Autonomous Underwater Environment Perceiving and Modeling: An Experimental Campaign With FeelHippo AUV for Forward Looking Sonar-Based Automatic Target Recognition and Data Association

Leonardo Zacchini, *Member, IEEE*, Alberto Topini, *Student Member, IEEE*, Matteo Franchi, Nicola Secciani, Vincenzo Manzari, Lorenzo Bazzarello, Mirko Stifani, and Alessandro Ridolfi, *Senior Member, IEEE*

***Abstract*—Seabed inspection is one of the most sought-after applications for autonomous underwater vehicles (AUVs). Acoustical sensors, such as side-scan sonars and forward-looking sonars (FLSs), are commonly favored over optical cameras to carry out such a task. Indeed, sonars are not influenced by illumination conditions and can provide high-range data. However, due to the lack of features and low resolution, acoustical images are often hard to interpret with conventional automatic techniques, forcing human operators to analyze thousands of collected images to identify the so-called objects of potential interest (OPIs). In this article, we report the development of an automatic target recognition (ATR) methodology to identify and localize OPIs in FLS imagery. Such detections have been then exploited to realize a virtual world model with the probabilistic multiple hypothesis anchoring data association and model tracking algorithm. Distinct models of convolutional neural networks have been trained with a data set acquired in May 2019 at the Naval Support and Experimentation Centre (Centro di Supporto e Sperimentazione Navale—CSSN) basin in La Spezia, Italy. The ATR strategy has been successfully validated offline with the data gathered in October 2019 in the same site where the seabed targets were replaced and relocated. As regards the world modeling technique, it has been preliminarily tested on a simulated scenario built upon unmanned underwater vehicle Simulator. Finally, both the ATR and world modeling systems were on-field tested in October 2020 at the CSSN basin in a multivehicle architecture by employing an acoustical channel between FeelHippo AUV and an autonomous moving buoy.**

***Index Terms*—Artificial intelligence (AI), automatic target recognition (ATR), autonomous underwater vehicles (AUVs), convolutional neural networks (CNNs), intelligent robotics, marine robotics, underwater surveillance.**

I. Introduction

THE recent advancement in autonomous vehicles aims to develop increasingly intelligent systems capable of interacting with the surrounding environment and independently deciding the best actions to fulfill specific tasks. In this context, modern robotics tries to integrate artificial intelligence (AI) concepts and technologies; this pattern especially arises in the underwater domain, where the poorness of communications and the total absence of global navigation satellite system (GNSS) signal force to give more autonomy to the vehicle. AI is currently used in robotic systems for different purposes, such as making autonomous decisions, planning paths, and extensive data processing, which are all fields where excellent results are being achieved. Since the tasks demanded to autonomous underwater vehicles (AUVs) have become more and more challenging [1], [2], researchers and scientists are investigating the use of AI technologies in the marine environment. Indeed, autonomous inspection strategies for underwater installations [3], exploration planning [4], and autonomous coverage strategies [5], [6] have become essential tools to execute complex and hazardous subsea operations in unknown scenarios.

Perceiving and understanding the environment is a fundamental hierarchical step to accomplish such complicated tasks. To this end, AUVs can be equipped with several payload sensors, such as optical cameras, multibeam echosounders, side-scan sonars (SSSs), forward-looking sonars (FLSs), sub bottom profilers, and so on. However, collecting raw data might not be enough for meaningfully understanding the environment; as a matter of fact, such data shall be processed online and fused

Manuscript received 7 December 2021; revised 27 June 2022; accepted 19 August 2022. Date of publication 18 November 2022; date of current version 14 April 2023. This work was supported in part by the European Project EU-MarineRobots, which received funding from the European Unions Horizon 2020 Research and Innovation Program under Grant 031103, and in part by the H2020 European Project PASSport, which received funding from the European GNSS Agency under Grant N0101004234 (H2020-SPACE-EGNSS-2020). An earlier version of this paper was presented at IEEE-OES Autonomous Underwater Vehicles Symposium 2020, St. Johns, Newfoundland, Sept. 30–Oct. 02, 2020. *(Corresponding author: Alessandro Ridolfi.)*

Associate Editor: F. Arrichiello.

Leonardo Zacchini, Alberto Topini, Matteo Franchi, Nicola Secciani, and Alessandro Ridolfi are with the Department of Industrial Engineering, University of Florence, 50121 Firenze, Italy, and also with the Interuniversity Center of Integrated Systems for the Marine Environment, 16145 Genova, Italy (e-mail: leonardo.zacchini@unifi.it; alberto.topini@unifi.it; matteo.franchi@unifi.it; nicola.secciani@unifi.it; a.ridolfi@unifi.it).

Vincenzo Manzari and Mirko Stifani are with the Naval Support and Experimentation Centre, Centro di Supporto e Sperimentazione Navale, 19126 La Spezia, Italy (e-mail: vincenzo.manzari@marina.difesa.it; mirko.stifani@marina.difesa.it).

Lorenzo Bazzarello is with the Naval Support and Experimentation Centre, Centro di Supporto e Sperimentazione Navale, 19126 La Spezia, Italy, and also with the Dipartimento di Ingegneria dell'Informazione, Università di Pisa, 56126 Pisa, Italy (e-mail: lorenzo.bazzarello@marina.difesa.it).

Digital Object Identifier 10.1109/JOE.2022.3209719

to identify obstacles, objects of interest, or hazardous targets. Free and occupied areas must be carefully identified [7] and 3-D occupancy maps must be created for exploring unknown areas and navigating into highly unstructured environments. Mechanical scanned imaging sonar, optical cameras, and FLSs are viable solutions [4], [8]. As regards autonomous interventions, vehicles shall correctly detect and localize objects of interest to interact with the external environment; more specifically, when dealing with structured areas, such as water tanks, augmented reality markers, and computer vision (CV), techniques represent a simple and extremely effective solution [9]. Nevertheless, robots have to face frequently nonstructured and unknown regions. For instance, in sea mining explorations, the AUV performs optical surveys to identify nodules and stones; once a nodule is detected, a visual-guided landing maneuver to collect the object is performed. Since environmental and light conditions change continuously and cannot be foretold, the performance of CV techniques is limited. Hence, modern convolutional neural networks (CNNs) shall be exploited to achieve satisfying results [10].

For what concerns seabed inspections, AUVs are commonly used for a wide variety of applications, ranging from geomorphological and biological analyses to port supervision in the view of ensuring the safety of the vessel traffic. Marine scientists use AUVs to study the seafloor morphology and the bathymetric changes or examine benthic habitats [11], [12]. In archaeological investigations, the seabed is carefully photographed [13] to classify historical finds; acquire high-quality data is of utmost importance. For underwater surveillance, AUVs exploitation is related to mine counter measure tasks and analogous operations [14], where potential hazardous targets must be identified keeping human operators far from the risks.

The aforementioned tasks are generally performed by exploiting optical sensors. However, optical cameras are affected by water turbidity and lighting conditions, and gathering satisfactory images does arise as a nontrivial task not feasible in several scenarios. As a consequence, acoustical sensors, e.g., SSSs, as well as FLSs, are commonly favored to carry out inspection and exploration tasks. In fact, sonars are not influenced by illumination conditions and can provide high-range data. In particular, FLSs can synthesize satisfactory resolution images and, more importantly, do not require the vehicle to move to create an image. Nevertheless, although high-grade but extremely expensive sonars can provide excellent images, FLSs generally present high noise and a lack of features that make the images hard to interpret by using conventional image processing techniques. As a consequence, a human operator is usually in charge of analyzing the thousands of acquired images to identify the so-called objects of potential interest (OPIs). Once identified, the targets shall also be localized; for this end, the AUV navigation data and the sonar characteristics are hence needed. An automatic target recognition (ATR) strategy that detects and localizes OPIs in FLS imagery, hence, represents an important tool that could help human operators in this demanding task. In this context, cutting-edge deep learning (DL) techniques, which have become the state of the art in the classification and object detection tasks [15], are being investigated in marine ATR applications [16], [17].

