Cross domain Persistent Monitoring for Hybrid Aerial Underwater Vehicles

Ricardo B. Grando[1,3], Victor A. Kich[2], Alisson H. Kolling[1],
Junior C. D. Jesus[1], Rodrigo S. Guerra[1], Paulo L. J. Drews-Jr[1]

arXiv:2602.21259v1 [cs.RO] 23 Feb 2026

***Abstract*— Hybrid Unmanned Aerial Underwater Vehicles (HUAUVs) have emerged as platforms capable of operating in both aerial and underwater environments, enabling applications such as inspection, mapping, search, and rescue in challenging scenarios. However, the development of novel methodologies poses significant challenges due to the distinct dynamics and constraints of the air and water domains. In this work, we present persistent monitoring tasks for HUAUVs by combining Deep Reinforcement Learning (DRL) and Transfer Learning to enable cross-domain adaptability. Our approach employs a shared DRL architecture trained on Lidar sensor data (on air) and Sonar data (underwater), demonstrating the feasibility of a unified policy for both environments. We further show that the methodology presents promising results, taking into account the uncertainty of the environment and the dynamics of multiple mobile targets. The proposed framework lays the groundwork for scalable autonomous persistent monitoring solutions based on DRL for hybrid aerial-underwater vehicles.**

I. Introduction

Hybrid Unmanned Aerial Underwater Vehicles (HUAUVs) have recently been a topic of research [1]–[4], given their unique set of tasks that leverage their air and water characteristics to operate effectively in both environments. They can help, for example, in improving inspection and mapping in oil platforms and submerged industrial facilities, besides allowing for better search and rescue tasks in harsh scenarios. However, autonomous operation in two different environments includes additional problems, such as wind and energy autonomy in the air and current and turbidity while underwater. These limitations have been incentivizing the research of novel approaches to improve the vehicle's capability to perform basic tasks such as navigation [4]–[6].

Several approaches have already been proposed for autonomous navigation of mobile robots based on traditional algorithms and, more recently, based on Deep Reinforcement Learning (DRL) [7]. However, for more complex tasks such as the persistent monitoring task, there is still a need for scientific contribution [8]. Specifically, considering a multimedia context such as air-water, cross-domain methodologies can be a way forward to enable this kind of vehicle to perform such tasks, improving its overall ability to be aware of its surroundings and providing scalability on its use for tasks not only in monitoring, but also in search and rescue and other related tasks. Persistent monitoring consists of tasks related to the continuous or repeated observation of an area or a set of targets. Also, since DRL algorithms have been successfully used for autonomous navigation-related tasks for mobile robots in general, this task could benefit from the methodology's ability to handle the additional complexity it may demand.

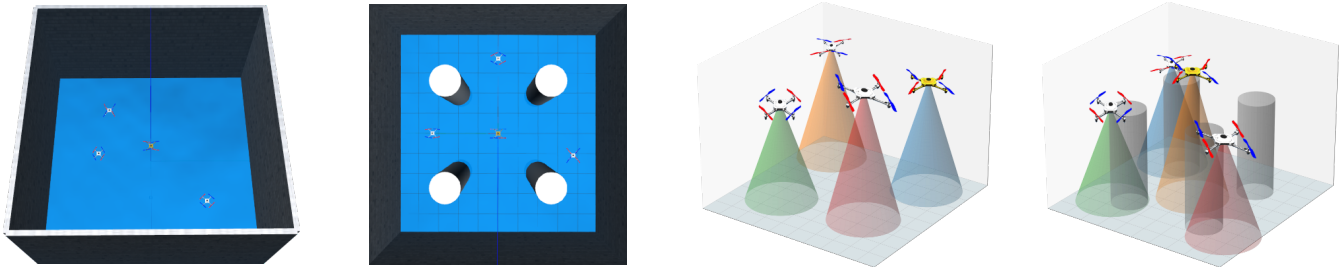
In this work, we introduce persistent monitoring tasks for HUAUV based on DRL. Considering that the vehicle needs to act in two different environments, we propose a methodology that can be applied in both environments independently of the differences between them. For that, we address the problem based on a ranging sensor point of view, considering a vehicle that has a Lidar for aerial sensing and a Sonar for underwater perception. Based on this kind of sensing, we demonstrate that DRL-agents with the same architecture can be transfer-learned to act in a different environment, providing a suitable monitoring capability while avoiding the use of visual or other positional information. We base our approach on a Parallel DRL architecture and on Transfer Learning technique, where the proof of concept based a single HUAUV and multiple moving targets can be improved and scaled up to more complex scenarios. Overall, our work provides the following contributions:

