# Autonomous Underwater Vehicle Link Alignment Control in Unknown Environments Using Reinforcement Learning

**Yang Weng**
Institute of Industrial Science
The University of Tokyo
Tokyo 153-8505, Japan
yangweng@iis.u-tokyo.ac.jp

**Sehwa Chun**
Institute of Industrial Science
The University of Tokyo
Tokyo 153-8505, Japan
sen0218@iis.u-tokyo.ac.jp

**Masaki Ohashi**
Institute of Industrial Science
The University of Tokyo
Tokyo 153-8505, Japan
ohashi24@iis.u-tokyo.ac.jp

**Takumi Matsuda**
School of Science and Technology
Meiji University
Kanagawa 214-8571, Japan
tmatsuda@meiji.ac.jp

**Yuki Sekimori**
Institute of Industrial Science
The University of Tokyo
Tokyo 153-8505, Japan
sekimori@iis.u-tokyo.ac.jp

**Joni Pajarinen**
Department of Electrical Engineering and Automation
Aalto University
Espoo 02150, Finland
joni.pajarinen@aalto.fi

**Jan Peters**
Computer Science Department
Technische Universität Darmstadt
Darmstadt 64289, Germany
jan.peters@tu-darmstadt.de

**Toshihiro Maki**
Institute of Industrial Science
The University of Tokyo
Tokyo 153-8505, Japan
maki@iis.u-tokyo.ac.jp

## Abstract

High-speed underwater wireless optical communication holds immense promise in ocean monitoring and surveys, providing crucial support for the real-time sharing of observational data collected by autonomous underwater vehicles (AUVs). However, due to inaccurate target information and external interference in unknown environments, link alignment is challenging and needs to be addressed. In response to these challenges, we propose a reinforcement learning-based alignment method to control the AUV to establish an optical link and maintain alignment. Our alignment control system utilizes a combination of sensors, including a depth sensor, Doppler velocity log, gyroscope, ultra-short baseline device, and acoustic modem. These sensors are used in conjunction with a particle filter to observe the environment and estimate the AUV's state accurately. The soft actor-critic algorithm is used to train a reinforcement learning-based controller in a simulated environment to reduce pointing errors and energy consumption in alignment. After experimental validation in simulation, we deployed the controller on the actual AUV Tri-TON. In experiments at sea, Tri-TON maintained the link and angular pointing errors within 1 m and 10°, respectively. Experimental results demonstrate that the proposed alignment control method can establish underwater optical communication between AUV fleets, thus improving the efficiency of marine surveys.

***Keywords***— Autonomous underwater vehicles (AUVs); Underwater wireless optical communication (UWOC); Deep reinforcement learning; Alignment control.

# 1 Introduction

As underwater mobile platforms, autonomous underwater vehicles (AUVs) can carry various sensors and instruments to perform special tasks in a wide range of marine environments, enhancing our ocean exploration (Sahoo et al., 2019). Taking the assessment of coral reefs as an example, the AUV can collect long-term visual data of the area where the reef is located and then upload the data in real-time to researchers to monitor the growth and populations of the reef (Modasshir et al., 2018). In recent years, joint surveys involving multiple AUVs have accelerated our underwater environment monitoring. Deploying a team of AUVs to conduct undersea survey missions in an area of interest can significantly increase the area covered and efficiency. Moreover, operating multiple AUVs is safer and more cost-effective than relying on a single AUV for a survey (Wang et al., 2023).

However, data sharing between AUV teams comes with trade-offs. On the one hand, there is a growing demand for high-resolution images and even video to understand the underwater world better, and these are much larger than traditional ocean data, such as temperature and salinity (Bongiorno et al., 2018). On the other hand, the AUV platform needs to share data in joint investigations. Data can be collected and uploaded to shore stations, or analyzed in real-time to determine the next step in the mission. Current underwater acoustic communication is limited by bandwidth, and the data rate is in the order of kilobit per second. This shortcoming in communication limits the efficiency of data collection in joint underwater investigations (Yang et al., 2021).

The development of underwater wireless optical communication (UWOC) can help alleviate this problem (Zhu et al., 2020) (Baiden et al., 2009). UWOC technique can create directional links between underwater platforms and transmit data at high rates. With data rates of up to a megabit or even gigabit per second, underwater optical communication is capable of real-time underwater video transmission, not to mention image data (Zhou et al., 2019) (Robertson et al., 2022). Although the effective transmission distance of UWOC is hardly 100 meters and cannot be improved in the short term, considering the mobility of AUVs, establishing short-distance optical links between platforms can still meet the communication requirements.

In order to apply underwater optical communications to AUV teams in the joint investigation, it is necessary to address how to maneuver the AUV to accomplish alignment control of the directional link for data transmission. Optical systems that can control the beam pointing together with scanning algorithms can be used for alignment control (Yang et al., 2020). Target search using such optical-only methods usually takes significant time and may lose links as the environment changes. Therefore, it is attractive to combine the motion of the AUV with optical link alignment control. The state measured by the sensors on the AUV can be used to estimate its relative relationship to the target. In 2021, Quintas *et al.* (Quintas et al., 2021) proposed a cooperative path-following method that allows two AUVs to maintain the same speed and move in parallel to complete link alignment. The alignment in the desired formation is not always stable, thus the optical link can be interrupted.

To solve the above problems, we propose a maneuvering method for underwater vehicles that allows stable alignment of optical links. During alignment control, depth, acoustic, and motion sensors are combined to observe the target, the environment, and the underwater vehicle's own state. In an unknown environment, a particle filter and a pointing error estimator are used to estimate the relative relationship between the transmitter and receiver of the optical communication system and are shared with the alignment controller and the optical system. An alignment controller based on the reinforcement learning algorithm soft actor-critic (SAC) is trained in a simulated environment to minimize pointing errors during alignment. The proposed alignment control method was tested in a simulated environment and then deployed in a real environment for validation. The experimental results show that the proposed method effectively scales down the pointing error compared to previous methods and enables AUVs to exchange data underwater using an optical communication system.

To summarize, the main contributions of the research are an alignment control method for underwater optical

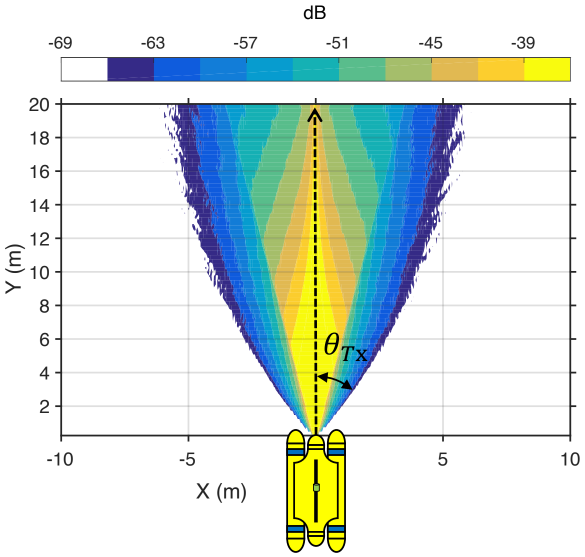


Figure 1: Simulated horizontal light field intensity distribution during underwater propagation of a beam emitted by an AUV.

communication that

1) uses a reinforcement learning-based controller to autonomously establish optical links for inter-fleet information sharing in underwater exploration;

2) uses a combination of sensors carried by AUVs to observe targets to overcome boresight and jitter in unknown environments;

3) uses acoustic ranging for fast link alignment control even at long distances, eliminating the need for time-consuming scanning;

4) can maintain an angular pointing error of less than 10° in actual machine experiments and can be used in conjunction with an optical pointing system to improve alignment further.

The rest of this article is organized as follows. The underwater optical link alignment control is described in Section 2. The reinforcement learning-based alignment method is presented in Section 3. Simulation verification is given in Section 4. Detailed experiments are conducted and discussed in Section 5. The conclusions are given in Section 6.

## 2 Problem Statement

### 2.1 Optical Link

UWOC is a communication technology that uses light to transmit data between two underwater platforms. It involves using modulated light waves to transmit data through the seawater medium. The detector at the receiving end acquires data based on light intensity. A condition for establishing a wireless optical link is that the beam from one AUV can partially cover the receiver from the other AUV. However, due to the attenuation of light in water, the beam can only be transmitted over short distances and cover part of the area. It is generally not possible to broadcast light signals. The receiving end of an optical communication system can be made omnidirectional.

Radiative transfer theory is applicable to describe the transfer of energy during underwater light propagation, taking

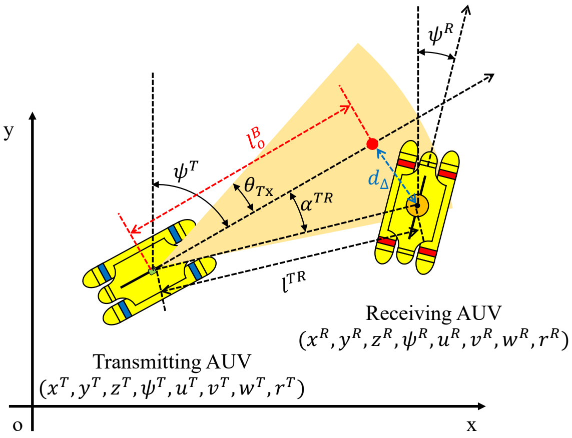


Figure 2: Alignment control for AUVs. The blue-yellow platform is the transmitting AUV, and the red-yellow platform is the receiving AUV. The orange sector is the transmitted beam.

into account the absorption, scattering, and reflection of light by water and other materials in the environment. The radiative transfer equation describes these interactions mathematically as follows:

$$dI_\lambda(s) = -c_\lambda I_\lambda(s)\, ds + c_\lambda J_\lambda(s)\, ds \tag{1}$$

where $I_\lambda(s)$ is the incident optical intensity for the wavelength $\lambda$. The source function $J_\lambda(s)$ contains the contributions of the emission and scattering processes. The attenuation coefficient $c_\lambda$ is the sum of the absorption coefficient $a_\lambda$ and scattering coefficient $b_\lambda$.

Combining the radiative transfer equation with Monte Carlo simulation allows analysis of light propagation and interactions with the underwater environment. A Monte Carlo-based simulation of the horizontal light field intensity distribution of the beam emitted by an AUV is shown in Fig. 1. The water type chosen for the simulation is clean water, and the volume scattering function is the Henyey-Greenstein function. Photons are emitted from an origin with a divergence half-angle of $\theta_{Tx}$. According to the radiative transfer equation and simulation results, the coverage area of the beam is a sector shape, and the light intensity is concentrated in the central region. The minimum conditions for establishing a link between two underwater mobile platforms are:

1) the relative angle between the two platforms is less than the divergence half-angle $\theta_{Tx}$;

2) the relative distance between the two platforms is less than the maximum distance of the optical link $l_{max}^B$.

