Reinforcement Learning Based Underwater Wireless Optical Communication Alignment for Autonomous Underwater Vehicles

Yang Weng, *Student Member, IEEE,* Joni Pajarinen, Riad Akrour, Takumi Matsuda, *Member, IEEE,* Jan Peters, *Fellow, IEEE,* and Toshihiro Maki, *Member, IEEE*

Abstract

With the developments in underwater wireless optical communication (UWOC) technology, UWOC can be used in conjunction with autonomous underwater vehicles (AUVs) for high-speed data sharing among the vehicle formation during underwater exploration. A beam alignment problem arises during communication due to the transmission range, external disturbances and noise, and uncertainties in the AUV dynamic model. We propose an acoustic navigation method to guide the alignment process without requiring beam directors, light intensity sensors, and/or scanning algorithms as used in previous research. The AUVs need stably maintain a specific relative position and orientation for establishing an optical link. We model the alignment problem as a partially observable Markov decision process (POMDP) that takes manipulation, navigation, and energy consumption of underwater vehicles into account. However, finding an efficient policy for the POMDP under high partial observability and environmental variability is challenging. Therefore, for successful policy optimization, we utilize the soft actor-critic (SAC) reinforcement learning algorithm together with AUV-specific belief updates and reward shaping based curriculum learning. Our approach outperformed baseline approaches in a simulation environment and successfully performed the beam alignment process from one AUV to another on the real AUV Tri-TON 2.

Yang Weng, Takumi Matsuda and Toshihiro Maki are with the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan (e-mail: yangweng@iis.u-tokyo.ac.jp; matsuda@iis.u-tokyo.ac.jp; maki@iis.u-tokyo.ac.jp). Takumi Matsuda is also with the School of Science and Technology, Meiji University, Japan.

Joni Pajarinen, Riad Akrour and Jan Peters are with Intelligent Autonomous Systems Laboratory, TU Darmstadt, Darmstadt, Germany (e-mail: joni.pajarinen@aalto.fi; riad.akrour@inria.fr; jan.peters@tu-darmstadt.de). Joni Pajarinen is also with the Department of Electrical Engineering and Automation, Aalto University, Finland.

Index Terms

Reinforcement learning, SAC, POMDP, AUV, Underwater wireless optical communication

I. INTRODUCTION

The use of AUVs has a prominent role in oceanic environment observation and investigation. That the AUVs operate in a team formation is desirable for challenging missions that require simultaneous coordinated surveying from multiple platforms [1]. The communication between formation members can improve the efficiency of joint investigations. One AUV can share the collected data with the other AUV immediately after finishing a measurement phase so that a single AUV can analyze all data and correspondingly adjust the following tasks without the onshore station's intervention. Suffering from low data rates and high latency [2], acoustic communication is not suitable for the transmission of complex observation data, such as seafloor images. The UWOC technology is gradually maturing, providing a high data rate and transmission bandwidth solution for information sharing between the AUV team [3] [4] . In 2019, the 1.2-m LED system by Shi *et al.* [5] achieved a data rate of 14.6 Gigabits per second (Gbps). The implementation of the UWOC across AUV teams is of great significance for underwater exploration.

However, the UWOC channel has limitations in transmission range and coverage area [6], which requires the establishment and maintenance of a line-of-sight (LOS) link for communication. In the vast ocean environment, it is hard to complete initial location identification for the link establishment. In involving AUVs scenarios, maintenance becomes challenging due to the external disturbances and uncertainties in the AUV dynamic model.

Some free-space-optical (FSO) communication systems have developed the modulating-retro-reflector (MRR) technique to solve the alignment problem [7]. This solution needs to be deployed on a stationary ground station for tracking, which is not feasible for the underwater mobile platform. Abadi *et al.* [8] proposed a low-complexity self-aligning FSO system that shares the global navigation satellite system (GNSS) coordinates, latitude, and longitude in radio link for localization. However, neither GNSS nor radio links are available in seawater environments. In underwater scenarios, Hardy *et al.* [9] referred to the FSO system and fabricated a real-time control system (RTCS) that can quickly and accurately steer the optical beam relative to the vehicle's frame of reference. The RTCS directs the beam director to move through a scan pattern in the acquisition process and engages its tracking loops to keep the received light

as close as possible to the center of the detector. Solanki *et al.* [10] also designed a beam pointing control system and proposed an effective maintenance algorithm based on received light intensity. The above methods attempt to eliminate the movement of the AUVs and the environmental disturbances with precise control of the optical devices. Cochenour *et al.* [11] utilizes the strong light scattering in seawater to establish a non-line-of-sight (NLOS) link to avoid the alignment problem. This optical channel is susceptible to noise and requires a more sophisticated receiver design, such as using the photomultiplier (PMT). The previous research did not discuss how to start the initial location identification in the sea area.

We propose to design a more practical way to solve the alignment problem between two AUVs. There is no additional servo in AUV to control the beam pointing and conduct scanning tasks for link acquisition, which is time-consuming. Two AUVs need stably maintain a specific relative position and orientation in the alignment process. Regardless of the use of an LED or laser based light source, the optical beam can be expanded to cover an area. As long as the position error between two AUVs is still within the coverage area of the beam, the LOS link can be successfully maintained. We choose acoustic positioning for the initial location identification due to its long effective transmission range [12] [13]. With acoustic positioning, AUVs can continuously observe the relative direction and position in the alignment process. We do not mount the camera for visual positioning that requires heavy image computations [14] and avoid the possible interference between a communication beam and a positioning light. Furthermore, we hope to reduce energy consumption during alignment, which is essential for underwater operation.

All the position errors, which are generated from the external disturbances, the noise in the environment, and the uncertainties in the dynamic model of the AUV, have a significant impact on the alignment process. The function approximation methods can be used to approximate the unknown robotic dynamics for more precise control [15] [16] [17]. Considering that the alignment task also requires combining multiple types of sensor data and optimizing multiple objectives, it is attractive to consider this complex alignment problem under a model-free reinforcement learning framework. The reinforcement learning algorithm is successfully utilized to restrain the impact of external disturbances and uncertainties in robotics motion planning [18] [19]. The position and orientation data of AUVs, which guide the alignment process, are partially observable in our navigation method. We consider the underwater optical communication alignment problem under a POMDP [20] model. To find an efficient policy for the POMDP under high partial observability

and environmental variability, we propose a reinforcement learning approach based on the SAC algorithm [21] together with particle filtering based state estimation to learn and optimize a policy from sample data. Due to task challenges, we use reward shaping based curriculum learning to shape the reward signal in simulation such that the task becomes progressively harder. By shaping the reward with respect to relative position, navigation, and energy consumption, optimization starts from an easy task until the end goal is reached – that of being able to solve the actual alignment task. Our approach outperformed baseline approaches in simulation experiments and successfully aligned the optical link between AUVs for wireless data transmission. The experimental facilities were prepared for real applications, and the AUV successfully aligned with the target in the water tank.

To summarize, the main contributions of this paper are an approach for the underwater alignment problem in AUVs that 1) does not require additional hardware such as fixed ground stations or beam directors; 2) can perform initial location identification regardless of the distance using acoustic communication; 3) incorporates a model-free algorithm which does not require knowledge of the ocean environment or the dynamic model of any vehicles; 4) uses a reinforcement learning approach to optimize the alignment of AUVs which realizes reducing battery and acoustic channel usages while maximizing alignment efficiency.

The remainder of this paper is organized as follows. Section II presents the acoustic navigation based optical beam alignment scheme. The alignment problem is modeled as POMDP in Section III. The reinforcement learning framework and SAC algorithm are utilized to solve the beam alignment problem in Section IV. Detailed learning studies, comparison with baseline approach, and transfer learning are discussed in Section V. The conclusions are given in Section VI.

II. Problem Formulation

This section presents how the AUVs complete the underwater optical beam alignment task with acoustic navigation. All states used in the alignment task and actions that the AUV needs to control are defined.

A. Autonomous Underwater Vehicles

The alignment task requires one AUV mounted with a directional transmitter to emit an optical beam to another AUV's receiver. An omnidirectional detector is usually designed as a receiver. In our method, the AUV that transmits optical signals is regarded as the transmitting AUV,

while the AUV that receives optical signals is called the receiving AUV. The receiving AUV can receive optical signals when the optical beam from the transmitting AUV at least partially covers its detector.

Note that the agent discussed in the reinforcement learning framework is the transmitting AUV, and all generated actions are completed by this AUV. As a mobile platform, the transmitting AUV needs to maintain a relative position and orientation to the receiving AUV. Conventionally, the motion of AUVs is described in six degrees of freedom (DOFs), including surge, sway, heave, yaw, roll, and pitch [22]. Depending on the configuration of the thrusters, AUVs are capable of propelling themselves in several of these DOFs.

The hovering AUV Tri-TON 2 is used as the transmitting AUV in the experiments as will be described henceforth. The performance and results from its previous sea experiments show that it can be deployed in this research [23] [24]. The mounted thrusters can stably control the surge, sway, heave, and yaw motions. One highly accurate pressure sensor allows the vehicle to cruise at the expected depth.

The underwater optical communication can be set to occur at a specific depth because the mounted pressure sensor can provide an accurate determination of absolute depth [9]. The alignment is considered on a horizontal plane with a horizontal position $[x, y]$, surge velocity u, sway velocity v, yaw orientation ψ, and yaw angular velocity r. The superscript R is used to indicate the variables belonging to the receiving AUV, and the superscript T shows the variables of the transmitting AUV. We ignore the motion in roll and pitch orientation because those are statically stable for the hovering AUV. In this task, the transmitting AUV needs to observe the position $[x^R, y^R]$, orientation ψ^R, and velocity information $[u^R, v^R, r^R]$ of the receiving AUV, and to control the surge velocity u^T and yaw angular velocity r^T in the movement.

B. Underwater Wireless Optical Communication

Unlike acoustic signals, the underwater optical beam has limited propagation distance and strong directivity. Generally, the effective range of the UWOC link is about 1 - 100 meters, which is affected by various factors such as absorption, scattering, turbulence, light source, and hardware configuration [6] [25] [26]. The limitations in optical signal transmission require maintaining the LOS link for underwater communication.