Moreover, as long as a large set of OPI detections and localizations is provided, the need for a world modeling methodology, containing a dynamically updated list of 3-D geolocalized objects, does arise as pivotal. Indeed, the major purpose of such a technique is constituted by providing a unique unified representation of the whole gamma of existing objects of interest in the surveyed area; in other words, if the same target is detected and localized several times, the world modeling algorithm aims at fusing the supplied information into a single world model object. Such a model represents a fundamental outcome provided by the AUV. In fact, the world model could help human operators analyze the data collected and plan additional surveys or intervention operations. Besides, it could be employed by the AUV to achieve complex tasks.

A. Contribution

In this work, an extension of the ATR strategy for FLS frames presented in [18] has been designed and implemented to detect and geolocalize potential targets of interest placed on the seabed. Besides, the several detections, supplied by the ATR system, have been employed to create a world model of 3-D-localized and labeled objects of interest.

The research activity has focused on developing and evaluating the aforementioned solution on FLS imagery; additionally, the system feasibility has been verified during real-time tests. First, selected CNN models have been trained by exploiting a custom gathered data set of heterogeneous images, acquired in May 2019 at the Naval Support and Experimentation Centre (Centro di Supporto e Sperimentazione Navale - CSSN) basin in La Spezia (Italy), and the open-source machine learning library TensorFlow [19]. Then, the trained neural networks have been incorporated into a custom ATR software, developed in the Robot Operating System framework [20]. Aiming to develop an onboard, pragmatically working ATR solution for compact AUVs, an NVIDIA Jetson Nano [21] has been selected as dedicated payload hardware for running the trained CNN models, and it was mounted on FeelHippo AUV [22], [23]. As a preliminary stage, the ATR strategy performance was assessed through hardware-in-the-loop offline validation using prerecorded data. Finally, the ATR with world modeling solution has been validated online with two vehicles both developed by the Department of Industrial Engineering of the University of Florence (UNIFI DIEF): FeelHippo AUV (equipped with a small towed buoy as Wi-Fi bridge), and an autonomous moving buoy [24], working as surface vessel and capable of Wi-Fi as well as acoustical communication. The two involved vehicles have been together employed during an experimental campaign performed within the activities of the SEALab, the joint research laboratory between the CSSN of the Italian Navy and the Interuniversity Center of Integrated Systems for the Marine Environment (ISME). In detail, the FLS images were acquired and processed online with the developed embedded ATR solution by FeelHippo AUV. To provide a visual feedback (not available from an acoustical link), the detected targets and their estimated positions, outcomes of the ATR strategy, were additionally streamed in real-time using the small towed buoy (phisically linked to FeelHippo AUV)

to a workstation where a human operator could supervise the process. It is worth highlighting that the towed buoy was only an additional asset for supervising the recognition process during the validation campaign. It is not necessary for the proposed methodology.

At the same time, FeelHippo AUV acoustically transmitted the same ATR results to the autonomous buoy, which also worked as a Wi-Fi bridge to a workstation running a data association algorithm to generate and maintain a consistent world state estimate. As the last stage, the developed data association solution allowed to model the perceived environment by deciding whether the detected OPIs had already been discovered or not and creating a world model symbolic representation of 3-D-localized, uniquely labeled objects.

The rest of this article is organized as follows. Section II reviews the most used CNN architectures for FLS-based ATR solutions and related works about environment modeling and data association strategies. Section III is dedicated to describing the proposed ATR methodology by accurately outlining the DL model selection and training processes. Section IV reports the reference frames used in this article and describes the FLS model exploited to localize the identified OPIs. In Section V, the developed data association algorithm to create a model of the environment is presented. Section VI reports and analyzes the offline validation results, while Section VII overviews the experimental scenario and the on-field results obtained by collecting data during a sea mission. Finally, Section VIII concludes this article.

II. Background

A. State-of-the-Art ATR Solutions

With the growing demand for intelligent systems capable of performing complex interactive tasks, reacting to the environment while inspecting areas, and cooperating meaningfully with human operators, object detection has become a fundamental feature of modern robots. Unmanned ground vehicles and unmanned aerial vehicles can rely on a large variety of sensors, ranging from optical cameras to light detection and ranging devices, to detect objects. Due to the wide use of modern cameras, several image-based target identification solutions have been developed. In particular, CNN-based approaches have shown outstanding results, becoming the golden standard in the image classification and target recognition tasks [15].

On the contrary, marine robots have limited recognition capabilities due to the underwater domain. Water turbidity, low-light conditions, and poor visibility degrade the quality of the optical images (see Fig. 1), making the subsea object detection hardly achievable in many cases. Acoustical sensors, such as FLS or SSS, represent a valid alternative. Indeed, these sensors provide high-range data that are not as affected by water conditions. Besides, even though recognizing object patterns in the high-noise acoustical sonar images can be challenging, FLS has the potential to be a functional device in underwater ATR tasks by providing decent resolution images (an example is provided in Fig. 1), at high frame rates, and not requiring the vehicle to move.

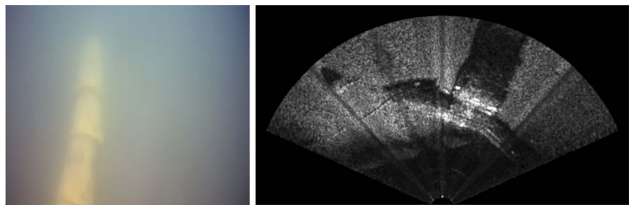


Fig. 1. On the left, an underwater pipeline structure in the corresponding optical image (native resolution of 704 × 576 pixels); on the right, the 2-D FLS acoustical frame (native resolution of 894 × 477 pixels).

Different template-matching-based object recognition approaches for FLS imagery have been developed and tested with different similarity measures and feature-trained classifiers [25], [26], [27], [28], [29]. Nonetheless, these techniques cannot generalize the template patterns; additionally, their performance degrades in the handling of multiscale objects. Therefore, these limitations led many marine researchers to investigate the use of CNN-based solutions also in acoustical imagery. In [16], custom CNN architectures to classify FLS images have been evaluated. The reported performance comparison with classical template matching solutions shown that CNNs could provide better performance while keeping a low number of parameters. Nevertheless, developing a custom CNN architecture is time expensive and requires plenty of images to train the network. Besides, AUVs usually have limited onboard computational power, and ATR should be performed in real-time to be effective during underwater missions. Thus, developing a CNN for onboard applications does emerge as a real challenge.

Turning to a more detailed overview, these solutions follow a common approach. The first network layers, called the backbone of the network, are in charge of extracting the dominant features, while the last layers classify those features and localize objects in the image. Generally, the backbone is tricky to train and requires a large data set. As regards the subsea environment, since gathering a large and heterogeneous data set in an underwater scenario is by no means straightforward, two common techniques do emerge as suitable solutions: data augmentation [30], which consists of applying several transformations to the previously acquired images to inflate the original data set with additional synthetically modified frames, as well as transfer learning [31], which involves the training of the last layers of pretrained models on a custom data set by fine-tuning higher order feature representations, speeding up the training phase. As a consequence of the aforementioned reasons, such deep neural network (DNN) approaches, relying on data augmentation and transfer learning, could be used to tackle object detection in FLS images.

Several remarkable DNN architectures have been proposed over the last few years for several and disparate fields of application; considering that for FLS the scientific literature highlights sparse applications, the most performing and promising DNN architectures have been taken into account to find which ones best fit our system. The you only look once (YOLO) network [32] was developed as an optimized end-to-end structure composed of 24 convolutional layers and 2 fully connected layers. This simple structure allows predicting bounding boxes and class

probabilities from full images in one evaluation. Thus, the network achieves real-time image processing with an extremely high frame per second (fps). YOLO was tested on FLS images ATR in [33], where the authors developed a system to detect divers.

A different approach was used to design the single shot multibox detector (SSD) [34], a convolutional network able to detect and classify objects at different scales at a high fps. Its native version used the visual geometry group network [35] as a backbone to extract the image features; however, the SSD strategy has been successfully incorporated with different feature extraction networks, such as Inception [36] and Mobilenets [37]. Small convolutional filters are then applied to different scale feature maps in the final layers to detect and classify objects. The network training aimed to optimize a multitask loss that took into account both the classification error and the bounding box coordinate error. This simple structure lets the SSD reach high-accuracy detections at high fps (up to 45). As the underwater domain is concerned, SSD was used to recognize objects in optical images [17], but as far as the authors known, it has not been tested on FLS imagery yet.

