high-speed UIoT network.

Nowadays, UWOC technology has been developed significantly, and its data rate as well as stability have been greatly improved (Zhu et al., 2020). However, how to establish an underwater link is still a significant challenge that limits the application of UWOC. Gabriel *et al.* (Gabriel et al., 2013) investigated the effect of misalignment on an underwater optical link and emphasized the difficulty of establishing a perfect alignment in practice. In general, there are several link establishment issues that need to be solved in the implementation of the UWOC across mobile platforms:

1) Directionality of light signals. Omnidirectional underwater optical communication is difficult and inefficient to achieve. The directivity of UWOC signals requires the establishment of a link between two platforms;

2) Target location. To establish a link, the underwater platform must complete the initial location identification and continuously observe the status of the target.

3) External disturbances. In mobile platforms scenarios, boresight and jitter effects caused by external disturbances and uncertainties in the vehicle dynamic model remain a problem (Yang et al., 2014).

One solution to the challenge of link establishment is to ease the requirements on the alignment. Arnon *et al.* (Arnon and Kedar, 2009) attempted to avoid the alignment problem and proposed a non-line-of-sight (NLOS) network concept in which the link is established by means of back-reflection of the propagating optical signal at the ocean-air interface. Sait *et al.* (Sait et al., 2019) utilized the diffused optical signal to create the NLOS link, reducing the requirement for the receiver location. NLOS links can be a good solution for establishing links between mobile platforms, but there are strict requirements for optical transceivers, environment,

*Corresponding author
yangweng@iis.u-tokyo.ac.jp (Y. Weng); tmatsuda@meiji.ac.jp (T. Matsuda); sekimori@iis.u-tokyo.ac.jp (Y. Sekimori); joni.pajarinen@aalto.fi (J. Pajarinen); jan.peters@tu-darmstadt.de (J. Peters); maki@iis.u-tokyo.ac.jp (T. Maki)
ORCID(s): 0000-0002-7976-2301 (Y. Weng); 0000-0002-5932-5315 (T. Matsuda); 0000-0003-4301-7999 (Y. Sekimori); 0000-0003-4469-8191 (J. Pajarinen); 0000-0002-5266-8091 (J. Peters); 0000-0001-6992-4672 (T. Maki)

and effective distance. Severe signal attenuation, or even complete loss of signal, can occur in communications, so a sensitive photodetector is required at the receiving terminal. At this point, the common background light from land and bioluminescence can limit the ultimate sensitivity of the receiver (Pontbriand et al., 2008). In addition, a seawater medium with a high scattering coefficient is required in light scattering based NLOS links, but this in turn affects the effective propagation distance.

To establish a line-of-sight (LOS) link, tracking and maintaining the relative position between two mobile platforms is unavoidable. Existing link establishment methods rely mainly on optical systems and scanning algorithms. Similar to free space optical communication systems, Hardy *et al.* (Hardy et al., 2019) developed a real-time beam director that can quickly steer the beam in a specific direction. The proposed spiral pattern scanning algorithm can converge the alignment and maintain the link based on the detected light intensity. Solanki *et al.* (Solanki et al., 2020) designed a beam pointing system that can quickly and precisely adjust the azimuth and elevation of the beam. This optical system searches for a target along a triangular path and estimates the center of the beam by a gradient ascent algorithm. The light intensity based scanning methods can locate the center of the beam and build a high quality communication link. However, such scanning methods are often limited to local searches within the coverage area of the optical beam. Most of the research has not discussed how to obtain the initial position information of each other in the vast ocean. Disturbances from the marine environment can affect the stability of the link but cannot be measured by the photodetector. In addition, previous studies have paid no attention to mobile platforms carrying optical communication devices.

We propose a method of establishing the LOS link by maintaining the relative position and orientation between AUVs. Establishing the LOS link by controlling the vehicles does not rely on a real-time beam director or a detected intensity-based beam-steering algorithm, as commonly used in previous studies. This method can also be combined with the optical system to further improve the LOS link's stability. The acoustic navigation is used to guide the establishment process. With acoustic communication, AUVs can identify the initial location even at a long distance, and continuously observe the relative relationship. In our prior work (Weng and Maki, 2021)(Weng et al., 2022), we used the reinforcement learning algorithm to optimize the link establishment process and verified it in a simulated environment. However, the discrepancy between the simulation and the real environment can affect the performance of the reinforcement learning policy in the sea environment. In this study, we improved the policy and demonstrated the link establishment method that can be applied in practical scenarios. To evaluate the performance of our proposed method in real scenarios, the field experiments were designed in this research. The hovering-type AUV and an autonomous surface vehicle (ASV) were prepared to validate the trained policy in the link establishment experiment. From the results of the water tank and at-sea experiments, the relative position and orientation maintained between underwater platforms were acceptable for establishing the LOS link. The experimental data showed that the AUV could further reduce the position fluctuation to improve the link stability. To benchmark the performance of our method, we compared it with an existing triangular scanning method (Solanki et al., 2020). The results showed the advantages of our approach in establishing and maintaining LOS links in the marine environment.

The rest of this article is organized as follows. Section 2 presents the link establishment task. The reinforcement learning algorithm is utilized for policy training in Section 3. The hardware and algorithm of the experimental setup are described in Section 4. Detailed experimental results from water tank and sea trials are discussed in Section 5. The conclusions are given in Section 6.

2. Model

2.1. LOS Communication Link

Underwater optical beams have strong directivity because of the angular distribution of the emitted light signal. The directivity of UWOC yields a sector-shaped coverage area instead of an omnidirectional area, which poses challenges to link establishment (Saeed et al., 2019). An underwater optical communication experiment between underwater platforms was conducted in a water tank, shown in Fig. 1. The optical transmitter in an underwater vehicle emits an optical beam and forms a sector-shaped coverage area. The optical receiver is considered as an omnidirectional device that can detect the light signal after entering this coverage area (Gabriel et al., 2013).

As shown in Fig. 2, the LOS configuration is a straightforward form of optical links where the transmitter and receiver devices communicate over an unobscured link. The basic idea to maintain the LOS link is the optical beam from the transmitter at least partially covers the corresponding detector (Hoeher et al., 2021). In our method, the receiver is expected to be located in the center of the optical beam and l_o meters away from the transmitter. The expected link distance l_o is affected by various factors such as absorption, scattering, oceanic turbulence, source power, and hardware configuration (Sahu and Shanmugam, 2018)(Weng et al., 2019). The distance between the receiver and the optimal point is defined as the pointing error d_Δ, which needs to be minimized during link maintenance. The maximum pointing error $d_{\Delta,max}$ is determined by the characteristics of the optical transceivers.

In AUVs scenarios, the pointing error d_Δ is affected by random disruptions caused by environmental disturbances, such as sea currents and oceanic turbulence. As a mobile platform, the receiving platform is expected to change its position due to its motion and environmental disturbances.

2.2. Link Establishment

We propose a method to establish the optical communication link by maintaining the relative position and orientation between AUVs. The LOS link can be maintained as long

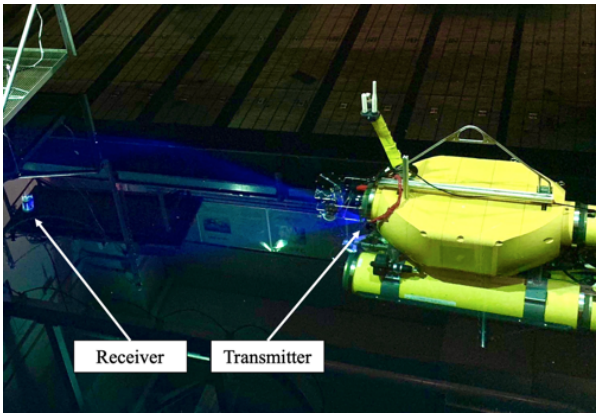


Figure 1: In the water tank, the receiver detected the optical beam emitted from the transmitter carried by an underwater vehicle. A LOS link is established.

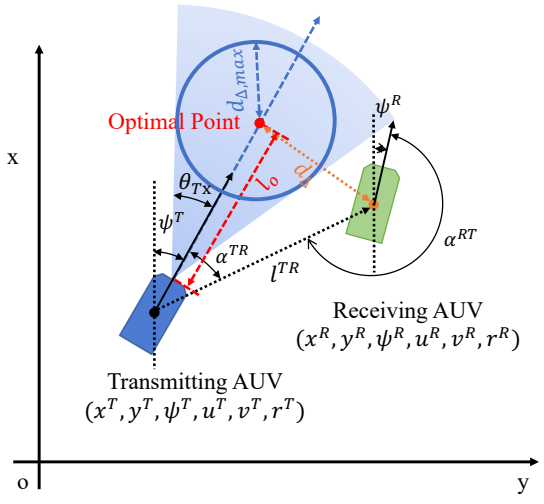


Figure 2: The establishment of a LOS link between two underwater platforms. The LOS link requires that the beam from one transmitter at least partially covers the receiver.

as two AUVs stably keep the required relationship. The AUV that transmits optical signals is defined as the transmitting AUV, while the AUV that receives optical signals is called the receiving AUV.