- We present a proof-of-concept methodology for persistent monitoring tasks with Hybrid Unmanned Aerial Underwater Vehicles (HUAUVs) based on Deep Reinforcement Learning (DRL) and Transfer Learning, showing that a single policy can be deployed across aerial and underwater domains using LiDAR and sonar range data.
- We propose a distributed parallel DRL training setup that accelerates policy learning and we show, in simulation, that the learned policy can manage the average uncertainty of multiple mobile targets more efficiently than a Bug2-style behaviour-based baseline, both in the original aerial domain and after zero-shot transfer to the underwater domain.
- We systematically analyse a simplified 2D setting with realistic vehicle dynamics and sensing, and we discuss the limitations of this formulation and directions towards more complex 3D scenarios and richer persistent monitoring objectives.

This work has the following structure: the related works are discussed in the following Section (Sec. II). We present our methodology in Sec. III and the results in Sec. IV. Finally, we present our contributions and future works in Sec. V.

II. Related Work

Persistent Monitoring tasks for mobile robots are still focused mostly on Unmanned Ground Vehicles (UGVs) [8]. As for HUAUVs, the literature is still mostly concerned



(a) First environment without obstacles, representing the baseline scenario for training the HUAUV's persistent monitoring policy.

(b) Second environment with four cylinders simulating drilling risers, used to evaluate obstacle-avoidance behavior.

(c) Visualization of the HUAUV agents in the obstacle-free environment, showing coverage cones and coordination for target uncertainty reduction.

(d) Visualization of the HUAUV agents in the environment with cylindrical obstacles, illustrating simultaneous coverage and obstacle avoidance.

Fig. 1: Simulated HUAUV persistent monitoring environments. (a) Aerial domain, Env 1 (obstacle-free). (b) Aerial domain, Env 2 (four cylindrical obstacles). (c)–(d) Example LiDAR scans in Env 1 and Env 2, respectively. Underwater scenarios use the same layouts but with sonar range measurements.

with mechanical design and modeling [9]–[15]. A few DRL methodologies have been already proposed for navigation in HUAUVs [4], [6] and for persistent monitoring with mobile robots [16], [17], [18], [19], [20], [21], [22].

Want et al. [18] introduced a spatiotemporal attention ANN built on *Transformer*-style attention layers that processes the agent's belief and the graph-structured environment to generate next landmark decisions. The problem was formulated as a UAV navigating in a 2D environment that must track moving ground targets. The vehicle was equipped with an RGB camera with a limited field of view. The metrics used to validate the models were visit count, estimation error, and target uncertainty. The methodology was compared with non-DRL-based algorithms. Overall, this work presents a DRL-based attention model based on visual information that learns when and where to navigate to reduce uncertainty about moving targets.

Mishra et al. [20] presented a DRL-based framework for managing persistent surveillance missions with energy-constrained UAVs. The task is to determine an optimal sequence of visits to a set of targets by a single UAV, minimizing the maximum time elapsed between successive visits to any target, while ensuring that the UAV does not run out of fuel or battery. The UAV is initially parked at a depot and must return periodically for refueling or battery recharge due to fuel or flight time constraints. A UAV was used, and the model was trained using the PPO algorithm with information related to the vehicle's position. The UAV is tasked with repeatedly visiting a set of targets with equal priority, ensuring persistent surveillance. Due to limited fuel or battery life, the UAV must return to the depot for refueling or recharge before depleting its energy reserves. In this way, the UAV seeks to minimize the maximum revisit time for all targets, ensuring that no target remains unobserved for long periods. The average number of visits was used to validate the methodology.

Sun et al. [19] presented a DRL-based framework for persistent coverage tasks using multiple UAVs. The UAVs are tasked with persistently covering a target region, ensuring that the area is continuously monitored over time. Each UAV uses the *Wonderful Life Utility* (WLU) function to evaluate the impact of its actions on overall system performance, fostering decisions that improve collective coverage efficiency. The WLU functions are used to guide the UAVs in making decisions that consider individual and collective benefits, promoting cooperative behavior. A total of eight UAVs were used to validate the methodology, trained following the actor-critic DRL algorithm based on the UAVs' position. The metrics used to validate the methodology were the rewards obtained during training, coverage, and navigation examples.