We calculate the optical beam's light field distribution on the horizontal plane to analyze the transmitting and receiving AUVs' optimal positional relationship for beam alignment. Radiative

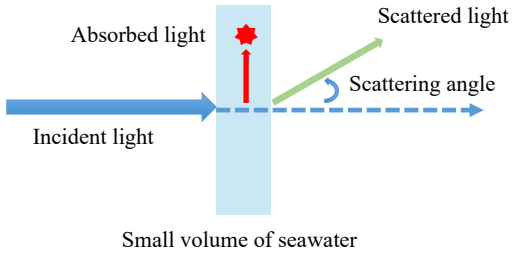


Fig. 1. Illustration of the penetration of photons into sea water. Part of the incident photons is absorbed by seawater, and the remaining part is scattered through different scattering angles. The scattering angle distribution can be given by the scattering phase function. This process happens continuously in the propagation.

transfer theory is the most accurate model for describing energy transfer in the process of optical propagation [27]. The transmission within seawater changes the optical beam both in power and direction. Some of the photons are lost due to absorption, and some are scattered out of the beam, which is illustrated in Fig. 1. The radiative transfer equation (RTE) describes these interactions mathematically [27]:

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

where $I_\lambda(s)$ is an incident optical intensity for the wavelength λ. $J_\lambda(s)$ is the source function that accounts for the contributions of emission and scattering processes. The attenuation coefficient c_λ defines what proportion of energy will change, which is the sum of the absorption coefficient a_λ and scattering coefficient b_λ. Typical coefficient values of four major water types usually considered in the literature are given in [6].

A Monte Carlo method is used to solve the RTE in realistic oceanic conditions [28]. Initially, each photon is launched at the origin position with random azimuth and polar angles. The Monte Carlo method needs to calculate all photons' trajectories and determine whether absorption and scattering are occurring during propagation, which is indicated by signal-scattering albedo b_λ/c_λ. The tracking of a photon will be stopped when absorption occurs. If a scattering process arises, the propagation direction is updated by new azimuth and polar scattering angles with respect to the incident direction. The new azimuth angle is generated by a random value drawn from $U[0, 2\pi]$, while the new polar angle depends on a scattering phase function called Henyey and

Greenstein function [6]. To obtain the horizontal distribution of the light field, we use the sensor deployed at a target position to determine if each photon's path crosses the sensor's area in an appropriate direction. Then, the received optical power at the position is derived from the cumulative number of photons.

The beam propagation from the transmitting AUV is simulated in clear ocean water with the absorption coefficient of 0.069 and the scattering coefficient of 0.08 [6]. As the light field distribution is shown in Fig. 2, a total of 10^8 photons are emitted from the origin and tracked. The number of photons received by the sensors placed at each position is compared with 10^8 photons at the origin and expressed in decibels (dB). The detection spacing of each sensor is 0.05 meters. The divergence angle of the beam is set to 30 deg. This angle is easy for an LED-based transmitter to achieve, but the spread is required for laser light sources. The asymmetry used parameter in Henyey and Greenstein function is 0.924, which shows a good approximation for most practical situations [29].

The expected positional relationship between the transmitting AUV and the receiving AUV in this research is derived from the above light distribution. It is a trade-off between link quality and communication range [30]. On the one hand, the receiving AUV is suggested to gather high-intensity signals, which can induce high Signal-to-Noise Ratio (SNR), to increase the data rate and limit the Bit Error Rate (BER) in communication. On the other hand, we hope the coverage radius of the beam is larger than the detector so that the communication link can still be maintained when a shift occurs in an AUVs' position. As the light field distribution indicates, the transmitting AUV is emitting an optical beam from the coordinates' origin. In this research, the receiving AUV is proposed to receive a high-intensity optical signal at the vicinity of point D, which is at the position of (5, 0) on the coordinates. As shown in Fig. 2, within 1 meter around point D, high-intensity signals are detected.

In this task, two AUVs are required to maintain this positional relationship for alignment: when the transmitting AUV emits optical signals, the receiving AUV must be within 1 meter of point D. The alignment distance d_Δ is used to represent the distance between the receiving AUV and point D:

$$d_\Delta = (x_\Delta^2 + y_\Delta^2)^{\frac{1}{2}} \tag{2}$$

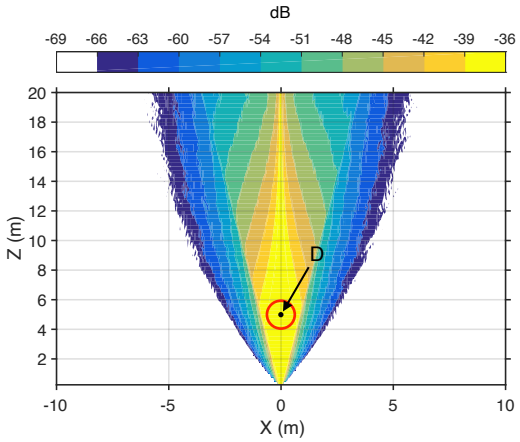


Fig. 2. The light field distribution of the emitted optical beam on a horizontal plane. The absorption and scattering coefficients of the seawater environment are 0.069 and 0.08, respectively. The number of photons received by the sensors placed at each position is compared with 10^8 photons at the origin and expressed in decibels (dB). The transmitter of the transmitting AUV is placed at the origin of coordinates, and the receiving AUV is proposed to receive the optical signal within 1 meter from point D (the area depicted as a red circle).

where

$$x_\Delta = x^R - (x^T + 5\cos\psi^T) \tag{3}$$

and

$$y_\Delta = y^R - (y^T + 5\sin\psi^T) \tag{4}$$

The optical beam is considered as successfully aligned when d_Δ is within 1 meter. Data transmission may fail even in the case of alignment considering the BER in the real situation. The probability of failure is set to 1% here [30].

It is not recommended to turn on the optical transmitter for a long time, although the power of the optical device is relatively low [31]. For saving energy, the transmitter only turns on when it is possible to establish a LOS link, which is considered an action, represented by a boolean variable i_{op}, in this alignment task. The transmitting AUV needs to learn when to turn on the optical device.

C. Acoustic Navigation

It is necessary to consider the issue that, during alignment, there may be an absence of GNSS signals underwater. The transmitting AUV needs to know the receiving AUV's position,

orientation, and velocity for tracking. Bounding the position error can improve the probability of alignment. This subsection introduces how the AUVs sense the environment.

The acoustic wave is the only media available for underwater long distance propagation, such as in the range of tens of kilometers [2]. With the aid of acoustic communication, completing initial location identification between two AUVs at any distance becomes possible. The acoustic time of flight (TOF) algorithm is implemented to estimate inter-vehicle range and direction. The position and velocity information of the receiving AUV $[x^R, y^R, \psi^R, u^R, v^R, r^R]$ can also be shared with the transmitting AUV through acoustic communication.

The one-way travel-time (OWTT) ranging offers the advantage that a single OWTT broadcast signal can serve many clients simultaneously [32]. We assumed that the clocks are not synchronized and only direction information ψ^T can be estimated. The receiving AUV can periodically broadcast the OWTT ranging signals in the alignment process. The interval of acoustic ranging is set to 5 seconds according to a prior experiment [33].

Another method is two-way travel-time (TWTT) ranging. The receiving AUV may receive the interrogating acoustic signal from the transmitting AUV and then reply [23]. The transmitting AUV can calculate the inter-vehicle range $[x^R - x^T, y^R - y^T]$ from the round-trip time and the direction ψ^T. However, the scalability of TWTT ranging is weak as the update rate of the position must be divided by the number of vehicles. We hope to avoid frequent use of the TWTT ranging to reduce the occupancy of the acoustic channel, although the estimation result is more accurate than unsynchronized OWTT. Compared with OWTT ranging, TWTT ranging takes a longer time to complete localization, which is set to 10 seconds in an actual implementation [23].

The transmitting AUVs can decide to take an action i_{twtt} to periodically listen to the OWTT ranging signals, or to request a higher performance method, TWTT ranging, to bound positional error during their alignment process.

The states in the alignment process with the aid of acoustic navigation are partially observable. Limited by the efficiency of acoustic transmission, the transmitting AUV cannot frequently obtain observation results. It is more convenient to choose the OWTT ranging, but the distance information is not available. The observation and estimation work becomes essential in the alignment task. The state estimator will be introduced later in this partially observable problem.

In summary, we have defined a total of four actions $[u^T, r^T, i_{op}, i_{twtt}]$ that the transmitting AUV needs to learn for alignment. In the alignment task, the transmitting AUV needs to observe

the relative distance $[x^{\Delta}, y^{\Delta}]$, the orientation ψ^R and velocity information $[u^R, v^R, r^R]$ of the receiving AUV, and then learn how to make decisions on four actions $[u^T, r^T, i_{op}, i_{twtt}]$. The goal is to shorten the alignment distance d_{Δ} for maintaining the LOS link while also reducing the use of TWTT acoustic ranging and optical transmitter.

III. POMDP MODELING

To utilize the reinforcement learning algorithm, we consider the beam alignment problem under a POMDP model and define the state space, the action space, and the reward function. The particle filter algorithm is presented as a state estimator to update the state belief in POMDP.

A. Partially Observable Markov Decision Process

The POMDP provides a model to describe how an agent, for example an AUV, interacts with a partially observable environment [34]. It can be described as a tuple $\langle S, A, T, R, \Omega, O \rangle$, where

- S is a set of states;
- A is a set of actions;
- $T : S \times A \to \Pi(S)$ is the state-transition function, such that $T(s', a, s)$ is defined as the probability of ending in state s', given that the agent starts in state s and takes action a;
- $R : S \times A \to \mathbb{R}$ is the reward function, where $r(s, a)$ is the reward for taking action a in state s;
- Ω is a set of observations;
- $O : S \times A \to \Pi(\Omega)$ is the observation function, such that $O(s', a, o)$ is the probability of making observation o given that the agent took action a and landed in state s'.