## 2.2 Alignment Control

The alignment control of the AUV is defined as maintaining the relative position and orientation of the two AUVs so that the receiver can detect a sufficiently high light intensity from the transmitted beam. The link alignment process takes place on a horizontal plane, as two AUVs can cruise to a set depth for link alignment when optical communication is required. In previous studies, the complexity of the link alignment process has often been reduced with the help of accurate depth sensors (Hardy et al., 2019). We distinguish between the two AUVs in the link alignment control process. The hovering AUV that actively transmits the beam and performs alignment is the transmitting AUV, while the AUV that detects the light signal and receives the data is the receiving AUV. As shown in Fig. 2, we represent the state of the two platforms with the superscript T and R respectively: position $[x, y, z]$, yaw orientation $\psi$, surge velocity $u$, sway velocity $v$, heave velocity $w$, and yaw angular velocity $r$. The motion model of AUV is as follows (Fossen, 2011):

$$\dot{\eta} = J_{\Theta}(\eta)\nu \tag{2}$$

and

$$M\dot{\nu} + C(\nu)\nu + D(\nu)\nu + g(\eta) + g_0 = \tau \tag{3}$$

and

$$J_{\Theta} = \begin{bmatrix} cos(\psi) & -sin(\psi) & 0 \\ sin(\psi) & cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

where $\nu = [u \; v \; r]^T$, $\eta = [x \; y \; \psi]^T$, and $\tau = [\tau_u \; 0 \; \tau_r]^T$ is the thruster control input. $\tau_u$ is the force in the surge direction and $\tau_r$ is the torque acting on yaw angle. $M$, $C(\nu)$, $D(\nu)$ represent inertia, Coriolis and damping, respectively.

According to (1) and the simulated light field distribution, the receiving AUV is best kept in the center of the beam coverage area. The optimal length of the optical link maintenance is defined as $l_o^B$, which is influenced by factors such as the environmental attenuation coefficient and the power of the light source. A link too close or too far will negatively affect the received light intensity. We define the pointing error $d_\triangle$ as the distance between the optimal point of the optical beam and the receiving AUV, and the angular pointing error $d_\theta$ as the angle at which the receiving AUV deviates from the center of the beam:

$$d_\triangle = [(x^R - x^T - l_o^B \cos\psi^T)^2 + (y^R - y^T - l_o^B \sin\psi^T)^2]^{\frac{1}{2}} \tag{5}$$

and

$$d_\theta = \alpha^{TR} \tag{6}$$

Therefore, the requirements for link alignment control are:

1) to reduce the angular pointing error $d_\theta$ is and make it less than the divergence half-angle $\theta_{Tx}$;

2) to reduce the pointing error $d_\triangle$ and ensure that it is less than the maximum distance of the optical link $l_{max}^B$.

To save energy and increase propagation distance, the divergence half-angle of the optical transmitter $\theta_{Tx}$ is usually not large. Reducing the angular alignment error can better improve the stability of the link.

### 2.3 Unknown Environment

Boresight and jitter are two essential concepts related to the alignment and stability of the underwater optical communication link in unknown environments (Li et al., 2019). Because of the limited coverage area of the underwater beam, the link between the two AUVs can be misaligned. According to the pointing error described by (5), the position and orientation of the two platforms play a decisive role in stabilizing the link. The boresight is the displacement between the center of the beam and the receiving end, usually caused by inaccurate target position information. Jitter refers to the small, rapid variations or fluctuations in the position, usually caused by unknown disturbances. The horizontal displacements can be assumed to be Gaussian distributed (Rahman et al., 2022):

$$\tau_x \sim \mathcal{N}(\mu_x, \sigma_x^2) \tag{7}$$

and

$$\tau_y \sim \mathcal{N}(\mu_y, \sigma_y^2) \tag{8}$$

where $\mu_x$ and $\mu_y$ quantify the boresight, and $\sigma_x^2$ and $\sigma_y^2$ are the variances of jitter.

Optical communication systems can use optical tracking or feedback mechanisms to observe boresight and jitter, but it is often difficult to directly observe effects from the environment. In scenarios where mobile platforms such as AUVs carry optical communication systems for information exchange, the delay in thruster control also contributes to jitter. Boresight and jitter are not equal in different directions ($\mu_x \neq \mu_y$, $\sigma_x^2 \neq \sigma_y^2$) and are difficult to measure in unknown environments.

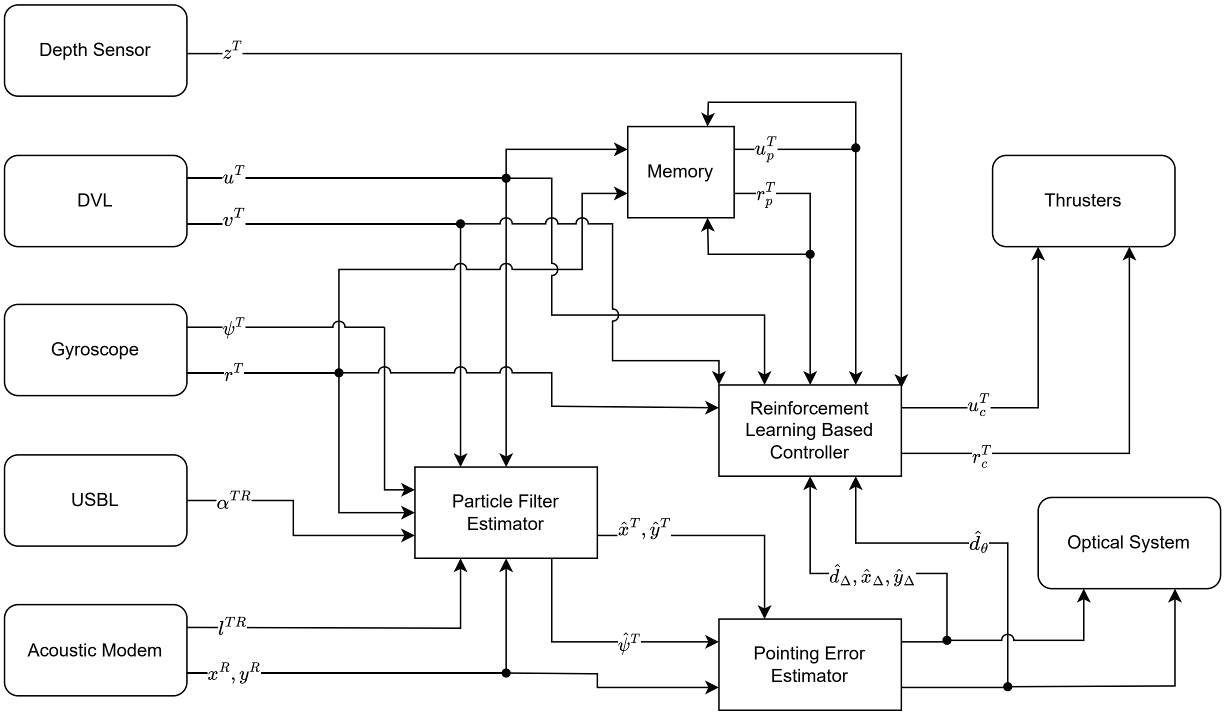


Figure 3: The diagram of the proposed alignment control method. Depth, acoustic, and motion sensors are combined to observe the target, the environment, and the underwater vehicle's state. A particle filter and a pointing error estimator are used to estimate the relative relationship and are shared with the reinforcement learning-base alignment controller and the optical system. Rounded rectangle symbols represent devices on the AUV, and rectangle symbols are used to represent processes. Labeled directional connectors indicate the corresponding data streams.

# 3 Alignment Control

## 3.1 Alignment Method

Previous methods that rely on optical systems have some limitations in observing, tracking, and controlling the alignment of targets in the sea environment. We propose to use the AUV platform for alignment control in a way that maintains the relative position and orientation relationships. The AUV platform can carry a variety of equipment to facilitate the observation of the parameters defined in pointing errors shown in (5) and (6). Then, the AUV's thruster system can adjust its position and orientation in time to track the target based on observations. The transmitting AUV uses the following devices in alignment control:

1) Depth sensor. The depth sensor can obtain accurate depth values by measuring seawater pressure. Alignment control can be simplified to the same horizontal plane when both platforms can be kept at the same depth with the help of depth information.

2) Doppler velocity log (DVL). DVL can measure the velocity of the AUV relative to the water in which it is operating. On the one hand, the AUV can observe its own velocity and estimate its position in the state estimator; on the other hand, with the help of the DVL, it can observe the jitter caused by the external disturbance as well as the delay of the thruster control, which is mentioned in (7) and (8).

3) Gyroscope. The gyroscope provides critical information about the AUV's attitude, heading, and rotation rate. This information is used for its own state estimation and allows the transmitting to reorient itself to minimize angular

**Algorithm 1** Alignment Control Algorithm

```
Initialize the particle filter estimator
Initialize acoustic ranging and communication
Initialize the pointing error estimator
Initialize the reinforcement learning controller
while alignment control is required do
    Update states $z_t^T$ through the depth sensor
    Maintain the same depth as the target
    Update states $u_t^T, v_t^T$ through the DVL
    Update states $\psi_t^T, r_t^T$ through the gyroscope
    if acoustic ranging results are available then
        Update states $\alpha_t^{TR}$
    end if
    if acoustic communication data is available then
        Update states $l_t^{TR}$, $x_t^R$, $y_t^R$
    end if
    Update states $x_t^T$, $y_t^T$, $\psi_t^T$ by the particle filter estimator
    Update states $d_{\Delta,t}$, $d_{\theta,t}$ by the pointing error estimator
    Generate actions $u_t^T$, $r_t^T$ by the reinforcement learning controller
    if $u_t^T$, $r_t^T$ are available then
        Calculate $\tau_u$ and $\tau_r$ by thrusters to perform actions $u_t^T$, $r_t^T$
    end if
    Update states in $u_{p,t}^T$ by [ $u_{t-5\Delta t}^T$, $u_{t-4\Delta t}^T$, $u_{t-3\Delta t}^T$, $u_{t-2\Delta t}^T$, $u_{t-1\Delta t}^T$ ] = [ $u_{t-4\Delta t}^T$, $u_{t-3\Delta t}^T$, $u_{t-2\Delta t}^T$, $u_{t-1\Delta t}^T$, $u_t^T$ ]
    Update states in $r_{p,t}^T$ by [ $r_{t-5\Delta t}^T$, $r_{t-4\Delta t}^T$, $r_{t-3\Delta t}^T$, $r_{t-2\Delta t}^T$, $r_{t-1\Delta t}^T$ ] = [ $r_{t-4\Delta t}^T$, $r_{t-3\Delta t}^T$, $r_{t-2\Delta t}^T$, $r_{t-1\Delta t}^T$, $r_t^T$ ]

end while
End the reinforcement learning controller
End acoustic ranging and communication
End the pointing error estimator
End the particle filter estimator
```

pointing errors.

