Journal of *Marine Science and Engineering*

MDPI

Article

Environment-Aware Optimal Placement and Dynamic Reconfiguration of Underwater Robotic Sonar Networks Using Deep Reinforcement Learning

Qiming Sang [1,2], Yu Tian [1,*], Jin Zhang [3], Yuyang Xiao [1], Zhiduo Tan [1], Jiancheng Yu [1] and Fumin Zhang [4]

[1] State Key Laboratory of Robotics and Intelligent Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; sangqiming@sia.cn (Q.S.); yjc@sia.cn (J.Y.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Liaoning University, Shenyang 110036, China
[4] Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong 999077, China
* Correspondence: tiany@sia.cn

Abstract

Underwater dynamic target detection, classification, localization, and tracking (DCLT) is central to maritime surveillance and monitoring and increasingly relies on distributed AUV-based robotic sonar networks operating in passive listening and, when required, cooperative multistatic modes. Achieving a robust performance in realistic oceans remains challenging, because sensor placement must adapt to time-varying acoustic conditions and target priors while preserving acoustic communication connectivity, and because frequent reconfiguration under dynamic currents makes classical large-scale planning computationally expensive. This paper presents an integrated deep reinforcement learning (DRL)-based framework for passive-stage sonar placement and dynamic reconfiguration in distributed AUV networks. First, we cast placement as a constructive finite-horizon Markov decision process (MDP) and train a Proximal Policy Optimization (PPO) agent to sequentially build a collision-free layout on a discretized surveillance grid. The terminal reward is formulated to jointly optimize the environment-aware detection performance, computed from BELLHOP-based transmission loss models, and global network connectivity, quantified using algebraic connectivity. Second, to enable time-critical reconfiguration, we estimate flow-aware motion costs for all AUV–destination pairs using a PPO with a Long Short-Term Memory (LSTM) trajectory policy trained for partial observability. The learned policy can be deployed onboard, allowing each AUV to refine its path online using locally sensed currents, improving robustness to ocean-model uncertainty. The resulting cost matrix is solved via an efficient zero-element assignment method to obtain the optimal one-to-one reassignment. In the reported simulation studies, the proposed Sequential PPO placement method achieves a final reward 16–21% higher than Particle Swarm Optimization (PSO) and 2–3.7% higher than the Genetic Algorithm (GA), while the proposed PPO + LSTM planner reduces average travel time by 30.44% compared with A*. The proposed closed-loop architecture supports frequent re-optimization, scalable fleet operation, and a seamless transition to communication-supported cooperative multistatic tracking after detection, enabling efficient, adaptive DCLT in dynamic marine environments.

Keywords: underwater acoustic sensor network; sensor placement; dynamic reconfiguration; path planning; deep reinforcement learning

Check for updates

Academic Editor: Rafael Morales

Received: 17 February 2026
Revised: 11 April 2026
Accepted: 13 April 2026
Published: 15 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Underwater dynamic target detection, classification, localization, and tracking (DCLT) encompass several fundamental tasks in underwater sensing. Specifically, detection determines whether a target is present, classification identifies the type of target, localization estimates its position, and tracking continuously updates its motion state over time. Together, these capabilities play a pivotal role in a wide range of maritime applications, including naval surveillance, environmental monitoring, and marine biological research. The complex and challenging underwater environment necessitates the development of specialized sensing technologies. Among these, sonar systems have become indispensable tools due to their capability to detect and localize underwater targets through sound waves, even in the absence of visual information. Sonar systems are generally classified into two categories: active sonar, which emits sound waves and measures the return echoes to detect objects, and passive sonar, which listens for sounds emitted by external sources. Recent advancements in Autonomous Underwater Vehicles (AUVs), which are lightweight, cost effective, and capable of long-duration operation [1–3], have significantly transformed the deployment of sonar systems. The integration of these AUVs with both active and passive sonar technologies to form underwater robotic sonar networks [4–7] enables the development of distributed sonar systems characterized by dynamic deployment, controllable maneuverability, networked cooperation, and intelligent autonomy. This synergy enhances flexible, adaptive, scalable, and efficient underwater operations, providing an improved DCLT performance in dynamic and complex marine environments.

Underwater robotic sonar networks can operate in two primary modes, with each tailored to specific mission requirements. The first is the multistatic mode, where one or multiple sonar sources collaborate with one or multiple receivers [6,8,9]. By fusing high-SNR and informative measurements from receivers positioned at optimal source–target–receiver multistatic geometries, the network can significantly improve its ability to robustly track even quiet or evasive targets. The second mode is passive listening [5], where AUVs detect sounds emitted by external sources, such as marine life or surface vessels. This mode is energy efficient and ideal for long-term surveillance, as it allows the network to remain covert while continuously monitoring large areas. In both operational modes, AUVs must adapt to environmental variations, such as changes in sound speed, ocean currents, and target distribution, to ensure optimal DCLT. To achieve this, the locations of the AUVs and their paths must be strategically optimized, which remains a key research issue in this field.

In multistatic sonar studies aimed at improving target tracking, the primary focus is optimizing ping sequences and AUV motion to create optimal temporal and spatial geometries between the sources, targets, and receivers. This optimization ensures that the receiver captures high-SNR and informative sonar measurement reflected by the target, thereby improving tracking performance aspects, such as accuracy. To achieve this, integrating AUV optimization with target tracking in a closed-loop system is essential. By modeling this system, the AUV's ping sequence or motion can be optimized using search methods. In the bistatic scenario, where a stationary acoustic source cooperates with a receiver-equipped AUV, solutions were proposed in [8,9] based on this approach. Later, this method was expanded in [6] to jointly optimize both ping sequences of multiple stationary sources and AUV motion. However, these methods rely on accurate modeling and face computational challenges. To address these, Ref. [10] recently proposed using deep reinforcement learning (DRL) to learn an optimal AUV motion planning strategy, demonstrating promising potential for overcoming these challenges.

In addition to the above-mentioned multistatic tracking orientated optimization, another key research issue is AUV placement optimization. The primary objective is to

determine the optimal locations for AUVs in the network to maximize performance metrics such as spatial coverage, detection probability, and network connectivity, while minimizing operational costs. For instance, in multistatic sonar networks, multiple sources and receivers must be positioned strategically to maximize coverage while minimizing the overlap. In passive sonar networks, the placement of AUVs needs to be optimized to maximize the probability of detecting external acoustic sources across the surveillance region while maintaining reliable communication links between the AUVs. Effective placement is crucial for improving network performance in both multistatic and passive sonar networks, and significant research efforts have been dedicated to this problem for both operational modes. Typically, sonar sensor placement is framed as a multi-objective optimization task [11,12], and a range of methods have been developed or employed to solve it. These include greedy and submodular approaches [13,14], mathematical programming techniques such as integer linear programming [15] and mixed-integer programming [16,17], as well as heuristic and metaheuristic algorithms like Genetic Algorithms (GAs) [18,19] and Particle Swarm Optimization (PSO) [11,20]. Additionally, a few studies have employed mixed or tailored approaches specifically designed to address the unique structures of the problem [21].

Recently, machine learning-based methods have also been increasingly applied to underwater and marine problems [22]. Among them, reinforcement learning (RL) has emerged as a promising approach for sensor placement optimization [23,24], demonstrating significant potential in underwater robotic sonar network applications. In the present study, RL is particularly attractive because sonar placement is formulated as a constructive sequential decision process in which the final network configuration is built through a sequence of placement actions. This makes the problem naturally compatible with a Markov decision process framework. Compared with conventional population-based optimization methods such as PSO and GA, which directly search for complete solutions in a high-dimensional combinatorial space, the RL-based formulation is better able to exploit the sequential structure of the placement process. Moreover, DRL is more suitable than traditional RL for the present setting, because traditional RL methods typically rely on tabular or low-dimensional representations and become impractical when the state-action space is large. By using deep neural networks to approximate policies and value functions, DRL enables scalable learning in large and structured placement problems. Our earlier work has shown that a DRL-based method can be effectively used for sensor placement optimization [25].

Early research in sonar placement optimization often modeled sonar coverage using simplistic range-based approaches, such as the disk model [17,26,27]. However, there has been a shift toward considering the environmental acoustic performance [28,29]. Environment-aware optimization ensures that sonar placement solutions are more practical and effective in real-world conditions, such as placing a sonar in a deep sound channel to extend its range or avoiding regions with severe shadow zones that would hinder detection. State-of-the-art approaches increasingly integrate acoustic propagation modeling into the sensor placement optimization loop. This can involve the use of numerical acoustic models, such as ray-tracing or parabolic equation-based models, to compute the transmission loss or detection performance from each candidate sensor location to various points within the monitored area. For example, Ref. [30] computed acoustic maps for an ocean region using temperature and salinity profiles from the South China Sea and the BELLHOP acoustic model [31]. Based on these maps, they optimized placement to maximize coverage leveraging the environment's influence on sonar performance. The work by [13] utilized a digital twin of the ocean off Victoria, Australia, using Copernicus Marine forecasts and the PyRAM acoustic model to simulate the detection probability for candidate sensor positions, followed by solving a submodular optimization problem.

Historically, sonar placement research has primarily focused on static sonars, such as sonobuoys, moor buoys, and seabed nodes, resulting in a fixed layout. In contrast, underwater robotic sonar networks allow for adaptive reconfiguration. For instance, when areas with a high probability of target presence shift due to new intelligence, changes in target behavior, or variations in ocean soundscapes, the network can dynamically adjust to maintain optimal coverage or acoustic connectivity. This capability is critical for responding to changing ocean acoustic conditions and evolving tactical situations. However, this introduces a time dimension, as sensor movement incurs a cost in terms of time or energy. Therefore, optimization in underwater robotic sonar networks must not only determine where to place sensors but also address when and how to move them. This necessitates path planning within dynamic ocean flow environments. Given the rapid changes in ocean conditions, optimization and reconfiguration need to be performed quickly to respond to these shifts. As a result, computationally efficient solutions are required for the frequent re-optimization of sonar placement and recalculation of paths for large-scale networks, which present significant challenges. Despite the advancements in sonar placement, a notable research gap remains in systematically addressing the integrated placement and dynamic reconfiguration problem within this context of underwater robotic sonar networks.

To fully realize the potential of underwater robotic sonar networks and address existing research gaps, this paper proposes an integrated sonar placement and dynamic reconfiguration framework, utilizing Proximal Policy Optimization (PPO) [32]-based DRL for both sonar placement optimization and AUV path planning. DRL offers the key advantage of quickly computing optimized solutions after training, with incremental learning allowing for the rapid adaptation to changing conditions and facilitating efficient re-optimization and replanning, which is essential for dynamic reconfiguration. The framework is specifically designed for the passive sonar mode, with the goal of maximizing the detection probability. Additionally, acoustic communication connectivity is incorporated into the sonar placement, enabling the network to transition to the cooperative multistatic tracking mode, which requires acoustic communication support, after the detection of potential targets, utilizing methods such as those proposed in [6,8–10] to enhance maneuvering target tracking and trailing.

The main contributions and novelties of this paper are threefold. First, whereas most existing studies treat sonar placement and dynamic reconfiguration as separate problems, we develop an integrated and environment-aware optimization framework for passive underwater robotic sonar networks that unifies both tasks within a single closed-loop architecture. The framework explicitly incorporates time-varying acoustic propagation, target priors, ocean currents, and network-level communication feasibility, thereby enabling adaptive operation in realistic and uncertain marine environments. Second, unlike conventional placement methods that directly search for a complete sensor configuration using population-based or heuristic optimization, we formulate the sonar placement problem as a constructive finite-horizon Markov decision process (MDP) and develop a PPO-based learning strategy that sequentially builds sensor configurations. Communication feasibility is enforced through a graph-theoretic connectivity constraint based on algebraic connectivity. This formulation preserves multi-hop network connectivity without requiring uniformly strong pairwise links, thereby balancing wide-area sensing coverage and cooperative operability. In addition, the constructive formulation reduces decision-space dimensionality, improves scalability on large spatial grids, and supports the direct optimization of complex, non-analytical objectives derived from physics-based acoustic propagation models. Third, in contrast to reconfiguration approaches that rely on geometric distance or computationally expensive repeated planning, we propose a flow-aware reconfiguration strategy that combines a PPO + LSTM-based trajectory planner under partial observability

with an efficient zero-element assignment method. This strategy enables the rapid and current-aware estimation of motion costs and optimal AUV reassignment for large-scale networks, while also supporting online trajectory adaptation using locally sensed flow information to improve robustness against flow-prediction uncertainty.

The rest of the paper is structured as follows. Section 2 states the problem and overview of the proposed framework. Section 3 details the MDP formulation and PPO-based sensor placement algorithm, including state/action definitions and reward design. Section 4 describes the assignment strategy and PPO–LSTM trajectory planner. Section 5 presents simulation evaluations, illustrating the performance gains of our approach. Finally, Section 6 concludes with a summary and future research directions.