At each timestep the agent executes an action, receives a reward, and transitions to the next state according to the state-transition function. In a POMDP, the agent tries to maximize the expected reward over a time horizon. Since the agent cannot observe the hidden state directly, the agent needs to decide on actions based on the history of previous actions and observations, a sufficient statistic for optimal decision making [34]. In place of the history, the agent can utilize the belief $b(s)$, a probability distribution over states $s \in S$. The belief state is updated by a state estimator that employs Bayes' rule [35]. Classical POMDP methods typically assume discrete actions and observations [36] [37]. Since our POMDP model has continuous actions and observations, we utilize a model-free reinforcement learning method with neural network based policies that allows for continuous values.

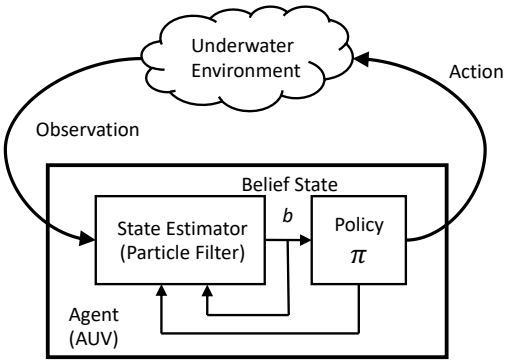


Fig. 3. The POMDP model of the alignment problem.

B. POMDP for Alignment Process

The POMDP model is used here to consider all state, action, and observation variables described in Section II. The state space, action space, and reward function are clarified. The state-transition function is unknown in the alignment process. There are some uncertainties in the dynamics model, which may relate to the specific structure of vehicles. The external disturbances are affected by the time-varying oceanic environment. The model-free method is proposed to solve this problem. The observation and its function are presented in the particle filter estimator.

The transmitting AUV needs to maintain the tracking based on the receiving AUV's position, orientation, and velocity, acquired from the acoustic signal. All the variables are considered in state space. The angular variables are replaced by their sine and cosine values to eliminate the ambiguity from angular periodicity [38]. The position information is represented by the alignment distance $[x_\Delta, y_\Delta]$ by which we denote how far point D is from the receiving AUV in Section II and calculate in Equation 2. The state space of the agent is as follows:

$$s = [\hat{x}_\Delta, \hat{y}_\Delta, \cos\psi^R, \sin\psi^R, u^R, v^R, r^R, \cos\hat{\psi}^T, \sin\hat{\psi}^T] \tag{5}$$

where the variables with hat symbols are updated using a particle filtering based estimator. In this research, all variables in the state space are one-dimensional and continuous. The unit of position variable is meter, and the unit of angle is degree. The unit of linear velocity and angular velocity is meter per second and degree per second, respectively. The range of the receiving AUV's linear velocity and angular velocity are $[-0.1, 0.1]$ and $[-5, 5]$.

The transmitting AUV can adjust the surge and yaw angular velocity to maintain the relative position and orientation. The sway velocity is not controllable based on the most straightforward propellers configuration. Furthermore, it can request the TWTT ranging to eliminate the estimation error. Another action is the transmitting AUV can power on the optical beam to establish a LOS link. All these four control inputs are defined as actions in this task:

$$a = [u^T, r^T, i_{twtt}, i_{op}] \tag{6}$$

where u^T and r^T are desired surge and yaw angular velocities. All variables in the action space are one-dimensional. The maximum value of u^T and r^T are set to 1 m/s and 5 deg/s. The u^T and r^T are continuous variables, and the ranges are $[-1, 1]$ and $[-5, 5]$, respectively. Both i_{twtt} and i_{op} are boolean variables that represent whether the transmitting AUV requests for TWTT ranging, and whether to turn on the optical transmitter in the current timestep, respectively. In training, actions i_{twtt} and i_{op} are considered as the continuous variables from $[-1, 1]$, and the transmitting AUV will take action if the variable is positive.

We consider four parts in the design of the reward function. The first term is for minimizing the alignment distance d_Δ, which is the most basic requirement for establishing a LOS link. The second term is to reduce the relative movement between two AUVs so that the link can remain stable for a longer period. It becomes significant when the alignment distance is close to 0. The third term is used to restrict the usage of TWTT ranging and the optical transmitter, especially when the two AUVs are far apart. The fourth item is related to the final reward that the agent can receive when the LOS link is successfully established. It promotes the agent to achieve the goal faster. We propose the reward function of the form:

$$\begin{aligned} r(s,a) = &- \rho_1(1+\rho_2 i_{twtt})(1+\rho_3 i_{op}) d_\Delta^{\frac{1}{2}} \\ &- \rho_4 u_\Delta - \rho_5 r_\Delta + \rho_6 i_{done} \end{aligned} \tag{7}$$

where the coefficients ρ_1 to ρ_6 are decided by the importance of the parameters for completing alignment, which will be discussed in reward shaping work later. Note that all coefficients are positive here. The alignment distance d_Δ represents the distance between the receiving AUV and point D of the optical beam. We compared the square root, exponent, and square functions to process value d_Δ in the reward function. When d_Δ is close to 0, the gradient of the square root function is the largest. The u_Δ and r_Δ represent the relative velocities in surge and yaw. A boolean variable i_{done} is used to indicate if the LOS link is successfully established. Excluding the final reward, the transmitting AUV can receive immediate rewards from the first three terms

of reward function at every timestep. The environment emits a reward ranging from $(-\infty, 0]$ on each transition. The AUV can receive the largest immediate reward when the alignment distance d_Δ is 0 and the relative velocity u_Δ and r_Δ are 0. The range of cumulative rewards that the AUV can receive in an episode is $(-\infty, \rho_6)$.

C. Particle Filter Estimator

The relative distance and orientation, which guide the alignment process, are partially observable. They are measured by acoustic ranging and updated by the state estimator in the alignment task, which is shown in Fig. 3. Several methods can be used to update the belief in POMDP: assuming a Gaussian distribution and using a Kalman filtering, using discrete values and vector for belief probabilities, and particle filtering. In previous studies, when the states are estimated by a linearized filter, such as the extended Kalman filter (EKF), information about the observability will be lost if repeated range updates come from the same relative bearing [39] [40] [41]. The previous experiments show that the computation performance of the on-board computer in AUV Tri-TON 2 is sufficient to support a real-time particle filter algorithm [23]. Therefore, we use particle filtering [42] for state estimation instead of Kalman filtering.

In this particle filter algorithm, the probability density of the states is expressed by a set of particles [43]. There are 1000 particles used to store the guesses of the position $[x^T, y^T]$ and orientation $[\psi^T]$ of the transmitting AUV. The estimator is established by randomly scattering these particles around the initial position of the transmitting AUV.

In the prediction phase, the particle filter updates the position and orientation from time t to $t + \Delta t$ by on-board sensors' data:

$$x_{t+\Delta t}^{T,i} = x_t^{T,i} + (u_t^{T,i} \cos \psi_t^{T,i} - v_t^{T,i} \sin \psi_t^{T,i})\Delta t \tag{8}$$

$$y_{t+\Delta t}^{T,i} = y_t^{T,i} + (u_t^{T,i} \sin \psi_t^{T,i} + v_t^{T,i} \cos \psi_t^{T,i})\Delta t \tag{9}$$

$$\psi_{t+\Delta t}^{T,i} = \psi_t^{T,i} + r_t^{T,i}\Delta t \tag{10}$$

$$u_t^{T,i} \sim \mathcal{N}(\widetilde{u}_t^T, (\sigma_{u,t}^T)^2) \tag{11}$$

$$v_t^{T,i} \sim \mathcal{N}(\widetilde{v}_t^T, (\sigma_{v,t}^T)^2) \tag{12}$$

$$r_t^{T,i} \sim \mathcal{N}(\widetilde{r}_t^T, (\sigma_{r,t}^T)^2) \tag{13}$$

where i indicates the *i*th particles in the estimator. $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian sampling with mean μ and standard deviation σ. $\widetilde{u}_t^T$, $\widetilde{v}_t^T$, and $\widetilde{r}_t^T$ are the measurement data of surge, sway, and yaw angular velocities from the transmitting AUV on-board sensors. $\sigma_{u,t}^T$, $\sigma_{v,t}^T$, and $\sigma_{r,t}^T$ are the standard deviations of measurements, which are set to 0.1, 0.1, and 1 in this research.

Once the acoustic observation results are available, the estimator calculates the difference between observed values and prediction values for weighting all particles:

$$W_{owtt}^i = \max\left\{\exp\left(\frac{k_\psi^2}{2} + \frac{-(\Delta\psi^i)^2}{2(\sigma_\psi)^2}\right), 1\right\} \tag{14}$$

$$W_{twtt}^i = \max\left\{\exp\left(\frac{k_d^2}{2} + \frac{-(\Delta d^i)^2}{2(\sigma_d)^2}\right)\exp\left(\frac{k_\psi^2}{2} + \frac{-(\Delta\psi^i)^2}{2(\sigma_\psi)^2}\right), 1\right\} \tag{15}$$

where W_{owtt}^i is the weight of *i*th particle after receiving OWTT ranging and W_{twtt}^i is the weight based on TWTT ranging results. Δd and $\Delta\psi$ are the difference in inter-range and direction between observation and prediction, respectively. σ_d and σ_ψ are standard deviations of inter-range and direction in the weighting process. σ_ψ uses 20 when AUV updates OWTT ranging results, while σ_d and σ_ψ are set to 0.5 and 4 in TWTT ranging. k_d and k_ψ are parameters for judging outliers [44]. Both of k_d and k_ψ are set to 2 in the weighting phase.

Then, the transmitting AUV can resample its particles based on weights to determine the current state [23]. The particles in the estimator represent the belief distribution of position $[x^T, y^T]$ and orientation $[\psi^T]$. The mean value of the particles is considered as state s to input to the policy π:

$$\hat{x}_\Delta = x^R - (\overline{x}^{T,i} + 5\cos\overline{\psi}^{T,i}) \tag{16}$$

$$\hat{y}_\Delta = y^R - (\overline{y}^{T,i} + 5\sin\overline{\psi}^{T,i}) \tag{17}$$

$$\hat{\psi}^T = \overline{\psi}^{T,i} \tag{18}$$

where the variables with overline symbols are averaged from all particles.

IV. Reinforcement Learning Solution

This section introduces the SAC algorithm to train the policy for the alignment task. The curriculum learning and reward shaping techniques are used in policy learning.