4) Ultra-short baseline (USBL) device. The USBL system uses acoustic signals to measure the distance and bearing between the transmitting AUV and the receiving AUV. The transmitting AUV can refer to the acoustic ranging results to maintain the desired relative position and orientation for optical communication. Boresight in link alignment is usually caused by inaccurate target information. With the help of a USBL device and continuous observation, the transmitting AUV can obtain the target state and effectively minimize the displacement caused by boresight.

5) Acoustic modem. An acoustic modem is a communication device that allows two AUVs to communicate with each other over an underwater acoustic channel. Through acoustic communication, the receiving AUV can share its own status and the results of acoustic ranging. Even if they are thousands of meters apart in the marine environment, the transmitting AUV can track the receiving AUV for alignment control. In addition, it can also be used to provide feedback on link quality.

The proposed alignment control method is shown in Algorithm 1. The transmitting AUV utilizes the mentioned devices to observe the state of the receiving AUV. The two-way travel-time (TWTT) ranging is performed periodically between the USBL devices on the two platforms at intervals $T_{ac}$ (Matsuda et al., 2019). Through the acoustic channel, the receiving AUV shares its position information and acoustic ranging results with the transmitting AUV. A particle filter on the transmitting AUV estimates its own state based on the observations. A pointing error estimator calculates the pointing error of the optical communication link and shares it with the controller and the optical system. The reinforcement learning controller generates control commands and outputs them to the thrusters.

The previous velocity values are stored in the memory and will be fed into the reinforcement learning-based controller for comparison with the generated commands to estimate external disturbances and delays in the thruster system. The data flow diagram in alignment control is shown in Fig. 3.

### 3.2 State Observation

In alignment control, the transmitting AUV utilizes TWTT ranging to observe the receiving AUV and combine it with data from onboard sensors to estimate its position and orientation. The particle filter method can estimate the current state based on the acoustic measurement results. A particle filter estimator is designed to be used in the alignment control, where each particle is defined as:

$$\xi^i = \{s^{PF,i}, W^i\} \tag{9}$$

$$s^{PF,i} = \begin{bmatrix} x^{T,i} & y^{T,i} & \psi^{T,i} \end{bmatrix}^{\mathrm{T}} \tag{10}$$

where $i = 1, 2, \cdots, \mathrm{N}$ is the identification, indicating the $i$th particles in the estimator. $s^{PF,i}$ is the estimated state, and $W^i$ is the weight.

The state transition from time $t$ to $t + \Delta t$ is as follows:

$$x^{T,i}_{t+\Delta t} = x^{T,i}_t + (u^{T,i}_t \cos\psi^{T,i}_t - v^{T,i}_t \sin\psi^{T,i}_t)\Delta t \tag{11}$$

$$y^{T,i}_{t+\Delta t} = y^{T,i}_t + (u^{T,i}_t \sin\psi^{T,i}_t + v^{T,i}_t \cos\psi^{T,i}_t)\Delta t \tag{12}$$

$$\psi^{T,i}_{t+\Delta t} = \psi^{T,i}_t + r^{T,i}_t \Delta t \tag{13}$$

$$u^{T,i}_t \sim \mathcal{N}(u^T_t, (\sigma^T_{u,t})^2) \tag{14}$$

$$v^{T,i}_t \sim \mathcal{N}(v^T_t, (\sigma^T_{v,t})^2) \tag{15}$$

$$r^{T,i}_t \sim \mathcal{N}(r^T_t, (\sigma^T_{r,t})^2) \tag{16}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian probability distribution with mean $\mu$ and standard deviation $\sigma$.

When the results of the acoustic observations are available, the transmitting AUV can update particle weights for resampling:

$$W^i = \max\left\{\exp\left\{(\frac{k_d^2}{2} + \frac{-(\Delta d^i)^2}{2(\sigma_d)^2})\right\}\exp\left\{(\frac{k_\alpha^2}{2} + \frac{-(\Delta \alpha^i)^2}{2(\sigma_\alpha)^2})\right\}, 1\right\} \tag{17}$$

where $\Delta\alpha$ is the absolute difference in relative bearing angle between observation and prediction, and $\Delta d$ is the difference in relative distance. $k_d$ and $k_\alpha$ are parameters for judging outliers (Maki et al., 2007a).

The current state can be determined from the generated particles:

$$\hat{x}^T_t = \frac{\sum_{i=1}^N x^{T,i}_t}{N} \tag{18}$$

**Algorithm 2** Soft Actor-Critic (Haarnoja et al., 2018)

```
Input:
  Initialize target network $\theta_1, \theta_2$
  Initialize policy network $\phi$
  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 $a_t$
      Sample reward $r(s_t, a_t)$
      Sample new state $s_{t+\Delta t}$ by $s_{t+\Delta t} \sim p(s_{t+\Delta t}|s_t, a_t)$
      Store samples by $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$
    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: $\theta_1, \theta_2, \phi$
```

$$\hat{y}_t^T = \frac{\sum_{i=1}^N y_t^{T,i}}{N} \tag{19}$$

$$\hat{\psi}_t^T = \frac{\sum_{i=1}^N \psi_t^{T,i}}{N} \tag{20}$$

where the variables with hat symbols are updated by the particle filter estimator.

Once the particle filter has estimated its own position and orientation based on the measured data, a pointing error estimator on the transmitting AUV can use its own estimated state and the received target state to calculate the pointing error for alignment control:

$$\hat{x}_\triangle = x^R - \hat{x}^T - l_o^B \cos \hat{\psi}^T \tag{21}$$

$$\hat{y}_\triangle = y^R - \hat{y}^T - l_o^B \sin \hat{\psi}^T \tag{22}$$

where $x_\triangle$ and $y_\triangle$ are pointing errors in the northern and eastern directions. $d_\triangle$ and $d_\theta$ are then calculated according to (5) and (6), respectively. Estimated pointing errors are output to the reinforcement learning controller for generating commands. In addition to this, pointing errors can be shared with the optical system. Previously, optical technology could not observe pointing errors due to boresight and jitter effects. We leave adjusting the link alignment in the optical system based on pointing errors as future work.

### 3.3 Reinforcement Learning Controller

After the AUV observes the environment and the target state, we use a reinforcement learning-based controller to generate commands for alignment control. The use of reinforcement learning algorithms has the following main considerations:

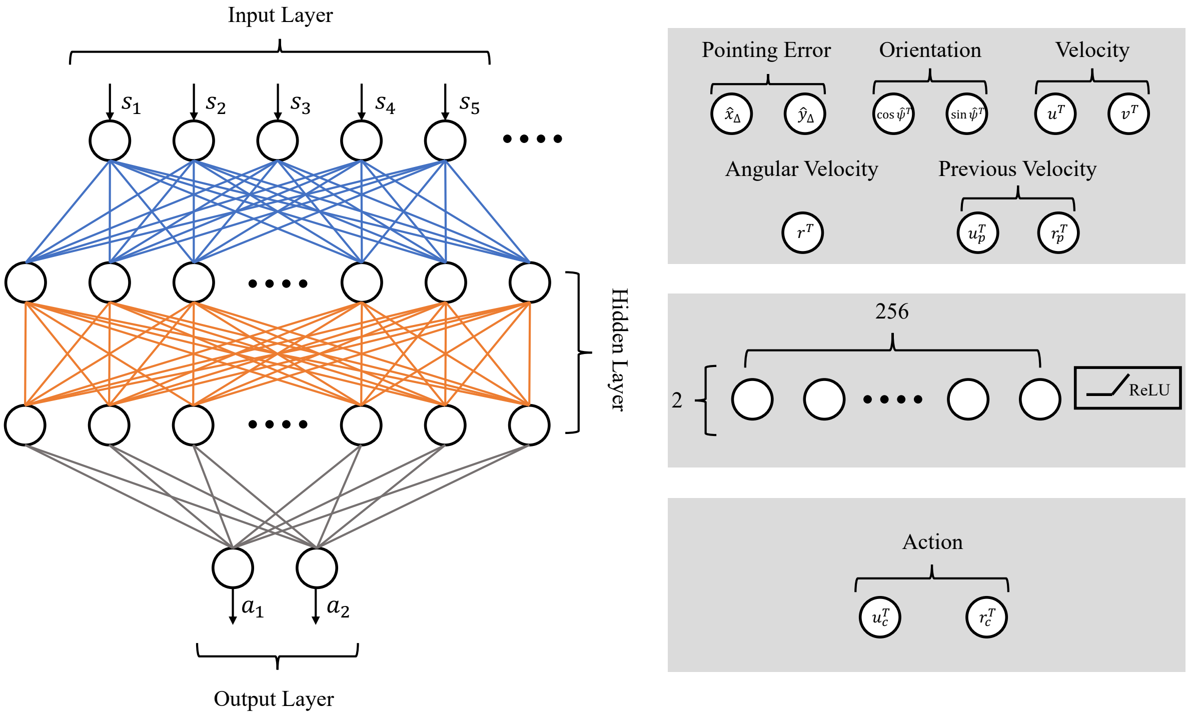


Figure 4: The structure of the reinforcement learning controller.

1) Reinforcement learning algorithms can fuse different data and make decisions. We use measurement and estimation data from DVL, acoustic devices, gyroscope, and particle filter in link alignment. In the future, we will include optical communication link quality to enhance the performance of alignment control. Reinforcement learning algorithms can fuse data and make complex decisions in high-dimensional and continuous spaces to find optimal or near-optimal solutions.

2) Reinforcement learning-based controller optimizes multiple objectives simultaneously. Compared to the guidance method, we expect the command output from the reinforcement learning-based controller to reduce pointing errors while considering energy consumption, action smoothing, and misalignment due to boresight and jitter.

3) As a model-free approach, we do not need to model the environment and the underwater vehicle explicitly. When AUVs share data using optical communications, the surrounding environment is usually dynamic and unknown. Misalignment due to boresight and jitter cannot be modeled accurately. Reinforcement learning algorithms can learn to map states directly to actions based on trial-and-error interactions with the environment, without the need to understand the underlying dynamics or rules.