During link establishment, two AUVs should remain at the same depth. The pressure sensor attached in the underwater vehicle can determine its absolute depth with an overall accuracy of about 0.01% of full scale (Kinsey et al., 2006). Without the accumulated errors, the AUV can cruise to a set depth. We consider the link establishment in the horizontal plane because the jitter of the AUV in the vertical plane is smaller than the coverage of the beam. We simplify the motion planning of AUVs by reducing the number of control state variables from the three-dimensional space to the two-dimensional plane.

The model of optical link establishment between AUVs is defined in Fig. 2. In this research, the horizontal position $[x, y]$, surge velocity u, sway velocity v, yaw orientation ψ, and yaw angular velocity r of AUVs are considered. The superscript T and R indicate the variables belonging to the transmitting AUV and the receiving AUV, respectively.

Two AUVs need to complete the initial location identification, and then the transmitting AUV moves to the position that meets the relative relationship requirements. During optical communication, the transmitting AUV continuously observes the relative relationship and maintains it through motion planning. The acoustic navigation is utilized for the initial location identification. The acoustic signal is the only media available for underwater propagation over distances of up to several kilometers (Stojanovic, 2007). To locate the position of the receiving AUV, the transmitting AUV can start the two-way travel time (TWTT) ranging (Kussat et al., 2005). The TWTT ranging can measure the relative distance l^{TR} and the relative bearing angle of the transmitting AUV's side α^{TR}. The estimates of position and orientation are updated by a particle filter estimator running in the transmitting AUV.

Once the receiving AUV's position is determined, the transmitting AUV can track the receiving AUV by controlling the surge velocity u^T and yaw angular velocity r^T, which is basically available for all types of AUVs. The pointing error d_Δ that needs to be minimized can be calculated as:

$$d_\Delta = [(x^R - x^T - l_o \cos\psi^T)^2 + (y^R - y^T - l_o \sin\psi^T)^2]^{\frac{1}{2}} \quad (1)$$

The transmitting AUV needs to keep relative position and orientation to maintain a stable LOS link. Acoustic navigation is constantly utilized for observing the relative position and orientation since the relative relationship is susceptible to vehicle motions, ocean turbulence, and ocean currents. The TWTT ranging (Matsuda, 2021) and the bearing only ranging (Fujita et al., 2019) are combined to guide the transmitting AUV during maintenance. Compared to TWTT ranging, the bearing only method is scalable for AUV formation and requires fewer acoustic channel resources. However, the observability of the bearing only ranging is weak and only relative angles α^{TR} and α^{RT} can be observed (Weng and Maki, 2021). The transmitting AUV needs to consider scalability and observability, and chooses TWTT ranging or bearing only ranging depending on the current state.

3. Reinforcement Learning

3.1. Establishment Policy

The policy trained by a reinforcement learning algorithm is utilized to control the transmitting AUV to establish the LOS link, considering:

1) the reinforcement learning algorithm succeeds in suppressing the effects of external disturbances and uncertainties in motion planning (Kober et al., 2013) (Wang et al., 2022a);

2) as a model-free method, the model of the ocean environment and vehicle dynamics are not essential;

3) reinforcement learning can fuse different types of data, such as vehicle motion states and acoustic ranging results, to make decisions.

To use the reinforcement learning, the state space of the agent is defined as:

$$s = [\hat{x}_\Delta, \hat{y}_\Delta, \cos\psi^R, \sin\psi^R, u^R, v^R, r^R, \cos\hat{\psi}^T, \sin\hat{\psi}^T] \quad (2)$$

where $[x_\Delta, y_\Delta]$ is the pointing error. The variables with hat symbols are updated by a particle filter estimator. All variables in the state space are one-dimensional and continuous.

The action space of the agent is as follows:

$$a = [u^T, r^T, i_{twtt}, i_{op}] \quad (3)$$

where the boolean variable i_{twtt} represents whether the transmitting AUV requests for TWTT ranging or bearing only ranging. The boolean variable i_{op} represents whether the transmitting AUV turns on the optical transmitter in the current timestep. These two actions are defined so that reinforcement learning can optimize the usage of acoustic channel resources and energy during the establishment process.

We propose the reward function of the form:

$$\begin{aligned} r(s,a) = &- \rho_1(1+\rho_2 i_{twtt})(1+\rho_3 i_{op}) d_\Delta^{\frac{1}{2}} \\ &- \rho_4 u_\Delta - \rho_5 r_\Delta + \rho_6 i_{done} \end{aligned} \quad (4)$$

where ρ_1 to ρ_6 are coefficients to determine the importance of different reward terms. The u_Δ and r_Δ represent the relative velocities in surge and yaw. A boolean variable i_{done} indicates if the link establishment is completed.

The flow of reinforcement learning policy is shown in Fig. 3. Sensors on underwater vehicles need to collect and observe the states defined in (2). When the current state is available, the reinforcement learning policy can decide the next action and output it to the executing device.

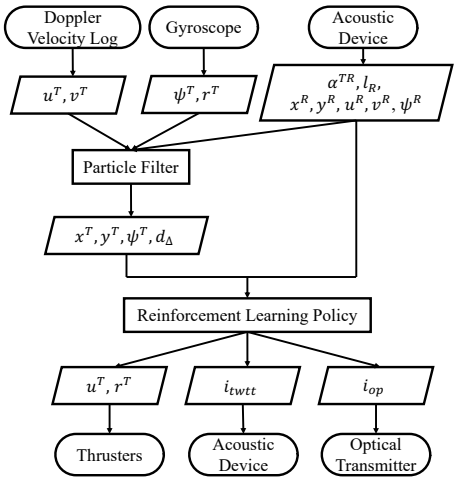


Figure 3: The flowchart of reinforcement learning policy in the link establishment process. It describes how the AUV observes the environment and generates action decisions through the reinforcement learning policy to complete the link establishment task. Rounded rectangle symbols represent devices on the AUV, and parallelogram symbols represent data available for input or output. Rectangle symbols are used to represent processes. Directional connectors represent the flow of data or instructions.

3.2. Policy Training

An optimal reinforcement learning policy that can perform the link establishment needs to be trained and then deployed on the AUV. In this study, the soft actor-critic (SAC) reinforcement learning algorithm presented by Haarnoja *et al.* (Haarnoja et al., 2018b) is used to search for an optimal policy π^*. The objective that the SAC algorithm wants to maximize is not only cumulative reward, but also entropy, which is a measure of randomness in the policy. The motivation for using the SAC algorithm in policy training is: 1) The combination of cumulative reward and entropy items in the objective function can enhance exploration ability and avoid convergence to a bad local optimum; 2) The maximum entropy requirement can minimize the need for hyperparameter tuning. We can avoid hyperparameter tuning when deploying on the actual machine; 3) Typical Gaussian exploration may cause the action to jitter at high frequency. The SAC algorithm can smooth the action by temporally correlating the exploration (Haarnoja et al., 2018a). For underwater vehicles, the thrusters cannot handle the high frequency of command changes and may be damaged.

We use the OpenAI Stable Baselines toolkit to implement the SAC algorithm in policy training (Hill et al., 2018). As listed in Algorithm 1, the SAC algorithm collects sample data to update the network by gradient descent. The hyperparameters used in the policy training are listed in Table 1. We use feedforward neural networks to initialize the policy and target networks. The neural network consists of two hidden layers, each with 64 neurons. Since no camera and images are utilized in this study and the number of observation dimensions is not large, this original structure is sufficient. Referring to some benchmark applications of the SAC algorithm (Haarnoja et al., 2018c), we set the discount factor γ and learning rate λ to 0.99 and 0.0003, respectively.

In order to search for the optimal policy, the SAC algorithm needs to learn from a substantial amount of sample data. Considering the sample efficiency of the reinforcement learning, we use the OpenAI Gym interface (Brockman et al., 2016) to build a simulated environment for data collection. As shown in Fig. 4, the transmitting AUV and the receiving AUV are randomly initialized on a horizontal plane. The reinforcement learning policy samples the actions based on the current state. After the action is generated, the simulator uses the kinematic model to calculate the next state of the underwater vehicle. The reward for each step is calculated according to (4).

The parameters in the simulator refer to the settings of previous simulation studies and are modified based on the real environment. Configuration of the coefficients in the

reward function is a trade off between different controlling objectives. In prior work (Weng et al., 2022), we trained the policy in a simulated environment and evaluated the configuration of the coefficients. When the coefficients ρ_1 to ρ_6 in the reward function are set to 0.01, 9, 1, 0.01, 0.002, and 10, respectively, the reinforcement learning policy can converge faster and perform well. This configuration allows the AUV to quickly approach the target and then adjust the attitude for link establishment.

In order to deploy the reinforcement learning policy in real scenarios, we improve the training of the policy. The end condition is removed so that the AUV can maintain a stable link for a long time. The maximum length of each episode in the simulator is increased to 1500 steps, and the duration of each step is reduced to 0.2 seconds. In this way, the policy can control the underwater platform more precisely. The optimization of the optical and acoustic devices for energy saving is not considered in the current experiments. To suppress the effects of external disturbance in the real environment, we enhance noise interference in vehicle motion. In each step, the surge, sway, and yaw angular velocities are affected by Gaussian noise with standard deviations of 0.1, 0.1, and 1, respectively. The maximum surge and yaw angular velocities are 0.2 m/s and 0.2 rad/s, respectively. The expected link distance l_o is set to 5 meters.

At each environmental step, the simulator calculates the states, actions, and rewards of the AUVs. The current state s_t, action a_t, reward $r(s_t, a_t)$, and the next state s_{t+1} are combined as a complete sample data. A total of 3×10^6 steps of sample data are collected in the simulator for policy training.