2. Problem Statement and Framework Overview

2.1. Problem Statement

In this study, we consider the dynamic optimization of a distributed passive sonar network composed of G AUVs. Each AUV is equipped both with an acoustic source and a passive sonar array. The passive sonar array is assumed to be omnidirectional, i.e., it exhibits a constant directivity index independent of AUV orientation, as can be achieved using circular array configurations. While the network is capable of operating in both multistatic active and passive sensing modes, this work focuses specifically on the passive surveillance stage, which is critical for covert, energy-efficient, and long-duration monitoring tasks. The abbreviations and symbols used in this paper are listed in Abbreviations Section.

The mission is conducted over a two-dimensional maritime surveillance region, within which underwater targets may appear, maneuver, and emit acoustic signals. The network operation is structured over a sequence of discrete decision epochs $k = 0, 1, \ldots, K$. At each epoch, the network configuration may be updated in response to newly available information. Specifically, environmental conditions such as acoustic propagation characteristics and ambient noise levels, as well as target prior distributions, are updated based on external data sources and accumulated AUV network observations. The objective at each decision epoch is to determine an optimal spatial configuration of the AUV network that maximizes the expected detection performance over the surveillance region, while satisfying constraints on acoustic communication connectivity. Between consecutive epochs, the network undergoes a reconfiguration process in which AUVs move from their current locations to newly assigned positions. These transitions incur time and/or energy costs, particularly under the influence of ocean currents that must be explicitly accounted for. As a result, the overall problem is inherently sequential and dynamic, requiring the joint consideration of sensing performance, communication reliability, and reconfiguration cost across time.

Specifically, the surveillance region is defined as a bounded two-dimensional domain $\Omega = [0, L_x] \times [0, L_y] \subset \mathbb{R}^2$, which is discretized into a uniform grid of $N_x \times N_y$ square cells to enable computationally tractable optimization. This discretization partitions the domain evenly along the length and width directions, with grid resolutions $\Delta x = L_x / N_x$ and $\Delta y = L_y / N_y$, such that each cell has an area of $\Delta A = \Delta x \cdot \Delta y$. Each grid cell represents a candidate deployment location for an AUV, and its center is denoted by $\mathbf{r}_{i,j} = (x_i, y_j)$, $i = 1, \ldots, N_x$; $j = 1, \ldots, N_y$. This discretized representation approximates the continuous surveillance domain by a finite set of candidate sensor locations, facilitating both acoustic performance evaluation and optimization over space. At each decision epoch k, the network configuration is represented as $\mathcal{S}_k = \left\{ \mathbf{s}_1^k, \mathbf{s}_2^k, \ldots, \mathbf{s}_G^k \right\}, \mathbf{s}_g^k \in \Omega$, where $\mathbf{s}_g^k$ represents the position of AUV g at epoch k. Each AUV occupies a unique grid cell, and no two AUVs occupy the same location. Based on the environmental information and target prior, the placement module determines a desired configuration for the next decision epoch, denoted

by $\mathcal{S}_{k+1}^{*} = \left\{\mathbf{s}_1^{k+1}, \mathbf{s}_2^{k+1}, \ldots, \mathbf{s}_G^{k+1}\right\}$. The reconfiguration problem is then to determine how the AUVs in $\mathcal{S}_k$ should be reassigned to the target positions in $\mathcal{S}_{k+1}^{*}$.

Given the current acoustic environment and a target prior distribution $P_T^k(i,j)$ defined over the discretized surveillance region, the expected sensing performance of a configuration $\mathcal{S}_k$ is quantified by an aggregate detection utility $J_d(\mathcal{S}_k)$, which captures the spatially weighted probability of detecting potential targets across the monitored area and reflects both acoustic propagation effects and target likelihoods. In addition, reliable network operation requires that inter-AUV acoustic communication links satisfy minimum quality constraints, which are incorporated through a connectivity cost or penalty term $J_{\text{conn}}(\mathcal{S}_k)$, which discourages configurations that result in weak or fragmented communication topologies.

When transitioning from $\mathcal{S}_k$ to $\mathcal{S}_{k+1}^{*}$, each AUV must be assigned from its current position $\mathbf{s}_i^k$ to one target position $\mathbf{s}_j^{k+1}$. Let c_{ij}^k denote the flow-aware motion cost of assigning AUV i to target position j, and let $\boldsymbol{C}^k = \left[c_{ij}^k\right]$ denote the corresponding cost matrix. The reassignment problem can be formulated as:

$$\min_{\vartheta_{ij}^k} \sum_{i=1}^{G} \sum_{j=1}^{G} c_{ij}^k \vartheta_{ij}^k,\ for\ \forall i, \sum_{j=1}^{G} \vartheta_{ij}^k = 1, and\ for\ \forall j,\ \sum_{i=1}^{G} \vartheta_{ij}^k = 1 \tag{1}$$

where $\vartheta_{ij}^k \in \{0,1\}$, $\vartheta_{ij}^k = 1$ indicates that AUV i is assigned to target position j, and $\vartheta_{ij}^k = 0$ otherwise. This formulation enforces a one-to-one assignment between the current AUV positions and the target configuration.

Based on the above definitions, the dynamic optimization problem is to determine a sequence of network configurations $\{\mathcal{S}_k\}_{k=0}^{K}$ that maximizes the long-term sensing performance while accounting for communication connectivity requirements and reconfiguration cost over time. A general form of this problem can be written as:

$$\max_{\{\mathcal{S}_k\}_{k=0}^{K}} \sum_{k=0}^{K} [J_d(\mathcal{S}_k) - J_{\text{conn}}(\mathcal{S}_k)] - \sum_{k=1}^{K} J_{\text{reconf}}(\mathcal{S}_{k-1}, \mathcal{S}_k) \tag{2}$$

where $J_{\text{reconf}}(\mathcal{S}_{k-1}, \mathcal{S}_k)$ denotes the minimum transition cost between two consecutive configurations, as determined by the assignment problem in Equation (1). This formulation explicitly captures the sequential and time-varying nature of the distributed AUV network optimization problem. Configuration decisions must be made online under evolving environmental and operational conditions, which naturally motivates the deep reinforcement learning-based framework introduced in Section 2.2.

2.2. Framework Overview

To address the dynamic, uncertain, and sequential nature of the AUV placement and reconfiguration problem, we propose an integrated framework that combines environment-aware sensing evaluation, DRL-based configuration optimization, and flow-aware path planning into a unified closed-loop architecture. The framework is designed to enable the adaptive, scalable, and computationally efficient operation of distributed underwater robotic sonar networks under time-varying environmental and operational conditions, while explicitly accounting for uncertainty in ocean flow prediction.

At each decision epoch, the framework first incorporates updated environmental information, including acoustic propagation characteristics, ambient noise levels, and ocean current fields, together with revised target prior distributions. These inputs are derived from a combination of numerical ocean models, acoustic propagation solvers, and accumulated observations. While numerical ocean circulation models provide large-scale flow predictions, it is well recognized that such predictions are subject to uncertainty and

may not fully capture local or rapidly evolving flow structures experienced by individual AUVs. Consequently, the framework is designed to tolerate imperfect environmental predictions and to leverage online adaptation during execution.

Based on the updated situational awareness, a PPO-based DRL agent determines the desired network configuration for the current epoch. The PPO agent operates on the discretized surveillance grid and outputs a target sensor layout that balances multiple objectives, including maximizing the expected detection performance and maintaining reliable acoustic communication connectivity. By learning a policy over configurations rather than solving a static optimization problem at each epoch, the framework enables fast decision-making and supports frequent re-optimization as conditions evolve.

Once a target configuration is selected, the framework proceeds to the reconfiguration execution stage, which tightly couples AUV-to-position assignment with flow-aware path planning. For each AUV–target location pair, a candidate transition cost is estimated based on the predicted travel time expenditure required to move between the two locations under the available ocean flow forecasts. These costs are obtained using a PPO-LSTM-based path planning module operating on the predicted flow field, and they populate the elements of the assignment cost matrix. In this way, the assignment problem reflects flow-aware motion cost rather than simple geometric distance. The resulting Linear Assignment Problem is solved using a zero-element method, yielding an optimal matching between AUVs and target locations that minimizes the estimated reconfiguration cost.

Following assignment, each AUV executes the reconfiguration using a PPO–LSTM-based path planning policy that has been trained to operate under partial observability and flow uncertainty. Rather than relying solely on precomputed paths, the learned policy itself can be deployed onboard the AUV. During execution, the AUV continuously incorporates locally sensed flow information obtained from onboard sensors to generate control actions in real time. This online policy execution allows the vehicle to adapt its trajectory to local flow variations that may deviate from the predicted ocean model, resulting in more accurate, robust, and efficient motion.

Overall, the proposed framework operates as a closed-loop process consisting of the following steps:

1: Update environmental information and target priors using ocean-model predictions, acoustic propagation analysis, and accumulated observations.
2: Determine the desired sensor-network configuration on the surveillance grid using the PPO-based placement policy.
3: Estimate flow-aware transition costs for all AUV–target location pairs using the PPO + LSTM-based path planning module.
4: Construct the assignment cost matrix and solve the resulting one-to-one assignment problem to obtain the optimal AUV-to-position matching.
5: Execute the reconfiguration online using the learned path planning policy and locally sensed flow information.
6: Update environmental estimates and target priors as new observations become available and then repeat the optimization cycle at the next decision epoch.

All components operate in a closed loop across decision epochs. As AUVs execute the planned motions and new observations become available, the environmental models and target priors are updated, and the optimization cycle repeats. Although the framework is developed for passive sonar surveillance, acoustic communication connectivity is explicitly maintained to support the seamless transition to cooperative multistatic tracking once a potential target is detected. In this way, the proposed framework provides a general and extensible foundation for adaptive underwater robotic sonar networks capable of operating across multiple sensing modes and mission phases.

3. PPO-Based Optimal Sonar Placement

3.1. Optimal Sonar Placement as a Markov Decision Process

The optimal placement of sonar-equipped AUVs over a discretized surveillance region is a combinatorial optimization problem defined on a finite spatial grid. The objective is to determine the locations of G AUVs on distinct grid cells such that the resulting network configuration maximizes the detection performance while maintaining reliable inter-AUV acoustic communication.

Recent studies have explored RL for sensor placement optimization. In some RL-based formulations, the entire sensor configuration is treated as the state, and actions modify this configuration through operations such as sensor relocation, swapping, or permutation operations [23]. Such formulations typically involve high-dimensional state and action spaces, which can severely limit scalability as the number of sensors and candidate locations increases. To address these challenges, this work adopts a constructive RL formulation for sonar placement. Instead of directly operating on complete layouts, the network configuration is constructed incrementally by placing one AUV at each decision step. This strategy reduces action space complexity, naturally enforces placement constraints (e.g., one sensor per location), and enables the learning agent to reason about spatial interactions as the configuration evolves. Moreover, this formulation aligns with classical incremental placement strategies while leveraging DRL to optimize complex, non-analytical objectives arising from environment-dependent acoustic propagation and communication constraints.

Specifically, the placement problem is modeled as a finite-horizon MDP with horizon length G, corresponding to the sequential placement of one AUV at each decision step. The MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where the components are specified as follows.

State space: At placement step $t \in \{0, 1, \ldots, G\}$, the state s_t is represented by a binary occupancy vector $\mathbf{z}_t = [z_1, z_2, \ldots, z_N]^\top \in \{0,1\}^N$, where $z_n = 1$ indicates that grid cell n is already occupied by an AUV, and $z_n = 0$ otherwise. Initially, $\mathbf{z}_0 = 0$, corresponding to an empty configuration. As the placement proceeds, the vector is updated to reflect the growing sensor layout. Optional auxiliary inputs, such as target prior distributions and environmental features, may be incorporated into the state representation; their integration is left for our future work.

Action space: At each step t, the agent selects an action a_t corresponding to one unoccupied grid cell. Although the nominal action space has cardinality N, a dynamic action mask is applied to exclude previously occupied cells, ensuring that only valid placements are selectable. This enforces the one-sensor-per-cell constraint.

State transition: The transition function is deterministic. After selecting action a_t, the next state is given by $z_{t+1}(a_t) = 1, z_{t+1}(n) = z_t(n), \forall n \neq a_t$, reflecting the placement of a new AUV at the selected location.

Reward and discount factor: The reward is defined only at the terminal step $t = G$ once all AUVs have been placed. Intermediate rewards are set to zero. Accordingly, the discount factor is set to $\gamma = 1$. The terminal reward evaluates the utility of the complete configuration, as detailed in Section 3.2.

3.2. Reward Function Design

The reward function is designed to steer the PPO agent toward configurations that are simultaneously acoustically effective for passive detection and operationally viable as a cooperative network. Accordingly, the terminal reward assigned at the final placement step $t = G$ consists of two components: (i) a detection performance term that measures the expected probability of detecting targets over the surveillance region under the current acoustic environment and target prior, and (ii) a network connectivity term that enforces

the feasibility of acoustic communication required for information sharing and cooperative behaviors. By integrating these terms, the learned policy is encouraged to deploy AUVs into high-value acoustic regions while maintaining a communication topology that supports coordinated sensing and, when necessary, a transition to communication-supported multistatic tracking.