When the detection accuracy shall be favored over the inference speed, region-based architectures, such as the Faster region-based convolutional neural network (R-CNN) [38], are the recommended choice. The Faster R-CNN's backbone is composed of a feature extractor network and a region proposal network to produce the regions of interest (ROIs) in the feature maps and predict the bounding boxes. Two fully-connected sibling layers take each ROI as input and classify possible objects and refine the bounding boxes. The loss function used to train the network was a tradeoff between the classification and the localization tasks. As accurately highlighted in the speed/accuracy tradeoff analysis proposed in [39], compared with the SSD, the Faster R-CNN is more accurate but cannot reach the exceptionally high inference speed.

Mask R-CNN [40] extended the Faster R-CNN. First, the backbone was improved through the feature pyramid network that can better represent objects at multiple scales. Besides, the authors added in the final layers a convolutional branch to generate a segmentation mask for the selected ROIs. The training loss also considered the segmentation tasks, improving the network performance. In fact, instance segmentation enables identifying object outlines at the pixel level, enhancing the localization precision. R-CNN architectures were tested on optical underwater images [41] and on FLS imagery [17]. However, in [17], an analysis of their performance on FLS images was not reported.

B. World Modeling State-of-the-Art

Creating an accurate world model of the scenario where the AUV is navigating is a crucial stage for understanding the surrounding environment. For this reason, the targets detected by the ATR architecture alongside their localized positions must be handled, selected, and filtered to get a symbolic representation of the underwater scenario. Three significant points require hence to be addressed [42]. First, the robot needs to link the features supplied by the ATR system (e.g., label, size, position, shape, etc.) to semantic objects (*anchoring* [43] [44]). Second, probabilistic methodologies have to be exploited to associate the ATR sensor measurements with the corresponding object in the world model (*data association* [45]). Finally, a mathematical model, describing the prior knowledge of the detected target motion, is employed to perform *object tracking* [45].

The scientific literature highlights a wide range of distinct alternatives for the combined data association and model tracking problem, generally renown as multiple target tracking. For instance, the probability hypothesis density (PHD) Filter [46] can be considered analogous to a constant gain Kalman filter (KF) in a multitarget scenario. On the other hand, the multiple hypothesis tracker (MHT) [47] simultaneously takes into account the whole gamma of possible explanations from sensor measurements by identifying each world state a hypothesis alongside their correctness probability. The joint probabilistic data association filter [48] and the global nearest neighbor [49] methodology arise as another possible approach. Finally, in [42], the probabilistic multiple hypotheses anchoring (PMHA) architecture is proposed by fusing the advantages of multiple model (MM) techniques along with the major features of the MHT procedure as well as an anchoring strategy.

As far as the specific underwater scenario is concerned, the examples of world modeling are still sparse. Indeed, data association methodologies are rather employed within simultaneous localization and mapping contexts [50], [51], [52]. Nevertheless, an underwater target mapping strategy for AUV is suggested in [53], where a PHD-based approach aims at clustering several detections of the same object in a single unique representation (or cluster).

III. CNN-Based Automatic Target Recognition

A. Model Selection

The presented work investigates the development of a DL ATR strategy for FLS imagery that can run real-time on compact AUVs with limited hardware capabilities. In particular, since the effectiveness of image-based state-of-the-art DNNs on FLS images was shown in previous works [17], [33], this research focuses on a practical application of such DNN techniques. Besides, even though gathering a large and heterogeneous underwater data set is time and cost consuming, the aforementioned state-of-the-art DNNs allow the use of both data augmentation, increasing the number of data set frames by ad hoc modifying the original data set and transfer learning, which speed up the ATR development by fine-tuning the final network layers, while the backbone is not modified. As a result, a network model does not require thousands of images to be trained on a custom data set.

When it comes to select the most appropriate network, some relevant points must be considered. First, the acoustical frames are acquired at a low frame rate (3 Hz) in this work; thus, a high inference rate is not required. Moreover, the ATR solution has to provide additional geolocalization of possible seabed objects; within this context, since the target 3-D positions are estimated from the 2-D DNN localization in the FLS frame, minor errors in

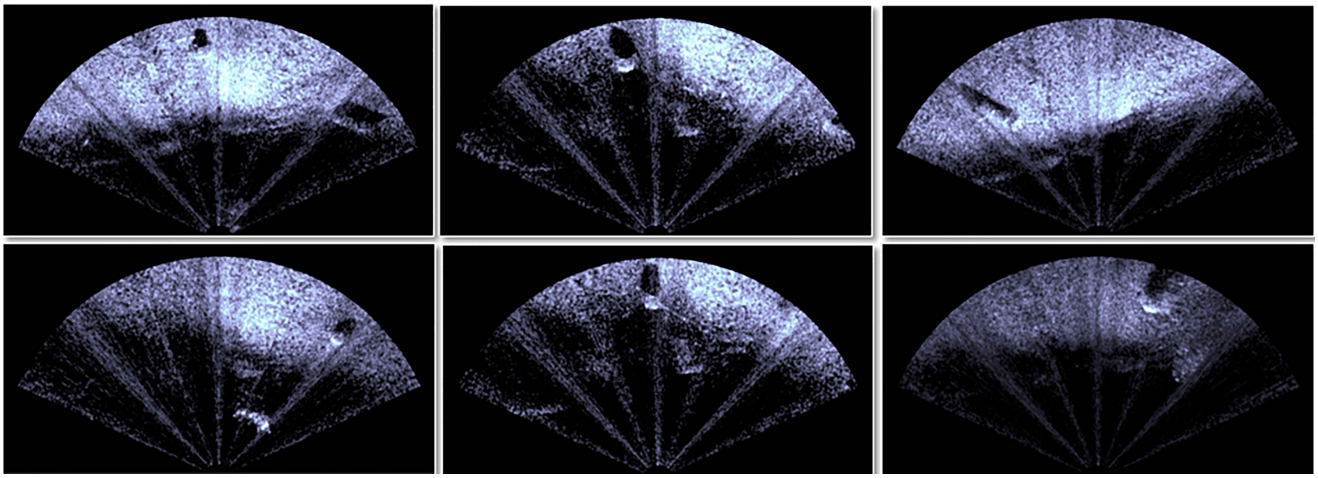


Fig. 2. Examples of FLS images used in the training data set. The selected DNN architectures were trained to detect and localize the depicted OPIs. It is worth noting that to get a heterogeneous ATR solution, the OPI forms and sizes vary while affecting their rendering, which also depends on the FLS viewpoint.

the bounding boxes at the pixel level could lead to large errors in meters in the 3-D localization. Therefore, the network accuracy is of utmost importance and shall favor the inference speed as the model selection parameter. As a consequence of these statements, region-based DNNs represent a functional selection for the developed ATR solution.

Faster R-CNN Inception V2 was employed in [17] to detect a massive underwater structure in FLS images showing satisfying results. However, in the context of this research activity, the primary goal has shifted to detect and localize objects of different sizes on the seabed, whose rendering is strictly related to both their form and the FLS viewpoint. Thus, the network multiscale object detection capability, together with the localization precision, plays a fundamental role. As a matter of fact, the Mask R-CNN, based on the Inception V2, fitted the above-described characteristics and has been selected to be tested within the hereafter suggested ATR strategy. Moreover, since this research activity proposes a preliminary investigation of an ATR solution for self-contained onboard applications, the required computational resources are of utmost importance, and the efficient SSD Mobilenet V2 network, designed for mobile and embedded devices, has also been tested.

B. Data Set Gathering

Within the context of this work, the training data set was acquired with FeelHippo AUV, whose main features and characteristics are illustrated in Section VII. The data set was gathered during on-field trials, performed in May 2019, at the CSSN basin, La Spezia, Italy. As depicted in Fig. 2, target OPIs have different shapes and dimensions, and their rendering on FLS images is strictly related to the sonar viewpoint.

Among the recorded FLS images, 175 frames, in a native resolution of 894 $\times$ 477 pixels and containing one or more detectable targets, had been selected. In particular, it is worth noting that this procedure has been fulfilled in the view of collecting a diversified heterogeneous data set. Consequently, images with different informative regions have been taken into account. Although it may be considered an evident and negligible pattern, this design guideline plays a fundamental role in providing the DNN architecture with an optimal generalization capability by avoiding overfitting during the training phase. Furthermore, coherently with the aim of building a robust data set, a data augmentation strategy has been employed; in particular, the data set has been augmented by randomly horizontally flipping the picked images and randomly varying their brightness and contrast.