A. Soft Actor-Critic Algorithm

The SAC algorithm presented by Haarnoja *et al.* [21] is used to search for an optimal policy π^* that can collect the maximum cumulative reward and entropy. The entropy is a measure of randomness in the policy, which encourages the policy to explore more widely and capture multiple modes of near-optimal behavior. Increasing entropy can also prevent the policy from prematurely converging to a bad local optimum. The SAC algorithm supports continuous spaces, which is essential for the alignment task. Another reason for choosing the SAC algorithm is its advantages in applying the reinforcement learning method in real AUVs, discussed later. The reinforcement learning objective that the SAC algorithm wants to maximize is:

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^{\infty} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[\sum_{l=t}^{\infty} \gamma^{l-t} \mathbb{E}_{s_l\sim p, a_l\sim\pi}[r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))|s_t,a_t]] \tag{19}$$

where $\mathbb{E}$ is the expectation operation. Note that the state s is the result of the particle filter in this research. α is defined as a temperature parameter that determines the relative importance of the entropy term versus the reward. γ is the discount factor used to determine the importance of future rewards, and $\mathcal{H}$ is the entropy term.

An action-value function $Q_(s_t, a_t)$ (also called Q-value function) with lower variance is introduced to replace the cumulative rewards to evaluate the performance of the policy:

$$Q(s_t,a_t) = r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1},a_{t+1}}[Q(s_{t+1},a_{t+1})] \tag{20}$$

The policy and value function are considered as the actor and critic in the reinforcement learning framework. Using the sampled trajectory data $(s_t, a_t, r(s_t, a_t), s_{t+1})$, the Q-value function can be approximated and the policy will be iterated by evaluation and improvement processes until the optimal policy is reached. Since the state and action spaces of the alignment problem are continuous, we propose to use a neural network to approximate the Q-values and a neural network to generate actions.

As listed in Algorithm 1, the SAC algorithm updates both the soft Q-function and policy with a stochastic gradient descent:

$$\begin{aligned}\hat{\nabla}_\theta J_Q(\theta) =& \nabla_\theta Q_\theta(a_t, s_t)(Q_\theta(s_t, a_t) - (r(s_t, a_t) \\ &+ \gamma(Q_{\overline{\theta}}(s_{t+1}, a_{t+1}) - \alpha \log(\pi_\phi(s_{t+1}|a_{t+1}))))\end{aligned} \tag{21}$$

and

$$\begin{aligned}\hat{\nabla}_\phi J_\pi(\phi) =& \nabla_\phi \alpha \log(\pi_\phi(s_t|a_t)) + \nabla_{a_t} \alpha \log(\pi_\phi(a_t|s_t)) \\ &- \nabla_{a_t} Q(s_t, a_t))\nabla_\phi f_\phi(\epsilon_t; s_t)\end{aligned} \tag{22}$$

where θ, $\overline{\theta}$ and ϕ are the network parameters of soft Q-function, target soft Q-function and policy, respectively. The algorithm also uses a neural network transformation $f_\phi(\epsilon_t; \boldsymbol{s}_t)$ to reparameterize the policy.

The SAC algorithm is implemented with the OpenAI Stable Baselines toolkit [45]. The Multilayer Perceptron (MLP) structure with 2 hidden layers of 64 neurons is used in training, which was successfully used for many similar tasks [21]. The discount factor γ uses 0.99 in the SAC algorithm. All networks, including policy and value function, set the learning rate λ to 0.0003. The buffer size and batch size are set to 50000 and 64, respectively.

B. Simulator Configuration

As the training requires a lot of sample data, it is unrealistic to collect real experiment data directly. We define the episode based experiments that can be efficiently repeated by the simulator. In practice, underwater optical communication may be used many times in underwater investigation, which is similar to the episode based experiments. When there is data that needs to be shared, the vehicle will request to establish a link and maintain alignment. The configuration here is consistent with those described in the previous sections. The trajectory data sampled from simulation experiments are used to train the actor and critic networks.

The maximum timestep of each episode is set to 1000, and one timestep is considered one second in the real world. All the parameters are initialized at the beginning. The alignment process is carried out on a horizontal plane. The transmitting AUV starts from the origin of coordinates while the receiving AUV is randomly placed at a point where it is 20 meters away from the origin. The initial yaw angles of the two AUVs are randomly generated. The transmitting AUV moves with the velocity commands in each timestep, while the receiving AUV moves randomly. The simulator constrains the surge velocities of the transmitting and receiving AUVs

Algorithm 1 Soft Actor-Critic [21]

Input:

Initial parameters of critic and actor networks θ_1, θ_2, ϕ

Initial weights of target networks $\overline{\theta}_1 \leftarrow \theta_1, \overline{\theta}_2 \leftarrow \theta_2$

Empty replay buffer $\mathcal{D} \leftarrow \emptyset$

for each iteration **do**

for each environment step **do**

Sample action by $a_t \sim \pi_\phi(a_t|s_t)$

Sample transition state by $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$

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 critic by $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$

Update policy 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 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

in the range of 1 and 0.1 m/s, respectively. The sway velocities induced by thrusters are both kept at 0 m/s, and their yaw angular velocities are limited to 0 to 5 deg/s. The trajectories of an episode are depicted in Fig. 4. Two AUVs are sailing from the starting point according to the above configuration. The position and velocity information of the receiving AUV can be listened to by the transmitting AUV every timestep. The OWTT ranging signals are broadcasted from the receiving AUV with the interval of 5 timesteps. The TWTT ranging results will be used for estimation when the variable i_{twtt} is true. This measurement will take 10 timesteps to complete due to round-trip transmission. All the observation results are input into the particle filter estimator. When the variable i_{op} is true, the simulator will check if the transmitting AUV keeps d_Δ within 1. Considering the BER in a real situation, even in the case of alignment, the LOS link may fail with the probability of 1% in this experiment. The LOS link is considered

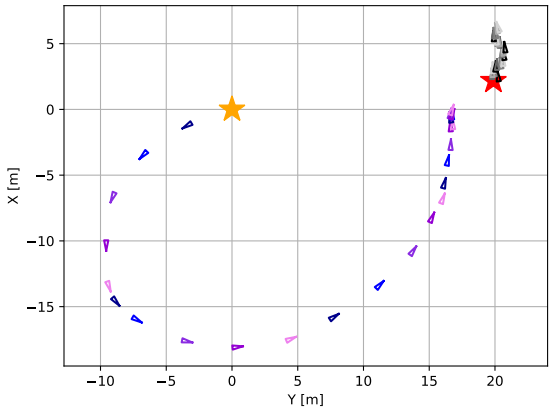


Fig. 4. The trajectories of the AUVs in an episode. The orange and red star markers are the starting points for the transmitting AUV and the receiving AUV. The positions of vehicles are plotted every 4 timesteps. The transmitting AUV is cyclically represented by dark-blue, blue, blue-violet, dark-violet, 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.

to be aligned at the current timestep when $d_{\triangle}$ meets the conditions. One option is to activate the end state when the maximum timestep of this episode is reached. Another option for the end state is that the transmitting AUV successfully maintains the LOS link for a period of time. After activating the end state in this way, the transmitting AUVs can receive the final reward. The required alignment duration of the LOS link, which is temporarily set to 1 timestep at the beginning of this training work, can be adjusted in learning.

Similar to the actual environment, the noises are introduced in all aspects of movements in the simulator. These may be caused by external disturbances and the uncertainties of vehicles. In each timestep, AUVs are sailing with planned velocities. The real surge, sway, and yaw angular velocities are derived from desired velocities (decided by action) mixed with Gaussian noises, whose standard deviations are 0.1, 0.1, and 1, respectively. The measured velocities are also the mixing results of real velocities and Gaussian noises with the standard deviations of 0.1, 0.1, and 1. The same type of noise occurs again when inputting the measured velocities into the particle filter algorithm.

The simulation environment defined above is built through the OpenAI Gym interface [46]. In the simulation, all the states, actions, and received rewards of the agent at each timestep are collected for policy training.

C. Curriculum Learning

Curriculum learning [47] is a continuous method that gradually changes the training rules from simple to complex during the agent learning process. The alignment is a complicated task because of several challenge requirements for the relative position, navigation, and energy consumption. Directly implementing all the task requirements on training may cause results to converge slowly.

To improve learning efficiency, the current research proposes a curriculum learning method in policy learning. The transmitting AUV first learns to track the target, align the beam, and finally extends the alignment duration. In the first step, the agent learns to track the target while ignoring the navigation and optical communication actions. The learning process is based on the reward function r_1. In the second step, the agent starts to receive rewards from navigation and optical communication actions based on reward function r_2. It is necessary for the AUV to consider the use of acoustic and optical devices to establish the LOS link. The final reward can be received when the AUV has activated the end state by successfully maintaining the link. The required alignment duration for end state i_{done} is 1 timestep. In the third step, the agent intends to maintain the LOS link for a longer time with the reward function r_2. The required alignment duration starts from 1 timestep and gradually grows to 10 timesteps. The reward functions r_1 and r_2 are as follows:

$$r_1(s,a) = -\rho_1 d_{\Delta}^{\frac{1}{2}} - \rho_4 u_{\Delta} - \rho_5 r_{\Delta} \tag{23}$$

and

$$\begin{aligned} r_2(s,a) = &- \rho_1(1+\rho_2 i_{twtt})(1+\rho_3 i_{op}) d_{\Delta}^{\frac{1}{2}} \\ &- \rho_4 u_{\Delta} - \rho_5 r_{\Delta} + \rho_6 i_{done} \end{aligned} \tag{24}$$

A comparison is presented to discuss the effect of the curriculum learning method on policy learning. The total training steps of two agents, the parameters used in learning, and the condition for activating the end state are the same. One agent directly learns the policy from 10^7 timesteps of sample data based on reward function r_2. Another agent first learns from 5×10^6 timesteps of sample data with reward r_1 and then learns from 5×10^6 timesteps of sample data with r_2. The coefficients ρ_1 to ρ_6 in reward functions use 0.01, 9, 1, 0.01, 0.002, and 0, respectively.