For detection performance, let $P_T^k(i,j) \geq 0$ denote the target presence prior over grid cell (i,j) at epoch k. Given the configuration $\mathcal{S}_k = \left\{\mathbf{s}_1^k, \ldots, \mathbf{s}_G^k\right\}$, the acoustic performance of AUV g for a hypothetical target at location $\mathbf{r}_{i,j}$ could be characterized through the signal-to-noise ratio, SNR (in dB), which could be calculated using the standard passive-sonar equation form:

$$SNR_g^k(i,j) = SL - TL\left(\mathbf{r}_{i,j}, \mathbf{s}_g^k\right) - NL + DI \tag{3}$$

where SL is the assumed target source level, $TL(\cdot)$ is the transmission loss between $\mathbf{r}_{i,j}$ and $\mathbf{s}_g^k$ estimated with the acoustic propagation model, NL is the ambient noise level, and DI is the directivity index (constant under the omnidirectional array assumption). In this work, transmission loss $L(\cdot)$ is computed using the BELLHOP ray-tracing acoustic propagation model, which accounts for range-dependent sound-speed profiles, boundary interactions, and environmental variability. This allows for the detection utility to explicitly reflect a realistic, environment-aware acoustic performance.

The probability that AUV g detects a target at cell (i,j) is then modeled as a function of SNR. As an example, we adopt the following model [33]:

$$P_{d,g}^k(i,j) = \mathrm{Q}\left(\mathrm{Q}^{-1}(P_{FA}) - \sqrt{2 \times 10^{SNR_g^k(i,j)/10}}\right) \tag{4}$$

where P_{FA} is the prescribed false-alarm probability, Q denotes the complementary cumulative distribution function (right-tail probability), and $Q^{-1}(P_{FA})$ is its inverse. It is emphasized that alternative detection models, such as empirically derived ROC-based mappings, or learned detection likelihoods may be substituted depending on sensor characteristics and mission requirements.

Assuming conditional independence across AUV detections, the network-level detection probability at (i,j) is:

$$P_{\mathrm{net}}^k(i,j) = 1 - \prod_{g=1}^{G}\left(1 - P_{d,g}^k(i,j)\right) \tag{5}$$

Finally, the overall detection utility is computed as the target prior-weighted spatial expectation:

$$J_d(\mathcal{S}_k) = \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P_T^k(i,j) P_{\mathrm{net}}^k(i,j) \tag{6}$$

Reliable cooperative operation does not require all AUV pairs to maintain strong direct acoustic links; multi-hop communication via intermediate vehicles is often sufficient. Enforcing uniformly strong pairwise links can therefore be overly restrictive and may unnecessarily limit spatial coverage. To capture network feasibility in a principled and physically meaningful manner, we adopt a graph-level connectivity formulation grounded in environment-aware acoustic modeling.

For the ordered AUV pair (p,q), the communication SNR at decision epoch k is modeled as:

$$\mathrm{SNR}_{pq}^k = SL_c - TL\left(\mathbf{s}_p^k, \mathbf{s}_q^k\right) - NL_c \tag{7}$$

where SL_c and NL_c denote the source level and effective noise level used for communications (not necessarily identical to the passive detection parameters), and $TL\left(\mathbf{s}_p^k, \mathbf{s}_q^k\right)$ is again obtained from the BELLHOP propagation model. As an example of physical-layer

mapping, we assume coherent binary phase shift keying (BPSK) transmission over an additive white Gaussian noise (AWGN) channel, giving an approximate bit error rate [34]:

$$BER_{pq}^{k} = Q\left(\sqrt{2 \cdot 10^{SNR_{pq}^{k}/10}}\right) \tag{8}$$

where $Q(\cdot) = \frac{1}{2} erfc(\frac{\cdot}{\sqrt{2}})$ is the complementary cumulative distribution function of the standard normal distribution. For a packet of length L bits, the corresponding packet success probability is:

$$p_{pq}^{k} = (1 - BER_{pq}^{k})^{L}. \tag{9}$$

As with the detection model, it is emphasized that sophisticated underwater acoustic communication models incorporating multipath, Doppler spread, or learned link quality mappings can be readily integrated into the proposed framework.

Based on the link-level communication reliability, we construct a weighted, undirected communication graph $\mathcal{G}_k = (\mathcal{V}, \mathcal{E})$, where each vertex in $\mathcal{V}$ corresponds to an AUV. To ensure conservative and bidirectional reliability, the weight of the edge between AUVs p and q is defined as $w_{pq}^{k} = \min\left(p_{pq}^{k}, p_{qp}^{k}\right)$. Let $\boldsymbol{W}^{k} = \left[w_{pq}^{k}\right]$ denote the resulting symmetric adjacency matrix, and let $\boldsymbol{D}^{k}$ be the corresponding degree matrix with diagonal entries $d_{p}^{k} = \sum_{q} w_{pq}^{k}$. The weighted graph Laplacian is then given by $L^{k} = \boldsymbol{D}^{k} - \boldsymbol{W}^{k}$. The algebraic connectivity, defined as the second smallest eigenvalue $\lambda_2\left(\boldsymbol{L}^{k}\right)$, provides a global measure of network robustness. A larger λ_2 implies stronger overall connectivity, guaranteed multi-hop reachability, and increased resilience to individual link degradation, making it well suited for characterizing cooperative feasibility in distributed AUV networks.

Rather than treating communication quality as a soft preference, we regard connectivity as an operational requirement that must be satisfied to enable cooperative sensing and information exchange. Accordingly, we adopt a constrained (Lagrangian) reward formulation. Let $\tau > 0$ denote the minimum acceptable algebraic connectivity. The terminal reward for a completed placement episode is defined as:

$$R(\mathcal{S}_k) = J_d(\mathcal{S}_k) - \lambda_k\left[\tau - \lambda_2\left(\boldsymbol{L}^{k}\right)\right]_{+} \tag{10}$$

where $[x]_{+} = \max(x, 0)$ penalizes only violations of the communication requirement, and $\lambda_k > 0$ is a Lagrange multiplier.

The multiplier λ_k is adjusted using a dual-ascent update, e.g.,

$$\lambda_{k+1} = [\lambda_k + \eta(\tau - \lambda_2(\boldsymbol{L}^{k}))]_{+} \tag{11}$$

with step size $\eta > 0$. This adaptive mechanism automatically tightens the connectivity constraint when learned configurations risk becoming weakly connected and relaxes it once the requirement is satisfied, avoiding brittle hard constraints and excessive manual tuning.

3.3. Constructive Sensor Placement via PPO

Given the above MDP formulation and reward definition, PPO is employed to learn the placement policy. PPO is adopted due to its stable training behavior, robustness to noisy reward signals, and suitability for constrained discrete action problems.

The placement policy is realized using an actor–critic architecture.

Actor: At step t, the actor outputs a masked categorical distribution over the remaining unoccupied grid cells. The action is sampled from this distribution and corresponds to placing a sensor at a valid location. To guarantee valid configurations, a binary mask is applied at each step to exclude already occupied cells. Specifically, logits corresponding to

selected cells are assigned a large negative value (e.g., -10^6) before the softmax, ensuring zero probability of reselection. This mechanism enforces the one-sensor-per-cell constraint throughout the placement process.

Critic: The critic estimates the expected return of the current partially constructed configuration, providing feedback to guide the actor's learning.

Policy optimization follows the PPO clipped surrogate objective:

$$\mathcal{L}_{\mathrm{PPO}}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)] \tag{12}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$ is the importance sampling ratio, A_t is the advantage estimate, and ϵ is the clipping threshold.

The overall PPO-based constructive placement procedure is summarized in Algorithm 1.

Algorithm 1 PPO-Based Sonar Placement Algorithm

The full PPO training procedure unfolds as follows:

1. Initialization:

Initialize policy network parameters θ and value network parameters ϕ.

Set the number of grid cells N, number of sensors G, target prior P_T, connectivity threshold τ, initial value of the Lagrange multiplier λ_0, Generalized Advantage Estimation (GAE) parameters, and optional environmental features $\mathbf{f}$.

Specify PPO hyperparameters: learning rate, discount factor $\gamma = 1$, clipping threshold ϵ, dual step size η, etc.

2. Episode Rollout:

Begin with an empty placement vector $\mathbf{z}_0 \in \{0,1\}^N$.

For each placement step $t = 0, \ldots, G-1$:

Construct the current state s_t from $\mathbf{z}_t$, optionally including P_T and $\mathbf{f}$.

Compute logits $\mathbf{l}_t = f_\theta(s_t)$ and apply the action mask.

Obtain placement probabilities $\pi_\theta(a_t \mid s_t)$ via softmax.

Sample a valid action a_t, update the placement vector $\mathbf{z}_{t+1}$.

Store (s_t, a_t) in the trajectory buffer $\mathcal{D}$.

3. Reward Evaluation:

Once G sensors have been placed, derive the final layout $\mathcal{S}_k = \mathcal{S}(\mathbf{z}_G)$.

Compute detection performance $J_d(\mathcal{S}_k)$ and algebraic connectivity $\lambda_2\left(\boldsymbol{L}^k\right)$.

Assign terminal reward: $r_G = J_d(\mathcal{S}_k) - \lambda_k\left[\tau - \lambda_2\left(\boldsymbol{L}^k\right)\right]_+$, and set $r_t = 0$ for all $t < G$.

4. Policy Update:

Compute return estimates R_t and advantages A_t, e.g., using GAE.

For multiple epochs, sample batches from $\mathcal{D}$ and:

Update policy by maximizing $\mathcal{L}_{\mathrm{PPO}}(\theta)$,

Update critic by minimizing $\mathcal{L}_V(\phi) = \mathbb{E}_t\left[\left(R_t - V_\phi(s_t)\right)^2\right]$.

5. Dual Variable Update:

Update the Lagrange multiplier using dual ascent $\lambda_{k+1} = [\lambda_k + \eta(\tau - \lambda_2(\boldsymbol{L}^k))]_+$

6. Repeat:

Repeat steps 2–5 until convergence.

Although the proposed placement objective is not strictly submodular due to nonlinear acoustic propagation effects and connectivity constraints, the constructive PPO formulation exhibits a behavior analogous to greedy marginal-gain placement. In particular, the learned policy tends to prioritize globally informative locations in early decision steps, while later placements provide diminishing refinement. As a result, truncating the placement sequence yields high-quality configurations for smaller sensor counts, even when the

policy is trained for a larger maximum fleet size. This prefix robustness mirrors a well-known property of greedy submodular optimization but is here achieved without requiring analytical submodularity or handcrafted marginal utility models and emerges naturally from data-driven policy learning in complex environments.

4. Optimal Assignment and Path Planning Under Dynamic Ocean Currents

4.1. Strategy Overview

In dynamic multi-AUV sonar network operations, sensor layouts must be periodically reconfigured to respond to evolving mission objectives, target priors, and environmental conditions. Given the current AUV configuration at decision epoch k, $\mathcal{S}_k = \left\{\mathbf{s}_1^k, \ldots, \mathbf{s}_G^k\right\}$, and a newly optimized target layout obtained from the placement module $\mathcal{S}_{k+1}^* = \left\{\mathbf{s}_1^{k+1}, \ldots, \mathbf{s}_G^{k+1}\right\}$, the network must determine how to reassign individual AUVs to their new target positions in a manner that is both efficient and feasible under ocean current influences.

As formulated in Section 2.1, this reconfiguration process is modeled as a one-to-one assignment problem between the current AUV positions and the target positions in $\mathcal{S}_{k+1}^*$. Let c_{ij}^k denote the flow-aware motion cost for AUV i to move from its current position $\mathbf{s}_i^k$ to target position $\mathbf{s}_j^{k+1}$, and let $\mathbf{C}^k = \left[c_{ij}^k\right] \in \mathbb{R}^{G\times G}$ denote the corresponding cost matrix. The key challenge is therefore to construct $\mathbf{C}^k$ efficiently and then solve the resulting assignment problem with low computational cost.

Unlike conventional assignment formulations that rely on Euclidean distance, the transition cost here is defined as the minimum travel time under ocean currents, $c_{ij}^k = T_{\min}\left(\mathbf{s}_i^k \to \mathbf{s}_j^{k+1}\right)$, which more accurately reflects the energetic and temporal constraints of underwater navigation. However, directly computing optimal travel-time trajectories for all G^2 AUV–destination pairs using classical planners is computationally prohibitive, especially under time-varying flows.

To address this challenge, we adopt a two-stage strategy. First, the transition costs in $\mathbf{C}^k$ are estimated rapidly using a learned PPO + LSTM-based trajectory planner that accounts for ocean current effects and partial observability. Second, once the flow-aware cost matrix has been constructed, the resulting one-to-one assignment problem is solved efficiently using the zero-element assignment method. After the assignment is obtained, each AUV executes the corresponding reconfiguration trajectory using the learned planning policy and locally sensed flow information.

This decoupled yet tightly integrated design enables a scalable and current-aware reconfiguration suitable for real-time multi-AUV operations. In the following subsections, Section 4.2 describes the zero-element assignment method, and Section 4.3 presents the PPO + LSTM-based trajectory planner used for flow-aware motion cost estimation and online path execution.