C. Training Details and Results

Regarding the training details, SSD and Mask R-CNN present distinct size configuration strategies; however, the former resizes the images on a fixed shape, the latter exploits a shorter edge-based image scaling procedure. More in detail, as far as the SSD network is concerned, a down-scaling process is achieved for a final image size of 300 $\times$ 300 pixels. On the other hand, since the acoustical sonar provides low-resolution and low-frame rate pictures, the Mask R-CNN training pipeline has been designed by maintaining the image dimensions, and, thus, the aspect ratio to focus on the classification performance rather than on the computational cost. The SSD model has been trained using RMSProp with batch sizes of 24, whereas Mask R-CNN has exploited stochastic gradient descent with momentum with batch sizes of 1. Finally, the learning rate schedules have been defined explicitly for each CNN architecture to accomplish optimal inference outcomes and a fast convergence timing. The whole training procedure has been performed on a PC fitted with 16-GB RAM, an Intel Core i7-8750H processor, and an Nvidia GeForce GTX 1070 Ti card. For sake of completeness, the training outcomes, in terms of training curves, have been reported in Fig. 3.

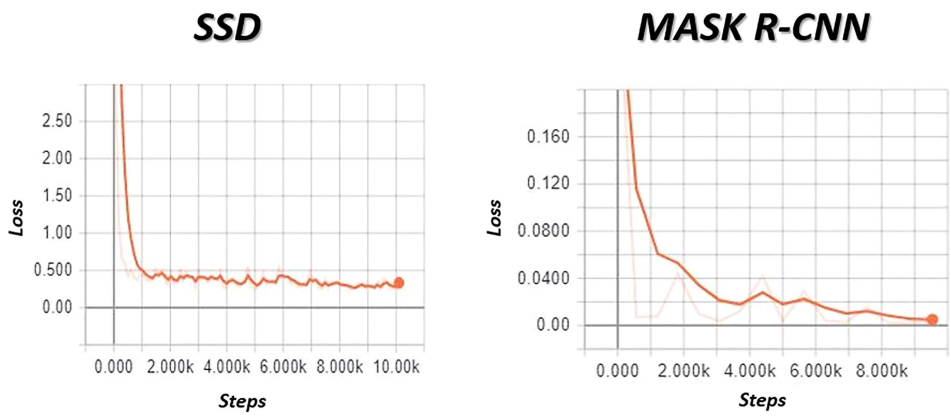


Fig. 3. Results of the training processes in terms of the loss curves. On the left, the SSD model training outcome is reported; conversely, on the right, the Mask R-CNN loss curve is reported.

IV. FLS Model Based Target Localization

As described in Section III, the proposed ATR solution uses image-based DNN architectures to identify OPIs in FLS images. Such models give the predicted classes as output, with the computed confidence, and the object bounding boxes. In particular, the bounding boxes are provided as the top-left and the bottom-right corners in the image reference system that uses pixels as the measurement unit. Therefore, the DNN detections shall be projected into an inertial reference frame to allow the AUV to correctly localize the OPIs and construct a model of the environment (see Section V).

First, the mathematical notation used in this research is introduced. Given a generic reference system $\{O^i x^i y^i z^i\}$, a vector $\boldsymbol{p} \in \mathbb{R}^3$ expressed in this frame will be denoted as ${}^i\boldsymbol{p}$. A rotation matrix $R \in \mathrm{SO}(3)$, for which it holds that $R \in \mathbb{R}^{3\times3}$, $RR^\top = I_3$, where I_3 is the 3×3 identity matrix, and $\det(R) = 1$ is referred as ${}^k R_i^j$; it rotates a unit vectors of the frame $\{O^i x^i y^i z^i\}$ in unit vectors of the frame $\{O^j x^j y^j z^j\}$, both expressed in the frame $\{O^k x^k y^k z^k\}$. If $k = j$, the three-indexes notation could be simplified in the following form: ${}^j R_i^j = R_i^j$. Introducing the transformation matrix $\boldsymbol{T}$ of the special Euclidean group in $\mathbb{R}^3$

$$\mathrm{SE}(3) := \left\{ \boldsymbol{T} = \begin{bmatrix} R & \boldsymbol{p} \\ \mathbf{0}^T & 1 \end{bmatrix} \mid R \in \mathrm{SO}(3), \boldsymbol{p} \in \mathbb{R}^3 \right\} \quad (1)$$

the relation between two reference frames can be described in a compact notation by using homogeneous transformations as well as the 4-D representation vector $\widetilde{\boldsymbol{p}}$. In particular, it holds that

$${}^j\widetilde{\boldsymbol{p}} = \begin{bmatrix} {}^j p \\ 1 \end{bmatrix} = \begin{bmatrix} R_i^j & \boldsymbol{t}_i^j \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} {}^i p \\ 1 \end{bmatrix} = \boldsymbol{T_i^j}\widetilde{\boldsymbol{p}} \quad (2)$$

where $\boldsymbol{t}_i^j$ is the translation vector between the center of the frames $< i >$ and $< j >$.

Then, the reference systems can be defined. The North–East–Down frame $\{O^{\mathrm{NED}} x^{\mathrm{NED}} y^{\mathrm{NED}} z^{\mathrm{NED}}\}$, denoted as $< N >$, is commonly used in marine robotics as the inertial reference system. It is a local Earth-fixed frame whose axes point, North, East, and Down (NED) respectively, and its center are placed on Earth's surface at a specific latitude and longitude pair, depending on the specific application [54], [55]. Attached to the vehicle, a reference system called body frame $< b >$, $\{O^b x^b y^b z^b\}$, is defined assuming the x-axis along the longitudinal axis of the vehicle, the z-axis pointing downwards, and the y-axis completes a right-handed system (see [56]). The vehicle used in this research work is FeelHippo AUV, described in Section VII, which estimates its pose with sufficient accuracy by using high-quality sensors and navigation strategies developed by the UNIFI DIEF. In detail, the nonlinear observer detailed in [57] is employed for estimating the vehicle attitude using inertial measurement unit (IMU) and fiber optic gyroscope (FOG) data. Then, the attitude estimate is combined with the Doppler velocity log (DVL) measurements in a dead reckoning algorithm to calculate the AUV position. The performance of this navigation methodology has been assessed in various works (e.g., [23] and [58]), resulting in an error of about 3% of the total traveled distance. For what concerns this works, the vehicle navigation strategy has been considered accurate for conducting inspection surveys, and the robot pose uncertainty was not considered for the target localization phase. Further information about the exploited navigation solutions can be found in [23], [57], and [58]. In conclusion, the relation between the $< N >$ frame and the $< b >$ frame, $\boldsymbol{T_b^N}$, is assumed completely known [56].

Regarding the FLS, it is rigidly attached in front of the AUV, and a right-handed reference system, denoted as $< F >$, $\{O^F x^F y^F z^F\}$, can be considered. The introduced $< F >$ frame center corresponds with the FLS center; its x-axis points forward while the z-axis points downwards. Since the FLS mounting position and orientation with respect to the AUV are known, the homogeneous transformation $\boldsymbol{T_F^b}$ is determined. Fig. 4 depicts the overall situation.