The structure of the reinforcement learning controller is shown in Fig. 4.To use reinforcement learning, we define the state space of the agent as:

$$s = [\hat{x}_\Delta, \hat{y}_\Delta, \cos\hat{\psi}^T, \sin\hat{\psi}^T, u^T, v^T, r^T, u_p^T, r_p^T] \tag{23}$$

where $u_p^T$ is the value of the five previously measured surge velocities and $r_p^T$ is the value of the five previously measured yaw angular velocities:

$$u_{p,t}^T = [u_{t-\Delta t}^T, u_{t-2\Delta t}^T, u_{t-3\Delta t}^T, u_{t-4\Delta t}^T, u_{t-5\Delta t}^T] \tag{24}$$

and

$$r_{p,t}^T = [r_{t-\Delta t}^T, r_{t-2\Delta t}^T, r_{t-3\Delta t}^T, r_{t-4\Delta t}^T, r_{t-5\Delta t}^T] \tag{25}$$

All variables in the state space are one-dimensional continuous variables. Previous states can be compared with generated actions to estimate external disturbances and delays in the thruster system. All variables can be observed or estimated in the real environment by the method shown in Fig. 3. We do not include pixels in the state space for image processing, which can reduce the complexity during the training or deployment of the reinforcement learning algorithm.

The action space of the agent is as follows:

$$a = [u_c^T, r_c^T] \tag{26}$$

where $u_c^T$ and $r_c^T$ are surge velocity and yaw angular velocity commands for alignment control, which are one-dimensional continuous.

The reward function is used to assess and provide feedback on the quality of the agent's actions to guide the learning process. In reinforcement learning, the basic goal of the agent is to maximize the sum of the cumulative rewards it receives. We consider three parts in designing the reward function. The first part aims to minimize pointing errors, which is a fundamental requirement in alignment control tasks. We set the first reward function $r_{po}$ as follows:

$$r_{po}(s,a) = \rho_1 d_\Delta^{\frac{1}{2}}(1 + \rho_2|d_\theta|) \tag{27}$$

where $\rho_1$ and $\rho_2$ are coefficients to determine the importance of different reward terms and are usually negative. As discussed in the previous section, the limitation of the beam divergence angle is more challenging than the propagation distance of the optical link. Therefore, we design a high negative reward for the angular pointing error to prevent the target from deviating from the center of the beam.

The second part is to optimize energy consumption. When the two platforms are far apart, the transmitting AUV can output larger velocity commands to track the target at the cost of a low negative reward. When the two platforms are close together, high-velocity commands are not recommended because of possible oscillations and instability. We set the second reward function $r_{ve}$ as follows:

$$r_{ve}(s,a) = (\rho_3|u_c^T| + \rho_4|r_c^T|)\rho_{dist} \tag{28}$$

and

$$\rho_{dist} = \min(\rho_{dist\_max}, 1 + \frac{1}{d_\Delta}) \tag{29}$$

where $\rho_{dist}$ is a coefficient related to the pointing error, and $\rho_{dist\_max}$ is the maximum value of this coefficient. $\rho_3$ and $\rho_4$ are coefficients to determine the importance of different reward terms.

The third part is used to calculate the difference with the previous velocities, preventing the output velocity command from changing at high frequencies. Because of the delay, the thrusters cannot handle the rapid jittering, which can also cause damage to the thrusters. We set the third reward function $r_{pr}$ as follows:

$$\begin{aligned} r_{pr}(s,a) =& \rho_5 \sum_{i=1}^{5}(|u_c^T - u_{t-i\Delta t}^T|\rho_{disc}^{i-1}) \\ &+ \rho_6 \sum_{i=1}^{5}(|r_c^T - r_{t-i\Delta t}^T|\rho_{disc}^{i-1}) \end{aligned} \tag{30}$$

where $\rho_{disc}$ is a discount factor for previous states. $\rho_5$ and $\rho_6$ are coefficients to determine the importance of different reward terms.

Finally, we propose the following reward function form for reinforcement learning:

$$r(s,a) = r_{po}(s,a) + r_{ve}(s,a) + r_{pr}(s,a) \tag{31}$$

We train the reinforcement learning controller using the SAC algorithm proposed by Haarnoja *et al.* (Haarnoja et al., 2018). As shown in Algorithm 2, the SAC algorithm searches for an optimization policy that collects the maximum cumulative reward and entropy. The SAC algorithm is used to train the controller because it has several properties when applied to real robots:

Table 1: Hyperparameters configuration

| Parameter | Symbol | Value |
|---|---|---|
| Layer of MLP | | 2 |
| Neuron of MLP | | 256 |
| Discount factor | $\gamma$ | 0.99 |
| Learning rate | $\lambda$ | 0.0003 |
| Buffer size | | 1000000 |
| Batch size | | 256 |
| Activation | | ReLU |

1) The SAC algorithm has good sampling efficiency, which reduces the difficulty of training. This algorithm can learn real-world tasks in a short period of time.

2) Introducing entropy in training encourages the agent to explore more extensively and prevents the policy from prematurely converging to a bad local optimum.

3) Including maximum entropy in the objective function provides a robust framework that minimizes the need for hyperparameter tuning when we deploy the controller on a real robot.

# 4 Simulation

## 4.1 Policy Training

A simulation environment was designed to sample data for reinforcement learning and then validate the proposed alignment method. The OpenAI Gym interface (Brockman et al., 2016) was used in the simulation environment so that the reinforcement learning algorithm could sample the state and actions of each step for learning.

In the simulation environment, each episode of the alignment task has a fixed duration and contains a total of 1500 time steps, each equivalent to 0.2 seconds. At the beginning of the alignment task, the transmitting and receiving AUVs are randomly placed on a horizontal plane. At each time step, the AUV moves according to the current command, and its position and direction are calculated based on (2). The range of surge and yaw angular velocities are set to -0.2 to 0.2 m/s and -0.2 to 0.2 rad/s, respectively. The optimal length of the optical link maintenance $l_o^B$ is set to 5 m. In order to make the simulated environment more similar to the real environment, we include external disturbances in the simulation. At each step, the surge, sway, and yaw motions of the AUV are affected by Gaussian noise with standard deviations of 0.1 m/s, 0.1 m/s, and 1 deg/s, respectively.

We sample data from the simulated environment to train the reinforcement learning controller for alignment control. Collecting data from simulated environments is more efficient and secure than sampling data from actual environments. As listed in Algorithm 2, the agent interacts with the environment in the simulation and collects data to optimize the controller by updating the policy. The OpenAI Stable Baselines3 toolkit was used to implement the SAC algorithm in controller training (Raffin et al., 2021). The hyperparameters for reinforcement learning are listed in Table 1. As shown in Fig. 4, the neural network consisting of two hidden layers with 256 neurons each was used to initialize the target and policy networks. We did not scale up the size of the neural network because no cameras and images were used in this study and the observation dimensions were small. Drawing on some benchmark applications of the SAC algorithm (Haarnoja et al., 2018), we set the discount factor $\gamma$ and the learning rate $\lambda$ to 0.99 and 0.0003, respectively.

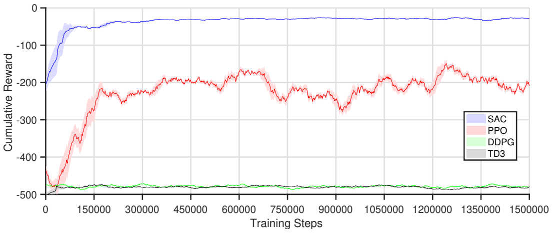


Figure 5: The training process of reinforcement learning controller. The curves represent the moving average (window of 60,000 steps) and standard deviation of the cumulative reward value during training for different reinforcement learning algorithms.

Table 2: Comparison of different reward functions

| **Case** | **Maximum alignment** | **Total alignment** | **Pointing error** | **Angular pointing error** | **Surge command** | **Yaw command** |
|---|---|---|---|---|---|---|
| Case 1 | 460 | 758 | 0.65 | 3.55 | 40.57 | 31.38 |
| Case 2 | 463 | 754 | 0.72 | 3.48 | 52.38 | 38.29 |
| Case 3 | 305 | 558 | 1.10 | 5.56 | 65.27 | 56.91 |
| Case 4 | 489 | 801 | 0.58 | 3.86 | 95.75 | 34.25 |

In training, the optimal length of the optical link maintenance $l_o^B$ is set to 5 m. The maximum surge and yaw angular velocities are set to 0.2 m/s and 0.2 rad/s, respectively. The coefficients $\rho_1$ to $\rho_6$ of the reward function are set to 0.01, 0.05, 0.01, 0.01, 0.01, and 0.01. The coefficients $\rho_{dist\_max}$ and $\rho_{disc}$ are set to 10 and 0.9. The agent needs to interact with the environment for 1000 episodes, sampling a total of $1.5 \times 10^6$ data for reinforcement learning.

## 4.2 Comparison

The goal of the policy training is to find an optimized controller that can allow the AUV to complete the alignment control of the optical communication underwater. During the training process, we used different reinforcement learning algorithms to train the controller, and also compared different reward functions.

We select four commonly used reinforcement learning algorithms for comparison. In addition to the SAC algorithm, there are Proximal Policy Optimization (PPO) (Schulman et al., 2017), Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015), and Twin Delayed DDPG (TD3) (Fujimoto et al., 2018) algorithm. All the above methods are applicable to continuous state space and action space. The four algorithms used the same reward function and sampled a total of $1.5 \times 10^6$ data for policy training. Algorithms are implemented using the OpenAI Stable Baselines3 toolkit (Raffin et al., 2021).

The learning curves for the four algorithms are shown in Fig. 5. The curve represents the moving average cumulative reward for the surrounding 60,000 steps, and the shaded patches around the curve represent the standard deviation. As shown in the results, the learning efficiency of DDPG and TD3 is low because both algorithms require complex tuning of the hyperparameters. The SAC algorithm possesses higher learning efficiency and stability than the PPO algorithm in alignment controller training. The performance of the SAC algorithm shows that it is more suitable for later deployment on real machines for practical experiments.

In the design of the reward function, we propose three components for optimizing the alignment control of an AUV. During training, we tried different reward functions to compare their performance. In the first case, the coefficients of the reward function $\rho_1$ to $\rho_6$ are set to 0.01, 0.05, 0.01, 0.01, 0.01, 0.01, and 0.01, respectively. In the second case, the energy consumption optimization part is ignored, and the coefficients are set to 0.01, 0.05, 0.01, 0, 0, 0.01, and 0.01. The third case does not include rewards from previous velocities, and the coefficient is set to 0.01, 0.05, 0.01, 0.01, 0.01, 0, and 0.