4.2. Optimal Assignment via the Zero-Element Assignment Method

Once the flow-aware cost matrix $\mathbf{C}^k = \left[c_{ij}^k\right] \in \mathbb{R}^{G\times G}$ is constructed, the AUV reassignment problem becomes a classical Linear Assignment Problem (LAP). While the Hungarian algorithm provides an exact solution, its $O\left(G^3\right)$ complexity can be limiting for large fleets or frequent reconfiguration cycles. To improve computational efficiency, we adopt a zero-element method proposed in our previous work [35,36].

First, the zero-element method begins by preprocessing the cost matrix $\mathbf{C}$ to rapidly increase the number of zeros. This preprocessing subtracts the minimum element of each column from that column. Then, the minimum element of each row is subtracted from that

row, accumulating the total subtracted amount into d_{sum}. The process is repeated until d_{sum} no longer increases. Second, the preprocessed matrix $\mathbf{C}$ is then searched for G-independent zeros (one per row and column) using a space-tree search with depth-first backtracking. To reduce search uncertainty, a root node is chosen from rows or columns with the fewest zeros. Cofactors are then formed by recursively deleting the corresponding row and column until a complete assignment is obtained. Third, if no complete zero matching is found after preprocessing, a postprocessing stage is invoked to alter the zero distribution via selective row/column operations. These operations are based on a reference constant c_{ref}, defined as the average of nonzero entries. The algorithm then returns to the space-tree search. Once G-independent zeros are identified, they define the optimal mapping $\mathbf{s}_i^k \rightarrow \mathbf{s}_{\pi_k^*(i)}^{k+1}$, $i = 1, \ldots, G$.

Compared with generic LAP solvers, this method offers a superior solving efficiency, which is ideal for the rapid, large-scale reconfiguration required in underwater robotic networks.

4.3. PPO + LSTM-Based Path Planning Under Partial Observability

Efficient reconfiguration of a distributed AUV network requires repeated evaluation of motion costs between all current AUV positions and candidate target locations. In a network of G vehicles, this entails computing up to $G \times G$ point-to-point trajectories at each reconfiguration epoch. Classical path planning methods, such as graph-search approaches (e.g., A* and Dijkstra), sampling-based planners (e.g., RRT and PRM), evolutionary algorithms, level-set methods [37], and optimal control-based solvers, are well studied [38]. However, their computational cost becomes prohibitive when invoked repeatedly at large scale. As a result, directly applying these methods to construct the full reconfiguration cost matrix in real time is challenging for large AUV fleets.

To address this challenge, we adopt a learning-based path planning approach that amortizes the planning cost through offline training and enables fast inference at runtime. Once trained, a neural planner can generate near-optimal trajectories and corresponding travel-time estimates significantly faster than traditional planners. Moreover, the learned policy can be deployed onboard each AUV, enabling online adaptation to local environmental conditions that may deviate from predicted ocean models. The overall network architecture of the PPO + LSTM path planning module is illustrated in the right sub-figure of Figure 1.

Each AUV path planning task is formulated as an MDP, in which the vehicle navigates from its current position to a designated target location under the influence of ocean currents. Unlike existing PPO + LSTM-based path planning approaches that assume access to a complete global flow field [39], this work adopts a different MDP formulation and operational setting. Specifically, we consider underwater AUV navigation under partial observability, where only locally sensed flow information is available to each vehicle, and the learned policy is designed to operate both as a fast cost estimator during network-level reconfiguration and as an onboard controller for real-time trajectory execution in dynamic ocean environments.

The state at time step t is defined as $\mathbf{s}_t^{\text{traj}} = [x_t, y_t, u_t, v_t, \psi_t, g_x, g_y, t]$, where (x_t, y_t) denotes the AUV position, (u_t, v_t) represents the locally measured ocean current velocity, ψ_t is the vehicle heading, and (g_x, g_y) specifies the target location. By relying on local flow measurements, the policy can respond directly to the true environmental forcing experienced by the vehicle rather than solely depending on potentially inaccurate forecasts.

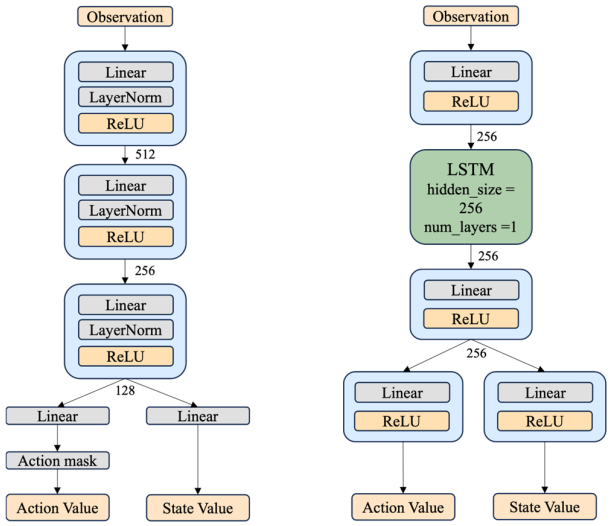

Figure 1. Illustration of the network architectures for optimal placement and path planning.

The action space consists of a discrete set of heading commands: $\mathcal{A}^{\text{traj}} = \{\psi_1, \psi_2, \ldots, \psi_K\}$, K uniformly spaced directions. This discretization provides a balance between control expressiveness and computational efficiency.

A hybrid reward function is designed to encourage both efficient exploration and precise goal convergence:

$$r_t = \begin{cases} \frac{\Delta d_t}{d_t + \epsilon} \xi_n - \lambda_1 \Delta t - \lambda_2 T_{\text{turn}}, & \xi_n > 1, \\ -\Delta t - T_{\text{turn}} - \lambda_3 |\psi_t|, & \xi_n \leq 1, \end{cases} \tag{13}$$

where Δd_t denotes progress toward the target, d_t is the remaining distance, ξ_n is a dynamic exploration factor, T_{turn} penalizes sharp heading changes, and ϵ avoids numerical singularities. This reward structure promotes exploration in the early stages of navigation and emphasizes smooth, time-efficient convergence as the AUV approaches the goal.

We employ PPO as the learning algorithm. To address the partial observability inherent in local flow-based navigation, the policy network is augmented with an LSTM module. The LSTM enables the policy to aggregate temporal information from sequential observations, implicitly infer unobserved flow patterns, and mitigate the effects of short-term sensing noise.

As shown in Figure 1, the network architecture consists of a feedforward feature extraction layer, followed by an LSTM layer and two output heads. The feedforward layer first transforms the instantaneous observation into a latent feature representation. The LSTM then processes the sequential feature stream and maintains a hidden state that captures historical context from local flow observations. Based on this temporally enriched representation, the actor head outputs action scores for the discrete heading commands, which are converted into a probability distribution for action selection, while the critic head estimates the state value. This actor–critic design supports stable learning and effective long-horizon decision-making under uncertainty.

Once trained, the PPO + LSTM policy serves as a neural path oracle. For each AUV–destination pair, the travel time is approximated as $c_{ij}^k \approx \sum_{t=0}^{T_{ij}} \Delta t$, where T_{ij} is the duration of the trajectory generated by the learned policy from AUV i to target location j.

Importantly, the learned policy itself can be downloaded onboard each AUV. During execution, the vehicle uses locally sensed flow information to perform online trajectory planning, continuously adapting its motion to actual environmental conditions. This allows

the realized trajectory to improve upon the initial plan derived from predicted flows, resulting in greater robustness and efficiency in practice.

5. Simulation Evaluation

This section presents a comprehensive evaluation of the proposed framework through a series of simulation studies. First, the simulation setup is introduced to establish realism and reproducibility. Next, the PPO-based constructive sonar placement method is evaluated independently. The PPO + LSTM-based path planning module is then examined separately to quantify its performance gains. Finally, an end-to-end dynamic reconfiguration demonstration is presented.

5.1. Simulation Setup

The simulation environment is designed to emulate a realistic, time-varying underwater operational scenario in which a distributed AUV-based passive sonar network performs adaptive placement and dynamic reconfiguration. All simulations were conducted on a laptop equipped with an Apple M4 Pro processor, and all algorithms were implemented in Python 3.9, with PyTorch 2.2.1 for RL components. For reproducibility, the main simulation settings, including surveillance-region discretization, environmental input data, acoustic and communication parameters, network architecture, and training-related configurations, are explicitly summarized in this subsection and in Table 1.

Table 1. Network architecture and training hyperparameters for PPO-based placement and PPO + LSTM path planning.

Category	Parameter	Value
Common PPO Settings	Discount factor (γ)	0.99 (placement), 0.995 (path planning)
	PPO clipping threshold (ϵ)	0.2
	Value loss coefficient	0.5
	Entropy regularization coefficient	0.01
	GAE parameter	0.95
	Optimizer	Adam
	PPO update epochs	10
	Mini-batch size	512
	Rollout length	4096 (placement), (256) (path planning)
Placement Network (PPO)	Input dimension	900 (binary occupancy vector)
	Shared hidden layers	2 fully connected layers
	Hidden layer width	256
	Activation function	ReLU
	Actor output	Masked categorical distribution over grid cells
	Critic output	Scalar state value
	Learning rate	3×10^{-4}
Path Planning Network (PPO + LSTM)	Input dimension	8 (position, local flow, heading, goal, time)
	Encoder hidden layers	1 fully connected layers
	Encoder width	128
	LSTM hidden size	256
	LSTM sequence length	256
	Actor output	Discrete heading set (K = 16)
	Critic output	Scalar state value
	Learning rate	3×10^{-4}
Training Strategy	Warm-up episodes (placement)	50
	Connectivity threshold (τ)	0.95
	Lagrange multiplier initial value	5
	Dual ascent step size (η)	0.5
Implementation	Framework	PyTorch 2.2.1
	Hardware	Apple M4 Pro (CPU)

The simulated surveillance region is modeled as a two-dimensional maritime area located in the South China Sea. The region spans a square area defined by latitudes and longitudes [118.1, 118.4] and is discretized into a uniform 30 × 30 grid of candidate deployment locations, with each grid cell having a length and width of 1.03 km. Each grid cell corresponds to a feasible AUV deployment position, and a strict one-node-per-cell constraint is enforced throughout all simulations. The network consists of $G = 40$ AUVs, representing a moderately large-scale robotic sonar network suitable for evaluating the proposed framework. The water depth in the simulated region varies up to 2000 m. And both the AUVs and the acoustic targets are assumed to operate at a constant depth of 200 m below the sea surface. All candidate deployment locations are restricted to grid cell centers within the surveillance region, and the deployment and reconfiguration processes are constrained to remain within this domain.

Environmental inputs, including temperature, salinity, and ocean current profiles, are obtained from the Copernicus Marine Environment Monitoring Service (CMEMS) over the period from 08:00 on 9 October 2025 to 08:00 on 19 October 2025. The CMEMS data provide a spatial resolution of 1/12° and a temporal resolution of 6 h, enabling the construction of a representative time-varying ocean environment. Based on the temperature and salinity fields, three-dimensional sound-speed profiles are computed using the Mackenzie empirical formula [40]. Figure 2 illustrates representative sound-speed fields at four different time instants, highlighting the spatiotemporal variability that directly influences underwater acoustic propagation.

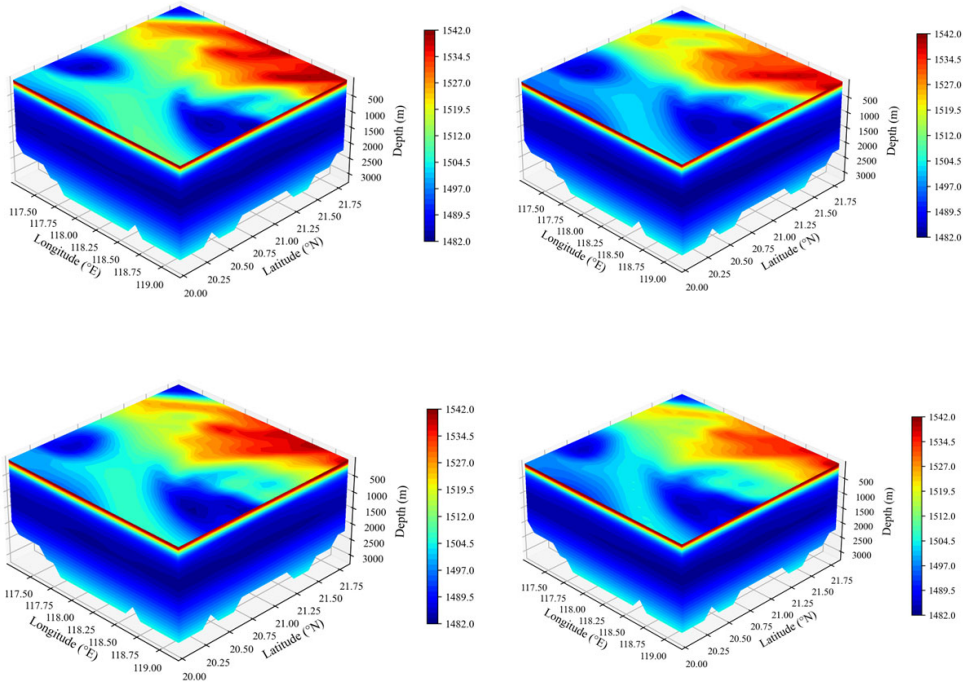

Figure 2. Illustration of the three-dimensional sound-speed field at 08:00 and 20:00 on 9 October and 8:00 and 20:00 on 10 October 2025.