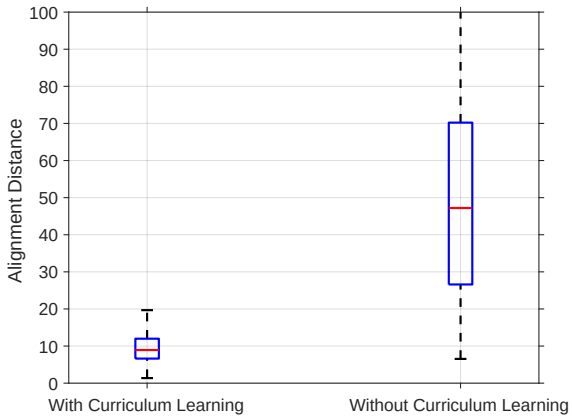


Fig. 5. Average alignment distance statistics of two policies in 1000 episodes. In box-and-whisker plots, the lower and upper boundaries of the box represent the 25th (Q1) and 75th (Q3) percentiles, respectively; the bottom and top ends of the whisker indicate the most extreme values within the lower limit $\mathrm{Q1}-1.5(\mathrm{Q3}-\mathrm{Q1})$ and the upper limit $\mathrm{Q3}+1.5(\mathrm{Q3}-\mathrm{Q1})$, respectively; the red line inside the box marks the median.

The reason for this reward configuration is given in reward shaping work. We tested two learned policies for 1000 episodes in the previously defined simulation environment. As shown in Fig. 5, the alignment distance statistics of two AUVs from 1000 trials is used to evaluate these two policies. The policy without curriculum learning cannot converge alignment distance while the agent using the curriculum learning method effectively maintains the relative distance with the target. Learning to keep tracking and reduce energy consumption at the same time makes the task too complicated. The mixed rewards from different terms slow down the learning efficiency at the beginning of training.

D. Reward Shaping

Reward shaping is a method that gives additional shaping to the reward function and guides the agent towards learning an optimal policy faster [48]. As mentioned in Equation (7), in addition to the final reward, there are three other terms in the proposed reward function: the term to minimize the alignment distance d_{Δ} , the term to reduce the relative movement between two AUVs, and the term to restrict the usage of TWTT ranging and the optical transmitter. All terms guide the transmitting AUV to learn the LOS link alignment faster.

The reward function provides a trade off between different controlling objectives through the coefficient. We propose the following coefficients in the reward function and discuss the reasons

in this subsection:

$$r_1(s,a) = -0.01 d_{\Delta}^{\frac{1}{2}} - 0.01 u_{\Delta} - 0.002 r_{\Delta} \tag{25}$$

and

$$\begin{aligned} r_2(s,a) = &- 0.01 d_{\Delta}^{\frac{1}{2}} (1 + 9 i_{twtt})(1 + i_{op}) \\ &- 0.01 u_{\Delta} - 0.002 r_{\Delta} + 10 i_{done} \end{aligned} \tag{26}$$

In the alignment process, the transmitting AUV needs to minimize the alignment distance and then keep the relative motion when the two AUVs are ready to establish a LOS link. It requires that the reward from term $\rho_1 d_{\Delta}^{\frac{1}{2}}$ is relatively larger than the rewards from $\rho_4 u_{\Delta}$ and $\rho_5 r_{\Delta}$ when the value of d_{Δ} is larger than 1. We accordingly set coefficients ρ_1, ρ_4, and ρ_5 to 0.01, 0.01, and 0.002.

The terms with coefficients ρ_2 and ρ_3, may reduce the usage of TWTT ranging and the optical transmitter while the transmitting AUV is far from the receiving AUV since these terms are multiplied in the first term of Equation (7). It is to save energy and channel resource consumption in unnecessary situations. The taking of actions i_{twtt} and i_{op} is encouraged when the transmitting AUV approaches receiving AUV. Considering that the acoustic channel resources are more precious than energy in practice, the coefficients ρ_2 and ρ_3 are set to 9 and 1, respectively.

The final reward received due to successful transmission is the only positive term in the reward function. The magnitude difference between this final reward and the other negative rewards can affect the learning performance. Several learning cases are conducted here where the coefficient ρ_6 is set to 0, 1, 10, 100, and 1000. The learning cases use the curriculum learning method, and 5×10^6 timesteps sample data are learned with reward functions r_1 and r_2. The coefficients ρ_1 to ρ_5 in reward functions use 0.01, 9, 1, 0.01, and 0.002. All learned policies are tested for 1000 episodes in the simulation environment. The received negative rewards (without final reward) per episode are shown in Table I. The learning case with the coefficient ρ_6 set to 10 accumulated the most rewards in the simulation experiments. The learned policy also reduces the usage of TWTT ranging and the optical transmitter.

TABLE I

STATISTICS OF LEARNING CASES WITH DIFFERENT FINAL REWARDS IN 1000 EPISODES

	Statistics in different cases				
	$\rho_6 = 0$	$\rho_6 = 1$	$\rho_6 = 10$	$\rho_6 = 100$	$\rho_6 = 1000$
Average negative rewards[*]	-3.33	-3.26	-3.18	-4.82	-4.34
TWTT ranging[**]	2.76	2.21	1.79	3.92	2.90
Optical transmitter[***]	21.89	24.48	25.37	34.42	31.43

[*] Received rewards without final reward

[**] Average number of TWTT requests per episode

[***] Average timestep of turning on the optical transmitter per episode

V. LEARNING RESULTS AND DISCUSSION

We train a policy for the alignment task and present the learning results in this section. A comparison experiment with other baseline approaches is designed. The application of the learned policy to real AUVs by transfer learning is discussed.

A. Policy Learning

We train an alignment policy for the transmitting AUV that can maintain the LOS link for 10 timesteps. The policy training uses a curriculum learning method defined in Section IV-C. The agent firstly learns 5×10^6 timesteps of sample data with reward r_1. Then, the agent starts to learn 2×10^6 timesteps of sample data with reward r_2. The required alignment duration for activating end state i_{done} is 1 timestep. We gradually improve the requirement of alignment duration from 1 timestep to 10 timesteps. The required duration will be increased by one timestep after every 2×10^6 sample data are learned. The coefficients of the reward function are given in Equation (25) and Equation (26). All the sample data are collected from the simulator defined in Section IV-B. The SAC algorithm and hyperparameter configurations are given in Section IV-A.

The learned policy is tested for 10000 episodes in the simulation environment. The evaluation statistics are shown in Table II. The learned policy has a 97.53% probability of maintaining the LOS link for 10 timesteps. It takes an average of 262.62 timesteps to complete alignment. The optical transmitter does not need to be turned on all the time. Decided by the learned policy, it only needs to be used for 168.28 timesteps. In each episode, the transmitting AUV requests

TABLE II

STATISTICS OF LEARNED POLICY IN 10000 EPISODES

	Statistics
Success rate	97.53%
Average episode duration*	262.62
Average rewards	-3.92
TWTT ranging**	12.13
Optical transmitter***	168.28

* The unit is timestep and one timestep is considered one second in the real world.

** Average number of TWTT requests per episode

*** Average timestep of turning on the optical transmitter per episode

12.13 TWTT ranging. In the rest of the time, the transmitting AUV listens to the OWTT ranging signal, which requires less acoustic channel resources. This combination of navigation can still ensure that the relative position of the two vehicles is maintained. As in the distribution shown in Fig. 6 and Fig. 7, the use of acoustic and optical devices is less than average in most episodes. The distribution in Fig. 6, Fig. 7, and Fig. 8 is similar. As the length of the episode increases, the vehicle will request more TWTT ranging, and the optical communication device is turned on more frequently. Compared with the episode length distribution as shown in Fig. 8, the tail of the distribution in Fig. 6 and Fig. 7 is lighter. It indicates our policy optimizes the use of acoustic and optical devices, which is a goal we propose in our research. As for the optimization of task duration, it needs to be compared with other methods.

B. Selected Episodes

Two episodes using this learned policy are presented here. With the trajectory and the estimated states, we discuss the AUV manipulation and navigation.

As the trajectory is shown in Fig. 9, the transmitting AUV marked in the blue triangle is started from the origin point, and the initial yaw angles of two AUVs are randomly generated. The transmitting AUV outputs high surge and yaw angular velocities initially, aiming to approach the receiving platform as soon as possible. The negative reward from the distance value is larger

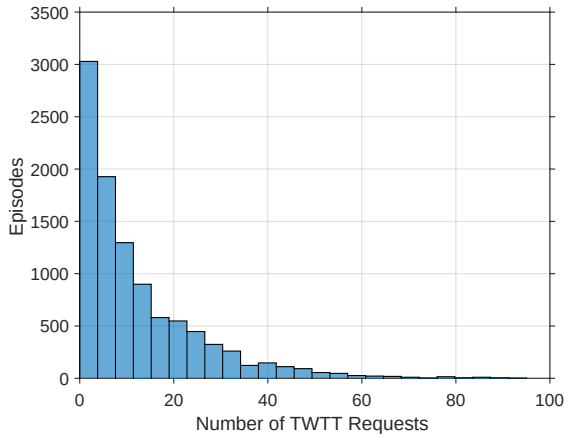


Fig. 6. The distribution of TWTT requests in a episode.