Vashisth et al. [22] demonstrated an adaptive DRL approach for persistent image-based tracking. The UAV seeks to find targets, avoid obstacles, and adapt to truly unknown environments in real time. The methodology employs an actor-critic neural network within the policy, trained to select landmarks using dynamic, locally sampled landmark graphs. A multi-rotor UAV equipped with a ground-pointing RGB-D camera was used. At each time step, candidate landmarks in the UAV's local free space are defined. Afterward, a dynamic graph of reachable waypoints and yaw combinations is constructed. Finally, the RL algorithm is used to select the next landmark. The methodology was tested in a photorealistic orchard environment using AirSim + Unreal Engine. Example trajectories and object detection rates were used to validate the methodology.

Our work differs from the previously discussed works by focusing on a cross-domain methodology with transfer learning, where the persistent monitoring task is assigned for the aerial and underwater domains. We also propose a novel methodology based on distributed parallel DRL, which provides faster learning and a scalable methodology for more complex scenarios when compared to non-parallel methods. Our methodology is also based on information from a ranging distance sensor to perform the persistent monitoring task and avoid obstacles, which present fewer limitations in both aerial and underwater domains when compared with image-based or GPS-based approaches.

III. Methodology

In this section, we discuss the proposed persistent monitoring problem and our proposed DRL methodology. We detail the monitoring system, method, and the simulation setup.

A. Problem Setup

We consider a persistent monitoring task in bounded 2D scenarios $\mathcal{E} \subset \mathbb{R}^2$, where a single autonomous agent, modeled as a continuous-control DRL policy, must reduce and manage the uncertainty of N mobile targets. Each target i moves according to a predefined Lissajous curve path. The physical HUAUV is simulated with full 6-DoF dynamics, but for the DRL formulation we use a 2D kinematic state:

$$s_t = [x_t, y_t]^\top, \quad a_t = [v_t, \omega_t]^\top$$

where v_t and ω_t are linear and angular velocity setpoints forwarded to the low-level controllers of the RotorS (air) and UUV (water) simulators. The resulting state evolution can be written abstractly as

$$s_{t+1} = s_t + \Delta t \cdot f(s_t, a_t)$$

where $f(\cdot)$ captures the closed-loop dynamics induced by the underlying vehicle model, controllers, and disturbances.

In the aerial and underwater domains we obtain different functions f_{air} and f_{water} because of the different vehicle dynamics and control loops. Thus, although the RL state is 2D, the policy experiences a genuine domain shift in its transition dynamics when transferred from the aerial to the underwater simulator.

B. Observations and Target Uncertainty

The agent observes the current 2D state and noisy estimates of each target's position:

$$o_t = \{s_t, \hat{g}_{1,t}, \hat{g}_{2,t}, \hat{g}_{3,t}\}, \tag{1}$$

where $\hat{g}_{i,t} \in E$ is the estimated position of target i at time t, obtained from range observations.

Each target i has an associated uncertainty $\sigma_{i,t} \in \mathbb{R}_+$ that evolves as

$$\sigma_{i,t+1} = \begin{cases} 0, & \text{if } \|s_t - g_{i,t}\| \le r_{\text{sense}}, \\ \sigma_{i,t} + \lambda\, \Delta t, & \text{otherwise}, \end{cases} \tag{2}$$

where $g_{i,t}$ is the true target position, r_{sense} is the sensing radius, and $\lambda > 0$ is the linear uncertainty growth rate.

The policy $\pi_\phi(a_t|o_t)$ is parameterized by a fully connected neural network trained using the DSAC algorithm. When resetting the uncertainty, the agent receives a positive reward r. This encourages the agent to reduce the uncertainty of all targets as efficiently and frequently as possible.

The proposed DSAC-based policy is trained in simulation operating in the aerial domain. The DSAC policy is trained in a 2D aerial environment $\mathcal{E}_{\text{air}} \subset \mathbb{R}^2$ where the HUAUV exhibits relatively fast, low-drag dynamics. After training, the learned policy π_ϕ is transferred to the underwater domain $\mathcal{E}_{\text{water}} \subset \mathbb{R}^2$. The underwater dynamics differ due to hydrodynamic drag and slower accelerations:

$$s_{t+1}^{\text{water}} = s_t + \Delta t \cdot f_{\text{water}}(s_t, a_t)$$

with $f_{\text{water}}(\cdot) \neq f_{\text{air}}(\cdot)$, representing a domain shift.

The same state and action spaces are shared across domains: $\mathcal{S}_{\text{air}} = \mathcal{S}_{\text{water}}$, $\mathcal{A}_{\text{air}} = \mathcal{A}_{\text{water}}$. The policy π_ϕ trained via DSAC in the aerial domain is deployed in the underwater domain without retraining (zero-shot transfer). The reward structure and target uncertainty model remain unchanged.