The trained policy is tested 100 times in the simulated environment, and the results compared to the prior work are presented in Fig. 5. Statistics show that the current policy can maintain the link continuously for 68.86 seconds and the total time for link maintenance is 120.47 seconds. In contrast, in the prior work, the transmitting AUV only needed to maintain the link for 10 seconds.

In summary, the advantages of the current policy and the differences from the prior work are 1) The current policy is trained in a new environment with high environmental perturbations that is more similar to the real environment; 2) Change the end condition to allow the reinforcement learning algorithm to explore the optimal policy that can maintain the link for longer periods; 3) Increase the frequency of action generation to enhance control efficiency, allowing the current policy to establish the link in less time. Therefore, the current policy is suitable for deployment in field experiments.

Algorithm 1 Soft Actor-Critic (Haarnoja et al., 2018c)

Input:
Initialize target network θ_1, θ_2
Initialize policy network ϕ
Initialize target network weights $\overline{\theta}_1 \leftarrow \theta_1, \overline{\theta}_2 \leftarrow \theta_2$
Initialize an empty replay buffer to store sample data $\mathcal{D} \leftarrow \emptyset$
for each iteration **do**
  for each environment step **do**
    Sample action according to the policy $a_t \sim \pi_\phi(a_t|s_t)$
    Execute action a_t and sample reward $r(s_t, a_t)$ and new state s_{t+1} by $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    Store samples by $\mathcal{D} \leftarrow \mathcal{D} \cup \left\{(s_t, a_t, r(s_t, a_t), s_{t+1})\right\}$
  end for
  for each gradient step **do**
    Update target network by $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    Update policy network by $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    Adjust temperature by $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$
    Update target network weights by $\overline{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\overline{\theta}_i$ for $i \in \{1, 2\}$
  end for
end for
Output: θ_1, θ_2, ϕ

Table 1
Hyperparameters configuration

Parameter	Symbol	Value
Layer of MLP		2
Neuron of MLP		64
Discount factor	γ	0.99
Learning rate	λ	0.0003
Buffer size		50000
Batch size		64

4. Implementation

4.1. Underwater Platforms

The hovering-type AUV Tri-TON acted as the transmitting AUV in actual machine experiments. The main concept of this platform is the ability to hover. The AUV has a payload space that allows the integration of different devices depending on the application. The specifications of the AUV Tri-TON are given in Table 2. As shown in Fig. 7, a total of 5 thrusters are mounted on the underwater platform. A thruster is configured in the transversal direction for controlling the sway motion, and two thrusters are installed symmetrically on both sides of the longitudinal axis to control the surge and yaw motion. The other two thrusters are installed in the vertical direction to maneuver the heave motion. The thruster configuration is sufficient to support the motion required in the link establishment task, including surge, yaw, and heave. The power of each thruster is 100 watts, and the maximum speed can reach 0.5 m/s.

There are three waterproof pressure hulls, including the main controller hull and two battery hulls. Four lithium-ion batteries in two battery hulls can provide the AUV to

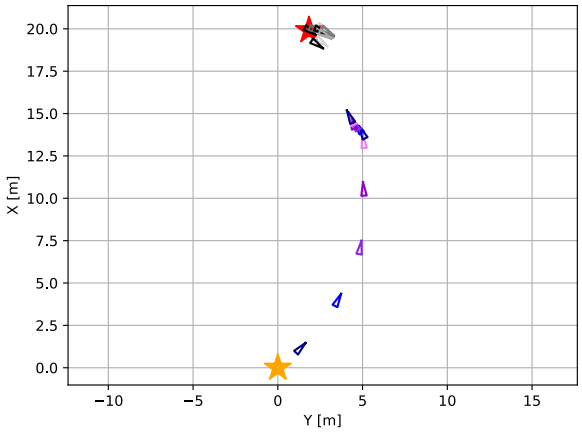


Figure 4: The trajectories of the AUVs in the simulation environment. The orange and red star markers are the starting points for the transmitting AUV and the receiving AUV. The positions of vehicles are plotted every 4 timesteps. The transmitting AUV is cyclically represented by dark-blue, blue, blue-violet, dark-violet, and violet triangles (every 20 timesteps), while the receiving AUV is represented by black, dim-grey, grey, dark-grey, and light-grey triangles. The sharp corner of the triangle is the head of the vehicle.

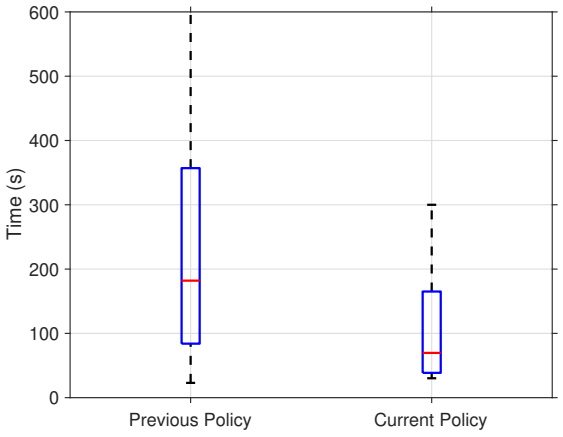


Figure 5: Comparison of the current policy with the policy trained in prior work. The plot indicates the time taken by the two policies to complete the link establishment task. The end condition is that the transmitting AUV keeps the pointing error within 1 meter and maintains it for 10 seconds. In box-and-whisker plots, the lower and upper boundaries of the box represent the 25th (Q1) and 75th (Q3) percentiles, respectively; the bottom and top ends of the whisker indicate the most extreme values within the lower limit $Q1-1.5(Q3-Q1)$ and the upper limit $Q3 + 1.5(Q3 - Q1)$, respectively; the red line inside the box marks the median.

operate for 8 hours. The main computer in the main hull is a single board UP Core with Intel Atom x5-z8350 CPU. The Ubuntu 20.04 operation system is installed in UP Core. The computing resources on the main computer can support the operation of the particle filter algorithm.

A Doppler velocity log (DVL) Teledyne RDI Navigator 1200 kHz, the red colored device in Fig. 7, is attached to the bottom of the vehicle. It provides the measurement of ground velocity. The SeaTrac X150 is used as the ultra-short baseline (USBL) device and underwater acoustic communication modem. The JAE JG-35FD is the fiber-optic gyroscope (FOG) device that measures orientation. When Tri-TON is on the sea surface, the WiFi antenna supports the wireless connection, and the global navigation satellite system (GNSS) service is available.

The ASV BUTTORI is used as the receiving AUV in real experiments. BUTTORI has a dynamic positioning ability and can deal with sea waves and strong winds. It provides the position and acoustic ranging for underwater platforms in the task. The SeaTrac X150 is the USBL device for acoustic ranging and communication. Three waterproof pressure hulls, including the main controller hull and two battery hulls, are equipped in the surface vehicle. There are 3 thrusters mounted on the vehicle to control the surge and yaw motion. It supports the wireless connection, and the GNSS compass service is available.

4.2. Algorithm

To integrate the trained reinforcement learning policy into the AUV control system, an open-source middleware suite called Robot Operating System (ROS) is used for

Figure 6: The hovering-type AUV Tri-TON is used as the transmitting AUV in real experiments.

building AUV applications. It provides a set of software libraries and tools for development. There are two basic concepts called nodes and topics. A node represents a single process running in the ROS system. It can take actions based on information received from other nodes, and/or send information to other nodes. Topics are named buses over which nodes send and receive messages (Koubaa, 2019). The topic defines the type of information that nodes want to share.

The ROS graph structure used in Tri-TON is given in Fig. 9. The USBL device, DVL device, FOG device, and depth sensor are connected with corresponding ROS driver

Table 2
AUV Tri-TON Specifications

Parameter	Value (Device)
Size	1.40 m (L) × 1.33 m (H) × 0.76 m (W)
Mass	230 kg
Max. Speed	0.5 m/s
Max. Depth	800 m
Duration	8 hours
Thruster	100 W thruster × 5
Battery	LiIon 26.6 V 25 Ah × 4
Main Computer	UP Core
DVL	Teledyne RDI Navigator 1200 kHz
USBL	SeaTrac X150
FOG	JAE JG-35FD
Depth Sensor	Mensor DPT6000

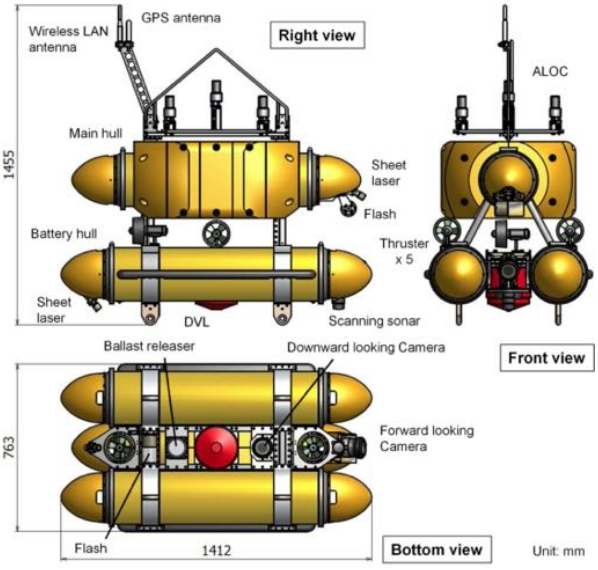


Figure 7: The specifications of AUV Tri-TON.

nodes through serial communication. All data collected by the aforementioned devices are published to the particle filter state estimator. The states defined in (2), estimated by the particle filter, are the input to the reinforcement learning node called RLpolicy. The actions defined in (3), generated by the reinforcement learning policy, are performed by thrusters, USBL devices, and an optical transmitter, respectively.