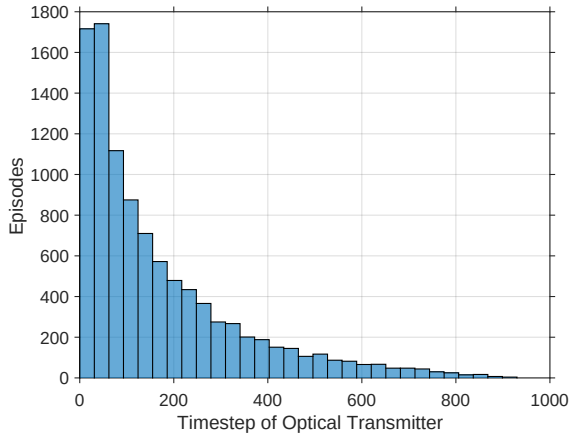


Fig. 7. The distribution of the usage of optical transmitter per episode. It counts the timestep of turning on the optical transmitter in one episode. The maximum timestep of each episode is set to 1000.

than that from velocity and angle differences. After arriving at the vicinity of the target, the transmitting AUV slows down and begins to adjust its attitude for alignment. Meantime, the optical transmitter remains open. It takes 326 timesteps for the transmitting AUV to complete the beam alignment of 10 timesteps with 5 requesting TWTT rangings. The duration of this episode is higher than average, which means that the vehicle took longer time for alignment. As shown in the trajectory, the transmitting AUV may keep a similar velocity to follow the receiving AUV if the alignment is not successful. Such path planning conforms to common sense in target tracking and alignment. It proved that this reward function's shaping and training

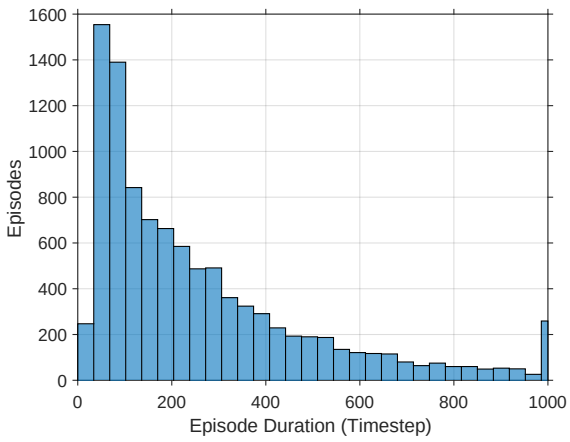


Fig. 8. The distribution of the length of the episode. It indicates how many timesteps the vehicle takes to complete the alignment in an episode. One timestep is considered one second in the real world.

rules are reasonable to manipulate the underwater vehicle.

Another episode with estimation data is plotted in Fig. 10. In this episode, the transmitting AUV used 225 timesteps to complete the alignment of 10 timesteps with 19 requests of TWTT ranging (at timestep of 41, 51, 61, 71, 81, 91, 101, 111, 121, 131, 141, 151, 161, 171, 181, 191, 201, 211, and 221). The optical communication device is on for 177 timesteps. The particle filter results are depicted by blue dots in different timesteps. The states related to estimation are listed in the bottom left corner of each figure. In the beginning, the negative reward of requesting TWTT ranging is relatively large. Without TWTT navigation, the transmitting AUV moves toward the receiving AUV since the accurate position is not essential at this stage. But in a few episodes, the vehicle may request TWTT ranging once at the beginning.

At the timestep of 41, the transmitting AUV requests TWTT ranging signals for the first time. Comparing the particle's distribution in step 41 with that in step 40, one can find that the estimation results begin to converge. In steps 51 and 61, the vehicle requests TWTT ranging again, making the particles further converge. After three TWTT ranging actions, the standard deviation of particle filter estimation results in position (x^T, y^T) being reduced from (1.84, 1.62) to (0.37, 0.41). At this time, the transmitting AUV has reached the vicinity of the receiving AUV. The transmitting AUV constantly tries to keep the alignment distance within 1 meter. In step 221, the vehicle requests TWTT ranging for the last time. The standard deviation of particle filter estimation results broadly maintains a similar value as before. The navigation request timing is

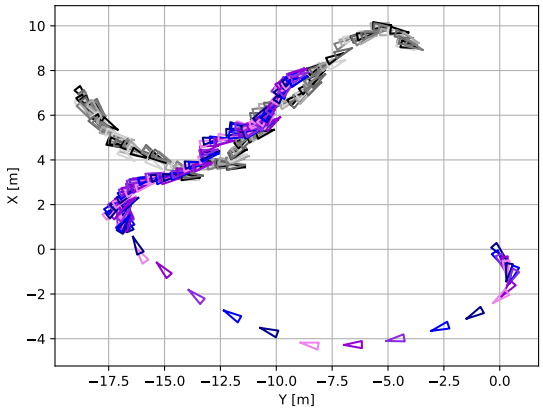


Fig. 9. The trajectories of the AUVs in an episode. The positions of vehicles are plotted every 2 timesteps. The transmitting AUV is cyclically represented by dark-blue, blue, blue-violet, dark-violet, and violet triangles (every 10 timesteps), while the receiving AUV is represented by black, dim-grey, grey, dark-grey, and light-grey triangles. The sharp corner of the triangle is the head of the vehicle. The transmitting AUV used 326 timesteps to complete the alignment of 10 timesteps with 5 TWTT rangings (at a timestep of 44, 54, 64, 174, and 212). The optical communication device is on for 257 timesteps.

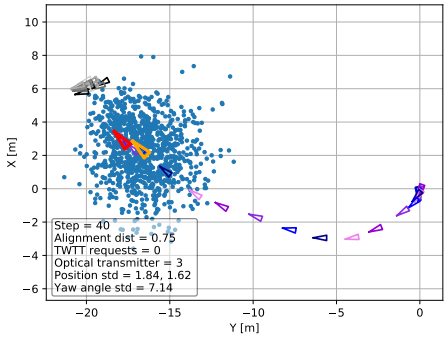
appropriate and helps the transmitting AUV finish this episode within 225 timesteps.

Under partial observability, the agent makes decisions based on the belief. In order to act optimally, the agent may need to perform actions that gather information about the current belief instead of trying to maximize immediate reward. In this experimental evaluation, the transmitting AUV performed TWTT ranging as the information gathering action. It shows that the agent has learned complex behavior that requires advanced exploration techniques. It also indicates a method that forces exploration is needed, such as the SAC algorithm.

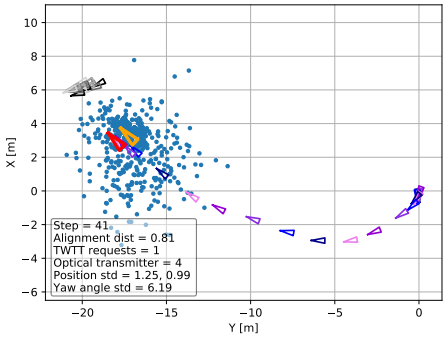
C. Heuristic Approach Comparison

To evaluate the significance of our reinforcement learning based method, a heuristic baseline approach derived from the motion planning method used in previous experiments is implemented for comparison [49]. The vehicle used in previous experiments is the same type of hovering AUV. The performance of this heuristic method is verified by the real field experiments for visual mapping of shallow vent fields.

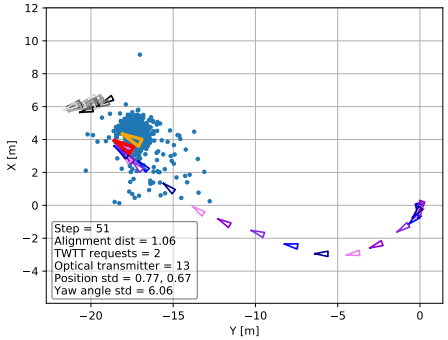
Unlike our proposed method, the transmitting AUV in the heuristic approach knows that it needs to move towards the receiving AUV and reduce the distance d_Δ to complete the alignment. The motion rules refer to the previous research [49]. The transmitting AUV adjusts its yaw angle



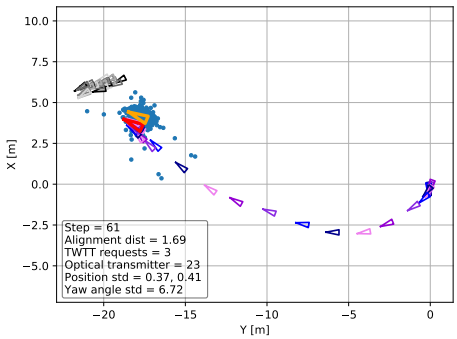
(a) Step 40



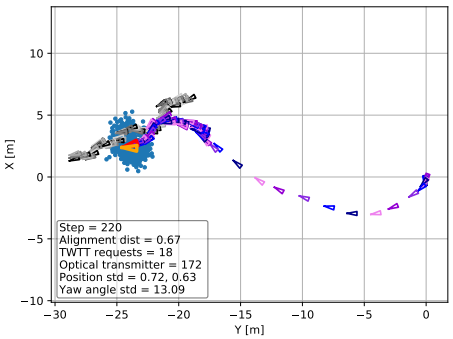
(b) Step 41



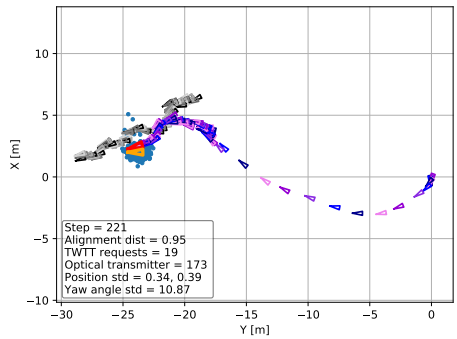
(c) Step 51



(d) Step 61



(e) Step 220



(f) Step 221

Fig. 10. The states of AUVs at the timestep of (a) 40, (b) 41, (c) 51, (d) 61, (e) 220, and (f) 221. The trajectories of vehicles are plotted every 2 timesteps. The transmitting AUV is cyclically represented by dark-blue, blue, blue-violet, dark-violet, and violet triangles (every 10 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. The particle filter estimation results of the transmitting AUV at the current timestep are depicted by blue dots. The current position estimated by a particle filter and the real position are indicated by orange and red triangles, respectively. The parameters listed in the bottom left corner are the current timestep in this episode, the alignment distance $d_{\triangle}$, the number of times to request TWTT ranging, the number of times to turn on the optical transmitter, and the standard deviation of particle filter estimation results in position (x^T, y^T) and yaw orientation (ψ^T).