C. Distributional Soft Actor–Critic (DSAC)

The Distributional Soft Actor–Critic (DSAC), proposed by [23], combines the maximum entropy reinforcement learning framework of SAC with principles from distributional reinforcement learning. Instead of learning scalar Q-values,

$$Q_\theta(s,a) \approx \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right], \tag{3}$$

DSAC models the full distribution over returns using parametric distributions, the Categorical distribution in our work,

$$Z_\theta(s,a) \approx \text{Cat}\left(\sum_{t=0}^{\infty} \gamma^t r_t\right). \tag{4}$$

Thereby, capturing uncertainty in value estimates more effectively. This approach enhances both the stability and performance of the learning process, especially in complex continuous control tasks.

Standard SAC estimates the expected value of return via scalar Q-functions, which can be sensitive to bootstrapping errors. DSAC, by contrast, learns the full return distribution using distributional critics, allowing better uncertainty modeling, reduced overestimation bias, and more stable policy gradients. DSAC is built on the same maximum entropy principle as SAC, encouraging stochastic exploration by maximizing both expected reward and policy entropy. DSAC maintains a stochastic policy $\pi(a|s;\phi)$, modeled as a Gaussian distribution (actor), and a pair of distributional Q-networks $Z_1(s,a;\theta_1)$, $Z_2(s,a;\theta_2)$, where each outputs the quantiles of a return Categorical distribution. It also maintains a corresponding set of target Q-networks $Z_{\theta_{1,2}^-}$ and soft target updates.

We parameterise the return distribution $Z_\theta(s,a)$ as N quantiles $\{z_i(s,a)\}_{i=1}^N$ with associated quantile fractions $\{\tau_i\}_{i=1}^N$, where $\tau_i \in (0,1)$ and $\sum_i \tau_i = 1$. In all experiments we use $N = 64$ quantiles. Here, τ_i denotes the quantile fraction of the i-th critic output and is *not* related to the target uncertainty $\sigma_{i,t}$ defined in the monitoring problem.

The learning process is based on the distributional Bellman equation:

$$Z(s,a) \overset{D}{=} r(s,a) + \gamma\, Z(s',a'), \quad a' \sim \pi_\phi(\cdot|s'), \tag{5}$$

where $\overset{D}{=}$ denotes equality in distribution.

In practice, $Z_\theta(s,a)$ is parameterized as a finite set of quantiles $\{z_i(s,a)\}_{i=1}^N$. For each sampled transition (s,a,r,s'), the target quantiles are computed as:

$$y_j = r(s,a) + \gamma\, z'_j(s',a'), \quad a' \sim \pi_\phi(\cdot|s'),\ z'_j \sim Z_{\bar\theta}(s',a'), \tag{6}$$

where $Z_{\bar\theta}$ denotes the target network for the distributional critic.

The critical parameters are optimized using the Quantile Huber Loss:

$$\mathcal{L}_{Z_{1,2}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \rho^{\kappa}_{\tau_i}\left(y_j - z_i(s,a)\right), \tag{7}$$

where τ_i is the quantile fraction and ρ^κ_τ is the quantile regression Huber loss defined as:

$$\rho^\kappa_\tau(u) = |\tau - \mathbb{I}\{u<0\}|\ \mathcal{L}_\kappa(u), \quad \mathcal{L}_\kappa(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \le \kappa, \\ \kappa(|u| - \frac{1}{2}\kappa) & \text{otherwise.} \end{cases} \tag{8}$$

The actor retains SAC's entropy-regularized objective, encouraging both high return and policy entropy:

$$J_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi_\phi}\left[\min_i Q_{\theta_i}(s,a) - \alpha \log \pi_\phi(a|s)\right]. \tag{9}$$

Here, Q_{θ_i} is obtained as the expectation of the predicted return distribution, $\mathbb{E}[Z_\theta(s,a)]$, although other distributional statistics (e.g., lower quantiles) can be used to incorporate risk-sensitive behavior. The temperature parameter α is adjusted automatically to balance exploration and exploitation.

The DSAC policy is trained only in the aerial domain using LiDAR observations. All underwater results are obtained by deploying the aerially trained policy in the underwater simulator without any further fine-tuning, in a transfer leaning or a zero-shot transfer setting. We do not train a separate underwater policy

D. Simulated Environments and Vehicle Description

Our simulated environments are built on the Gazebo simulator and integrated with ROS. It makes use of the RotorS framework [24], which supports the simulation of aerial vehicles with various control modes, such as angular rates, attitude, and position control. The simulation also includes a wind simulation using an Ornstein-Uhlenbeck noise model.