We also compare the reinforcement learning-based control with the existing guidance method. Based on the line-of-sight (LOS) guidance method, the transmitting AUV directly estimates and minimizes the point error $d_\Delta$ and the angular pointing error $d_\theta$ (Oh and Sun, 2010) (Fossen et al., 2003) (Maki et al., 2007b). Since there is no requirement for a specific path and there is only one target point, the surge velocity $u^T$ and yaw angular velocity $r^T$ are given by:

$$u^T = \begin{cases} K_{p,l}d_\Delta, & K_{p,l}d_\Delta \leq u^T_{max} \; and \; l^{TR} \geq l^B_o \\ u^T_{max}, & u^T_{max} < K_{p,l}d_\Delta \; and \; l^{TR} \geq l^B_o \\ -K_{p,l}d_\Delta, & K_{p,l}d_\Delta \leq u^T_{max} \; and \; l^{TR} < l^B_o \\ -u^T_{max}, & u^T_{max} < K_{p,l}d_\Delta \; and \; l^{TR} < l^B_o \end{cases} \tag{32}$$

$$r^T = \begin{cases} K_{p,\theta}d_\theta, & K_{p,\theta}d_\theta \leq r^T_{max} \; and \; d_\theta \geq 0 \\ r^T_{max}, & r^T_{max} < K_{p,\theta}d_\theta \; and \; d_\theta \geq 0 \\ -r^T_{max}, & K_{p,\theta}d_\theta < -r^T_{max} \; and \; d_\theta < 0 \\ K_{p,\theta}d_\theta, & -r^T_{max} \leq K_{p,\theta}d_\theta \; and \; d_\theta < 0 \end{cases} \tag{33}$$

where $u^T_{max}$ and $r^T_{max}$ are the maximum surge and yaw angle velocities of the transmitting AUV in alignment control, $K_{p,d}$ and $K_{p,\psi}$ are the proportional gains of surge and yaw angle velocities.

The first three cases are trained using the SAC algorithm, with a total of $1.5 \times 10^6$ samples of data. After policy training, each of the three reinforcement learning controllers is tested in the simulation environment for 1000 episodes. The fourth guidance method is also tested 1000 times in the simulation environment. We set both $K_{p,d}$ and $K_{p,\psi}$ to 0.25. The statistics for each case are shown in Table 2: maximum alignment time step maintained by AUV (pointing error less than 1 m), time step in an episode in which the AUV maintains alignment, pointing error, angular pointing error, the sum of surge velocity commands in an episode, and the sum of yaw angular velocity commands in an episode.

The comparison of the first three cases shows that our proposed reward functions effectively optimize each objective. Although the LOS method maintains a longer link duration than the reinforcement learning method, the LOS method has disadvantages in terms of angular pointing error and energy consumption. The reinforcement learning-based controller with the proposed reward function can optimize the pointing error while maintaining the stability of the link and reducing energy consumption.

### 4.3 Simulation Evaluation

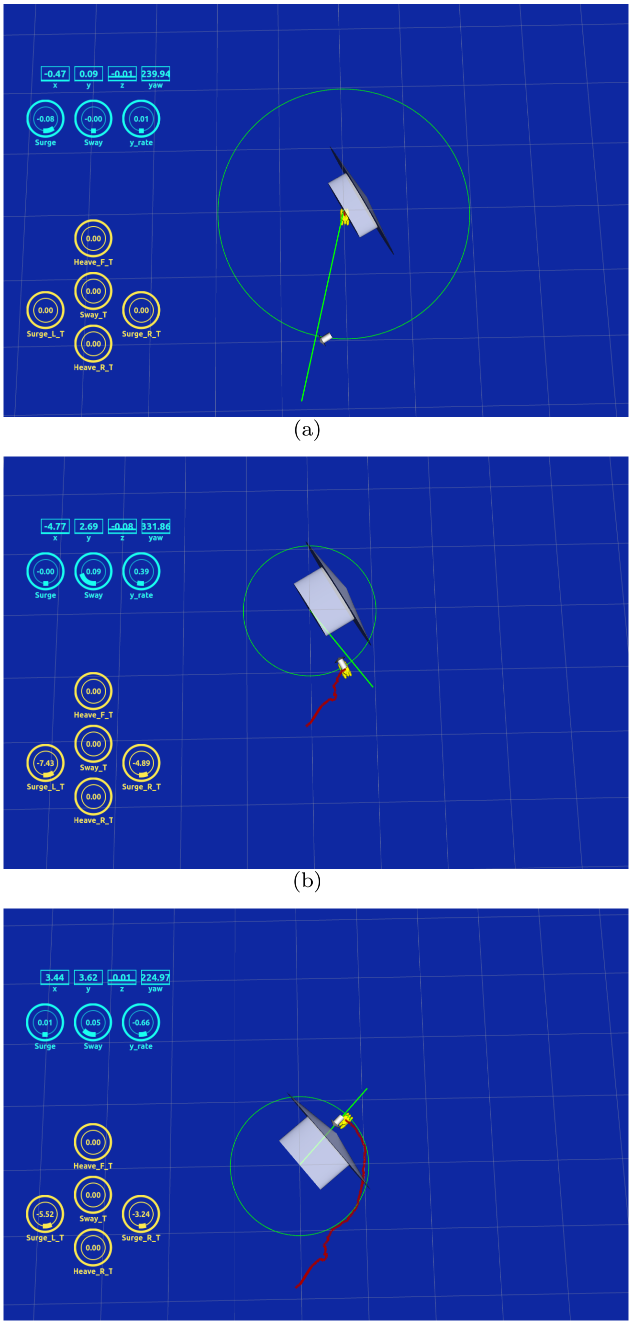
The reinforcement learning algorithm does not consider the environment and the AUV's modeling in policy training. The sample data used for policy training is collected from a simulated environment. The gap between simulation and real environments can degrade the performance of the reinforcement learning controller in the real environment. Before deploying the trained policy on a real machine, we need to evaluate it in a simulated test environment that is closer to the real environment. The simulation test is designed to evaluate and discuss the following:

1) to test whether our proposed alignment control method can be implemented on the robotic system and manipulate the actual machine;

2) to verify whether the particle filter combined with the acoustic ranging method can accurately observe the pointing error and estimate its own state in unknown environments;

3) to evaluate whether the reinforcement learning-based controller can accomplish alignment control even in the presence of environmental disturbances.

The simulation environment for testing is built on Robot Operating System (ROS) (Stanford Artificial Intelligence Laboratory, 2020). The same system is later used to manipulate the AUV to perform the alignment task in sea



(a)

(b)

(c)

Figure 6: A test experiment in the simulated environment. The blue patterns on the left side show the position $(x^T, y^T, z^T)$, yaw angle $\psi^T$, surge $u^T$, sway $v^T$, and yaw angular velocity $r^T$ of the transmitting AUV, and the yellow patterns represent the action of the thrusters. The length of the grid in the simulation environment is 5 m.

Table 3: Parameters of the test environment

| **Parameter** | **Value** | **Description** |
|---|---|---|
| $L_{init}$ | 10 | Initial relative distance |
| $T_{ac}$ | 6 | Period of acoustic ranging |
| $\sigma_{DVL}$ | $0.01+0.038\sqrt{\vert v\vert}$ | Standard deviation (SD) of DVL |
| $\sigma_{\phi}$ | 5/3600 | SD of yaw angle measurements |
| $\sigma_{r}$ | 0.01 | SD of yaw angular velocity measurements |
| $\sigma_{z}$ | 0.05 | SD of depth measurements |
| $\sigma_{d}^{USBL}$ | 0.5 | SD of USBL measurements (distance) |
| $\sigma_{\alpha}^{USBL}$ | 5 | SD of USBL measurements (angle) |

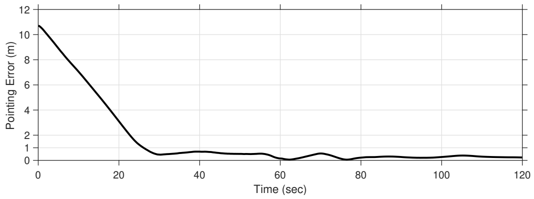


Figure 7: The pointing error $d_{\Delta}$ in the test experiment.

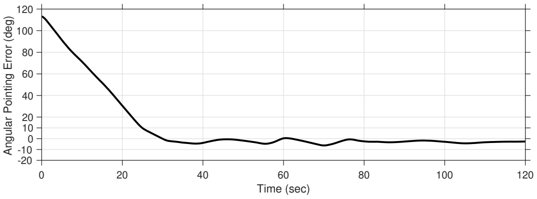


Figure 8: The angular pointing error $d_{\theta}$ in the test experiment.

experiments. The devices and algorithms mentioned in Fig. 3 are used and implemented in the test environment. The dynamic model presented by Fossen (Fossen, 2011) is implemented in this test environment. We plan to deploy the AUV Tri-TON in real sea experiments, so the specifications and parameters related to the actual machine are measured in real sea experiments (Matsuda, 2021)(Matsuda et al., 2022) and applied to the test environment. The parameters in the test environment are listed in Table 3. We tested the performance of the USBL carried by AUV Tri-TON in measuring the relative bearing angle and distance in sea experiments. We build the test environment using results obtained from previous USBL performance test experiments, as discussed in detail in our previous study (Matsuda et al., 2024) (Sekimori et al., 2024). Random currents are introduced in the test environment, interfering with the optical link alignment control. The velocity of the currents varies with time and is expressed as follows:

$$v_x^W = \rho_1^W + \rho_2^W \sin t_{sim} \tag{34}$$

$$v_y^W = \rho_3^W + \rho_4^W \cos t_{sim} \tag{35}$$

where $t_{sim}$ is the time in the simulated test environment. The coefficients $\rho_1^W$ to $\rho_4^W$ are used to adjust the velocity of the currents.

We deployed the trained reinforcement learning policy on the ROS system and tested the alignment control method. One of the test experiments in the simulated environment is shown in Fig. 6. The large grey arrow represents the currents, the small grey arrow indicates the real position of the transmitting AUV, and the yellow vehicle represents the position of the transmitting AUV as estimated by the particle filter. The green circle and line represent the relative distance and angle measured by acoustic ranging. The trajectory of the transmitting AUV is presented by the red line. We placed the receiving AUV at the origin position and kept it dynamically positioning, while the transmitting AUV was randomly placed on the same horizontal plane and maintained an initial relative distance of 10 m from the receiving AUV. The parameters for the currents $\rho_1^W$ to $\rho_4^W$ were set to 0.10, 0.05, 0.10, and 0.05, respectively.

In the test environment, we need to verify whether the proposed particle filter combined with the acoustic ranging can accurately estimate the relative distance and bearing angle to the target. The reinforcement learning-based controller can generate commands to accomplish alignment control only when the particle filter provides confident results. As shown in Fig. 6 (a), the position of the yellow vehicle (estimated position) was not close to the grey arrow (real position). After the alignment task is requested, the particle filter needs to estimate the position of the transmitting AUV in combination with the measured relative distance and bearing angle. In Fig. 6 (b), the transmitting AUV had estimated its position and the relative relationship with the receiving AUV. At this point, the yellow vehicle coincided with the gray arrow, indicating that the particle filter provided accurate estimates under the errors of the real environment. The reinforcement learning-based controller generated action commands based on the observations, and then the decision was executed by the thruster through the ROS system. Due to the sea currents in the simulated environment, the transmitting AUV gradually adjusted its position to the external disturbances. Fig. 6 (c) shows that the transmitting AUV adjusted the yaw angle against the current to keep the link aligned. The pointing error and angular pointing error in the alignment control are shown in Fig. 7 and 8, respectively. The reinforcement learning controller we deployed maintained the link pointing error and angular pointing error within 1 m and 10°.