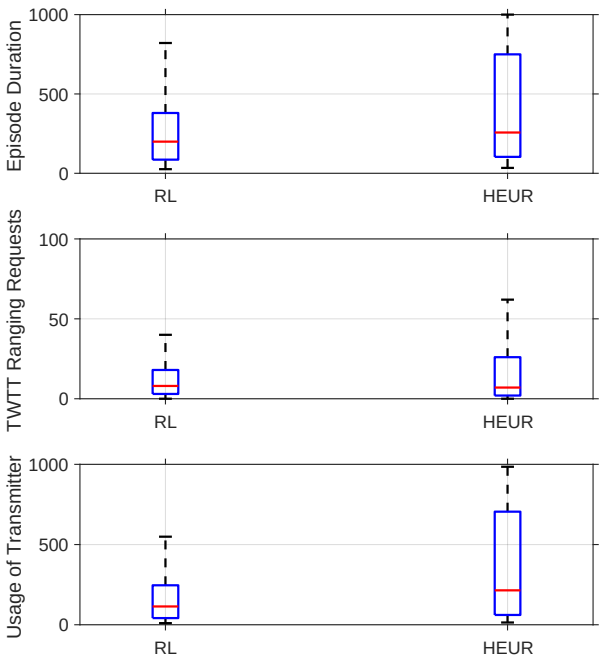


Fig. 11. The comparison of our proposed reinforcement learning based approach (RL) and the heuristic approach (HEUR). The three plots indicate the timesteps needed to complete the episode, the number of times to request TWTT ranging, and the number of times to open the optical transmitter, respectively. 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 $\mathrm{Q1} - 1.5(\mathrm{Q3} - \mathrm{Q1})$ and the upper limit $\mathrm{Q3} + 1.5(\mathrm{Q3} - \mathrm{Q1})$, respectively; the red line inside the box marks the median.

to point to the receiving AUV. The vehicle starts to move forward when the yaw angle deviation from the expected value is less than the set maximum value ψ_{max}:

$$\psi_{\Delta} = \psi^{T} - \psi^{TR} \leq \psi_{max} \tag{27}$$

where ψ_{Δ} is the yaw angle deviation and ψ^{TR} is the relative angle from the receiving AUV to the transmitting AUV with respect to North.

The yaw angular velocity r^{T} is given by:

$$r^{T} = \begin{cases} K_{p,\psi}\psi_{\Delta}, & K_{p,\psi}\psi_{\Delta} \leq r^{T}_{max} \\ r^{T}_{max}, & r^{T}_{max} < K_{p,\psi}\psi_{\Delta} \end{cases} \tag{28}$$

where r^{T}_{max} is the maximum yaw angle velocity of the transmitting AUV, which we defined in the simulator configuration. $K_{p,\psi}$ is the proportional gain for yaw angle velocity.

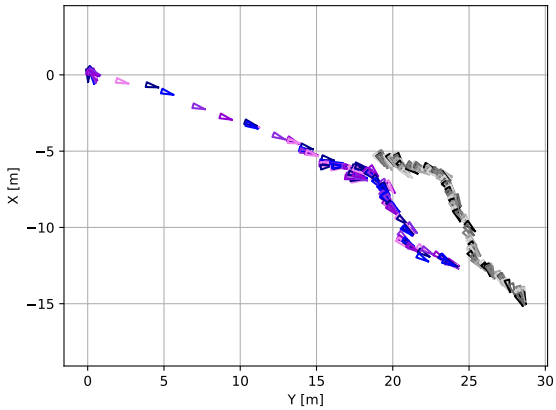


Fig. 12. The trajectories of the AUVs with the HEUR approach. The positions of vehicles are plotted every 2 timesteps. The transmitting AUV is cyclically represented by dark-blue, blue, blue-violet, dark-violet, and violet triangles (every 10 timesteps), while the receiving AUV is represented by black, dim-grey, grey, dark-grey, and light-grey triangles. The sharp corner of the triangle is the head of the vehicle. The transmitting AUV used 254 timesteps to complete the alignment of 10 timesteps with 14 TWTT rangings. The optical communication device is on for 221 timesteps.

The surge velocity u^T is decided by:

$$u^T = \begin{cases} 0, & \psi_{max} < \psi_\Delta \\ K_{p,d} d_\Delta, & K_{p,d} d_\Delta \leq u^T_{max} \; and \; \psi_\Delta \leq \psi_{max} \\ u^T_{max}, & u^T_{max} < K_{p,d} d_\Delta \; and \; \psi_\Delta \leq \psi_{max} \end{cases} \tag{29}$$

where u^T_{max} is the maximum surge velocity of the transmitting AUV, which we defined in the simulator configuration. $K_{p,d}$ is the proportional gain for surge velocity.

The heuristic approach, called HEUR, is deploying the above motion rules to AUV for comparison. The value ψ_{max} is set to 5 degrees. The proportional gain $K_{p,\psi}$ and $K_{p,d}$ use 0.5 and 0.5, respectively. The optical transmitter in the transmitting AUV turns on when d_Δ is less than 5 meters. The HEUR approach uses navigation methods, including OWTT and TWTT acoustic ranging, in the movement. Whether to request for TWTT ranging is determined by our reinforcement learning based policy.

The reinforcement learning trained policy (RL), HEUR are tested in the simulation environment for 1000 episodes, respectively. The transmitting AUV is required to maintain the LOS link for 10 timesteps. The success rate in HEUR is 82%, while RL has an 97.6% success rate. As the first plots are shown in Fig. 11, the AUV using the RL approach completes the alignment

in fewer timesteps. The AUVs in RL approach and HEUR approach use the same policy to determine when to request TWTT ranging. The average number of TWTT ranging requests in the HEUR approach is higher than in the RL approach. A trajectory of AUVs using the HEUR approach is shown in Fig. 12. The AUV guided by the heuristic approach also keeps the alignment distance at a small value. But the motion planning decided in Equation (28) and Equation (29) does not help the AUV to maintain the LOS link well. According to the third plot in Fig. 11, our reinforcement learning method also significantly reduces the use of the optical transmitter, which saves the energy for AUV. It concludes that our proposed reinforcement learning approach outperforms the heuristic approach.

D. Simulation to Reality

The learned policy needs to be applied to the real world AUV system. The real environment is more complicated since unknown situations may occur. The gap between the simulated environment (source domain) and the real world (target domain) may degrade the adaptability of the reinforcement learning policy in the real environment [50]. The following technologies are considered for simulation to reality transfer.

Domain randomization is usually utilized in the training process to encourage policies to be robust to the different environments [51] [52]. We randomize the initial conditions of the simulator, aiming to cover the real distribution in the target domain. Besides, perturbations are introduced in many aspects of the simulation environment. It can enhance the stability of the AUVs in the real environment against environmental disturbances. If observations differ between simulation and the real world, we can also learn a mapping between those.

The excellent performance of the SAC algorithm in transferring to real robots is one of the reasons we choose it [21]. The maximum entropy reinforcement learning method can provide a robust framework that minimizes the need for hyperparameter tuning [53]. When we need to adjust parameters and shape the reward function for a new environment, the already collected data can be reused since the SAC is an off-policy algorithm. Meanwhile, the good sample efficiency of SAC can speed up the transfer to the real world AUV.

We prepared experimental facilities to deploy the learned policy in the real environment. The hovering AUV Tri-TON 2 is used as the transmitting AUV in real experiments, while the autonomous surface vehicle BUTTORI is used as the receiving AUV. The DVL Teledyne RDI Navigator 1200 kHz is attached for measuring the ground velocity. The SeaTrac X150 is used

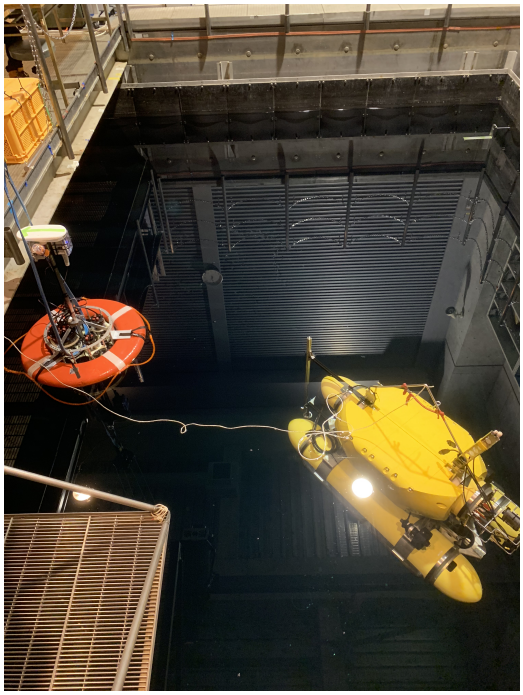


Fig. 13. Experimental facilities in the water tank. The AUV Tri-TON 2 and the autonomous surface vehicle BUTTORI are deployed in the water tank.

as the ultra-short baseline (USBL) device for acoustic ranging. The JAE JG-35FD is utilized as the FOG device for measuring orientation.

As shown in Fig. 13, we implemented the policy learned from the simulation environment on the AUV Tri-TON 2. The AUV successfully aligned with the static platform in the tank. We regard the alignment task in the real environment as a more difficult task than in the simulation environment. The policy we have trained needs to be further improved from the real environment. The AUV Tri-TON 2 that implemented our learned policy can repeatedly conduct alignment tasks in water tank and seawater. The AUV is randomly initialized at a random location, and then it starts to track and align with the target. The collected data will be used to retrain the policy in the future.

VI. Conclusion

Optical beam alignment between underwater vehicles is of great significance to implement high data rate UWOC technology in ocean exploration. With the help of acoustic navigation, this research provides a solution to maintain a relative position to establish LOS links between AUVs. The previous method of using additional servos to control the beam pointing and scanning for

link acquisition is no longer needed. Integrating acoustic communication makes it possible to identify the initial location in the vast ocean environment. The acoustic ranging based observation and particle filter estimator effectively bound the position error in the alignment process. When the position error between the two AUVs remains within the coverage area of the optical beam, the LOS link can be successfully maintained.

The SAC reinforcement learning algorithm, together with reward shaping based curriculum learning, succeeds in policy optimization. It takes less time to complete the alignment than the heuristic approach. The learned policy reduces the use time of the optical transmitter, which saves valuable energy for the underwater investigation of the AUV. We also decrease the TWTT ranging requests and reduce the occupancy of the acoustic communication channel. Saving channel resources is favorable to popularize the UWOC alignment method in multiple AUV operations. Multiple vehicles in the formation can use optical communication to share survey data simultaneously.

Furthermore, our proposed alignment method is practical and straightforward. The SAC algorithm shows excellent performance in simulation to reality works. The model-free reinforcement learning framework makes it easier to apply a learned policy to different AUVs. There is no requirement for light intensity detection. In addition to the beam director, the alignment method does not use the camera for visual positioning. On the one hand, it avoids heavy image computations. On the other hand, it prevents interference in wireless optical communication.