Alongside RotorS, the UUV Simulator [25] was used for underwater simulation. This framework enables the simulation of hydrostatic and hydrodynamic forces, thrusters, sensors, and external disturbances. Using this framework, the underwater model of the vehicle was configured with parameters including volume, added mass, center of buoyancy, and other relevant physical properties, along with the characteristics of the underwater environment.

We consider two planar environments E_1 and E_2 in both the aerial and underwater domains. Environment E_1 is an obstacle-free $10\,\text{m}\times 10\,\text{m}$ tank, used as a baseline for learning the monitoring policy. Environment E_2 adds four cylindrical obstacles representing drilling risers. In both environments we evaluate the policy first in the aerial domain (LiDAR) and then in the underwater domain (sonar) without retraining. Figure 1 highlights the scenarios used to validate the proposed methodology.

The vehicle used in our experiments is based on the model introduced by Drews-Jr et al. [9], Neto et al. [10], and further refined by Horn et al. [26]. It was modeled according to its real mechanical specifications, including parameters such as inertia, motor coefficients, mass, rotor velocity, among others. The onboard sensing system was configured to reflect real-world LIDAR and sonar capabilities. The simulated LIDAR is based on the Hokuyo UST-10LX model, offering a sensing range of up to 10 meters with a 270° field of view and an angular resolution of 0.25°. It was implemented using Gazebo's ray-based sensor plugin. For sonar simulation, a forward-looking sonar (FLS) was modeled using the plugin developed by Cerqueira et al. [27]. This sonar features a maximum range of 20 meters, with 1000 bins and 256 beams. The beam's angular dimensions were set to 90° horizontally and 15° vertically.

The three moving targets used in both scenarios have a predefined Lissajous curve, with a random linear velocity to follow the path. The vehicles follow the paths in a persistent manner, having a linear uncertainty function associated with them and sharing their position with the DRL agent.

The Lissajous curve has the form

$$g_{i,t} = \big(A_i \sin(a_i t + \phi_i),\ B_i \sin(b_i t + \psi_i)\big), \tag{10}$$

with amplitudes $A_i, B_i \in [2,5]$ m and integer frequencies a_i, b_i of 2. For each episode, we sample the phases ϕ_i, ψ_i uniformly in $[0, 2\pi)$ and draw the linear speed along the curve from a uniform distribution $v_i \sim \mathcal{U}(0.1, 0.25)$ m/s.

In all experiments, the DRL agent only observes the reduced 2D state s_t and range data, while the underlying RotorS and UUV Simulator instances evolve according to the full six-degrees-of-freedom hybrid vehicle model. The air and water dynamics therefore differ in the internal mapping from the commanded setpoints (v_t, ω_t) to the realised planar motion.

TABLE I: Average Time and Uncertainty.

Model	t_{air} (s)	t_{water} (s)	τ_{1air}	τ_{2air}	τ_{3air}	τ_{1und}	τ_{2und}	τ_{3und}
DSAC (Env1)	35.40 ± 30.61	**97.95 ± 83.62**	**9.91**	**9.66**	**9.08**	**27.75**	**27.45**	**23.84**
BUG2 (Env1)	**36.90 ± 16.98**	127.43 ± 56.85	10.32	10.20	12.51	35.54	36.69	42.12
DSAC (Env2)	**49.34 ± 49.45**	**133.67 ± 142.21**	**12.98**	**10.41**	**13.39**	**26.62**	**32.87**	**27.72**
BUG2 (Env2)	93.63 ± 119.55	155.73 ± 151.44	30.06	37.82	26.09	36.92	45.20	35.27

E. Behaviour-Based Baseline (BUG2)

As a behaviour-based baseline we adopt the minimalistic quadrotor navigation strategy of Marino et al. [28], which is conceptually similar to a Bug2 algorithm and has been used previously for indoor aerial navigation. In our implementation, the baseline follows Bug2-style obstacle avoidance and target-seeking behaviour using range measurements only. We evaluate it in the same single-agent, multi-target monitoring task as DSAC.

IV. Experimental Results

A generated path was created for each moving target i during the training and validation phase, with a randomly generated velocity at each step for each target i. During training, each episode starts with three targets initialised on their Lissajous trajectories with uncertainties $\sigma_{i,0} \approx 0$. Whenever the agent enters the sensing radius of a target, the corresponding uncertainty $\sigma_{i,t}$ is reset to zero, the agent receives a positive reward, and the next target to pursue is selected as the one with highest current uncertainty.