In the test experiment, the proposed method shown in Fig. 3 successfully observed the target and performed alignment control in an unknown environment. The results verify that our proposed alignment control method can be deployed on an actual machine, and that the same system setup will be used in sea experiments.

# 5 Sea Experiments

## 5.1 Preparation

We need to deploy two platforms in sea experiments to validate our proposed alignment control algorithm. The hovering-type AUV Tri-TON is the transmitting AUV, while the autonomous surface vehicle BUTTORI is the receiving AUV.

The specifications of the AUV Tri-TON are given in Table 4. As shown in Fig. 9, the transmitting AUV Tri-TON is configured to meet the requirements for performing alignment control. One thruster is configured in the transverse

Table 4: 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 |
| Gyroscope | JAE JG-35FD |
| Depth Sensor | Mensor DPT6000 |

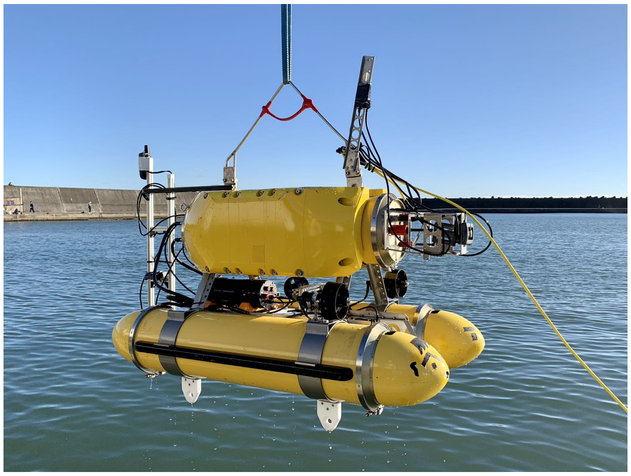


Figure 9: Hovering-type AUV Tri-TON

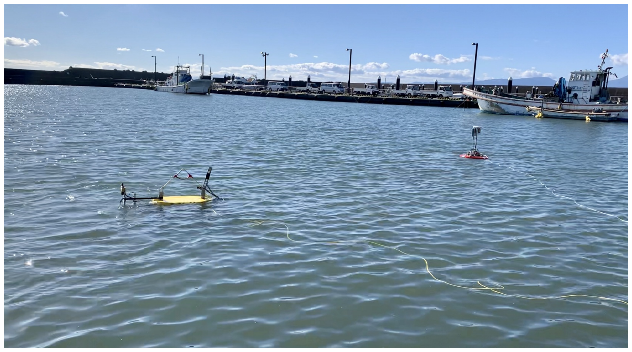


Figure 10: We deployed Tri-TON and BUTTORI for alignment control experiments in Hiratsuka, Japan. The yellow platform on the left side of the figure is Tri-TON, and the red platform on the right side is BUTTORI.

direction for sway motion, and two thrusters are mounted symmetrically on both sides of the longitudinal axis to control the surge and yaw motions. The other two thrusters are mounted in the vertical direction for controlling the heave motion. The equipment required for alignment listed in Fig. 3 is mounted on Tri-TON. The Mensor DPT6000 is used as the depth sensor, providing accurate depth information. Teledyne RDI Navigator 1200 kHz and JAE JG-35FD are used as the DVL and gyroscope, respectively, to measure the platform's motion. SeaTrac X150 is used as a USBL device and an acoustic communication modem to detect the status of the receiving AUV.

BUTTORI also carries SeaTrac as the USBL device and the acoustic communication modem to complete acoustic ranging and communication with Tri-TON. BUTTORI has a dynamic positioning capability to cope with sea currents. The platform supports wireless connectivity and global navigation satellite systems in experiments.

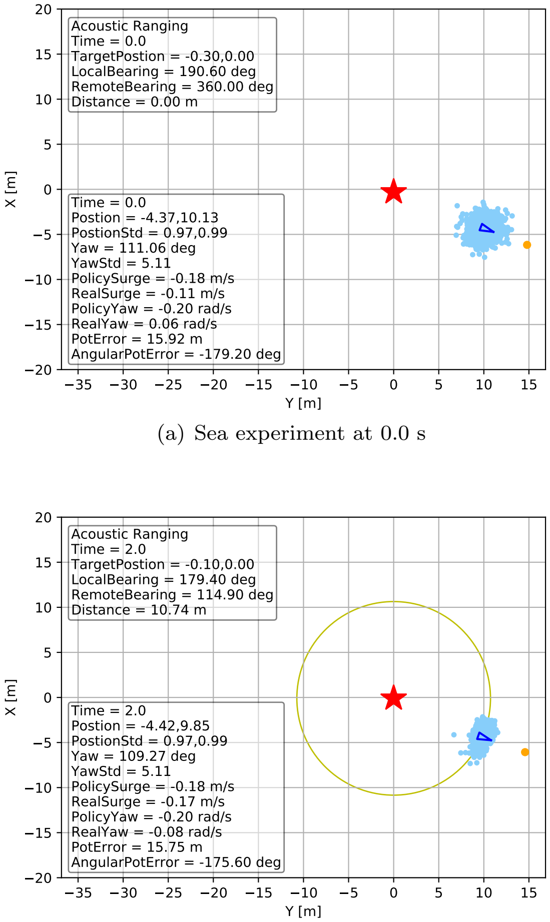
### 5.2 Experimental Results

As shown in Fig. 10, we deployed two platforms at the Hiratsuka port to test our alignment control method. The yellow platform on the left is Tri-TON for transmitting light signals, and the red platform on the right is BUTTORI for receiving light signals.

In the experiments, the trained reinforcement learning controller was deployed on Tri-TON. The maximum surge and yaw angular velocities were set to 0.2 m/s and 0.2 rad/s, respectively. The optimal length of the optical link maintenance $l_o^B$ was set to 5 m. The period of acoustic ranging between Tri-TON and BUTTORI was set to 6 seconds. One of the alignment control experiments is presented in Fig. 11 and Fig. 12. The AUV Tri-TON is represented by the blue triangle, while the red star marker is BUTTORI. The particle filter on Tri-TON estimates its position and orientation, with each estimate indicated by a blue dot in the figure. The orange point in front of the blue triangle is the optimal point for underwater optical communication. A yellow circle represents the relative distance between the two platforms obtained in acoustic ranging.

The top left and bottom left of the figure list the states measured in the experiment. The parameters listed in the bottom left corner are the time in the experiment, the position of Tri-TON $(x^T, y^T)$, the standard deviation of particle filter estimation in the position, the yaw orientation of Tri-TON $\psi^T$, the standard deviation of particle filter estimation in yaw, the surge velocity command $u_c^T$ generated by reinforcement learning controller, the surge velocity measured by DVL, the yaw angular velocity command $r_c^T$, the yaw angular velocity measured by gyroscope, the pointing error $d_\Delta$, and the angular pointing error $d_\theta$. The parameters listed in the top left corner are the time when the latest acoustic measurement results are received, the position of the BUTTORI $(x^R, y^R)$, the relative bearing angle measured by the USBL device in Tri-TON, the relative bearing angle measured by the USBL device in BUTTORI, and the relative distance.

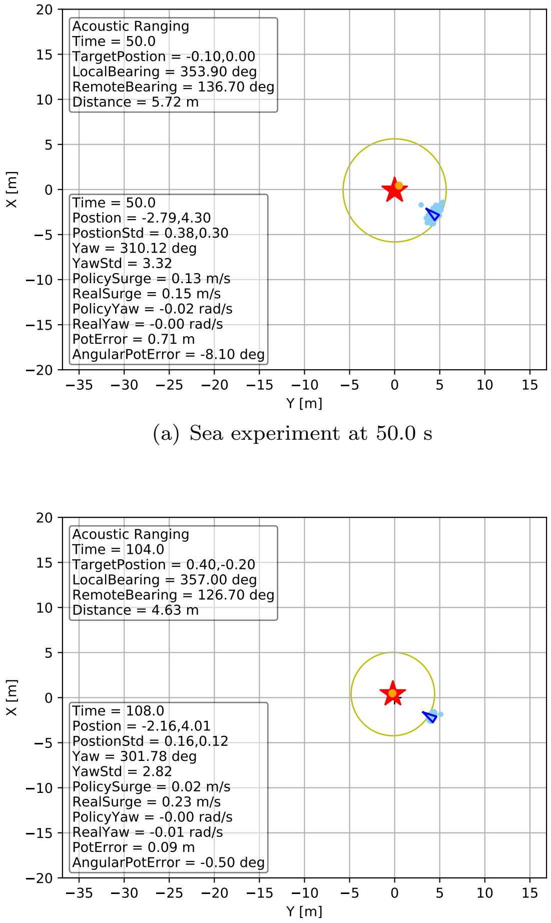
As shown in Fig. 11, Tri-TON was randomly placed at position (-4.37, 10.13) m and kept stationary, at which pointing error and angular pointing error are 15.92 m and -179.20°, respectively. At this point, the particle filter estimator, acoustic communication, pointing error estimator, and reinforcement learning controller had been initialized. When



(a) Sea experiment at 0.0 s

(b) Sea experiment at 2.0 s

Figure 11: The states of AUV in the sea experiment (a) 0.0 s, and (b) 2.0 s. The blue triangle represents Tri-TON, and the triangle's sharp corner is the vehicle's head. The red star marker is BUTTORI.



(a) Sea experiment at 50.0 s

(b) Sea experiment at 108.0 s

Figure 12: The states of AUV in the sea experiment (a) 50.0 s, and (b) 108.0 s. The blue triangle represents Tri-TON, and the triangle's sharp corner is the vehicle's head. The red star marker is BUTTORI.

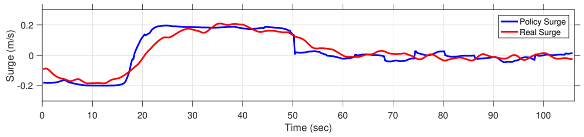


Figure 13: The surge velocity of AUV Tri-TON in the sea experiment (Dive 1). The blue curve is the surge velocity generated by the reinforcement learning-based controller, while the red curve is the real surge velocity (moving average of 8 seconds) measured by the DVL device.