For the calculation of the reward in the sonar placement, the transmission loss is computed using the BELLHOP ray-tracing acoustic propagation model, driven by the time-varying oceanographic inputs from the CMEMS and the bathymetric data from the ETOP2 global relief model. Based on the resulting TL fields, both detection probability and acoustic communication link reliability are computed using the corresponding models described in Section 3. The acoustic parameters employed in the sonar and communication equations are selected as hypothetical values, intended solely for algorithmic evaluation

rather than system-specific performance prediction. For passive detection, the source frequency is set to 1 kHz, with a source level of SL = 154 dB, ambient noise level of NL = 63 dB, directivity index DI = 0 dB, and false alarm $P_{fa} = 10^{-3}$. Under this parameterization, an individual passive sonar array is capable of detecting underwater targets within an effective range of approximately 3 km. The target prior distribution p_t is prescribed as a line-shaped high-probability region, emulating a likely transit corridor and serving as a representative scenario to spatially weight the detection performance. And additional scenarios with shifted or redistributed priors are considered to evaluate the adaptivity of the proposed method, as demonstrated in subsequent subsections. For acoustic communication modeling, the communication source level and the noise level are set to $SL_c = 145$ dB and $NL_c = 63$ dB, respectively, assuming omnidirectional transmission. And the minimum acceptable connectivity threshold is set as $\tau = 0.95$.

The PPO-based sonar placement policy is implemented using an actor–critic architecture with shared feature extraction layers, as shown in the left sub-figure in Figure 1. At each placement step, the actor outputs a masked categorical distribution over unoccupied grid cells, while the critic estimates the expected terminal reward of the partially constructed configuration. The PPO + LSTM path planning module is illustrated in the right sub-figure in Figure 1. Both networks are trained offline prior to deployment, and all hyperparameters—including learning rates, batch sizes, clipping thresholds, and LSTM hidden dimensions—are kept fixed across experiments to ensure fair and reproducible comparisons. For the PPO + LSTM planner, candidate start and goal locations are selected from feasible positions within the same surveillance region, and both training and testing are conducted under the same domain constraints. Training is performed on sampled start–goal navigation tasks under time-varying flow fields, while evaluation is conducted on separate test scenarios. To stabilize training of the PPO in the optimal placement, a warm-up strategy is adopted in which the connectivity constraint is temporarily disabled during the first 50 training episodes, allowing the placement policy to initially focus on learning effective sensing layouts. After this phase, the Lagrange multiplier associated with the connectivity constraint is activated and updated online using a dual-ascent rule. This approach prevents premature penalization during early exploration and improves convergence robustness. The detailed network architectures and training parameters used in the simulations are illustrated in Figure 1 and summarized in Table 1.

The parameters in Table 1 were selected from three main sources. First, the general PPO-related hyperparameters, such as the discount factor, clipping threshold, value-loss coefficient, entropy regularization coefficient, GAE parameter, optimizer, and PPO update epochs, were chosen according to commonly used settings in PPO-based reinforcement learning. Second, several architecture-related parameters were determined directly by the problem formulation, including the input and output dimensions of the placement and path planning networks, which follow from the corresponding state and action definitions. Third, the remaining parameters, such as the mini-batch size, rollout length, hidden dimensions, learning rate, and training strategy parameters, were selected through empirical tuning based on training stability, convergence behavior, optimization performance, and computational cost. Our tuning strategy was to begin with parameter combinations that produced stable training and a strong performance and then gradually adjust the most sensitive parameters to reduce the training cost while maintaining satisfactory results.

5.2. Evaluation of the PPO-Based Sonar Optimal Placement

This subsection presents a comprehensive evaluation of the proposed Sequential PPO placement strategy, focusing exclusively on the sensor placement problem and isolating it from path planning and execution effects. The evaluation proceeds in four stages:

(i) qualitative and quantitative assessment of the placement results under a realistic acoustic environment, (ii) comparison with an alternative RL formulation to analyze the benefit of the constructive decision structure, (iii) benchmarking against a classical heuristic method, and (iv) evaluation of generalization and partial-deployment capability.

We first examine the placement results produced by Sequential PPO under a representative oceanographic and acoustic environment. The results correspond to conditions at 08:00 on 9 October 2025), with sound-speed profiles derived from CMEMS data using the Mackenzie empirical function (as shown in Figure 2) and transmission loss computed using the BELLHOP ray-tracing model at a carrier frequency of 1 kHz. A line-shaped target prior distribution is prescribed to emulate a likely transit corridor. The total computation time for running Bellhop over the three-dimensional environment at multiple time instances is 32 min.

Figure 3 provides a comprehensive illustration of how the proposed Sequential PPO method exploits environment-dependent acoustic information to generate effective and well-connected sonar deployments. The three-dimensional sound-speed field shown in Figure 2 leads to heterogeneous acoustic propagation, as reflected in the transmission loss distribution shown in Figure 3a, for a source located at the region boundary center. The target prior distribution in Figure 3b defines the surveillance objective by assigning higher importance to specific regions of interest within this acoustically complex environment. Specifically, Figure 3b is generated by first defining a straight line to control the overall trend. Then, a Gaussian kernel is used to model the attenuation in the direction perpendicular to the line, while another Gaussian kernel is applied to model the attenuation along the projection direction on the line. For each point, these two factors are computed and then combined with an appropriate scaling factor to obtain the final field distribution. In this figure, the slope and intercept of the line are set to -2 and 13, respectively, and the kernel width of both Gaussian functions is set to 0.5. For comparison, the random placement result in Figure 3c is obtained under the same acoustic environment, target prior, and network-size setting as the optimized case. As shown in Figure 3c, random sensor placement produces a fragmented cumulative detection probability distribution that fails to exploit acoustic propagation or the target prior structure, yielding an average detection utility of only 14.6. In contrast, the optimized configuration produced by Sequential PPO in Figure 3d strategically places sensors in regions where favorable transmission conditions coincide with a high target likelihood, while simultaneously preserving a connected network backbone. This results in a substantially higher R of 35.98, corresponding to a relative improvement of 120.68%. At the same time, the algebraic connectivity is $\lambda_2 = 0.9$, which exceeds the prescribed connectivity threshold.

Importantly, the gain in detection performance is not achieved at the expense of network feasibility. The cumulative detection probability map shows that sensing improvements are concentrated in high-priority regions, while the connectivity metric confirms that multi-hop communication across the network remains reliable. This demonstrates that the learned policy internalizes the trade-off between sensing effectiveness and communication robustness, yielding configurations that are simultaneously acoustically efficient and operationally viable. Overall, these results indicate that Sequential PPO learns environment-aware placement principles that integrate acoustic propagation physics, mission priors, and graph-level connectivity constraints into a coherent and resilient deployment strategy.

Having established that Sequential PPO produces high-quality, environment-aware sonar deployments, we next examine why this formulation is more effective from a learning perspective. To this end, we compare Sequential PPO with an Exchange-based PPO baseline, in which the state encodes a complete binary deployment vector and each action swaps an occupied grid cell with an unoccupied one. This comparison isolates the effect of the

decision structure while keeping the learning algorithm (PPO), reward definition, and environmental conditions identical.

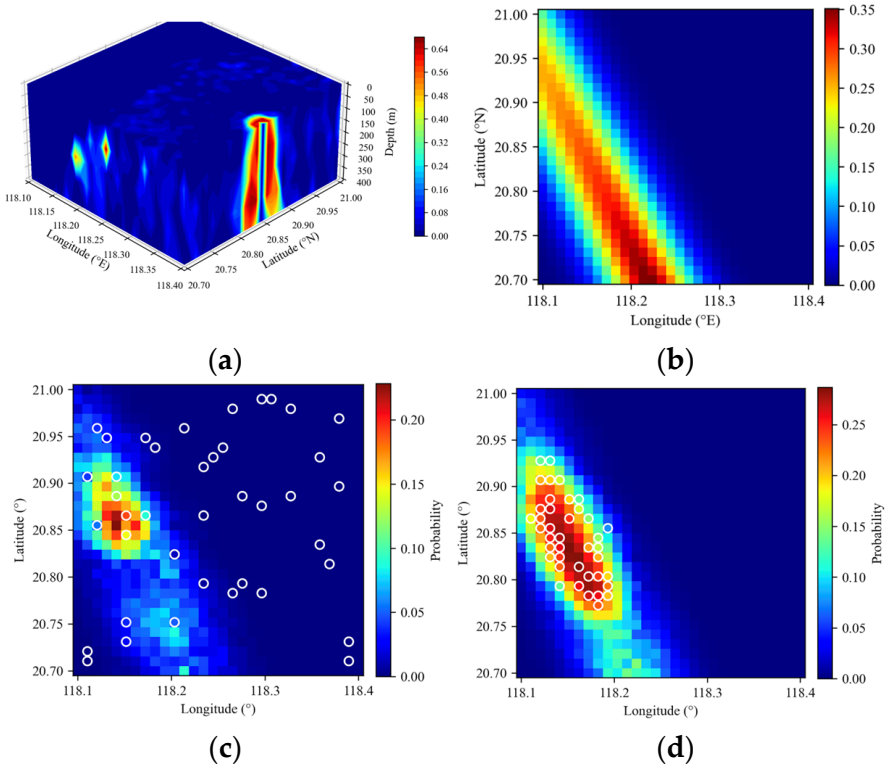

Figure 3. Environment-aware sonar placement under realistic acoustic conditions. (**a**) TL map computed using the BELLHOP model for a source located at a boundary center with depth of 200 m and frequency of 1 kHz. (**b**) Prescribed target prior distribution used to weight detection utility. (**c**) Cumulative detection probability map obtained under random sensor placement, where the white circles indicate the sensor locations. (**d**) Optimized sensor placement produced by the proposed Sequential PPO method and the resulting cumulative detection probability, where the white circles indicate the sensor locations.

Figure 4 presents the training reward curves of the two PPO formulations. Sequential PPO exhibits substantially faster reward improvement during early training and converges to a significantly higher final reward. Although both methods require a comparable wall-clock training time (approximately 250.0 s for Sequential PPO versus 240.9 s for Exchange PPO), the constructive formulation achieves markedly higher learning efficiency. Quantitatively, Sequential PPO attains a normalized final reward of 35.98, whereas Exchange PPO converges to 29.63, indicating a clear and consistent performance gap. This difference persists across multiple random seeds, suggesting that the observed advantage is structural rather than incidental.

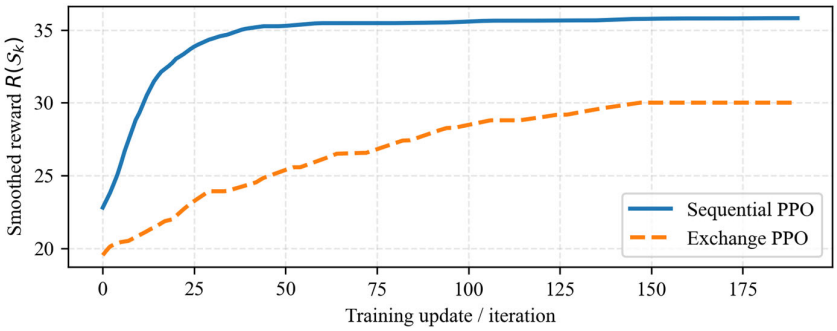

Figure 4. Training reward curves of Sequential PPO and Exchange PPO.

The underlying reason lies in the alignment between the decision structure and the combinatorial nature of the placement problem. Sequential PPO mirrors the incremental construction of a sensor set, allowing the policy to explicitly learn marginal detection gains and higher-order interactions among sensors as each new AUV is placed. This structure supports efficient credit assignment, since each action has a direct and interpretable contribution to the final configuration. In contrast, Exchange PPO operates through indirect local modifications of a high-dimensional binary vector, where the impact of a single swap on the terminal reward is often subtle and entangled with many other placement decisions, leading to noisier gradients and slower convergence. From an optimization standpoint, this comparison shows that the performance gains of Sequential PPO stem not merely from the RL itself but from a problem formulation that respects the inherent structure of sensor placement.

We further benchmark the proposed method against PSO and GA, which are widely used population-based heuristic approaches for static sensor placement problems. For fairness, all three methods are evaluated under the same acoustic environment, target prior distribution, connectivity threshold setting, and network size. PSO is implemented with 30 particles, 2000 iterations, an inertia weight of 0.9, a cognitive coefficient of 0.5, and a social coefficient of 0.3. Under fixed environmental conditions, the PSO has a relatively low computational cost, requiring 26.8 s per optimization run, and is capable of producing reasonable deployment solutions. However, despite its efficiency, PSO consistently underperforms Sequential PPO in final placement quality, particularly when detection performance and communication connectivity are jointly considered. GA is implemented with 500 generations, a mutation rate of 0.15, a crossover rate of 0.8, and an elitism ratio of 10%. Under the same acoustic environment and target prior, Sequential PPO achieves a final reward that is 16–21% higher than PSO and 2–3.7% higher than GA. These results indicate that the proposed method not only improves the solution quality relative to conventional population-based baselines but also better maintains performance when sensing and connectivity objectives must be optimized simultaneously.