The agents train until they reset at least one of the target's uncertainty or until they collide with an obstacle or with the tank border. A learning rate of 10^{-3} was used, alongside a minibatch of 256 samples and the Adam optimizer in the proposed ANN. We limited the number of episodes to 2500, where each episode terminates either when a fixed horizon of $T = 5000$ simulation steps is reached or when the agent collides. We consider that the $\lambda > 0$ linear uncertainty $\bar{\sigma}_i$ grows by 0.001 each simulated step, starting with a random value for each target i close to 0.

For evaluation, we use the same setting and let the agent run for the full horizon T. We then compute the average uncertainty $\bar{\sigma}_i$ of each target i and the average time-to-visit of each target. We denote by t_{air} and t_{water} the average times to first visit in the aerial and underwater domains, respectively, averaged over 100 trials.

We collected the statistics for our methodology and compared them with the performance of a behavior-based algorithm [28]. The evaluation was performed for 100 trials, and the average amount of uncertainty $\bar{\sigma}_i$ was recorded. Also, the average time for both underwater (t_water) and aerial (t_air) along navigation with their standard deviations was recorded. Specifically for the underwater evaluation, the evaluation was done using the models trained in the aerial setup. Table I shows statistics collected during the 100 trials in the aerial setup and the underwater setup, showing the average time and the average uncertainty $\bar{\sigma}_i$ of each target i.

Figure 2 shows the evolution of the target uncertainties $\sigma_{i,t}$ during evaluation for both environments and both domains. Each curve corresponds to one of the three targets under either the proposed DSAC policy or the Bug2 baseline. We observe that the DSAC policy resets target uncertainties more frequently and keeps their peaks lower than Bug2, particularly in Env 2 where obstacles induce longer detours. This behaviour is reflected in the lower average uncertainties $\bar{\sigma}_i$ reported in Table I. To accelerate convergence, we run DSAC in a distributed fashion with $K = 5$ parallel simulation workers. Each worker interacts with an independent copy of the environment, collects experience, and shares the replay buffer with the others. The policy and critic networks are shared across workers and updated centrally.

Overall, we can see that our methodology was able to perform the monitoring of a persistent task for its trained aerial scenario and to generalize and perform it in an underwater setup through a transfer learning methodology. We can also observe that when compared with a behavior-based algorithm, it was able to present better robustness both in terms of task time completion and average uncertainty $\bar{\sigma}_i$. We highlight the DRL agent's transfer learning capability, where the agent was able to perform the task underwater, having learned only in the aerial domain and presenting a similar overall uncertainty $\bar{\sigma}_i$ in all targets in the first scenario (27.75, 27.45, 23.84) and the second scenario (26.62, 32.87, 27.72). This shows that HUAUVs have the potential to benefit from DRL, considering a persistent monitoring task.

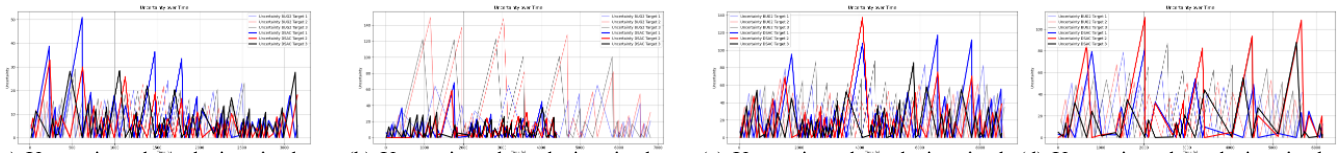
The present study uses a 2D abstraction of the HUAUV state and a simplified linear target-uncertainty model with three targets and noiseless target estimates. This formulation allows us to isolate the effect of cross-domain dynamics and range-based sensing on the learned policy, but it does not yet capture several important aspects of real HUAUV deployments, such as full 3D motion, sensor noise, occlusions, and more complex coverage objectives. In particular, the uncertainty dynamics are linear and local, and we do not consider global coverage metrics or long-term temporal-logic constraints.

From the algorithmic perspective, extending the formulation to 3D requires augmenting the state with altitude/depth and extending the action space to include vertical velocity or thrust commands. The DSAC framework and the range-based observation model remain applicable; however, the increased dimensionality will likely require more expressive policy networks and a larger amount of training data. Extending the proposed methodology to 3D hybrid motion, richer uncertainty models, and more realistic monitoring objectives is an important direction for future work.