According to [23], [59], and [60], in the $< F >$ frame, a point $\boldsymbol{P} \in \mathbb{R}^3$ represented in Cartesian coordinates ${}^F\boldsymbol{P} = [X, Y, Z]^\top$ can be expressed in spherical coordinates ${}^F\boldsymbol{P} = [\bar{r}, \alpha, \phi]^\top$, where $\bar{r}$ is the FLS delivering range, α is the azimuth angle,

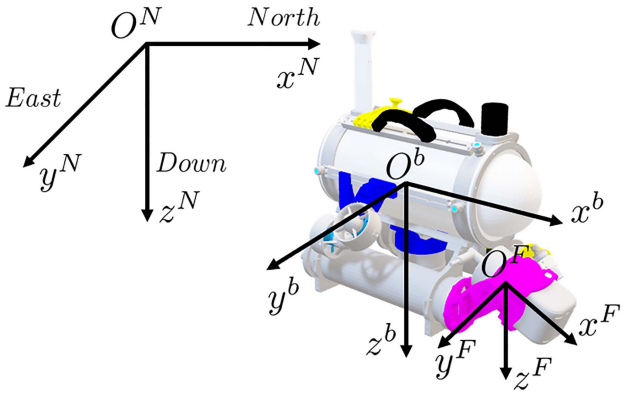


Fig. 4. Representation of the NED frame $< N >$, the body frame $< b >$, and the FLS frame $< F >$.

and ϕ the elevation angle. It holds that

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \bar{r} \begin{bmatrix} \cos\phi\cos\alpha \\ \cos\phi\sin\alpha \\ -\sin\phi \end{bmatrix}$$

$$\begin{bmatrix} \bar{r} \\ \alpha \\ \phi \end{bmatrix} = \begin{bmatrix} \sqrt{X^2+Y^2+Z^2} \\ \tan^{-1}(Y/X) \\ \tan^{-1}\left(-Z/\sqrt{X^2+Y^2}\right) \end{bmatrix}. \quad (3)$$

FLS devices natively use the spherical system in the imaging process: for each beam that composes the FOV, at every range interval, the average power of the reflected waves is measured and used to produce the corresponding pixel intensity in the image. However, the 3-D to 2-D image formation process leads to a loss of the information about the elevation angle ϕ [60]. In fact, as depicted in Fig. 5, the 3-D point ${}^F\boldsymbol{P}(\bar{r},\alpha,\phi) \in \mathbb{R}^3$ is projected on the FLS image plane, as depicted in red in Fig. 5 and denoted in the following as $< F_I >$, in a point $\boldsymbol{p}$ along the arc defined by the elevation angle ϕ [61]. Hence, given an FLS image, only the azimuth angle α and the range $\bar{r}$ of point can be computed. FLSs typically have limited vertical beamwidth ϕ (see [62], [61]) and are mounted with a small elevation angle γ, i.e., the angle between the horizontal plane and the insonifiying direction, which determines the FLS image plane. Besides, vehicles such as FeelHippo AUV, considered in this work, have the roll and pitch dynamics hydrostatically stable, and seabed inspection surveys do not excite these degrees of freedom (DOFs). Hence, the AUV navigates with roll and pitch angles almost zero with negligible variations. As a consequence of these considerations, a point ${}^F\boldsymbol{P}$ can be localized through its projection $\boldsymbol{p}$ in the FLS image plane. Thus, the approximation ${}^F\hat{\boldsymbol{P}}$ of ${}^F\boldsymbol{P}$ can be computed as

$${}^F\hat{\boldsymbol{P}} = \begin{bmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \end{bmatrix} = \bar{r} \begin{bmatrix} \cos\alpha \\ \sin\alpha \\ 0 \end{bmatrix}. \quad (4)$$

As previously discussed, in (4), there is a loss of the information about the point elevation angle ϕ that could lead to an error in the localization process. However, considering the problem tackled in this work, additional assumptions to enhance the localization accuracy can be drawn. First, the considered OPIs to be identified and localized are on the seabed. Then, the sea bottom imaged within a frame is assumed dominantly flat. That is, the detected bounding boxes lie on the seafloor at an altitude h from the AUV (see Fig. 5). Under these additional hypotheses, a point on the sea bottom can be accurately localized from an FLS image. To this end, the point elevation angle $\bar{\phi}$ shall be calculated. When the aforementioned assumptions hold, using altimeter data, the elevation of the point ${}^F\boldsymbol{P}$ can be retrieved. In fact, according to the local flat seafloor hypothesis, for a point ${}^F\boldsymbol{P}$ on the seabed, it holds

$$\bar{r}\sin(\gamma+\bar{\phi}) = h \quad (5)$$

where γ denotes the angle between the horizontal plane and the FLS insonifiying direction. Equation (5) allows to calculate the elevation angle $\bar{\phi}$, and thus, the point ${}^F\boldsymbol{P}$ can be localized accurately

$${}^F\boldsymbol{P} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \bar{r} \begin{bmatrix} \cos\alpha\cos\bar{\phi} \\ \sin\alpha\cos\bar{\phi} \\ \sin\bar{\phi} \end{bmatrix}. \quad (6)$$

Therefore, (5) and (6) are utilized to geolocalize the ATR findings. In fact, as discussed above, image-based DNN architectures produce as output the objects' bounding boxes whose coordinates are referred to the classic image reference frame $< I >$ that has the center in the image top-left corner and the x- and y-axis along the image width and height, respectively. The relation between the $< I >$ frame and the $< F_I >$ frame is known, as depicted in Fig. 6. Thus, for each OPI identified by the trained neural network, the bounding box can be projected from the $< I >$ frame to the $< F_I >$ frame; then the target can be localized: by using (6), its position is estimated in the $< F >$ frame, and since the transformation $\boldsymbol{T_F^N} = \boldsymbol{T_b^N}\boldsymbol{T_F^b}$ is accurately known, it is localized in the inertial reference frame $< N >$, and consequently, in the World Geodetic System (WGS84), which uses latitude, longitude, and altitude as coordinates (see [54] for more details).

V. World Modeling

Due to its feature of incorporating into an anchoring algorithm the MHT data association capability alongside the MM tracking characteristics, the PMHA methodology [42] has been selected as suitable to perform the world modeling task. Indeed, PMHA, coherently with its developers' purpose, has outlined the ability to extend MHT within a probabilistic anchoring framework where the targets to be semantically anchored are associated as well with a tracking mathematical model. As far as PMHA is concerned, the overall algorithm will be summarily presented in Section V-A, whereas in Section V-B, the proposed implementation will be highlighted by pointing out the distinctive features of this specific scenario.

A. PMHA Algorithm Description

As far as the anchoring process is concerned, the world objects are represented by anchors, whose attributes are updated over time by using the measurements provided by the sensors. More

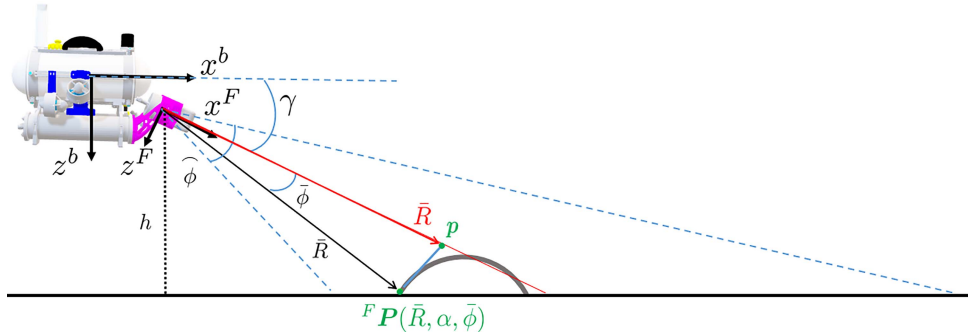


Fig. 5. FLS imaging process: given a range $\bar{r}$, points on the arc defined by the angle $\widehat{\phi}$ are projected in the FLS image plane $< F_I >$ (in red) in the point $\boldsymbol{p}$ [61].