Except for the reinforcement learning node RLpolicy, the rest of the nodes are running on the main computer UP Core.

Figure 8: The ASV BUTTORI is used as the receiving AUV in real experiments.

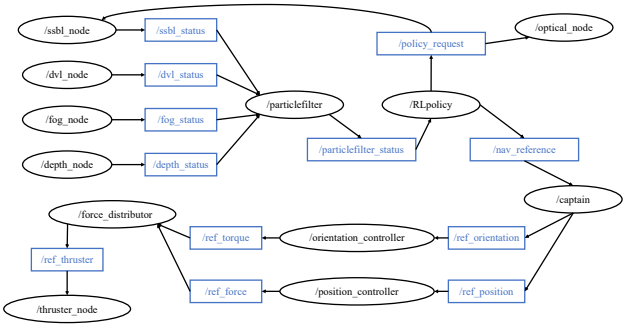


Figure 9: The ROS graph structure in Tri-TON. The black circle represents the node, and the blue square represents the topic in the ROS system.

In order to ensure that the reinforcement learning algorithm has sufficient computing resources, a separate board UP Board is used to compute the reinforcement learning policy. The UP Board has Intel Atom x5-z8350 CPU and Ubuntu operation system, and is physically connected to the UP Core. The entire ROS system runs on two machines, where the main computer is the master. In the RLpolicy node, the reinforcement learning policy is operated using the OpenAI Stable Baselines toolkit and Tensorflow. As shown in the algorithm flow chart in Fig. 3, the reinforcement algorithm computes actions when the observed state is available and publishes the actions on the ROS system. The proposed link establishment algorithm for deployment in AUVs is listed in Algorithm 2.

4.3. Experiments

The water tank and sea experiments were prepared to conduct the LOS link establishment task between the underwater platforms. In the field experiments, the trained reinforcement learning policy was deployed on Tri-TON. The actions i_{op} and i_{twtt} used by the agent optimize energy and acoustic channel resources in simulations, but are not currently used in real environments.

Algorithm 2 Link Establishment Algorithm

Initialize the particle filter
Move to the same depth
Initialize the reinforcement learning policy
while data sharing is needed **do**
  Maintain the same depth as the target
  Update states u_t^T, v_t^T through the DVL
  Update states ψ_t^T, r_t^T through the FOG
  if acoustic ranging results are available **then**
    Update states $\alpha_t^{TR}, l_{R,t}, x_t^R, y_t^R, u_t^R, v_t^R, \psi_t^R$
  end if
  Update states $x_t^T, y_t^T, \psi_t^T, d_{\Delta,t}$ by the particle filter
  Input states $x_{\Delta,t}$, $y_{\Delta,t}$, $\cos\psi_t^R$, $\sin\psi_t^R$, u_t^R, v_t^R, r_t^R, $\cos\psi_t^T$, $\sin\psi_t^T$ to the reinforcement learning policy
  Generate actions u_t^T, r_t^T, $i_{twtt,t}$, $i_{op,t}$ by the reinforcement learning policy
  if u_t^T, r_t^T are available **then**
    Execute actions u_t^T, r_t^T by thrusters
  end if
  if $i_{twtt,t}$ is available **then**
    Execute actions $i_{twtt,t}$ by the acoustic device
  end if
  if $i_{op,t}$ is available **then**
    Execute actions $i_{op,t}$ by the optical transmitter
  end if
end while
End the reinforcement learning policy

In the experiments, the AUV Tri-TON and the ASV BUTTORI were randomly deployed on the sea surface. Tri-TON needs to align with BUTTORI and shorten the pointing error d_Δ. No GNSS service or radio communications were used by Tri-TON in the experiments as they are not available in underwater environments. The TWTT ranging between Tri-TON and BUTTORI was performed every 6 seconds, and the states of BUTTORI $[x_R, y_R, \psi_R, u_R, v_R, r_R]$ was transmitted to Tri-TON through the acoustic signals.

During the experiments, the DVL recorded the ground velocity, and the FOG measured the orientation angular velocity of the vehicle. The position and relative relationship of Tri-TON and BUTTORI were measured by TWTT ranging and estimated by particle filter. The ROS system recorded the decisions generated by the reinforcement learning policy. The experiments were designed to evaluate and discuss the following:

1) whether the underwater vehicle can execute decisions generated by reinforcement learning policy;

2) whether the AUV can establish a LOS link and maintain the relative relationship;

3) the impact of external disturbances on alignment disruptions, and whether the vehicle can handle the disturbances;

4) whether the AUV can establish a LOS link with a moving target;

5) transfer of policy trained in simulated environments to real scenarios.

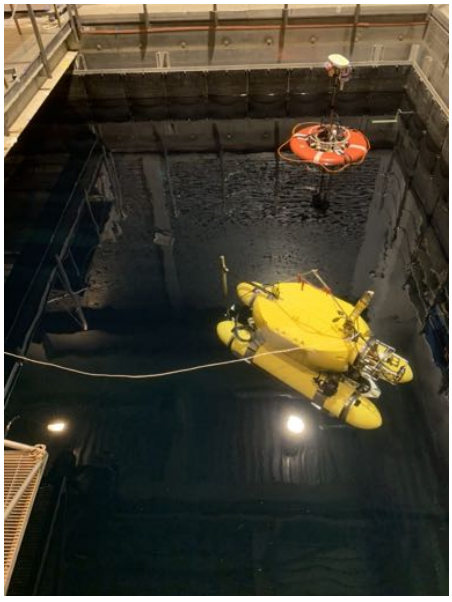


Figure 10: A photo of the water tank experiment at the Institute of Industrial Science, the University of Tokyo.

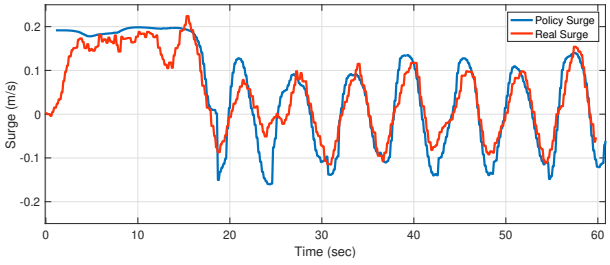


Figure 11: The surge velocity of Tri-TON in the water tank experiment. The blue curve is the surge velocity generated by the reinforcement learning policy, while the red curve is the real surge velocity measured by the DVL device. The blue curve in the figure is delayed by 1.1 seconds.

5. Results and Discussion

5.1. Water Tank Experiments

The purpose of water tank experiments was to verify that the reinforcement learning policy trained in the simulation environment can be deployed on real platforms. Compared to the sea environment, external disturbances of the water tank environment were not prominent. As shown in Fig. 10, we conducted the experiments in the water tank at the Institute of Industrial Science, the University of Tokyo. The size of the water tank is 8 meters long, 8 meters wide, and 8 meters deep. Tri-TON was controlled by a reinforcement learning policy and needed to maintain a relative relationship with BUTTORI. Due to the size of the water tank, the expected link distance l_o was set to 1 meter. The maximum surge and yaw angular velocities of the Tri-TON were set to 0.2 m/s and 0.2 rad/s, respectively. The BUTTORI kept stationary in the water tank.

During link establishment, the deployed reinforcement learning policy generated commands based on the current

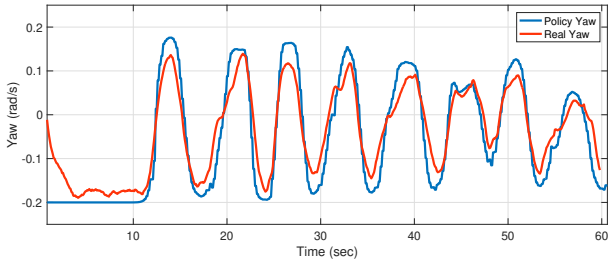


Figure 12: The yaw angular velocity of Tri-TON in the water tank experiment. The blue curve is the yaw angular velocity generated by the reinforcement learning policy, while the red curve is the yaw angular velocity measured by the FOG device. The blue curve in the figure is delayed by 0.8 seconds.

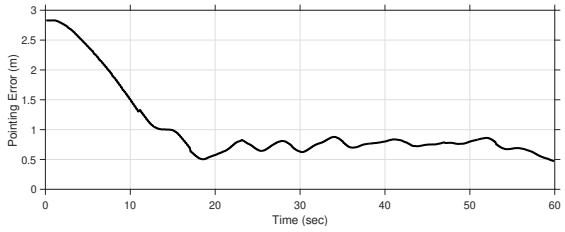


Figure 13: The pointing error d_Δ in the water tank experiment.

state. The commands for surge and yaw angular velocity are represented as blue curves in Fig. 11 and 12. The actual surge velocity of Tri-TON that DVL measured is the red curve in Fig. 11, while the actual yaw angular velocity measured by FOG is the red curve in Fig. 12. The same trend in the red and blue curves shows that the AUV's thrusters can execute the motion commands from the reinforcement learning policy. During motion, Tri-TON had 1.1 seconds and 0.8 seconds delays in implementing the surge and yaw angular velocity commands from the trained policy.