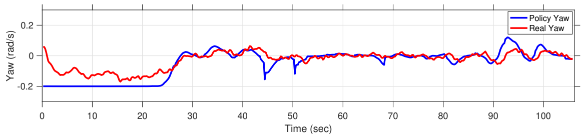


Figure 14: The yaw angular velocity of AUV Tri-TON in the sea experiment (Dive 1). The blue curve is the yaw angular velocity generated by the reinforcement learning-based controller, while the red curve is the yaw angular velocity measured by the gyroscope device.

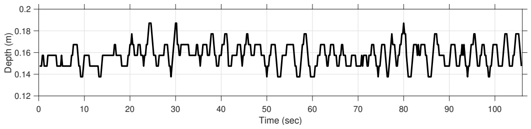


Figure 15: The depth of AUV Tri-TON in the sea experiment (Dive 1).

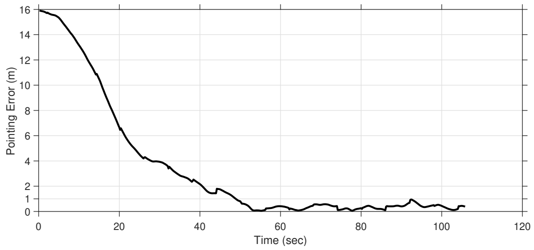


Figure 16: The pointing error $d_\Delta$ in the sea experiment (Dive 1).

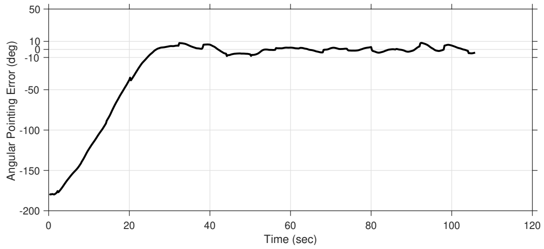


Figure 17: The angular pointing error $d_\theta$ in the sea experiment (Dive 1).

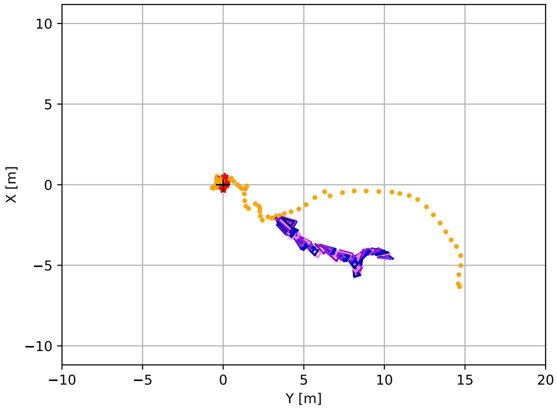


Figure 18: Trajectories of Tri-TON during link alignment control (Dive 1). The positions of platforms are plotted every 1 second. Tri-TON is cyclically represented by dark-blue, blue, blue-violet, dark-violet, and violet triangles.

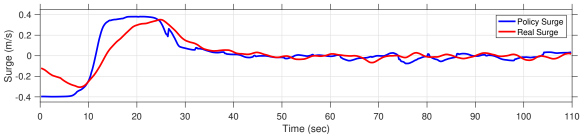


Figure 19: The surge velocity of AUV Tri-TON in the sea experiment (Dive 2). The blue curve is the surge velocity generated by the reinforcement learning-based controller, while the red curve is the real surge velocity (moving average of 8 seconds) measured by the DVL device.

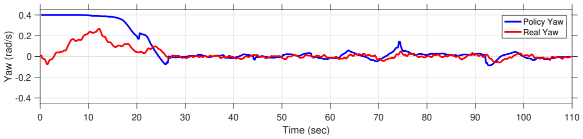


Figure 20: The yaw angular velocity of AUV Tri-TON in the sea experiment (Dive 2). The blue curve is the yaw angular velocity generated by the reinforcement learning-based controller, while the red curve is the yaw angular velocity measured by the gyroscope device.

alignment control was requested, Tri-TON started to execute the commands generated by the reinforcement learning controller. Acoustic ranging and communication between Tri-TON and BUTTORI continuously updated while tracking the target. At 2.0 seconds, the relative distance measured by acoustic ranging was 10.74 m, with a relative angle of 179.40° on Tri-TON side and 114.90° on BUTTORI side. The acoustic measurements were used in the particle filter estimator to update the position of Tri-TON. In alignment control, the velocity commands generated by the reinforcement learning controller and the actual velocity measured by AUV are shown in Fig. 13 and 14. Due to variations in Tri-TON's pitch motion, the fluctuation is introduced into the DVL measurements. We use the moving average method to filter out the noise in the measured surge velocity. At 50.0 seconds, the position of Tri-TON was (-2.79, 4.30) m, at which point the pointing error was 0.71 m, less than 1 m. After this, Tri-TON kept the optical link aligned until the end of the task at 108.0 s. The final pointing error and angular pointing error were 0.09 m and - 0.50°.

In the alignment control task, the depth of AUV Tri-TON is shown in Fig. 15. The pointing error and angular pointing error are shown in Fig. 16 and 17. The trajectory of the whole experiment is presented in Fig. 18. At the beginning, the value of the pointing error was 15.92 m, and the value of the angular pointing error was -179.20°. At 24.4 seconds, the angular pointing error was -9.24°, less than 10° for the first time, and remained in this range until the end of the task. At 48.8 seconds, the pointing error was 0.97 m, less than 1 m for the first time, and remained in that range until the end.

On another dive, we increased the speed of Tri-TON with the expectation of reducing the time spent on the tracking process. The maximum surge and yaw angular velocities were changed to 0.4 m/s and 0.4 rad/s, respectively. We similarly randomly placed AUVs and initialized the particle filter estimator, acoustic communication, pointing error estimator, and reinforcement learning controller. In the experiment, the velocity commands generated by the reinforcement learning controller and the actual velocity measured by AUV are shown in Fig. 19 and 20. The real surge velocity reached 0.4 m/s at the start of the alignment, but the yaw angular velocity failed to approach 0.4 deg/s

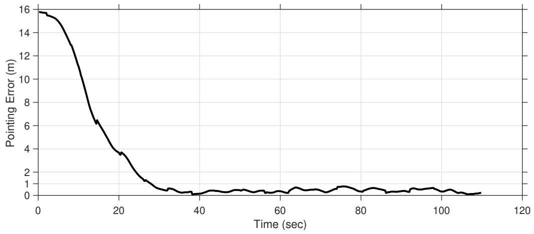


Figure 21: The pointing error $d_{\Delta}$ in the sea experiment (Dive 2).

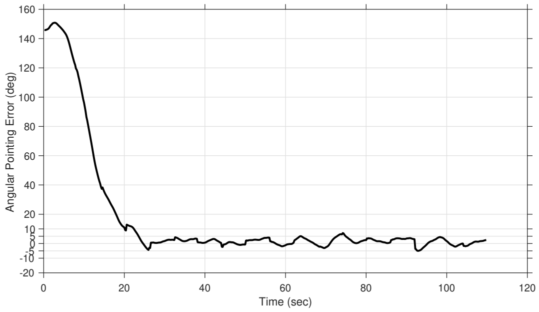


Figure 22: The angular pointing error $d_\theta$ in the sea experiment (Dive 2).

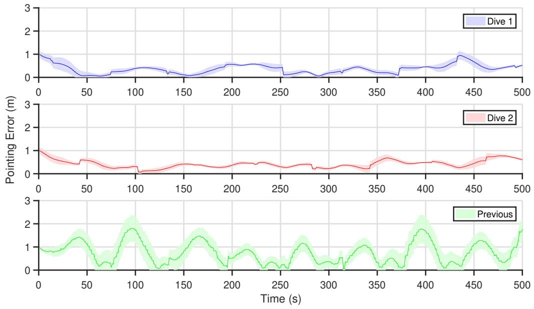


Figure 23: Pointing errors during alignment control with different methods. The shaded patches around the curve represent the standard deviation (window of 5 seconds) of the pointing error.

because of the thruster's performance.

With the greater velocity command, Tri-TON completed the link alignment control in a shorter period of time. The pointing error and angular pointing error are shown in Fig. 21 and 22. At the beginning, the value of the pointing error was 15.77 m, and the value of the angular pointing error was -145.91°. At 22.2 seconds, the angular pointing error was -9.92°, less than 10°, and remained in this range until the end of the task. At 27.8 seconds, the pointing error was 0.98 m, less than 1 m for the first time, and remained in that range until the end.

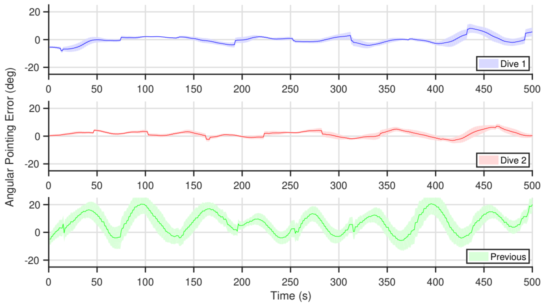


Figure 24: Angular pointing errors during alignment control with different methods. The shaded patches around the curve represent the standard deviation (window of 5 seconds) of the angular pointing error.

### 5.3 Comparison

In order to benchmark the performance of our method, we compare the experimental results with existing alignment control methods. The alignment method proposed by Weng *et al.* (Weng et al., 2022) used an underwater vehicle to reduce the pointing error and was tested in real sea experiments. The results were better than previous search methods based on optical signal intensity regarding efficiency and stability. However, in this approach, previous state as well as angular pointing errors are not taken into account in the decision-making, and instability of link alignment due to environmental disturbances cannot be avoided.

We compared the results of the current experiments with those of the existing alignment control method. Both methods were tested on the same actual machine in sea experiments. We assume that the optical link is established when the pointing error is less than 1 m or the angular pointing error is less than 10° in the alignment control. After the link is established, the pointing error and angular pointing error of the different methods are shown in Fig. 23 and 24. The blue curve represents our proposed alignment control method, and the red curve is the same method but with the maximum surge velocity and yaw angular velocity of the AUV increased to 0.4 m/s and 0.4 deg/s, respectively. The green curve is the experimental data based on the existing method. Experimental results demonstrate that our method improves the stability of alignment control. The previous method had a pointing error greater than 2 m and an angular pointing error greater than 20°, whereas our method stayed within 1 m and 10°. The standard deviations of the pointing errors for the three experiments are 0.20, 0.17, and 0.46, and the standard deviations of their angular pointing errors are 2.79, 2.07, and 6.91. Smaller pointing errors indicate that higher light-intensity signals can be collected to enhance the quality of optical communications. When the alignment controller can be stabilized to maintain smaller angular pointing errors, the design of the optical communication system can be more relaxed, and a large divergence half-angle is no longer required to keep the link aligned.