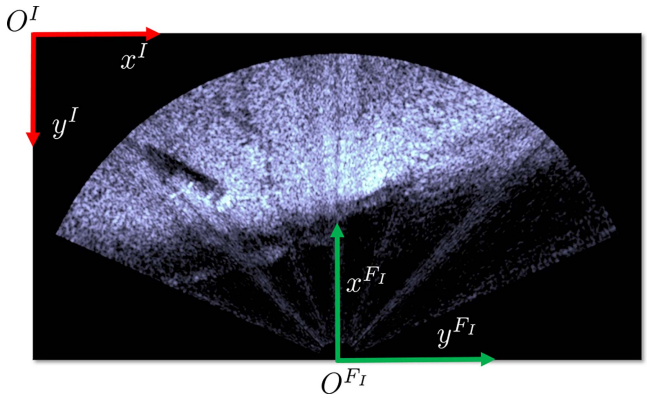


Fig. 6. 2-D FLS frame with respect to the image reference system $< I >$.

specifically, an anchor α_a is described as a tuple

$$\alpha_a = \left(\iota, z_i^k, \mathcal{M}_a^k\right) \tag{7}$$

where ι is the individual symbol constituting a physical object in the world, z_i^k is the measurement provided at the k-time and linked to this symbol with $i = 1, \ldots, n_{\text{meas}}$ representing the measurement index, $\mathcal{M}_a^k$ describes the whole set of anchor behavior models, whereas $a = 1 \ldots .n_{\text{obj}}$ represents the anchor index (with maximum value the object number in the exploited environment). Indeed, given

$$\mathcal{M}_i^k = \begin{cases} p\left(M_{a,1}^k\right) & : \quad M_{a,1}^k \\ \vdots & \\ p\left(M_{a,n_{\text{mod},a}}^k\right) & : \quad M_{a,n_{\text{mod},a}}^k \end{cases} \tag{8}$$

$M_{a,m}$ reflects an m-indexed behavior model, $n_{\text{mod},a}$ defines the total number of behavior models in the single anchor, and $p(M_{a,m})$ denotes the probability mass function associated with a single behavior model. Furthermore, the resulting attribute estimate Γ_a^k is computed by a weighted sum of the behavior model probability mass functions $p(M_{a,1}^k)$ and the probability density functions of the attribute estimated $\gamma_{a,m}^k$

$$\Gamma_a^k = \sum_{m=1}^{n_{\text{mod},a}} p\left(M_{a,m}^k\right) \gamma_{a,m}^k. \tag{9}$$

For instance, an AUV may observe a static OPI in an area described by a set of two behavior models: a fixed-position KF, where, in this case, $\mathcal{M}_a^k$ is the whole filter along with the motion model, and a uniform distribution behavior model highlighting the possibility for the OPI to be randomly moved, e.g., by a sea current, over the selected area.

As long as the anchor implementation is provided, the PMHA algorithm requires achieving the data association task by correctly selecting the right anchor attributes to be updated once new measurements are supplied by the sensors. As mentioned above, this specific problem is addressed by exploiting an MHT data association technique.

From a qualitative perspective, for every new set of measurements supplied by some sensors, three possible data association states are taken into account: First, a measurement can reflect the observation of a new object (i.e., not incorporated in the world model yet); second, a measurement may be descriptive of an object already included in the world model; finally, the last state represents a measurement originating from a false detection. Consequently, coherently with the MHT methodology, a hypothesis tree is built upon the assumption that each leaf constitutes a hypothesis Θ describing one admissible world state (i.e., containing a list of anchor-described objects). For each hypothesis, its corresponding correctness probability is evaluated and the highest value one is considered as the truthful world state model.

Quantitatively speaking, the correctness of the hypotheses (containing the listed anchors describing the world model) is computed by means of the Bayes' law as

$$p\left(\Theta_h^k \mid Z^k\right) = \frac{p\left(Z(k)|\Theta_h^k, Z^{k-1}\right) p\left(\theta_h^k|\Theta_{p(h)}^{k-1}, Z^{k-1}\right) p\left(\Theta_{p(h)}^{k-1}|Z^{k-1}\right)}{p(Z(k)|Z^{k-1})} \tag{10}$$

with $p(\Theta_h^k \mid Z^k)$ being the posterior probability of the h-indexed hypothesis Θ_h^k up to the time step k, the likelihood described by $p(Z(k) \mid \Theta_h^k, Z^{k-1})$, $p(\theta_h^k \mid \Theta_{p(h)}^{k-1}, Z^{k-1})$ evaluates the prior probability of the associations θ_h^k, $p(\Theta_{p(h)}^{k-1} \mid Z^{k-1})$ represents the posterior probability of the parent hypothesis, and $p(Z(k) \mid Z^{k-1})$ is introduced to normalize the probability value; in the following lines, each term will further be analyzed so as to provide a thorough understanding of their computing process.

By assuming independent measurements, the likelihood is computed as

$$p\left(Z(k) \mid \Theta_h^k, Z^{k-1}\right) = \prod_{i=1}^{n_{\text{mas}}} p\left(z_i^k \mid \alpha_{h,a_i}^k, Z^{k-1}\right) \tag{11}$$

with a_i indicating the anchor index to which the z_i^k measurement is linked in the specific hypothesis Θ_h^k. The second term of the previous equation, $p(z_i^k \mid \alpha_{h,a_i}^k, Z^{k-1})$, is specifically calculated in two distinct ways depending on whether a measurement is associated with a new/false detection or with an already existing anchor. In the first case, a uniform distribution around the measurement volume V is taken into account (with $n_{N,h}^k$ and $n_{F,h}^k$ representing the number of new and false measurements relative to the corresponding hypothesis) as follows:

$$p\left(z_i^k \mid \alpha_{h,a_i}^k, Z^{k-1}\right) = V^{-n_{N,h}^k - n_{F,h}^k}. \tag{12}$$

Conversely, Γ_a^k can be employed if the provided measurement is associated with an already established anchor by considering respectively the correctness probability of the behavior model $M_{h,a_i,m}^k$ and the probability $p(z_i^k \mid \gamma_{h,q_1,m}^k)$ that the measurement is supplied by the object linked to the anchor

$$\begin{aligned} p\left(z_i^k \mid \alpha_{h,a_i}^k, Z^{k-1}\right) &= p\left(z_i^k \mid \Gamma_a^k\right) \\ &= \prod_{m=1}^{n_{\text{mod}\,a}} p\left(M_{h,a_i,m}^k\right) p\left(z_i^k \mid \gamma_{h,a_i,m}^k\right). \end{aligned} \tag{13}$$

Turning to the prior probability $p(\theta^k \mid \Theta_{p(h)}^{k-1}, Z^{k-1})$, its evaluation is supplied by

$$\begin{aligned} p\left(\theta^k \mid \Theta_{p(h)}^{k-1}, Z^{k-1}\right) &= \frac{n_{N,h}^k! n_{F,h}^k!}{n_{\text{meas}}^k!} p_N\left(n_{N,h}^k\right) p_F\left(n_{F,h}^k\right) \\ &\times \prod_{a=1}^{n_{\text{obj},h}} \left(p\left(D_{h,a}^k\right)\right)^{\delta_a} \left(1 - p\left(D_{h,a}^k\right)\right)^{1-\delta_a} \end{aligned} \tag{14}$$

with $p_N(n_{N,h}^k)$ and $p_F(n_{F,h}^k)$, respectively, describing the probability mass functions of the new object and false detection numbers, $p(D_{h,a}^k)$ indicating the detection probability of the anchor α_a in the hypothesis Θ_h^k, whereas δ_a is calculated as

$$\delta_a = \begin{cases} 1, & \text{if the anchor } \alpha_a \text{ in } \Theta_{p(h)}^{k-1} \text{ is detected at time } k \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

With regard to the term $p(D_{h,a}^k)$, the computation is achieved by exploiting $p(V_{h,a}^k)$ (i.e., the probability that the object is actually visible)

$$\begin{aligned} p\left(D_{h,a}^k\right) &= p\left(D_{h,a}^k \mid V_{h,a}^k\right) p\left(V_{h,a}^k\right) \\ &= p\left(D_{h,a}^k \mid V_{h,a}^k\right) \sum_{m=1}^{n_{\text{mod},a}} p\left(V_{h,a}^k \mid M_{h,a,m}^k\right) p\left(M_{h,a,m}^k\right) \end{aligned} \tag{16}$$

where $p(D_{h,a}^k \mid V_{h,a}^k)$ clearly represents the detection probability conditioned to its visibility and $p(V_{h,a}^k \mid M_{h,a,m}^k)$ evaluating visibility probability given the model of the object itself.

Finally, the resulting world state is selected by checking the whole gamma of possible world states over the tree and selecting the most probable hypothesis.

B. Specific Implementation

Turning to the specific adaption of the aforementioned theoretical concepts for practical use of the PMHA algorithm, several parameters require to be tuned to achieve accurate results; as a matter of fact, a working implementation needs the definition of the behavior models $\mathcal{M}$ alongside the whole set of related parameters, the setting of the prior probabilities for new and existing objects as well as false detections.