ACKNOWLEDGMENT

This research is the result of the academic exchange program between the Institute of Industrial Science, The University of Tokyo and Technische Universität Darmstadt funded by the Continental Automotive Corporation. The authors would like to thank them for their generous support of this transnational research during the epidemic of COVID-19.

REFERENCES

[1] E. Petritoli, M. Cagnetti, and F. Leccese, "Simulation of autonomous underwater vehicles (auvs) swarm diffusion," *Sensors*, vol. 20, no. 17, p. 4950, 2020.

[2] J.-g. Huang, H. Wang, C.-b. He, Q.-f. Zhang, and L.-y. Jing, "Underwater acoustic communication and the general performance evaluation criteria," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 8, pp. 951–971, 2018.

[3] Z. Zeng, S. Fu, H. Zhang, Y. Dong, and J. Cheng, "A survey of underwater optical wireless communications," *IEEE communications surveys & tutorials*, vol. 19, no. 1, pp. 204–238, 2016.

[4] F. Leccese and G. S. Spagnolo, "Led-to-led wireless communication between divers," *ACTA IMEKO*, vol. 10, no. 4, pp. 80–89, 2021.

[5] J. Shi, X. Zhu, F. Wang, P. Zou, Y. Zhou, J. Liu, F. Jiang, and N. Chi, "Net data rate of 14.6 gbit/s underwater vlc utilizing silicon substrate common-anode five primary colors led," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 2019, pp. 1–3.

[6] C. Gabriel, M.-A. Khalighi, S. Bourennane, P. Léon, and V. Rigaud, "Monte-carlo-based channel characterization for underwater optical communication systems," *Journal of Optical Communications and Networking*, vol. 5, no. 1, pp. 1–12, 2013.

[7] G. G. Peter, S. R. William, R. Mahon, L. M. James, S. F. Mike, R. S. Michele, R. S. Walter, B. X. Ben, R. B. Harris, I. M. Christopher *et al.*, "Modulating retro-reflector lasercom systems at the naval research laboratory," in *2010-MILCOM 2010 MILITARY COMMUNICATIONS CONFERENCE*. IEEE, 2010, pp. 1601–1606.

[8] M. M. Abadi, M. A. Cox, R. E. Alsaigh, S. Viola, A. Forbes, and M. P. Lavery, "A space division multiplexed free-space-optical communication system that can auto-locate and fully self align with a remote transceiver," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[9] N. D. Hardy, H. G. Rao, S. D. Conrad, T. R. Howe, M. S. Scheinbart, R. D. Kaminsky, and S. A. Hamilton, "Demonstration of vehicle-to-vehicle optical pointing, acquisition, and tracking for undersea laser communications," in *Free-Space Laser Communications XXXI*, vol. 10910. International Society for Optics and Photonics, 2019, p. 109100Z.

[10] P. B. Solanki, S. D. Bopardikar, and X. Tan, "Active alignment control-based led communication for underwater robots," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1692–1698.

[11] B. Cochenour and L. Mullen, "Channel response measurements for diffuse non-line-of-sight (nlos) optical communication links underwater," in *OCEANS'11 MTS/IEEE KONA*. IEEE, 2011, pp. 1–5.

[12] J. González-García, A. Gómez-Espinosa, E. Cuan-Urquizo, L. G. García-Valdovinos, T. Salgado-Jiménez, and J. A. E. Cabello, "Autonomous underwater vehicles: Localization, navigation, and communication for collaborative missions," *Applied Sciences*, vol. 10, no. 4, p. 1256, 2020.

[13] S. Smith and D. Kronen, "Experimental results of an inexpensive short baseline acoustic positioning system for auv navigation," in *Oceans' 97. MTS/IEEE Conference Proceedings*, vol. 1. IEEE, 1997, pp. 714–720.

[14] T. Maki, R. Shiroku, Y. Sato, T. Matsuda, T. Sakamaki, and T. Ura, "Docking method for hovering type auvs by acoustic and visual positioning," in *2013 IEEE international underwater technology symposium (UT)*. IEEE, 2013, pp. 1–6.

[15] L. Kong, W. He, C. Yang, Z. Li, and C. Sun, "Adaptive fuzzy control for coordinated multiple robots with constraint using impedance learning," *IEEE transactions on cybernetics*, vol. 49, no. 8, pp. 3052–3063, 2019.

[16] L. Kong, W. He, Y. Dong, L. Cheng, C. Yang, and Z. Li, "Asymmetric bounded neural control for an uncertain robot by state feedback and output feedback," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019.

[17] L. Kong, W. He, W. Yang, Q. Li, and O. Kaynak, "Fuzzy approximation-based finite-time control for a robot with actuator saturation under time-varying constraints of work space," *IEEE Transactions on Cybernetics*, 2020.

[18] R. Cui, C. Yang, Y. Li, and S. Sharma, "Adaptive neural network control of auvs with control input nonlinearities using reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 6, pp. 1019–1029, 2017.

[19] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[20] G. E. Monahan, "State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms," *Management science*, vol. 28, no. 1, pp. 1–16, 1982.

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

[22] T. I. Fossen, *Handbook of marine craft hydrodynamics and motion control*. John Wiley & Sons, 2011.

[23] T. Matsuda, T. Maki, and T. Sakamaki, "Accurate and efficient seafloor observations with multiple autonomous underwater vehicles: theory and experiments in a hydrothermal vent field," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2333–2339, 2019.

[24] T. Matsuda, T. Maki, K. Masuda, and T. Sakamaki, "Resident autonomous underwater vehicle: Underwater system for prolonged and continuous monitoring based at a seafloor station," *Robotics and Autonomous Systems*, vol. 120, p. 103231, 2019.

[25] Y. Guo, A. Trichili, O. Alkhazragi, I. Ashry, T. K. Ng, M.-S. Alouini, and B. S. Ooi, "On the reciprocity of underwater turbulent channels," *IEEE Photonics Journal*, vol. 11, no. 2, pp. 1–9, 2019.

[26] Y. Weng, Y. Guo, O. Alkhazragi, T. K. Ng, J.-H. Guo, and B. S. Ooi, "Impact of turbulent-flow-induced scintillation on deep-ocean wireless optical communication," *Journal of Lightwave Technology*, vol. 37, no. 19, pp. 5083–5090, 2019.

[27] S. Chandrasekhar, "Dover books on intermediate and advanced mathematics," 1960.

[28] R. A. Leathers, T. V. Downes, C. O. Davis, and C. D. Mobley, "Monte carlo radiative transfer simulations for ocean optics: a practical guide," Naval Research Lab Washington Dc Applied Optics Branch, Tech. Rep., 2004.

[29] C. F. Bohren and D. R. Huffman, *Absorption and scattering of light by small particles*. John Wiley & Sons, 2008.

[30] C. Shen, Y. Guo, H. M. Oubei, T. K. Ng, G. Liu, K.-H. Park, K.-T. Ho, M.-S. Alouini, and B. S. Ooi, "20-meter underwater wireless optical communication link with 1.5 gbps data rate," *Optics express*, vol. 24, no. 22, pp. 25 502–25 509, 2016.

[31] Y. Guo, M. Kong, M. Sait, S. Marie, O. Alkhazragi, T. K. Ng, and B. S. Ooi, "Compact scintillating-fiber/450-nm-laser transceiver for full-duplex underwater wireless optical communication system under turbulence," *Optics Express*, vol. 30, no. 1, pp. 53–69, 2022.

[32] Z. J. Harris and L. L. Whitcomb, "Preliminary evaluation of cooperative navigation of underwater vehicles without a dvl utilizing a dynamic process model," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.

[33] K. Fujita, T. Matsuda, and T. Maki, "Bearing only localization for multiple auv with acoustic broadcast communication," *International Conference on Control, Automation and Systems*, pp. 1371–1376, 2019.

[34] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[35] V. Myers and D. P. Williams, "A pomdp for multi-view target classification with an autonomous underwater vehicle," in *OCEANS 2010 MTS/IEEE SEATTLE*. IEEE, 2010, pp. 1–5.

[36] J. Pajarinen and V. Kyrki, "Robotic manipulation of multiple objects as a pomdp," *Artificial Intelligence*, vol. 247, pp. 213–228, 2017.

[37] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online pomdp planning for autonomous driving in a crowd," in *2015 ieee international conference on robotics and automation (icra)*. IEEE, 2015, pp. 454–460.

[38] H. Wu, S. Song, K. You, and C. Wu, "Depth control of model-free auvs via reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 12, pp. 2499–2510, 2018.

[39] A. S. Gadre, "Observability analysis in navigation systems with an underwater vehicle application," Ph.D. dissertation, Virginia Tech, 2007.

[40] M. F. Fallon, G. Papadopoulos, J. J. Leonard, and N. M. Patrikalakis, "Cooperative auv navigation using a single maneuvering surface craft," *The International Journal of Robotics Research*, vol. 29, no. 12, pp. 1461–1474, 2010.

[41] G. Antonelli, F. Arrichiello, S. Chiaverini, and G. S. Sukhatme, "Observability analysis of relative localization for auvs based on ranging and depth measurements," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 4276–4281.

[42] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[43] S. Thrun, W. Burgard, and D. Fox, "Probalistic robotics," *Kybernetes*, 2006.

[44] T. Maki, H. Kondo, T. Ura, and T. Sakamaki, "Positioning method for an auv using a profiling sonar and passive acoustic landmarks for close-range observation of seafloors," in *OCEANS 2007-Europe*. IEEE, 2007, pp. 1–6.

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

[46] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[47] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[48] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *ICML*, vol. 99, 1999, pp. 278–287.

[49] T. Maki, H. Mizushima, H. Kondo, T. Ura, T. Sakamaki, and M. Yanagisawa, "Real time path-planning of an auv based on characteristics of passive acoustic landmarks for visual mapping of shallow vent fields," in *OCEANS 2007*. IEEE, 2007, pp. 1–8.

[50] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 737–744.

[51] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," *arXiv preprint arXiv:2004.00784*, 2020.

[52] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[53] T. Haarnoja, V. Pong, K. Hartikainen, A. Zhou, M. Dalal, and S. Levine, "Soft actor critic—deep reinforcement learning with real-world robots," 2018, https://bair.berkeley.edu/blog/2018/12/14/sac.