The pointing error d_Δ that the Tri-TON needs to shorten is presented in Fig. 13. Under the control of the reinforcement learning policy, Tri-TON gradually approached the target. Within 20 seconds, the pointing error d_Δ was reduced to within 1 meter and successfully maintained. From 20 to 60 seconds in this water tank experiment, the mean of the pointing error was 0.74 meters and the standard deviation was 0.08 meters.

The results of the water tank experiments demonstrated that the experimental platforms we prepared could be used to deploy and evaluate the trained reinforcement learning policy in the real environment. AUV can stably maintain the pointing error within 1 meter.

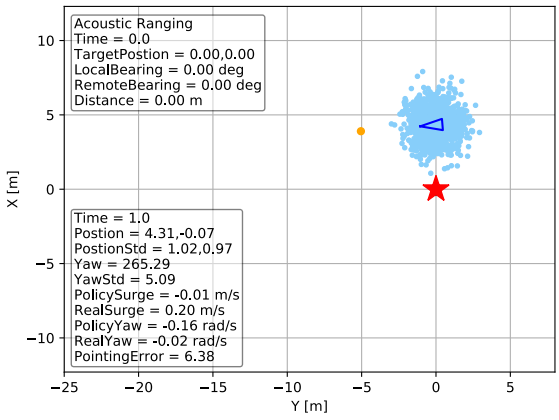
5.2. Sea Experiments - Dive 1

The sea experiments were implemented to evaluate the performance of our method in a turbulent environment. As shown in Fig. 14, we conducted sea experiments at Hiratsuka Port, Japan. The depth of the port was about 3-5 meters. The maximum surge and yaw angular velocities were set to 0.2 m/s and 0.2 rad/s, respectively. The expected link distance l_o was set to 5 meters.

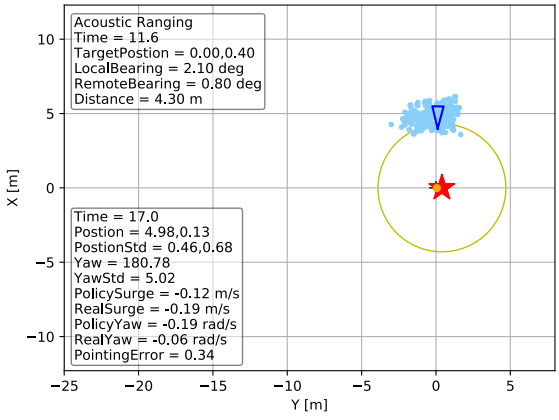
Figure 14: Sea experiments at Hiratsuka Port, Japan.

One of the sea experiments is presented in Fig. 15. The collected experimental data are plotted based on a fixed coordinate system. The AUV Tri-TON is represented by the blue triangle, while the red star marker is the ASV BUTTORI. The particle filter estimation results of the Tri-TON are depicted by blue dots. The orange point is the optimal point for optical communication. The yellow circle represents the result of acoustic ranging. The parameters listed in the bottom left corner are the time when reinforcement learning policy starts in the experiment, the position of Tri-TON (x^T, y^T), the standard deviation of particle filter estimation results in the position, the yaw orientation of Tri-TON ψ^T, the standard deviation of particle filter estimation results in yaw, the surge velocity command generated by reinforcement learning policy, the surge velocity measured by DVL, the yaw angular velocity command generated by reinforcement learning policy, the yaw angular velocity measured by FOG, and the pointing error d_Δ. The parameters listed in the top left corner are the time when the latest ranging results are received, the position of the BUTTORI, the relative bearing angle measured by the USBL device in Tri-TON, the relative bearing angle measured by the USBL device in the ASV BUTTORI, and the relative distance.

Tri-TON was randomly initialized to establish the LOS with BUTTORI. The initial position of Tri-TON was (4.31, -0.07) meters, and the pointing error d_Δ was 6.38 meters. We did not move BUTTORI manually, and the platform was affected by the ocean waves. The particle filter estimator in the vehicle was initialized, and the standard deviation of estimation results in position was (1.02, 0.97) meters. The particles gradually converged as the acoustic ranging results were continuously updated. At the 17.0 seconds of the experiment, the pointing error d_Δ was 0.34 meters. The standard deviation of estimation results in position was (0.46, 0.68) meters. The relative relationship between Tri-TON and BUTTORI was suitable for establishing the LOS link.



(a) Sea experiment at 1.0 s



(b) Sea experiment at 17.0 s

Figure 15: The states of AUV in the sea experiment (a) 1.0 s, and (b) 17.0 s. The AUV Tri-TON is represented by the blue triangle, and the sharp corner of the triangle is the head of the vehicle. The red star marker is the ASV BUTTORI.

As shown in Fig. 18, the vehicle successfully maintained the pointing error within two meters. The stability of the link was worse than in the case of the water tank due to more external disturbances in the sea environment. From 20 to 60 seconds in this sea experiment, the mean of the pointing error was 0.69 meters. The standard deviation was 0.44 meters, which is obviously larger than the water tank experiment.

The effect of perturbations was also reflected in the AUV motion. Fig. 16 and 17 show that thrusters can execute the command generated by the reinforcement learning policy but are disturbed by the environment. The execution of the surge and yaw angular velocity commands delayed by 3.2 and 1.0 seconds respectively from the commands of the trained policy, which was longer than the delays in the water tank experiment.

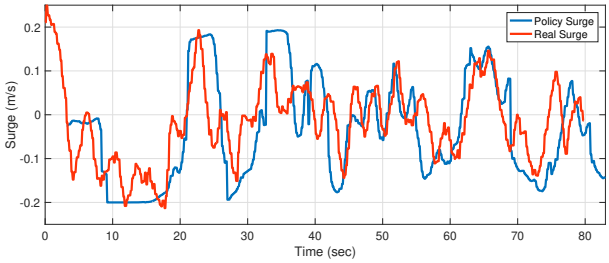


Figure 16: The surge velocity of Tri-TON in the sea experiment (Dive 1). The blue curve is the surge velocity generated by the reinforcement learning policy, while the red curve is the real surge velocity measured by the DVL device. The blue curve in the figure is delayed by 3.2 seconds.

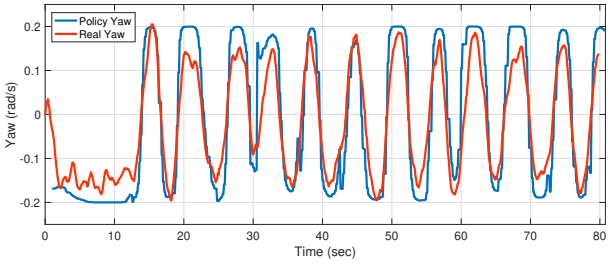


Figure 17: The yaw angular velocity of Tri-TON in the sea experiment (Dive 1). The blue curve is the yaw angular velocity generated by the reinforcement learning policy, while the red curve is the yaw angular velocity measured by the FOG device. The blue curve in the figure is delayed by 1 second.

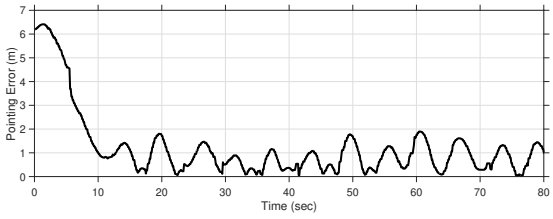


Figure 18: The pointing error d_Δ in the sea experiment (Dive 1).

5.3. Sea Experiments - Dive 2

Another experiment was conducted to test whether our method can continuously track and align with the moving target. The maximum surge and yaw angular velocities of Tri-TON remained at 0.2 m/s and 0.2 rad/s, and the expected link distance l_o was 5 meters.

We manually moved BUTTORI and then stopped it during the task. The trajectory of BUTTORI is given in Fig. 19. As shown in Fig. 20, the position of BUTTORI at the 7.0 seconds was (-0.70, -6.40) meters, which was already deviated from the original position. With the acoustic ranging, the AUV knew that the target had deviated and started to track the moving BUTTORI. During the movement, the pointing error became large, reaching 2.44 meters at 37.0 seconds. When we stopped controlling BUTTORI, the pointing error started to decrease. At 57.0 seconds of the experiment, the

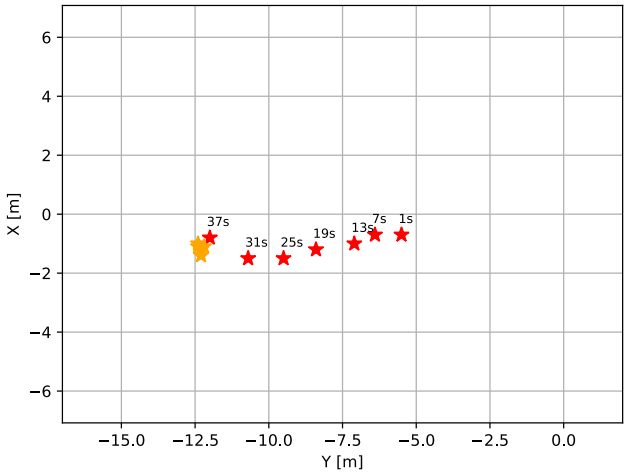
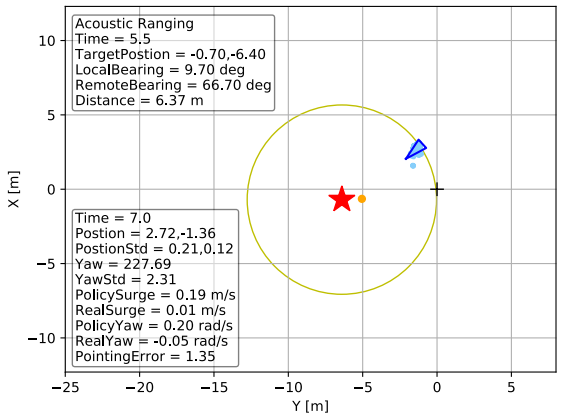