Figure 5 further illustrates the sensitivity of the three methods to the connectivity threshold τ. As τ increases, Sequential PPO maintains a remarkably stable reward level, whereas PSO and GA show more noticeable performance degradation under stricter connectivity requirements. For example, when τ increases from 0.93 to 0.97, the reward of PSO decreases by 3.2%, that of GA decreases by 1.6%, whereas Sequential PPO experiences only a 0.02% degradation. This result suggests that the proposed method is more robust to increasingly strict connectivity constraints than the two heuristic baselines.

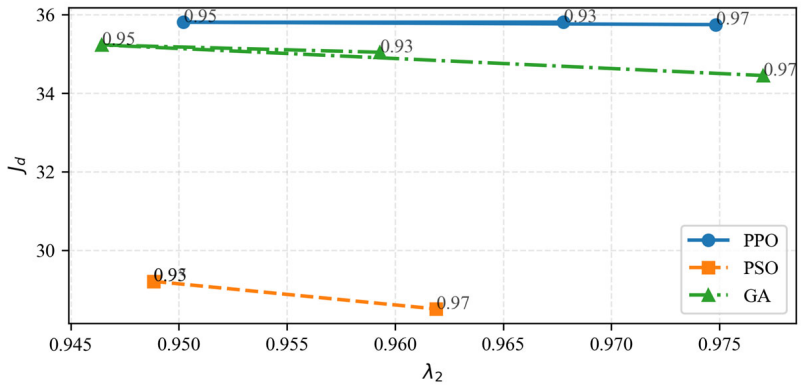

Figure 5. Final placement reward as a function of the connectivity threshold τ for Sequential PPO, PSO, and GA.

This behavior highlights a fundamental difference between the two approaches. PSO optimizes a static objective through stochastic search and tends to converge to solutions that only satisfy constraints marginally. As constraints tighten or environmental conditions vary, these solutions become fragile. In contrast, Sequential PPO internalizes the connectivity requirement during training, enabling it to learn deployment patterns that preserve a connected communication backbone while still allocating sensors to high-value sensing regions.

We further evaluate the generalization capability of the proposed Sequential PPO placement strategy in terms of scalability with respect to network size and robustness under partial deployment.

Because the Sequential PPO policy constructs the deployment as an ordered sequence of sensor placements, it naturally supports partial execution when fewer sensors are available. Figure 6 plots the detection utility achieved by Sequential PPO against the number of executed placement steps. The results demonstrate graceful performance degradation: deploying only 70% of the planned sensors retains approximately 85% of the full-deployment detection utility. This behavior indicates that the early placement decisions capture the most informative sensing locations, while subsequent placements primarily refine coverage and connectivity. Such behavior closely mirrors the diminishing-returns property of submodular optimization and provides a theoretical explanation for the robustness of the constructive formulation under resource constraints.

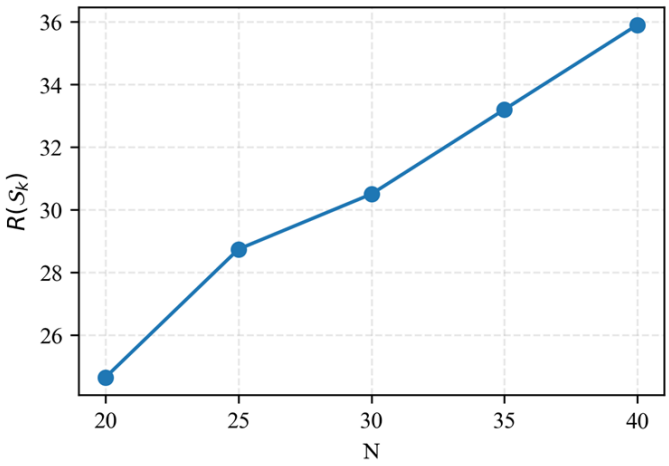

Figure 6. The detection utility achieved by Sequential PPO against the number of executed placement steps.

Building on this property, we next examine generalization across network scales by applying a placement policy trained with $G = 50$ AUVs directly to scenarios with fewer available vehicles (e.g., $G = 40$) without retraining. Figure 7 compares three cases: (i) a Sequential PPO policy trained with $G = 50$ sensors and executed using only the first 40 placement steps, (ii) a PSO policy trained and evaluated directly with $G = 40$ sensors, and (iii) an Exchange-based PPO policy trained and evaluated with $G = 40$ sensors. The results show that the policy trained with $G = 50$ sensors consistently achieves the highest reward when deployed with $G = 40$ sensors, outperforming both policies trained directly on the smaller scale. From an operational perspective, this result is significant: it implies that a single Sequential PPO policy trained for a larger fleet can be reliably reused across missions with varying numbers of available AUVs, enabling flexible adaptation to fleet attrition or resource constraints without requiring retraining.

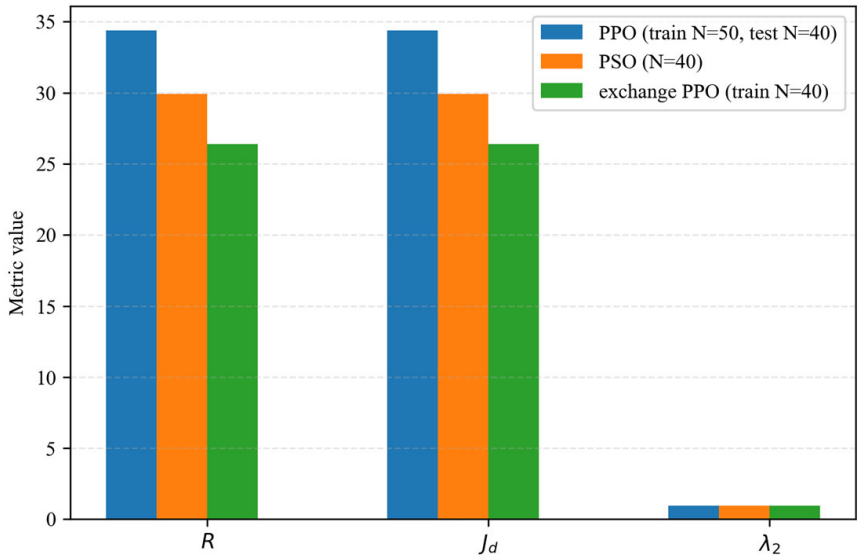

Figure 7. Comparison of deployment performance under reduced network size.

Collectively, these results demonstrate that the proposed Sequential PPO formulation offers substantial advantages over both alternative RL formulations and classical heuristics. By aligning the learning structure with the combinatorial nature of sensor placement, the method achieves faster convergence, higher solution quality, robust constraint satisfaction, and strong generalization.

5.3. Evaluation of the PPO + LSTM Method

This subsection evaluates the performance of the proposed PPO + LSTM approach for underwater path planning under time-varying ocean currents and clarifies its role as a computationally efficient enabler for large-scale network reconfiguration. In contrast to Section 5.2, which focuses on sensor placement, the analysis here isolates the path planning component, emphasizing its suitability for rapid transition cost evaluation during online AUV reassignment.

For the RL path planning task, 30 candidate start locations and 30 goal locations are predefined at feasible positions within the surveillance region. These locations are selected to provide diverse spatial coverage of the domain while respecting the same in-domain constraints used in the placement and reconfiguration experiments. At the beginning of each episode, one start–goal pair is randomly sampled from these sets, yielding a diverse task distribution that spans different spatial configurations and flow conditions. The PPO + LSTM agent receives only locally observed flow information along its trajectory and outputs discrete heading commands, consistent with partial observability constraints encountered in real underwater operations. Through interaction with the dynamic environment, the policy learns to exploit favorable currents while avoiding adverse flow regions, directly optimizing time-efficient navigation. During development, the convergence and generalization behavior of the planner are monitored on unseen navigation instances. Final performance evaluation is then conducted on separate test scenarios not used during training.

To quantitatively assess planning performance, the PPO + LSTM planner is compared against a classical grid-based baseline, A*, which is widely used in marine robotics. For fairness, the A baseline is evaluated using the same start–goal sets and the same domain constraints as the PPO + LSTM planner. In the implementation of the A*, a plan–execute evaluation protocol is adopted. Specifically, A* is first applied on a discretized grid to generate a geometric path between a start and a goal location. This path is then converted

into a sequence of waypoints and executed by an AUV kinematic model. By numerically integrating the vehicle motion, the corresponding travel time is obtained. This evaluation protocol reflects realistic deployment conditions in which high-level planning outputs must be translated into feasible trajectories through low-level control. In contrast to A*, the PPO + LSTM planner optimizes the path execution time directly.

Figure 8 presents a representative comparison between PPO + LSTM and A* for a selected start–goal pair after training, which is not used during training. Compared with the A*, the PPO + LSTM planner generates paths with more adaptive heading choices and smoother geometries in the presence of ocean currents, leading to shorter effective paths and significantly reduced travel times. In the illustrated example, the replayed A* trajectory requires approximately 97.5 h to reach the goal, whereas PPO + LSTM completes the same task in only 77.5 h. The AUV maintains a speed of 0.23 m/s relative to the still water. These improvements persist across multiple randomly sampled start–goal pairs. Averaged over the test set, which includes 10 members, A* yields a mean travel time of approximately 106.75 h, while PPO + LSTM achieves an average of about 74.25 h, corresponding to a reduction of 30.44%. All test cases involve comparable start–goal distances, confirming that the observed gains arise from the improved exploitation of flow dynamics rather than geometric bias. The performance advantages of PPO + LSTM stem from two fundamental factors. First, the learning agent interacts directly with a dynamical environment that explicitly incorporates vehicle kinematics and ocean current disturbances, allowing the policy to optimize time-related trajectory costs end to end. Second, the LSTM module provides temporal memory and hidden-state fusion of sequentially observed flow information. This enables anticipatory decision-making under partial observability and significantly enhances robustness in non-stationary environments.

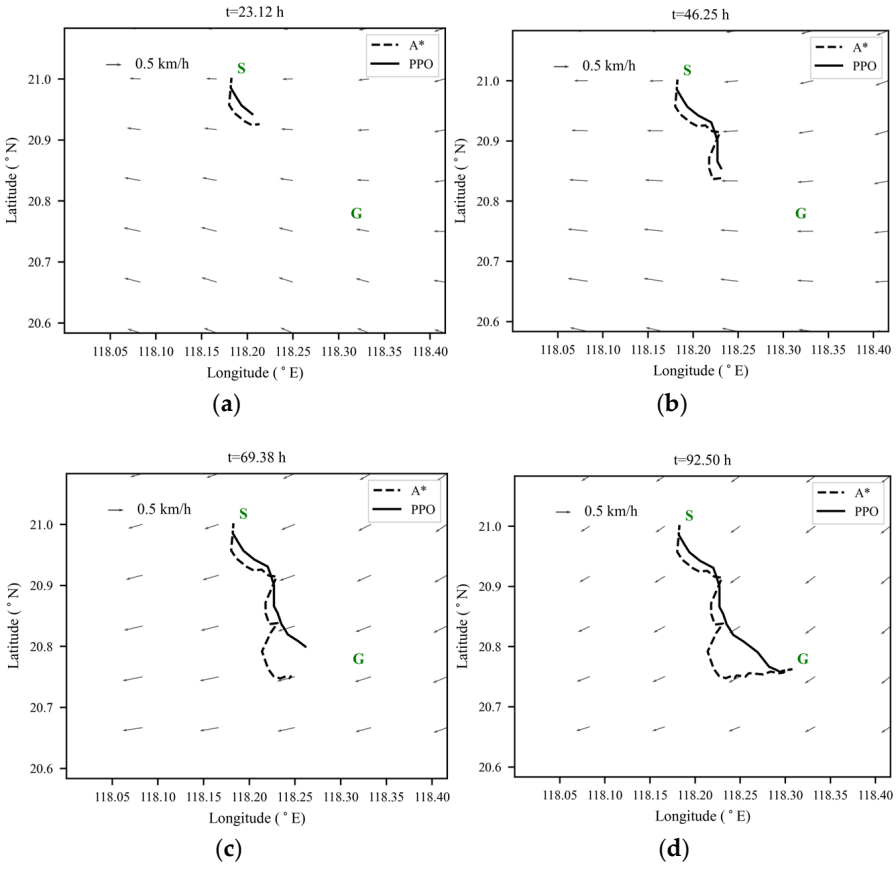

Figure 8. Comparison between PPO + LSTM and A* for path planning under time-varying ocean currents, where panes (**a**–**d**) correspond to flow field at 200 m depth of the surveillance region.