In particular, the FLS-based detection and localization methodology provides two major features: the class label classification alongside the probability of the target to actually represent the previously labeled object and the target position estimate; in this view, both these features have been selected as predicates of the PMHA algorithms. The FLS-supplied position estimate is, first, evaluated by using an NED frame with origin the first valid position provided by the robot; such a vector is exploited to define a multidimensional Gaussian density function with mean the NED-referenced position itself and covariance a previously defined 3-D identity matrix. Conversely, the class label feature is described by a probability mass function with probability computed by the DL-model prediction.

As far as the behavior models $\mathcal{M}$ are concerned, due to the intrinsic diversity of the class label and position attributes, two distinct behavior models have been considered. Specifically, the class label model has been selected as a probability mass function over two different classes: *target* and *not-target*; it is worth noting that while the first one evaluates the real class label property, the second one is required as a complementary class. Indeed, by assuming, for instance, that the DL-based ATR system provides a 90% probability detection (i.e., the recognized object is a *target* with the aforementioned probability), the *not-target* class will simply get a 10% probability value. On the other hand, the target position behavior model is represented by a model including both a constant position KF and a Gaussian fixed-uncertainty characterization. As a matter of fact, the target position estimate is propagated by employing the KF, whereas localization measurements are provided; conversely, if the localization measurements are not provided for an apriostically-defined arc of time, the target position propagation switches to a multidimensional Gaussian with mean the current position values and covariance previously set matrices. Within this context, the design guideline of providing the target position with a 3-D identity matrix covariance has been adopted. Since the targets of interest are stationarily placed on the seabed, a constant position kinematic model has been considered; in particular, the 3-D NED position vector ${}^N\boldsymbol{P}$ has been defined as the state vector x

$${}^N\boldsymbol{P} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = x \tag{17}$$

with a discrete-time state equation

$$x(k+1) = Fx(k) + v(k) \tag{18}$$

where

$$F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{19}$$

due to the constant-position design guideline and $v(k)$ defined as a 3-D identity matrix. Within this perspective, the pattern of defining several covariance matrices of the target position predicate arises from the idea (argued by both simulated and on-field tests) of a maximum error of 1 m in the target localization process. It is worth noting that such a parameter has been established by taking into account just the target localization error in the sonar frame without considering the vehicle navigation uncertainty; further research activities will focus on analyzing the effect of the AUV pose covariance within this context. Additionally, to supplement the information supplied so far, since the measurement vector does coincide with the system state, the measurement model [usually indicated as $H(k)$] results as a 3-D identity matrix as well. Furthermore, this behavior model relies on multidimensional uniform distributions for the new object case as well as false detections; namely, the probability of achieving both new object or false detections is equally distributed over the volume in which the robot navigates. Finally, as reported in [42], a typical design guideline is to establish the probability of a measurement representing a false detection higher than the probability of the measurement representing a new object. This way the false detections that arise in a standalone way can be actually "filtered" by the world modeling strategy; nevertheless, occasional (i.e., not repeated) correct detections are filtered as well. As far as the specific implementation is concerned, since the FLS sensor may provide quite noise frames and lead the ATR system toward false detections, the authors have decided to be coherent with the cautious design guideline provided in [42].

VI. Preliminary Offline Validation

A. Automatic Target Recognition and Localization on Prerecorded Data

The trained networks were validated offline with a prerecorded data set (hereinafter called "validation data set") acquired in October 2019 at the CSSN basin. First, a preliminary, qualitative analysis was performed. The validation data set was processed with the trained SSD and Mask R-CNN models to assess their performance. The ATR solution was run on the PC used for the networks training (see Section III). As illustrated in Figs. 7 and 8, the developed strategy has guaranteed to fulfill the detection task; both the CNN models manage to recognize the underwater targets.

In the light of providing a more comprehensive overview, the achieved results have also been quantitatively evaluated in terms of precision and recall for the Mask R-CNN as well as the SSD networks. As can be seen from Table I, the SSD model has achieved a higher recall value (0.86) by outlining a superior

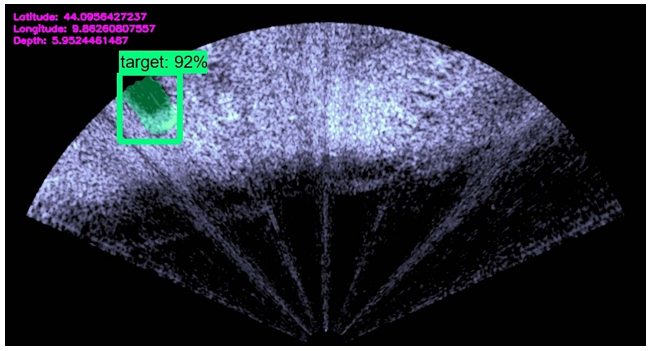


Fig. 7. Example of target recognition in a 2-D FLS acoustical image. The target was detected and classified by means of Mask R-CNN. The predicted bounding box is used to geolocalize the target through the FLS acquisition model.

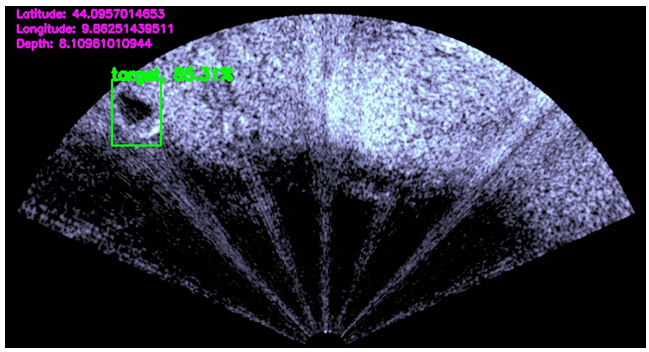


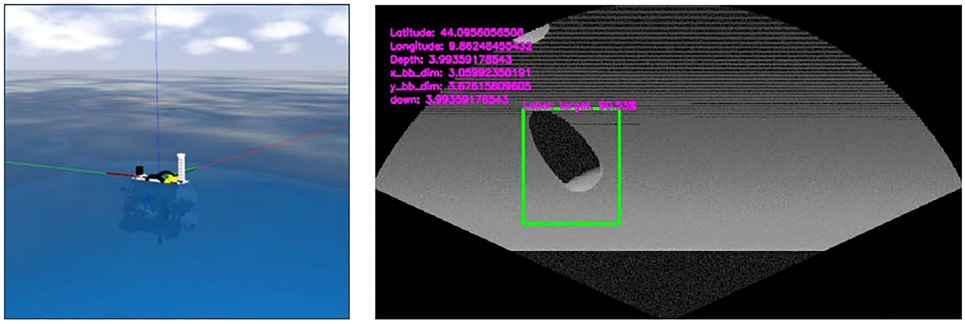
Fig. 8. Result of the ATR in a 2-D FLS acoustical image by means of SSD.

TABLE I
CNN Performance Indicator Score

Model	Precision	Recall
SSD MobileNet V2	0.77	0.86
Mask R-CNN Inception V2	0.95	0.82

detection capability of true-positive targets. Nonetheless, this pattern is counterbalanced by a significantly lower precision value (0.77), describing a nonnegligible aptitude in supplying several false positives. Concerning the Mask R-CNN, despite a slightly inferior recall (0.82), the excellent precision value (0.95) outlines the potential to accomplish adequate ATR performance. Indeed, within the scenario of developing a proactive intelligent system, enabling FeelHippo AUV to detect unknown targets and actively perform replanning for an inspection task, the Mask R-CNN has emerged as the most suitable architecture to avoid undesired reactive motions due to false-positive detections.

Since this research work focuses on developing a self-contained ATR methodology capable of running on compact AUVs, the CNN models were tested offline on FeelHippo AUV hardware to verify whether the developed CNN-based ATR solution could be used on hardware with limited computational resources. To this end, both the aforementioned ATR strategies were run on the NVIDIA Jetson Nano mounted on FeelHippo AUV, while its main computer was used to stream the training data set. It is worth noting that the NVIDIA Jetson Nano was



(a) (b)

Fig. 9. (a) The digital replica of FeelHippo AUV within the context of the UUV Simulator. (b) An example of the FLS simulated sensor alongside a detection of an OPI.