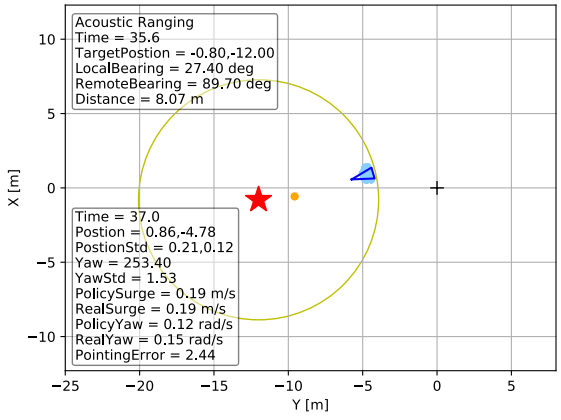
Figure 19: The trajectory of the ASV BUTTORI. The position of BUTTORI is plotted every 6 seconds. During the movement, the position of BUTTORI is represented by a red star. The time is marked above the stars. When BUTTORI remains stationary, the position is represented by an orange star.

pointing error was 0.34 meters. The surge and yaw angular velocities during the moving phase are as shown in Fig. 22 and Fig. 23, respectively. Tri-TON had 2.5 seconds and 0.9 seconds delays in executing the surge and yaw angular velocity commands.

The pointing error during the task is shown in 24. The pointing error increased when we moved BUTTORI, and decreased when it was dynamically positioned. From 0 to 50 seconds in this sea experiment, the mean of the pointing error was 1.75 meters, and the standard deviation was 0.50 meters. Between 50 and 80 seconds, the mean of the pointing error decreased to 0.54 meters, with a standard deviation of 0.36 meters.



(a) Sea experiment at 7.0 s



(b) Sea experiment at 37.0 s

Figure 20: The states of AUV in the sea experiment (a) 7.0 s and (b) 37.0 s. The AUV Tri-TON is represented by the blue triangle, and the sharp corner of the triangle is the head of the vehicle. The red star marker is the ASV BUTTORI.

5.4. Discussion

The results of water tank and sea experiments verify that the link establishment method we proposed can be deployed on real underwater vehicles. The developed platform supported reinforcement learning policy and particle filter estimator operations. A suite of sensors on the AUV collected the data needed for the state space. The thruster system performed the commands generated by the reinforcement learning policy. All data and commands were shared through the topics in the ROS structure.

The LOS link was successfully established in both water tank and sea environments. Through the guidance of the reinforcement learning policy, Tri-TON tracked the BUTTORI and kept the relative position and orientation for link establishment. In sea experiments, the maximum pointing error in yaw angle is about 20 degrees. This error is acceptable for LED-based optical communication systems, since the half-angle of the optical transmitter can easily be larger than 20 degrees (Rong et al., 2021). The pointing error results show that the relative relationship between the two platforms can be maintained, indicating that the LOS link can be stable for a long time.

The external disturbance from the sea environment had an impact on link establishment. The experimental results show that the actual surge and yaw angular velocities of Tri-TON were similar to the commands generated by the policy, but the differences still existed. The differences between the Hiratsuka Port and the water tank environments were obvious, which was reflected in Fig. 11, 12, 16, and 17. The commands of surge and yaw angular velocity indicate the pointing error can be further reduced. It shows that the reinforcement learning policy does not handle the delay from the platform's thrusters well. With the statistics and delay analysis, the AUV can further reduce the position fluctuation during link establishment.

We conducted experiments to discuss the ability of establishing the LOS link with a moving target. The pointing error was increased to about 2 meters due to the inaccurate

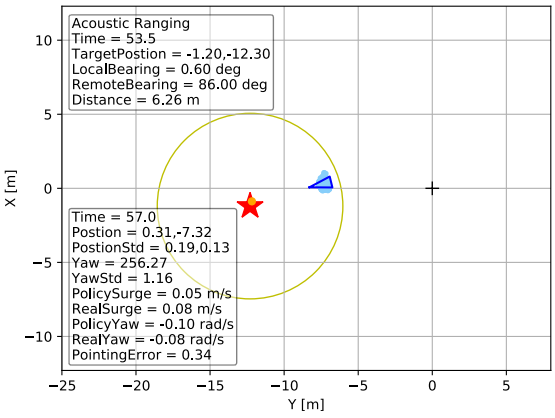


Figure 21: The states of AUV in the sea experiment at 57.0 s. The AUV Tri-TON is represented by the blue triangle, and the sharp corner of the triangle is the head of the vehicle. The red star marker is the ASV BUTTORI.

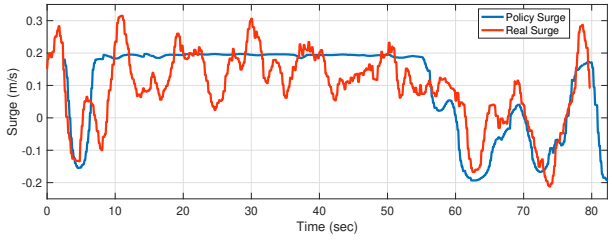


Figure 22: The surge velocity of AUV Tri-TON in the sea experiment (Dive 2). The blue curve is the surge velocity generated by the reinforcement learning policy, while the red curve is the real surge velocity measured by the DVL device. The blue curve in the figure is delayed by 2.5 seconds.

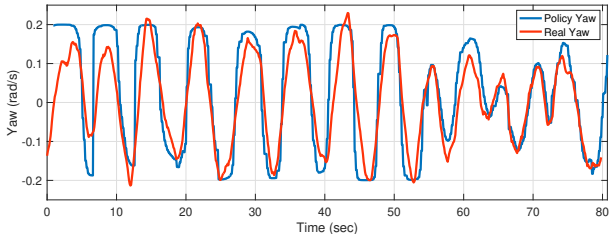


Figure 23: The yaw angular velocity of AUV Tri-TON in the sea experiment (Dive 2). The blue curve is the yaw angular velocity generated by the reinforcement learning policy, while the red curve is the yaw angular velocity measured by the FOG device. The blue curve in the figure is delayed by 0.9 seconds.

receiver location information. This situation could get better if the frequency of acoustic ranging is increased and the acoustic transmission delay is taken into account in the particle filter. However, underwater acoustic channel resources are very scarce, which is why we consider reducing the usage of acoustic channel resources in reinforcement learning optimization.

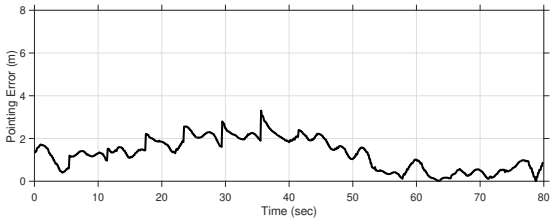


Figure 24: The pointing error d_Δ in the sea experiment (Dive 2).

The policy implemented on Tri-TON was trained in a simulated environment. The policy adaptability in field experiments was affected by the gap between simulation and the real environment. We randomized the initial conditions in the simulated environment and introduced the perturbations and noises in AUV motion. The performance of the SAC algorithm in real-world applications is one reason for choosing it (Haarnoja et al., 2018c). Maximum entropy exploration allows us to avoid hyperparameter tuning when transferring to the real environment. Action smoothing can prevent the thruster from jittering at high frequency during performing surge and yaw commands. The field experiments proved that training the reinforcement learning policy in a simulation environment and transferring the knowledge to real AUVs are available and attractive for underwater optical communication applications.

The reinforcement learning algorithm requires large amounts of data to optimize the policy. Sampling data in simulations is efficient, while conducting field experiments and collecting data in the real environment is time-consuming and costly. In addition, sampling data in a simulation environment does not need to worry about the safety problem of the underwater platform, which is difficult to avoid in the sea environment. The collected experimental data can be used in the future to retrain the policy and gradually reduce the gap between the simulation and the real world. The delay problem of the thrusters in experiments will be considered in simulations to improve the stability of the link.

5.5. Comparison

To benchmark the performance of our method, we compared it with an existing link establishment method proposed by Solanki *et al.* (Solanki et al., 2020). The basic concept of the previous methods for LOS link establishment is to quickly adjust the beam direction and search according to the set path. The method presented by Solanki *et al.* is used to solve the link establishment of LED-based underwater optical communication, and the application scenario is similar to our proposed method. Through the results of the experiments and the comparison in the simulated environment, we discuss the triangular exploration method and our approach in the following aspects: 1) applicable conditions of the algorithm; 2) maximum initial distance between two platforms; 3) time cost to establish the link; 4) stability under disturbance; 5) detected light intensity. A summary of the comparison is shown in Table 3.