Beyond trajectory quality, an important advantage of the PPO + LSTM approach lies in its computational structure for large-scale reconfiguration. The PPO + LSTM planner is trained offline, requiring approximately 178 s for convergence on the evaluation platform. After training, trajectory generation is reduced to lightweight forward inference, enabling the rapid evaluation of motion costs between arbitrary start–goal pairs. This property is critical for network reconfiguration. Given a set of current AUV positions and a set of target deployment locations, the trained PPO + LSTM policy can be used to efficiently compute a $G \times G$ transition cost matrix, where each entry corresponds to the minimum travel time between an AUV and a candidate destination. For example, with $G = 40$, the cost matrix requires 1600 trajectory rollouts, which can be computed rapidly using policy inference.

Overall, the results demonstrate that PPO + LSTM is not only superior to classical planners such as A* in terms of trajectory execution efficiency under dynamic ocean currents but also serves as a scalable and computationally efficient backbone for large-scale, real-time network reconfiguration. By combining RL with recurrent temporal modeling, the proposed approach enables fast, flow-aware motion cost estimation and robust navigation under partial observability, providing a crucial link between optimal placement decisions and their physical realization in dynamic underwater environments.

5.4. Simulation Demonstration of the Entire Framework

This subsection presents an end-to-end evaluation of the proposed framework, demonstrating how environment-aware placement, flow-aware assignment, and learning-based path planning operate jointly to enable adaptive reconfiguration of a large-scale underwater robotic sonar network.

We consider a dynamic surveillance scenario in which a fleet of 40 AUVs enters the monitored region and repeatedly reconfigures its formation in response to evolving target priors and environmental conditions. The simulation spans four consecutive placement–reconfiguration cycles, each corresponding to one of the four time instants of the three-dimensional sound-speed fields introduced in Section 4.1. The decision interval is set to 6 h, matching the temporal resolution of the CMEMS oceanographic data and the sound-speed snapshots shown in Figure 2. Thus, each cycle uses one environmental snapshot and one corresponding target prior field, and the network is re-optimized once per environmental update.

At each cycle, the framework operates in a closed loop: the current acoustic environment and target prior (as shown in Figure 9 are first used to generate an environment-aware optimal placement; the resulting target configuration then triggers flow-aware reassignment via the assignment module and the PPO + LSTM planner, which evaluates transition costs under the corresponding predicted flow field for that cycle. All placements, start locations, and target locations are restricted to feasible positions within the same surveillance region defined in Section 5.1. The placement and path planning results for the four time steps are summarized in Figure 10.

The end-to-end results demonstrate that the proposed framework consistently maintains high detection performance while preserving network connectivity throughout the reconfiguration process. From a computational perspective, the combined execution time of optimal placement, flow-aware assignment at each reconfiguration step remains within 300 s (for the three cycles, the computation times are 239 s, 246 s and 252 s, respectively), indicating that the framework is capable of near-real-time operation at the network scale considered.

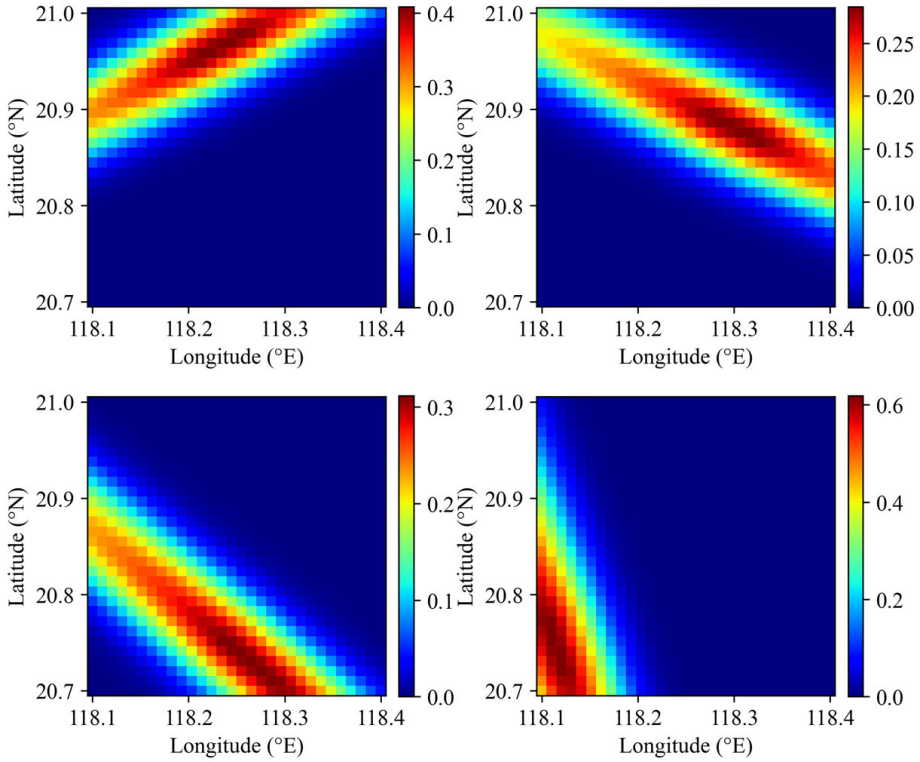

Figure 9. Target prior distributions.

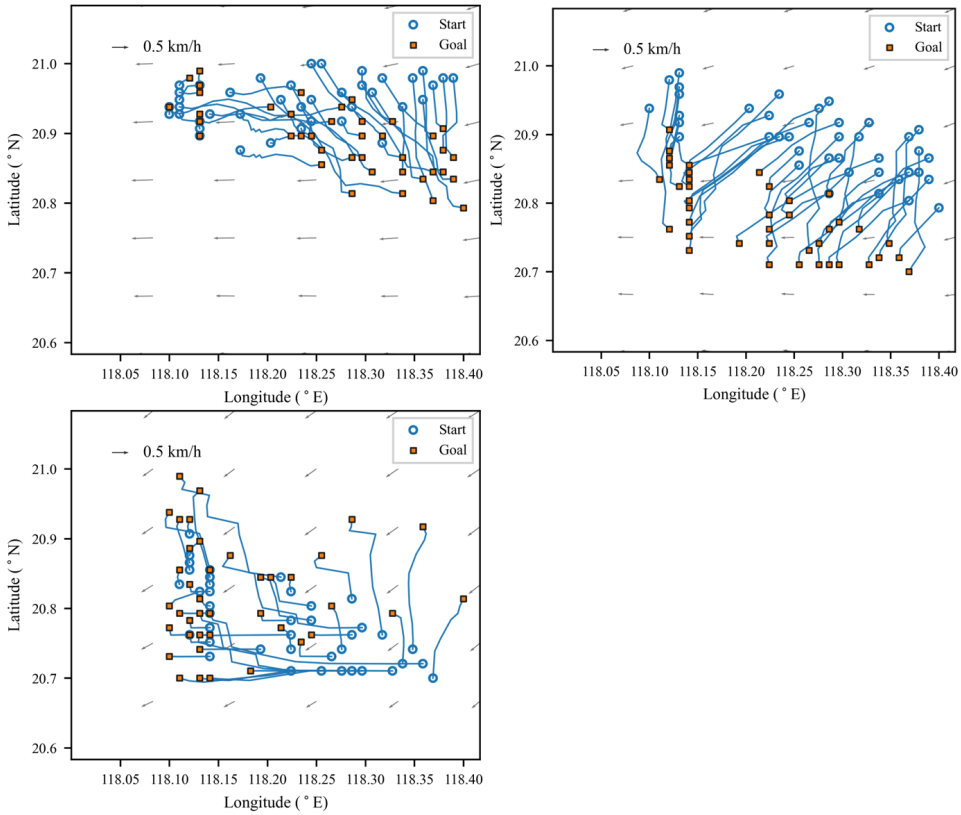

Figure 10. Optimal placement and corresponding path planning results at the three reconfiguration time steps.

6. Conclusions

This paper investigated the integrated problem of sonar placement optimization and dynamic reconfiguration in underwater robotic sonar networks operating in complex and time-varying ocean environments. Motivated by the limitations of static deployment strategies and computationally expensive replanning methods, we proposed a unified deep reinforcement learning-based framework that jointly addresses adaptive sensor placement, flow-aware reassignment, and path planning for large-scale AUV networks.

Specifically, the proposed framework formulates passive sonar placement as a constructive Markov decision process and employs PPO to learn placement policies that maximize detection probability while maintaining acoustic communication connectivity. By explicitly incorporating environment-aware acoustic propagation modeling and communication constraints into the reward design, the learned policies produce practically effective and operationally viable network configurations. To enable scalable reconfiguration, a flow-aware assignment strategy based on a zero-element method was introduced, leveraging learned travel-time estimates rather than simple geometric distances. Furthermore, a PPO + LSTM-based trajectory planner was developed to address navigation under dynamic ocean currents and partial observability, enabling both fast cost estimation for assignment and adaptive onboard execution using locally sensed flow information.

Despite these encouraging results, the present study still has several limitations. First, the proposed framework has only been validated in a simulation environment and has not yet been demonstrated on a specific real-world AUV platform. Although the simulated environment is constructed using CMEMS oceanographic data, bathymetric information, and BELLHOP-based acoustic propagation modeling, and therefore reflects realistic large-scale spatial and temporal environmental variability to a certain extent, it still differs from real-ocean operation in several important respects. In particular, ocean-model-based environmental predictions are subject to forecast errors, limited spatial and temporal resolution, and possible mismatch with local small-scale variability. Second, although acoustic propagation is modeled using a physics-based approach, the sonar-related parameters, detection model, and communication model adopted in this study are assumed for algorithmic evaluation and are not calibrated to a specific, real sonar system. As a result, the simulated sensing and communication performance cannot fully represent the actual performance of a real underwater robotic sonar network. Third, the proposed placement strategy is formulated on a discretized surveillance grid, and thus, the resulting deployment quality may depend on the chosen grid resolution. Continuous-space placement optimization is not explicitly addressed in the current framework. Fourth, the current framework does not yet explicitly include several platform-specific and operational factors, such as actuation uncertainty, localization error, hardware constraints, vehicle endurance, and long-range reachability. Therefore, the present results mainly validate the effectiveness of the proposed optimization and reconfiguration strategy at the algorithmic level rather than the absolute operational performance of a real deployed system.

Future work will focus on improving both the robustness and the practical deployment ability of adaptive underwater sonar networks under environmental uncertainty. One important direction is to incorporate ensemble-based environmental prediction and planning methods so that sensor placement and path planning decisions can explicitly account for multiple possible ocean realizations and thereby improve robustness to forecast uncertainty. Another promising direction is data-driven acoustic modeling, in which deep learning techniques are used to learn environment-specific acoustic performance directly from in situ measurements collected by the network, such as received-signal statistics, ambient noise levels, and communication quality indicators. In addition, future studies will consider system-calibrated sonar models, platform-specific vehicle constraints, and

real-ocean experiments on physical AUV platforms so as to assess the practical performance and deployment feasibility of the proposed framework under real operational conditions. Together, these developments would enable more resilient and self-improving underwater robotic sonar networks capable of long-term autonomous operation in highly dynamic marine environments.

Author Contributions: Conceptualization, Q.S. and Y.T.; methodology, Q.S., Y.T., J.Y. and F.Z.; software, Q.S. and J.Z.; validation, Q.S. and Y.T.; formal analysis, Q.S. and Y.T.; writing—original draft preparation, Y.T.; writing—review and editing, Y.T., J.Y. and F.Z.; visualization, Q.S., Y.X., Z.T. and Y.T.; supervision, Y.T., J.Y. and F.Z.; project administration, Y.T.; funding acquisition, Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China (Grant No. 52271353), the Program of the State Key Laboratory of Robotics and Intelligent Systems at Shenyang Institute of Automation, Chinese Academy of Sciences (Grant No. 2024-Z15, 2025-Z08), the Fundamental Research Program of Shenyang Institute of Automation, Chinese Academy of Sciences (Grant No. 2023JC3G03).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations and Nomenclature are used in this manuscript:

Symbol	Description
Abbreviation	
AUV	Autonomous Underwater Vehicle
DRL	Deep Reinforcement Learning
PPO	Proximal Policy Optimization
LSTM	Long Short-Term Memory
MDP	Markov Decision Process
RL	Reinforcement Learning
SNR	Signal-to-Noise Ratio
BELLHOP	Bellhop Acoustic Propagation Model
ROC	Receiver Operating Characteristic
BPSK	Binary Phase Shift Keying
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
LAP	Linear Assignment Problem
GAE	Generalized Advantage Estimation
PRM	Probabilistic Roadmap Method
RRT	Rapidly Exploring Random Tree
PPO-LSTM	PPO with LSTM
Nomenclature	
Ω	Two-dimensional surveillance region
L_x, L_y	Length and width of the surveillance region
N_x, N_y	Number of grid cells along the x- and y-directions
$\Delta x,\ \Delta y$	Grid resolutions along the x- and y-directions
ΔA	Area of each grid cell
$r_{i,j}$	Center of grid cell (i, j)
G	Number of AUVs in the network
$\mathcal{S}_k$	Network configuration at decision epoch k
s_g^k	Position of AUV g at epoch k
$\mathcal{S}_{k+1}^*$	Desired/target configuration for epoch $k+1$