V. Conclusions

In this work, we presented a methodology for persistent monitoring tasks using Hybrid Unmanned Aerial Underwater Vehicles (HUAUVs) operating both aerial and underwater domains. Based on Deep DRL and Transfer Learning, we demonstrated that a single DRL architecture can be effectively trained to perform in two distinct domains, using Lidar for aerial sensing and Sonar for underwater perception. Our results show that the proposed framework is capable of adapting to the different challenges imposed by each environment while maintaining robust monitoring capabilities. These findings indicate that cross-domain learning approaches can provide scalable and efficient solutions for complex monitoring tasks, not only in inspection and mapping but also in search and rescue missions.

Future work will focus on extending the proposed architecture to handle more heterogeneous sensor configurations,



(a) Uncertainty through time in the first environment (Air).
(b) Uncertainty through time in the second environment (Air).
(c) Uncertainty through time in the first environment (Underwater).
(d) Uncertainty through time in the second environment (Underwater).

Fig. 2: Evolution of target uncertainties $\sigma_{i,t}$ during evaluation for Env 1 and Env 2 in the aerial and underwater domains. The vertical axis shows the uncertainty, and the horizontal axis is the simulation time steps. DSAC reduces the peaks and frequency of high-uncertainty intervals compared to the Bug2 baseline.

integrating additional perception modalities, and testing in real-world HUAUV deployments. Moreover, we plan to explore multi-agent coordination strategies to improve scalability and resilience in large-scale monitoring operations.

Acknowledgment

The authors would like to thank the VersusAI team. This work was partly supported by the CAPES, FINEP, CNPq, PRH-ANP 22, and the Technological University of Uruguay.

References

[1] P. M. Pinheiro, A. A. Neto, D. G. Macharet, and P. L. Drews, "Energy-efficient trajectory planning with media transition for a hybrid unmanned aerial-underwater vehicle," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5424–5429.

[2] J. Costa de Jesus, V. A. Kich, A. H. Kolling, R. B. Grando, R. da Silva Guerra, and P. L. J. Drews-Jr, "Image-based mapless navigation of a hybrid aerial-underwater vehicle using prioritized deep reinforcement learning," *Journal of Intelligent & Robotic Systems*, vol. 111, no. 1, p. 27, 2025.

[3] P. J. Dias de Oliveira Evald, S. Loi Brião, V. Misaki Aoki, C. Bastos da Silva, D. dos Santos Cardoso, P. Miranda Pinheiro, F. Gomes de Oliveira, M. André Barbosa Cunha, and P. L. Jorge Drews-Jr, "Hydrone: Modeling of a hybrid unmanned aerial underwater vehicle considering trans-media dynamics," *Available at SSRN 4992978*.

[4] R. B. Grando, J. C. de Jesus, V. A. Kich, A. H. Kolling, N. P. Bortoluzzi, P. M. Pinheiro, A. Alves Neto, and P. L. J. Drews-Jr, "Deep reinforcement learning for mapless navigation of a hybrid aerial underwater vehicle with medium transition," in *IEEE ICRA*, 2021, pp. 1088–1094.

[5] J. C. de Jesus, V. A. Kich, A. H. Kolling, R. B. Grando, R. S. Guerra, and P. L. Drews, "Depth-cuprl: Depth-imaged contrastive unsupervised prioritized representations in reinforcement learning for mapless navigation of unmanned aerial vehicles," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 579–10 586.

[6] R. B. Grando, J. C. de Jesus, V. A. Kich, A. H. Kolling, P. M. Pinheiro, R. S. Guerra, and P. L. Drews, "Mapless navigation of a hybrid aerial underwater vehicle with deep reinforcement learning through environmental generalization," in *2022 Latin American Robotics Symposium (LARS), 2022 Brazilian Symposium on Robotics (SBR), and 2022 Workshop on Robotics in Education (WRE)*. IEEE, 2022, pp. 1–6.

[7] K. Zhu and T. Zhang, "Deep reinforcement learning based mobile robot navigation: A review," *Tsinghua Science and Technology*, vol. 26, no. 5, pp. 674–691, 2021.

[8] N. Basilico, "Recent trends in robotic patrolling," *Current Robotics Reports*, vol. 3, no. 2, pp. 65–76, 2022.

[9] P. L. Drews-Jr, A. A. Neto, and M. F. Campos, "Hybrid unmanned aerial underwater vehicle: Modeling and simulation," in *IEEE/RSJ IROS*, 2014, pp. 4637–4642.

[10] A. A. Neto, L. A. Mozelli, P. L. Drews-Jr, and M. F. Campos, "Attitude control for an hybrid unmanned aerial underwater vehicle: A robust switched strategy with global stability," in *IEEE ICRA*, 2015, pp. 395–400.