Table 3
Summary of comparison with benchmark approach

	Triangular exploration	**Our approach**
Applicable conditions	Light field distribution is unimodal	Model-free
Effective initial distance	Within the coverage of the optical signal	Within the coverage of the acoustic signal
Time to establish link	21 seconds	16.4 seconds
Stability	Error increases significantly when $\mu > 0.16$	Error increases when $\mu > 0.20$
Light intensity	High	Low

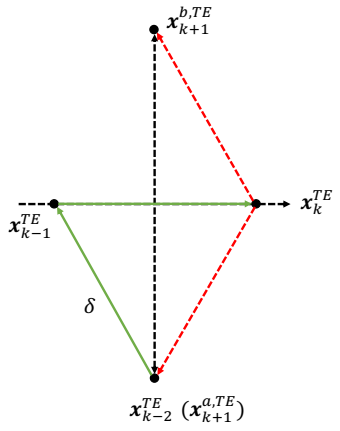


Figure 25: Illustration of the triangular-exploration algorithm (Solanki et al., 2020). The green arrow is the scanning path. Based on the light intensity of the three vertices $x_{k-2}^{TE}, x_{k-1}^{TE}, x_{k}^{TE}$, the triangular-exploration algorithm decides the next path to explore. Possible paths are indicated by red dashed arrows.

Solanki *et al.* designed an active transceiver module and a triangular-exploration mode to track the optical signals. Motors on the module are used to adjust the azimuth and elevation of the LOS link. As shown in Fig. 25, the photodiode in the module measures the light intensity of the three vertices $x_{k-2}^{TE}, x_{k-1}^{TE}, x_{k}^{TE}$ on the equilateral triangle and compares them. The increase or decrease in light intensity can determine whether the next measurement point is $x_{k+1}^{a,TE}$ or $x_{k+1}^{b,TE}$, and the latest three points $x_{k-1}^{TE}, x_{k}^{TE}, x_{k+1}^{TE}$ form a new triangle for comparison. Repeating this triangular-exploration path, the method can approach and track the optimum point in the beam. There is no device to adjust the relative distance on the module. The step size δ for triangular-exploration algorithm is $\sqrt{3}$ degrees. The sampling frequency f_s is set to 100 Hz, so the angular speed of the transceiver module ω^{TE} is 173.2 deg/s, or 3.02 rad/s.

The applicable condition of the triangular-exploration method is that the light field intensity distribution is unimodal. Light intensity based scanning methods all follow the gradient of light intensity to estimate the optimal point where the maximum intensity can be detected. Our approach uses a model-free reinforcement learning algorithm to maneuver the underwater vehicle. In the policy training, no special environment model or AUV Tri-TON's parameters are included.

The triangular-exploration method requires a quasi-static light source and the detector is covered by the optical beam. Its maximum distance to identify the initial location depends on the distance of the optical link, usually no more than tens of meters. In the swimming pool experiment, the initial relative distance between the two platforms was 1 meter. Triangular-exploration method has not discussed how to obtain the initial position information in the sea environment. In our sea experiments at Hiratsuka Port, we randomly placed Tri-TON and BUTTORI, and the initial relative distance between them was 4.31 meters. Considering the effective propagation distance of the acoustic signal, the transmitting AUV can observe the states of the receiving AUV at a distance of kilometers.

Both the experiments of the triangular-exploration method and our method recorded the time used to establish the link, but the initial conditions of the experiments were not the same. In the triangular-exploration experiment, the two platforms were one meter apart and the detector was covered by a light source. The link was created in about 21 seconds. In our sea experiment, the two platforms were initially 4.31 meters apart and had a pointing error of 6.38 meters. The pointing error of the yaw angle was 86.16 degrees, which is greater than the general beam-divergence angle. As shown in Fig. 15, Tri-TON used 16.4 seconds to approach BUTTORI and establish a link. The pointing error of the yaw angle was 0.46 degrees. As shown in Fig. 26, we also test the trained policy in the simulated environment with the random initial bearing angle. The average time cost to establish the link in these 1000 episodes is 16.68 seconds. It shows that our approach can establish the link much faster.

The stability of the LOS link is affected by external disturbances such as ocean currents and turbulence. In the triangular-exploration method, the researchers designed a scenario in which the center of the beam is kept fluctuating in a fixed path (circle) to test the algorithm's ability to track the beam under disturbances. The angular speed of the optimal point during fluctuation is set to:

$$\omega_o^{TE} = 10\sqrt{3}\pi f_M \tag{5}$$

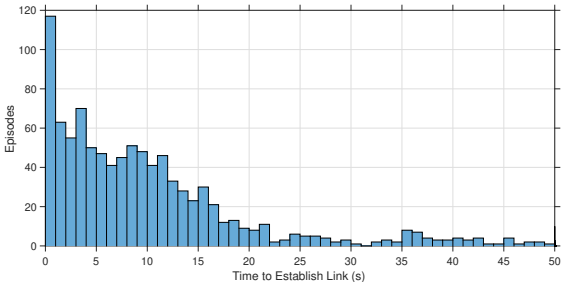


Figure 26: Distribution of time cost per episode to establish the LOS link. A total of 1000 episodes are tested, and the few episodes larger than 50 seconds are not presented.

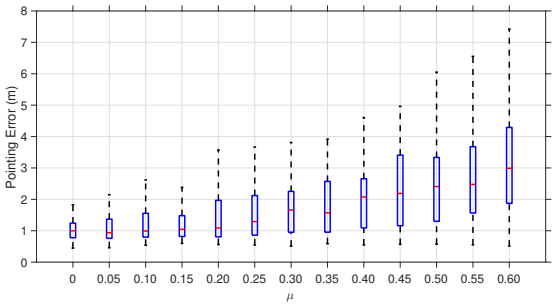


Figure 27: Pointing error of the trained policy in several scenarios with different μ values. In box-and-whisker plots, the lower and upper boundaries of the box represent the 25th (Q1) and 75th (Q3) percentiles, respectively; the bottom and top ends of the whisker indicate the most extreme values within the lower limit Q1 − 1.5(Q3 − Q1) and the upper limit Q3 + 1.5(Q3 − Q1), respectively; the red line inside the box marks the median.

where f_M is the frequency of fluctuation. Experiments show that the triangular algorithm failed to converge when the frequency f_M reached 0.5 Hz. The corresponding angular velocity ω_o^{TE} was 27.20 deg/s. For comparison with our approach, we define a relative velocity coefficient μ equal to the ratio of the velocity of the target motion to its own maximum motion velocity:

$$\mu = \frac{\omega_o^{TE}}{\omega^{TE}} \approx 0.31 f_M \tag{6}$$

According to the results presented in the study (Solanki et al., 2020), the triangular-exploration algorithm started to diverge in the tracking when the coefficient μ was 0.16 (f_M=0.5 Hz). We test the trained policy in several scenarios with different μ values. The maximum surge and yaw angular velocities of the transmitting AUV are set to 0.2 m/s and 0.2 rad/s. The receiving AUV moves continuously at a surge velocity of 0.2μ m/s, with a yaw angular velocity that takes random values in the range of 0.2μ rad/s to 0.2μ rad/s. The average pointing error of the transmitting AUV in different scenarios is shown in Fig. 27. When the μ value reaches 0.20, the pointing error starts to increase, but the increase is not as significant as that of the triangular-exploration algorithm.

The last concern is the light intensity detected by the platform. For underwater optical communication, if the received light intensity becomes higher, the signal-to-noise ratio can be improved. The triangular-exploration method outperforms our proposed method in maximizing the received light intensity. One way to improve is to combine our approach with scanning algorithms to improve the detected light intensity in the future. Since we establish the LOS link by maneuvering the AUV platform, no real-time beam control system and light intensity-based maintenance algorithm are used. According to experimental results, Tri-TON bounded the pointing error to within 2 meters. The measured pointing error can be shared with the optical system. On this basis, using the detected light intensity to adjust the direction of the optical beam can better maintain the LOS link.

6. Conclusion

The LOS link establishment between AUVs is of great significance for realizing the high data rate of optical communication in ocean exploration. With underwater optical communication, connecting mobile platforms to UIoT systems will enhance real-time data acquisition in underwater monitoring. To establish the LOS link, we propose an acoustic navigation-based solution to maintain the relative position and orientation between AUVs. The reinforcement learning algorithm is utilized to search for the optimal link establishment policy. We successfully deployed the trained reinforcement learning policy on a real AUV and completed field experiments. The experimental results show that Tri-TON successfully tracked the target and maintained the LOS link. The initial location identification for link establishment is no longer restricted to the optical coverage area.

Our research could pave the way for future applications of wireless optical communication in multiple AUVs. The implementation in the actual machine validated the performance of this new solution. The delay problem found in field experiments can be solved in the future to improve the stability of the LOS link. Compared with the previous optical methods, this link establishment method does not require additional beam pointing and scanning control. Combining our proposed method with scanning techniques may be a better way to address the challenges of underwater link establishment in the future.