Our proposed alignment method is significant in terms of energy savings, as the previous method consumed much action in maintaining the stability of the link. In the three experiments, the sum of commands for the surge velocity was 10.41, 12.97, and 53.64, and the sum of commands for the yaw angular velocity was 10.02, 9.45, and 74.50, respectively. The previous method did not consider previous action commands and states, and thus suffered from delays and disturbances in real-world environments. Therefore, in the process of reducing the pointing error, a large number of actions were used. In our approach, we include the previous states and actions in the reinforcement learning algorithm and introduce unknown perturbations in the training environment for the alignment controller to enhance stability. This performance is confirmed in simulation test experiments and compared in real experiments.

### 5.4 Discussion

The results of the sea experiment confirm that our proposed alignment control method can be deployed on actual machines. Depth, acoustic, and motion sensors on the AUV continuously observe the environment and, after passing through the particle filter and the pointing error estimator, provide the data to the reinforcement learning-based controller to make decisions. The generated commands maneuver the platform to reduce alignment errors and maintain link alignment progressively.

The reinforcement learning policy trained in the simulated environment is validated in the simulation test and then used in the sea experiments. Hyperparameters are not adjusted when transferring the policy from the simulated environment to the real environment. In real experiments, the reinforcement learning controller is still able to perform the link alignment task. As shown in Fig. 7, 8, 16, and 17, the pointing errors are similar in the test experiment and in the sea experiment. The SAC algorithm introduces maximum entropy in training, which provides a robust framework for the trained policy and reduces the need to adjust hyperparameters when deployed in real environments. In addition, the introduction of random disturbances in the simulated environment can narrow the gap between the simulated and real environment, which can improve the performance of the reinforcement learning controller on actual machines.

During the training of the reinforcement learning policy, no model of the environment or the specifications of Tri-TON and BUTTORI is provided. However, the controller successfully manipulated Tri-TON for alignment, confirming that it can be used on all types of hovering AUVs and various unknown environments. Continuous observations by depth sensor, DVL, gyroscope, USBL device, and acoustic modem allow AUV to keep track of the relative relationship to the target. In an unknown environment, it is difficult for an optical system to observe and predict the effects of boresight and jitter caused by external disturbances. Using sensors carried by AUV allows more data to be collected during alignment control than the previous optical method. Due to the long effective propagation distance of the acoustic signal, the effective range of the proposed alignment control method increases, and no further beam scanning search is required. Previous actions and the corresponding states can reflect disturbances in the external environment, and taking these into account in a reinforcement learning policy can optimize the manipulation of AUV in alignment tasks.

From the angular pointing errors shown in Fig. 17 and 22, Tri-TON preferentially reduced the angular error before moving closer to the target to reach the desired position. From the simulated optical field in Fig. 1 and the actual experiments (Zhou et al., 2022), in underwater optical communication, as long as the relative distance between the two platforms is less than the link length $l^B_{max}$, the receiver can still detect the optical signal. However, the optical communication link can easily be interrupted if the receiving end is off-center of the beam. In the reward function, we give a greater negative reward for angular pointing errors than pointing errors. Tri-TON maintained the angular pointing error in sea experiments within $10^\circ$, whereas the previous method had an angular pointing error of more than $20^\circ$. For optical communication devices based on LED light sources (Zhu et al., 2020), a divergence half-angle of $10^\circ$ is easy to achieve. From this point of view, our proposed alignment control method is applicable to most underwater platforms and optical communication devices. When AUV can maintain a small angular pointing error, the optical communication system does not need to transmit a beam covering a large area to ensure the stability of the link, which reduces the energy loss of the underwater system. In addition, the estimated pointing error can be shared with the optical pointing system for further accurate alignment.

# 6 Conclusion

Alignment control between AUV fleets is a fundamental requirement for deploying high-speed underwater optical communications. This work solves the alignment problem between AUVs by a reinforcement learning-based alignment control method. Combined with a depth sensor, DVL, gyroscope, USBL device, and acoustic modem, the AUV minimizes the effects of boresight and jitter by estimating environmental disturbances and its relative relationship to the target through particle filters. The reward function containing the previous states optimizes the pointing error and energy consumption in alignment control during the training of the reinforcement learning policy. With the performance of the SAC algorithm, the alignment controller is trained in a simulated environment where external disturbances and noise were introduced, and is successfully deployed on a real machine in sea experiments.

In the future, the quality of the optical communication link, the absorption and scattering coefficients of the environment, and the delay of the acoustic communication will be considered in the training of the reinforcement learning policy to improve the stability of the optical link further. This technology can enhance real-time data transmission in underwater exploration.

## References

Baiden, G., Bissiri, Y., and Masoti, A. (2009). Paving the way for a future underwater omni-directional wireless optical communication systems. *Ocean Engineering*, 36(9):633–640.

Bongiorno, D. L., Bryson, M., Bridge, T. C., Dansereau, D. G., and Williams, S. B. (2018). Coregistered hyperspectral and stereo image seafloor mapping from an autonomous underwater vehicle. *Journal of Field Robotics*, 35(3):312–329.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Fossen, T. I. (2011). *Handbook of marine craft hydrodynamics and motion control*. John Wiley & Sons.

Fossen, T. I., Breivik, M., and Skjetne, R. (2003). Line-of-sight path following of underactuated marine craft. *IFAC proceedings volumes*, 36(21):211–216.

Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. (2018). Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Hardy, N. D., Rao, H. G., Conrad, S. D., Howe, T. R., Scheinbart, M. S., Kaminsky, R. D., and Hamilton, S. A. (2019). Demonstration of vehicle-to-vehicle optical pointing, acquisition, and tracking for undersea laser communications.

In *Free-Space Laser Communications XXXI*, volume 10910, page 109100Z. International Society for Optics and Photonics.

Li, Y., Zhang, Y., and Zhu, Y. (2019). Capacity of underwater wireless optical links with pointing errors. *Optics Communications*, 446:16–22.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Maki, T., Kondo, H., Ura, T., and Sakamaki, T. (2007a). Positioning method for an auv using a profiling sonar and passive acoustic landmarks for close-range observation of seafloors. In *OCEANS 2007-Europe*, pages 1–6. IEEE.

Maki, T., Mizushima, H., Kondo, H., Ura, T., Sakamaki, T., and Yanagisawa, M. (2007b). Real time path-planning of an auv based on characteristics of passive acoustic landmarks for visual mapping of shallow vent fields. In *OCEANS 2007*, pages 1–8. IEEE.

Matsuda, T. (2021). Low-cost high-performance seafloor surveying by multiple autonomous underwater vehicles. *Applied Ocean Research*, 117:102762.

Matsuda, T., Fujita, K., Hamamatsu, Y., Sakamaki, T., and Maki, T. (2022). Parent–child-based navigation method of multiple autonomous underwater vehicles for an underwater self-completed survey. *Journal of Field Robotics*, 39(2):89–106.

Matsuda, T., Maki, T., and Sakamaki, T. (2019). Accurate and efficient seafloor observations with multiple autonomous underwater vehicles: theory and experiments in a hydrothermal vent field. *IEEE Robotics and Automation Letters*, 4(3):2333–2339.

Matsuda, T., Weng, Y., Sekimori, Y., Sakamaki, T., and Maki, T. (2024). One-way-signal-based localization method of multiple autonomous underwater vehicles for distributed ocean surveys. *Journal of Robotics and Mechatronics*, 36(1):190–200.

Modasshir, M., Rahman, S., Youngquist, O., and Rekleitis, I. (2018). Coral identification and counting with an autonomous underwater vehicle. In *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 524–529. IEEE.

Oh, S.-R. and Sun, J. (2010). Path following of underactuated marine surface vessels using line-of-sight based model predictive control. *Ocean Engineering*, 37(2-3):289–295.

Quintas, J., Petroccia, R., Pascoal, A., Cruz, J., Gois, P., Morlando, L., and Stipanov, M. (2021). Hybrid acoustic-optical underwater communication networks for next-generation cooperative systems: the eumr experience. In *OCEANS 2021: San Diego–Porto*, pages 1–7. IEEE.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.

Rahman, Z., Tailor, N. V., Zafaruddin, S. M., and Chaubey, V. K. (2022). Unified performance assessment of optical wireless communication over multi-layer underwater channels. *IEEE Photonics Journal*, 14(5):1–14.

Robertson, E., Free, J., Dai, K., Wiley, J., Miller, J., and Johnson, E. (2022). 25 gbit/s underwater optical communication through turbidity using constant-envelope modulation of coherently coupled oam beams. In *OCEANS 2022, Hampton Roads*, pages 1–4. IEEE.

Sahoo, A., Dwivedy, S. K., and Robi, P. (2019). Advancements in the field of autonomous underwater vehicle. *Ocean Engineering*, 181:145–160.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sekimori, Y., Noguchi, Y., Matsuda, T., Weng, Y., and Maki, T. (2024). Bearing, elevation, and depth difference passive inverted acoustic navigation for an auv fleet. *Applied Ocean Research*, 144:103897.

Stanford Artificial Intelligence Laboratory (2020). Robotic operating system. Accessed: May 1, 2023.

Wang, L., Zhu, D., Pang, W., and Zhang, Y. (2023). A survey of underwater search for multi-target using multi-auv: Task allocation, path planning, and formation control. *Ocean Engineering*, 278:114393.

Weng, Y., Matsuda, T., Sekimori, Y., Pajarinen, J., Peters, J., and Maki, T. (2022). Establishment of line-of-sight optical links between autonomous underwater vehicles: Field experiment and performance validation. *Applied Ocean Research*, 129:103385.

Yang, F., Zhang, X., and Chen, S. (2020). Multi-laser based cooperative scanning for auv acquisition over uwoc networks. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE.

Yang, Y., Xiao, Y., and Li, T. (2021). A survey of autonomous underwater vehicle formation: Performance, formation control, and communication capability. *IEEE Communications Surveys & Tutorials*, 23(2):815–841.

Zhou, H., Zhang, M., Wang, X., and Ren, X. (2022). Design and implementation of more than 50m real-time underwater wireless optical communication system. *J. Lightwave Technol.*, 40(12):3654–3668.

Zhou, Y., Zhu, X., Hu, F., Shi, J., Wang, F., Zou, P., Liu, J., Jiang, F., and Chi, N. (2019). Common-anode led on a si substrate for beyond 15&#x2009;&#x2009;gbit/s underwater visible light communication. *Photon. Res.*, 7(9):1019–1029.

Zhu, S., Chen, X., Liu, X., Zhang, G., and Tian, P. (2020). Recent progress in and perspectives of underwater wireless optical communication. *Progress in Quantum Electronics*, 73:100274.