[11] R. T. da Rosa, P. J. Evald, P. L. Drews-Jr, A. A. Neto, A. C. Horn, R. Z. Azzolin, and S. S. Botelho, "A comparative study on sigma-point kalman filters for trajectory estimation of hybrid aerial-aquatic vehicles," in *IEEE/RSJ IROS*, 2018, pp. 7460–7465.

[12] M. M. Maia, D. A. Mercado, and F. J. Diez, "Design and implementation of multirotor aerial-underwater vehicles with experimental results," in *IEEE/RSJ IROS*, 2017, pp. 961–966.

[13] D. Lu, C. Xiong, Z. Zeng, and L. Lian, "A multimodal aerial underwater vehicle with extended endurance and capabilities," in *IEEE ICRA*, 2019, pp. 4674–4680.

[14] D. Mercado, M. Maia, and F. J. Diez, "Aerial-underwater systems, a new paradigm in unmanned vehicles," *Journal of Intelligent & Robotic Systems*, vol. 95, no. 1, pp. 229–238, 2019.

[15] A. C. Horn, P. M. Pinheiro, R. B. Grando, C. B. da Silva, A. A. Neto, and P. L. Drews-Jr, "A novel concept for hybrid unmanned aerial underwater vehicles focused on aquatic performance," in *IEEE LARS/SBR*, 2020, pp. 1–6.

[16] H. Kurunathan, K. Li, W. Ni, E. Tovar, and F. Dressler, "Deep reinforcement learning for persistent cruise control in uav-aided data collection," in *2021 IEEE 46th Conference on Local Computer Networks (LCN)*. IEEE, 2021, pp. 347–350.

[17] A. S. Asarkaya, D. Aksaray, and Y. Yazıcıoğlu, "Temporal-logic-constrained hybrid reinforcement learning to perform optimal aerial monitoring with delivery drones," in *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2021, pp. 285–294.

[18] Y. Wang, Y. Wang, Y. Cao, and G. Sartoretti, "Spatio-temporal attention network for persistent monitoring of multiple mobile targets," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3903–3910.

[19] Z. Sun, N. Wang, H. Lin, and X. Zhou, "Persistent coverage of uavs based on deep reinforcement learning with wonderful life utility," *Neurocomputing*, vol. 521, pp. 137–145, 2023.

[20] M. Mishra, H. Bana, S. Sarkar, S. Sanjeevi, P. Sujit, and K. Sundar, "Deep reinforcement learning-based approach for a single vehicle persistent surveillance problem with fuel constraints," *arXiv preprint arXiv:2404.06423*, 2024.

[21] M. Mishra, P. Poddar, R. Agrawal, J. Chen, P. Tokekar, and P. Sujit, "Multi-agent deep reinforcement learning for persistent monitoring with sensing, communication, and localization constraints," *IEEE Transactions on Automation Science and Engineering*, 2024.

[22] A. Vashisth, J. Rückin, F. Magistri, C. Stachniss, and M. Popović, "Deep reinforcement learning with dynamic graphs for adaptive informative path planning," *IEEE Robotics and Automation Letters*, vol. 9, no. 9, pp. 7747–7754, 2024.

[23] X. Ma, J. Chen, L. Xia, J. Yang, Q. Zhao, and Z. Zhou, "Dsac: Distributional soft actor-critic for risk-sensitive reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[24] F. Furrer, M. Burri, M. Achtelik, and R. Siegwart, "Rotors—a modular gazebo mav simulator framework," in *Robot Operating System (ROS)*, 2016, pp. 595–625.

[25] M. M. M. Manhães, S. A. Scherer, M. Voss, L. R. Douat, and T. Rauschenbach, "UUV simulator: A gazebo-based package for underwater intervention and multi-robot simulation," in *MTS/IEEE OCEANS*, 2016, pp. 1–8.

[26] A. C. Horn, P. M. Pinheiro, C. B. Silva, A. A. Neto, and P. L. Drews-Jr, "A study on configuration of propellers for multirotor-like hybrid aerial-aquatic vehicles," in *19th ICAR*, 2019, pp. 173–178.

[27] R. Cerqueira, T. Trocoli, G. Neves, S. Joyeux, J. Albiez, and L. Oliveira, "A novel gpu-based sonar simulator for real-time applications," *Computers & Graphics*, vol. 68, pp. 66–76, 2017.

[28] R. Marino, F. Mastrogiovanni, A. Sgorbissa, and R. Zaccaria, "A minimalistic quadrotor navigation strategy for indoor multi-floor scenarios," in *Intelligent Autonomous Systems 13*. Springer, 2016, pp. 1561–1570.