CRediT authorship contribution statement

Yang Weng: Methodology, Software, Investigation, Writing - Original Draft. **Takumi Matsuda:** Software, Investigation. **Yuki Sekimori:** Software, Investigation. **Joni Pajarinen:** Software, Formal analysis, Conceptualization. **Jan Peters:** Conceptualization, Supervision. **Toshihiro Maki:** Conceptualization, Investigation, Supervision.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Arnon, S., Kedar, D., 2009. Non-line-of-sight underwater optical wireless communication network. JOSA A 26, 530–539. doi:https://doi.org/10.1364/JOSAA.26.000530.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W., 2016. Openai gym. arXiv preprint arXiv:1606.01540 doi:https://doi.org/10.48550/arXiv.1606.01540.

Fujita, K., Matsuda, T., Maki, T., 2019. Bearing only localization for multiple auv with acoustic broadcast communication, in: 2019 19th International Conference on Control, Automation and Systems (ICCAS), IEEE. pp. 1371–1376. doi:https://doi.org/10.23919/ICCAS47443.2019.8971524.

Gabriel, C., Khalighi, M.A., Bourennane, S., Léon, P., Rigaud, V., 2013. Misalignment considerations in point-to-point underwater wireless optical links, in: 2013 MTS/IEEE OCEANS-Bergen, IEEE. pp. 1–5. doi:https://doi.org/10.1109/OCEANS-Bergen.2013.6607990.

Haarnoja, T., Pong, V., Hartikainen, K., Zhou, A., Dalal, M., Levine, S., 2018a. Soft actor critic—deep reinforcement learning with real-world robots. https://bair.berkeley.edu/blog/2018/12/14/sac ,(accessed 2022 5-1).

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018b. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: International conference on machine learning, PMLR. pp. 1861–1870. doi:https://doi.org/10.48550/arXiv.1801.01290.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al., 2018c. Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905 doi:https://doi.org/10.48550/arXiv.1812.05905.

Hardy, N.D., Rao, H.G., Conrad, S.D., Howe, T.R., Scheinbart, M.S., Kaminsky, R.D., Hamilton, S.A., 2019. Demonstration of vehicle-to-vehicle optical pointing, acquisition, and tracking for undersea laser communications, in: Free-Space Laser Communications XXXI, SPIE. pp. 205–214. doi:https://doi.org/10.1117/12.2511178.

Heesemann, M., Insua, T.L., Scherwath, M., JUNIPER, K.S., Moran, K., 2014. Ocean networks canada: From geohazards research laboratories to smart ocean systems. Oceanography 27, 151–153. URL: http://www.jstor.org/stable/24862165.

Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., Wu, Y., 2018. Stable baselines. https://github.com/hill-a/stable-baselines. Accessed: May 1, 2022.

Hoeher, P.A., Sticklus, J., Harlakin, A., 2021. Underwater optical wireless communications in swarm robotics: A tutorial. IEEE Communications Surveys & Tutorials doi:https://doi.org/10.1109/COMST.2021.3111984.

Huang, J.g., Wang, H., He, C.b., Zhang, Q.f., Jing, L.y., 2018. Underwater acoustic communication and the general performance evaluation criteria. Frontiers of Information Technology & Electronic Engineering 19, 951–971. doi:https://doi.org/10.1631/FITEE.1700775.

Jahanbakht, M., Xiang, W., Hanzo, L., Azghadi, M.R., 2021. Internet of underwater things and big marine data analytics—a comprehensive survey. IEEE Communications Surveys & Tutorials doi:https://doi.org/10.1109/COMST.2021.3053118.

Kinsey, J.C., Eustice, R.M., Whitcomb, L.L., 2006. A survey of underwater vehicle navigation: Recent advances and new challenges, in: IFAC conference of manoeuvering and control of marine craft, Lisbon. pp. 1–12. URL: https://www.whoi.edu/cms/files/jkinsey-2006a_20090.pdf.

Kober, J., Bagnell, J.A., Peters, J., 2013. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32, 1238–1274. doi:https://doi.org/10.1177/0278364913495721.

Kong, M., Guo, Y., Alkhazragi, O., Sait, M., Kang, C.H., Ng, T.K., Ooi, B.S., 2022. Real-time optical-wireless video surveillance system for high visual-fidelity underwater monitoring. IEEE Photonics Journal doi:https://doi.org/10.1109/JPHOT.2022.3147844.

Koubaa, A., 2019. Ros/tutorials/understandingtopics. http://wiki.ros.org/ROS/Tutorials/UnderstandingTopics, (accessed 2022 5-1).

Kussat, N.H., Chadwell, C.D., Zimmerman, R., 2005. Absolute positioning of an autonomous underwater vehicle using gps and acoustic measurements. IEEE Journal of Oceanic Engineering 30, 153–164. doi:https://doi.org/10.1109/JOE.2004.835249.

Matsuda, T., 2021. Low-cost high-performance seafloor surveying by multiple autonomous underwater vehicles. Applied Ocean Research 117, 102762. doi:https://doi.org/10.1016/j.apor.2021.102762.

Matsuda, T., Takizawa, R., Sakamaki, T., Maki, T., 2020. Landing method of autonomous underwater vehicles for seafloor surveying. Applied Ocean Research 101, 102221. doi:https://doi.org/10.1016/j.apor.2020.102221.

Pontbriand, C., Farr, N., Ware, J., Preisig, J., Popenoe, H., 2008. Diffuse high-bandwidth optical communications, in: OCEANS 2008, IEEE. pp. 1–4. doi:https://doi.org/10.1109/OCEANS.2008.5151977.

Rong, Y., Nordholm, S., Duncan, A., 2021. On the capacity of underwater optical wireless communication systems, in: 2021 Fifth Underwater Communications and Networking Conference (UComms), IEEE. pp. 1–4. doi:https://doi.org/10.1109/UComms50339.2021.9598156.

Saeed, N., Celik, A., Al-Naffouri, T.Y., Alouini, M.S., 2019. Underwater optical wireless communications, networking, and localization: A survey. Ad Hoc Networks 94, 101935. doi:https://doi.org/10.1016/j.adhoc.2019.101935.

Sahu, S.K., Shanmugam, P., 2018. A theoretical study on the impact of particle scattering on the channel characteristics of underwater optical communication system. Optics Communications 408, 3–14. doi:https://doi.org/10.1016/j.optcom.2017.06.030.

Sait, M., Sun, X., Alkhazragi, O., Alfaraj, N., Kong, M., Ng, T.K., Ooi, B.S., 2019. The effect of turbulence on nlos underwater wireless optical communication channels. Chinese Optics Letters 17, 100013. doi:https://doi.org/10.3788/COL201917.100013.

Solanki, P.B., Bopardikar, S.D., Tan, X., 2020. Active alignment control-based led communication for underwater robots, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 1692–1698. doi:https://doi.org/10.1109/IROS45743.2020.9341442.

Stojanovic, M., 2007. On the relationship between capacity and distance in an underwater acoustic communication channel. ACM SIGMOBILE Mobile Computing and Communications Review 11, 34–43. doi:https://doi.org/10.1145/1347364.1347373.

Wang, D., Shen, Y., Wan, J., Sha, Q., Li, G., Chen, G., He, B., 2022a. Sliding mode heading control for auv based on continuous hybrid model-free and model-based reinforcement learning. Applied Ocean Research 118, 102960. doi:https://doi.org/10.1016/j.apor.2021.102960.

Wang, F., Liu, Y., Jiang, F., Chi, N., 2018. High speed underwater visible light communication system based on led employing maximum ratio combination with multi-pin reception. Optics Communications 425, 106–112. doi:https://doi.org/10.1016/j.optcom.2018.04.073.

Wang, Y., Thanyamanta, W., Bose, N., 2022b. Cooperation and compressed data exchange between multiple gliders used to map oil spills in the ocean. Applied Ocean Research 118, 102999. doi:https://doi.org/10.1016/j.apor.2021.102999.

Weng, Y., Akrour, R., Pajarinen, J., Matsuda, T., Peters, J., Maki, T., 2022. Reinforcement learning based underwater wireless optical communication alignment for autonomous underwater vehicles. IEEE Journal of Oceanic Engineering doi:https://doi.org/10.1109/JOE.2022.3165805.

Weng, Y., Guo, Y., Alkhazragi, O., Ng, T.K., Guo, J.H., Ooi, B.S., 2019. Impact of turbulent-flow-induced scintillation on deep-ocean wireless optical communication. Journal of Lightwave Technology 37, 5083–5090. doi:https://doi.org/10.1109/JLT.2019.2928465.

Weng, Y., Maki, T., 2021. Observability analysis of underwater wireless optical communication alignment between auvs, in: OCEANS 2021: San Diego–Porto, IEEE. pp. 1–6. doi:https://doi.org/10.23919/OCEANS44145.2021.9738730.

Yang, F., Cheng, J., Tsiftsis, T.A., 2014. Free-space optical communication with nonzero boresight pointing errors. IEEE Transactions on Communications 62, 713–725. doi:https://doi.org/10.1109/TCOMM.2014.010914.130249.

Zhang, R., Yang, S., Wang, Y., Wang, S., Gao, Z., Luo, C., 2020. Three-dimensional regional oceanic element field reconstruction with multiple underwater gliders in the northern south china sea. Applied Ocean Research 105, 102405. doi:https://doi.org/10.1016/j.apor.2020.102405.

Zhu, S., Chen, X., Liu, X., Zhang, G., Tian, P., 2020. Recent progress in and perspectives of underwater wireless optical communication. Progress in Quantum Electronics 73, 100274. doi:https://doi.org/10.1016/j.pquantelec.2020.100274.