$P_T^k(i,j)$	Target prior probability at grid cell (i,j) and epoch k
$J_d(\mathcal{S}_k)$	Detection utility of configuration $\mathcal{S}_k$
$J_{conn}(\mathcal{S}_k)$	Connectivity cost/penalty of configuration $\mathcal{S}_k$
$J_{reconf}(\mathcal{S}_{k-1}, \mathcal{S}_k)$	Reconfiguration cost between two consecutive configurations
c_{ij}^k	Flow-aware motion cost for assigning AUV i to target position j at epoch k
C^k	Assignment cost matrix at epoch k
ϑ_{ij}^k	Binary assignment variable
$P_{d,g}^k(i,j)$	Detection probability of AUV g for a target at cell (i,j)
$P_{\text{net}}^k(i,j)$	Network-level detection probability at grid cell (i,j)
$\text{SNR}_g^k(i,j)$	Detection SNR of AUV g for a target at grid cell (i,j)
SL	Source level in passive detection model
$TL(\cdot)$	Acoustic transmission loss
NL	Ambient noise level in passive detection model
DI	Directivity index
P_{FA}	False-alarm probability
$\mathcal{G}_k = (\mathcal{V}, \mathcal{E})$	Communication graph at epoch k
w_{pq}^k	Weight of the communication link between AUVs p and q
$\mathbf{W}^k$	Weighted adjacency matrix
$\mathbf{D}^k$	Degree matrix
$\mathbf{L}^k$	Weighted graph Laplacian matrix
$\lambda_2\left(\mathbf{L^k}\right)$	Algebraic connectivity
τ	Minimum acceptable algebraic connectivity threshold
λ_k	Lagrange multiplier for connectivity constraint
η	Dual ascent step size
$(\mathcal{S}, \mathcal{A}, P, R, \gamma)$	MDP tuple
s_t	Placement state at step t
$\mathbf{z}_t$	Binary occupancy vector at placement step t
a_t	Placement action at step t
r_t	Reward at step t
γ	Discount factor
$\pi_\theta(a_t \mid s_t)$	PPO policy
$V_\phi(s_t)$	Value function
A_t	Advantage estimate
R_t	Return estimate
ϵ	PPO clipping threshold
$\mathcal{L}_\mathcal{V}(\phi)$	Value loss
$\mathcal{D}$	Trajectory buffer
$\mathbf{s}_t^{\text{traj}}$	State of the trajectory-planning MDP at time step t
$\mathcal{A}^{\text{traj}}$	Discrete action space for trajectory planning
ψ_t	AUV heading at time step t
(u_t, v_t)	Locally measured ocean current velocity
(g_x, g_y)	Target location in trajectory planning
$T_{min}\left(\mathbf{s_i^k} \rightarrow \mathbf{s_j^{k+1}}\right)$	Minimum travel time from current to target position

References

1. Qiao, J.; Yu, J.; Huang, Y.; Cui, J.; Wang, B.; Wang, Z. Sea-Whale Series AUV-Extending the range to 4000 kilometers. In Proceedings of the OCEANS 2022, Hampton Roads, VA, USA, 17–20 October 2022.
2. Hobson, B.W.; Bellingham, J.G.; Kieft, B.; McEwen, R.; Godin, M.; Zhang, Y. Tethys-class long range AUVs-extending the endurance of propeller-driven cruising AUVs from days to weeks. In Proceedings of the 2012 IEEE/OES Autonomous Underwater Vehicles (AUV), Southampton, UK, 24–27 September 2012.
3. Furlong, M.E.; Paxton, D.; Stevenson, P.; Pebody, M.; McPhail, S.D.; Perrett, J. Autosub long range: A long range deep diving AUV for ocean monitoring. In Proceedings of the 2012 IEEE/OES Autonomous Underwater Vehicles (AUV), Southampton, UK, 24–27 September 2012.

4. Zheng, F.; Tian, Y.; Zhan, W.; Yu, J.; Liu, K. An IMM-BP based algorithm for tracking maneuvering underwater targets by multistatic marine robot networks. In Proceedings of the 2022 WRC Symposium on Advanced Robotics and Automation (WRC SARA), Beijing, China, 20–20 August 2022.
5. Ferri, G.; Tesei, A.; Stinco, P.; LePage, K.D. A Bayesian Occupancy Grid Mapping method for the control of passive sonar robotics surveillance networks. In Proceedings of the OCEANS 2019-Marseille, Marseille, France, 17–20 June 2019.
6. Zhan, W.; Tian, Y.; Zheng, F.; Yu, J. Integrated tracking and control framework for robotic multistatic sonar networks with IMM-BP and distributed MCTS. *Control Eng. Pract.* **2025**, *163*, 106403. [CrossRef]
7. Ferri, G.; Munafò, A.; Tesei, A.; Braca, P.; Meyer, F.; Pelekanakis, K.; Petroccia, R.; Alves, J.; Strode, C.; LePage, K. Cooperative robotic networks for underwater surveillance: An overview. *IET Radar Sonar Navig.* **2017**, *11*, 1740–1761. [CrossRef]
8. Ferri, G.; Munafo, A.; LePage, K.D. An Autonomous Underwater Vehicle Data-Driven Control Strategy for Target Tracking. *IEEE J. Ocean. Eng.* **2018**, *43*, 323–343. [CrossRef]
9. Ferri, G.; Petroccia, R.; De Magistris, G.; Morlando, L.; Micheli, M.; Tesei, A.; LePage, K. Cooperative autonomy in the CMRE ASW multistatic robotic network: Results from LCAS18 trial. In Proceedings of the OCEANS 2019-Marseille, Marseille, France, 17–20 June 2019.
10. Zhan, W.; Tian, Y.; Zheng, F.; Sang, Q.; Yu, J. AUV decision-making in bistatic sonar target tracking via deep reinforcement learning. *Robot. Auton. Syst.* **2026**, *197*, 105262. [CrossRef]
11. Ngatchou, P.N.; Fox, W.L.; El-Sharkawi, M.A. Multiobjective multistatic sonar sensor placement. In Proceedings of the 2006 IEEE International Conference on Evolutionary Computation, Vancouver, BC, Canada, 16–21 July 2006.
12. Fu, L.; Zhou, M.; Xu, L.; Dong, X.; Kou, Z.; Kang, C. Effective deployment strategies for optimizing area coverage in multistatic sonar detection based on Cassini oval approximation and a virtual force algorithm. *Ain Shams Eng. J.* **2024**, *15*, 103147. [CrossRef]
13. Clark, E.; Schoof, E.; Manzie, C.; Hunter, A. A Submodular Approach to Acoustic Sensor Placement Using the MARLIN Digital Twin. Available online: https://www.uaconferences.org/docs/2025_program_papers/2025_05_31_17_24_57_Clark_Edward.pdf (accessed on 5 March 2025).
14. Powers, T.; Krout, D.W.; Bilmes, J.; Atlas, L. Constrained robust submodular sensor selection with application to multistatic sonar arrays. *IET Radar Sonar Navig.* **2017**, *11*, 1776–1781. [CrossRef]
15. Craparo, E.M.; Fügenschuh, A.; Hof, C.; Karatas, M. Optimizing source and receiver placement in multistatic sonar networks to monitor fixed targets. *Eur. J. Oper. Res.* **2019**, *272*, 816–831. [CrossRef]
16. Ghafoori, A.; Altiok, T. A mixed integer programming framework for sonar placement to mitigate maritime security risk. *J. Transp. Secur.* **2012**, *5*, 253–276. [CrossRef]
17. Fügenschuh, A.R.; Craparo, E.M.; Karatas, M.; Buttrey, S.E. Solving multistatic sonar location problems with mixed-integer programming. *Optim. Eng.* **2019**, *21*, 273–303. [CrossRef]
18. Bi, W.; Zhou, J.; Shen, J.; Zhang, A. Optimization method of passive omnidirectional buoy array in on-call anti-submarine search based on improved NSGA-II. *Ocean. Eng.* **2024**, *293*, 116655. [CrossRef]
19. Yan, Z.; Cai, S.; Hou, S.; Zhang, M. Multi-sensor system deployment planning method for underwater surveillance based on formation characteristics. *Ad Hoc Netw.* **2025**, *170*, 103763. [CrossRef]
20. Kim, J.S.; Lee, D.H.; Yang, W.; Kim, Y.S.; Choi, J.W.; Kwon, H.; Park, J.; Son, S.-U.; Bae, H.S.; Park, J.-S. Optimal deployment of bistatic sonar using particle swarm optimization algorithm. *J. Acoust. Soc. Korea* **2024**, *43*, 437–444.
21. Karatas, M.; Craparo, E.; Akman, G. Bistatic sonobuoy deployment strategies for detecting stationary and mobile underwater targets. *Nav. Res. Logist. (NRL)* **2018**, *65*, 331–346. [CrossRef]
22. Panda, J.P. Machine learning for naval architecture, ocean and marine engineering. *J. Mar. Sci. Technol.* **2023**, *28*, 1–26. [CrossRef]
23. Wang, Z.; Li, H.X.; Chen, C. Reinforcement Learning-Based Optimal Sensor Placement for Spatiotemporal Modeling. *IEEE Trans. Cybern.* **2020**, *50*, 2861–2871. [CrossRef]
24. Jabini, A.; Johnson, E.A. A Deep Reinforcement Learning Approach to Sensor Placement under Uncertainty. *IFAC-PapersOnLine* **2022**, *55*, 178–183. [CrossRef]
25. Sang, Q.; Tian, Y.; Yu, J.; Zhang, F. A dynamic data driven application system for ocean environment prediction using reduced-order modeling and adaptive glider sensing networks. *Ocean Eng.* **2026**, *353*, 124812. [CrossRef]
26. Hof, C. Optimization of Source and Receiver Placement in Multistatic Sonar Environments. Ph.D. Thesis, Naval Postgraduate School, Monterey, CA, USA, 2015.
27. Craparo, E.; Karatas, M. Optimal source placement for point coverage in active multistatic sonar networks. *Nav. Res. Logist. (NRL)* **2019**, *67*, 63–74. [CrossRef]
28. Strode, C.; Mourre, B.; Rixen, M. Decision support using the Multistatic Tactical Planning Aid (MSTPA). *Ocean Dyn.* **2011**, *62*, 161–175. [CrossRef]
29. Wang, Y.; Fang, Z.; Li, J. Environment-Based Performance Optimization for Passive Sonar Network. In Proceedings of the OCEANS 2024-Singapore, Singapore, 15–18 April 2024.

30. Zhu, X.; Wang, Y.; Fang, Z.; Cheng, L.; Li, J. Strategic deployment in the deep: Principled underwater sensor placement optimization with three-dimensional acoustic map. *J. Acoust. Soc. Am.* **2024**, *156*, 2668–2685. [CrossRef] [PubMed]
31. Porter, M.B. *The Bellhop Manual and User's Guide: Preliminary Draft*; Heat, Light, and Sound Research Inc.: La Jolla, CA, USA, 2011; Volume 260.
32. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**. [CrossRef]
33. Kay, S.M. *Fundamentals of Statistical Signal Processing: Estimation Theory*; Prentice-Hall, Inc.: Hoboken, NJ, USA, 1993.
34. Proakis, J.G.; Salehi, M. *Digital Communications*; McGraw-Hill: New York, NY, USA, 2001; Volume 4.
35. Zhang, J.; Li, W.; Kang, S.; Yu, J.; Chen, S. Assigning Multiple AUVs to Form Arrays Under Communication Range Limitations Based on the Element Zero Method. *IEEE Syst. J.* **2021**, *15*, 1664–1673. [CrossRef]
36. Zhang, J.; Kang, S.; Yu, J.; Liu, S.; Li, W.; Chen, K. Assign multiple AUVs to form a row efficiently based on a method of processing the cost matrix. *Appl. Ocean Res.* **2020**, *101*, 102177. [CrossRef]
37. Lolla, T.; Ueckermann, M.P.; Yiğit, K.; Haley, P.J.; Lermusiaux, P.F. Path planning in time dependent flow fields using level set methods. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012.
38. Narayanan, V.L.; Gupta, A.; Dhaked, D.K.; Tewary, M.; Mondal, S.K.; Gopal, R.; Sarma, P. A systematic review on recent trends in path planning techniques for autonomous underwater vehicles. *Int. J. Intell. Robot. Appl.* **2025**, *9*, 936–976. [CrossRef]
39. Luo, W.; Wang, X.; Han, F.; Zhou, Z.; Cai, J.; Zeng, L.; Chen, H.; Chen, J.; Zhou, X. Research on LSTM-PPO Obstacle Avoidance Algorithm and Training Environment for Unmanned Surface Vehicles. *J. Mar. Sci. Eng.* **2025**, *13*, 479. [CrossRef]
40. Mackenzie, K.V. Nine-term equation for sound speed in the oceans. *J. Acoust. Soc. Am.* **1981**, *70*, 807–812. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.