TABLE II
TELEDYNE BLUEVIEW M900 MAIN SPECIFICS

FLS main characteristics	
Field-of-View [°]	130
Beam Width [°]	1×20
Max Range [m]	100
Beam Spacing [°]	0.18
Range Resolution [in.]	1

favored over other solutions since it could fit into the limited available space on FeelHippo AUV. This setup allowed the authors to simulate real experimental working conditions. The SSD Mobilenet V2 network, designed for mobile and embedded devices, managed to analyze up to 4 fps. On the other hand, the more accurate Mask R-CNN Inception V2 has a higher computational load and resulted slower than SSD Mobilenet V2 network, reaching just 0.5 fps. **The acoustical frames were captured by a Teledyne Blueview M900 2D FLS (see Table II) at 3 Hz. Hence, the Mask R-CNN network could not fulfill the online recognition requirements. In fact, it was able to analyze just one image out of six, suggesting its usage in nonstreaming postprocessing sensors as well as degrading the ATR methodology's performance by leading to a nonaccurate environment model obtained through the data association algorithm. Therefore, only the SSD Mobilenet V2 network, which can run at a satisfying frame rate, was considered for the sea trials described in the next section.

B. World Modeling Quantitative Evaluation on Simulated Data

Since the ground-truth location of OPIs, located on the seabed of the real on-field scenario, was not provided, a simulated environment has been realized upon the standard implementation of the unmanned underwater vehicle (UUV) simulator [63] to achieve a quantitative validation of the above-illustrated world modeling methodology. As a matter of fact, the selection of this framework has been justified by its major feature of supplying the users with the capability of replicating both the robot dynamics and software along with a useful representation of the underwater world. The FeelHippo AUV dynamics has been modeled by using a previously performed identification process, the basic FLS simulated sensor has been modified so as to be coherent with the one equipped by FeelHippo AUV (see Fig. 9), whereas the OPIs have been represented by semispheres laying on the simulated seabed. Additionally, it is worth outlining that with the aim of minimizing the so-called "reality gap," the digital replica of FeelHippo AUV replicates also the whole software architecture of the real robot.

A lawnmower path has been considered to be indeed consistent with the typical FeelHippo AUV motion during an OPI target search mission. Besides, the whole list of targets (see Table III) has been defined in a NED frame with origin as a previously selected available latitude–longitude–depth position, whereas a specific SSD model has been trained to detect the simulated targets of interest. By means of the estimated position values of Table III, a quantitative discussion is provided by computing the Euclidean distance error between the ground-truth and the estimated OPI NED positions and, then, evaluating the average of the whole set of computed distance values; more specifically, this metric value results as 0.73 m. Such an outcome (see Fig. 11), even though achieved in a simulated scenario, does represent an argument to validate the correctness of the world model process and emerges as a key preliminary step for the further on-field test stage.

VII. EXPERIMENTAL SETUP AND RESULTS

The proposed detection strategy was validated online during sea trials performed in October 2020 at the CSSN basin.

In this context, the testing framework was constituted of three components: FeelHippo AUV (see Fig. 12) and an autonomous moving buoy (see Fig. 13) both developed by the UNIFI DIEF, and a workstation comprised of a laptop PC and a Wi-Fi bullet.

FeelHippo AUV, which will be described in detail in Section VII-A, was equipped with a Teledyne BlueView M900 2D FLS (see Table II) and performed an autonomous lawnmower path at a constant altitude with a desired longitudinal speed of 0.5 m/s. Raw FLS images were online processed onboard using the developed ATR solution. The image-based DNN ATR algorithm was run on the NVIDIA Jetson Nano mounted on FeelHippo AUV. An acoustical channel was created between FeelHippo AUV, equipped with an EvoLogics S2CR 18/34

TABLE III
OPIs PREDEFINED AND PHMA ESTIMATED NED POSITIONS

OPIs	Ground-truth (m)	Estimated Position (m)	Distance (m)
OPI1	[0.0, 10.0, 4.0]	[-0.38, 9.19, 4.27]	0.93
OPI2	[0.0, 20.0, 4.0]	[-0.38, 19.32, 4.32]	0.84
OPI3	[0.0, 25.0, 4.0]	[-0.38, 24.29, 4.32]	0.87
OPI4	[-8.0, 10.0, 4.0]	[-8.01, 9.26, 4.31]	0.80
OPI5	[-7.0, 15.0, 4.0]	[-7.06, 14.72, 4.32]	0.43
OPI6	[-8.0, 22.0, 4.0]	[-8.01, 21.61, 4.32]	0.50
Average Distance (m)			0.73

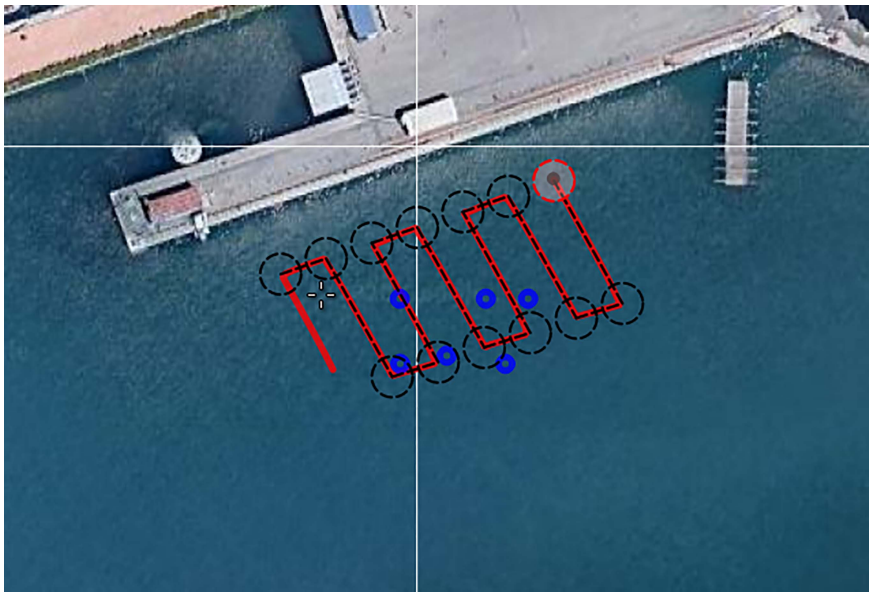


Fig. 10. Example of the world modeling process achieved by employing the PMHA algorithm over the simulated UUV Simulator world geolocalized at the CSSN, La Spezia, Italy. In a *red* line, the lawnmower path achieved by FeelHippo in a standard survey mission; every path waypoint is illustrated with a *black* dotted circle, where the current reference waypoint is highlighted in *red*; finally, the OPIs representing the world model objects are described by small *blue* circles.

acoustic modem, and the autonomous moving buoy (described in Section VII-B), fitted with an EvoLogics S2CR 18/34 ultra-shortbaseline (USBL), which acted as a modem. Each detection was associated with structured data incorporating the reliability level (i.e., the detection probability), the absolute time, and the estimated position of the target (latitude, longitude, and depth that were computed starting from the 2-D bounding box in the FLS frame and the AUV position (see Section IV), and it was communicated from FeelHippo AUV to the autonomous moving buoy. Besides, the autonomous buoy was connected through a Wi-Fi channel to the workstation. The use of this arrangement was twofold. First, it permits running the data association algorithm into the workstation (see Section V), where an operator could supervise the recognition and modeling process; this was deemed necessary during the preliminary validation. Besides, to enable the recognition process's complete supervision, a towed buoy connected to FeelHippo AUV was also used to stream the ATR processed FLS images, i.e., the detected OPIs bounding boxes and the computed positions painted over the images, to the workstation.

Second, in light of future development, the autonomous buoy could enable a multivehicle cooperation strategy. In fact, the buoy could act as a communication node that coordinates multi-inspections (e.g., a second AUV could perform an accurate optical inspection) or considers the data acquired from several AUVs for modeling the environment.

For the sake of clarity, the overall architecture is depicted in Fig. 14.

A. *FeelHippo AUV*

FeelHippo AUV is a compact and lightweight autonomous vehicle developed by the UNIFI DIEF. Its main features are summarized in Table IV.

The list of the primary electronic components and the sensor sets with which FeelHippo AUV is equipped are reported as follows:

1) Intel i7-based LP-175-Commel motherboard (main computer);
2) NVIDIA Jetson Nano (payload computer);

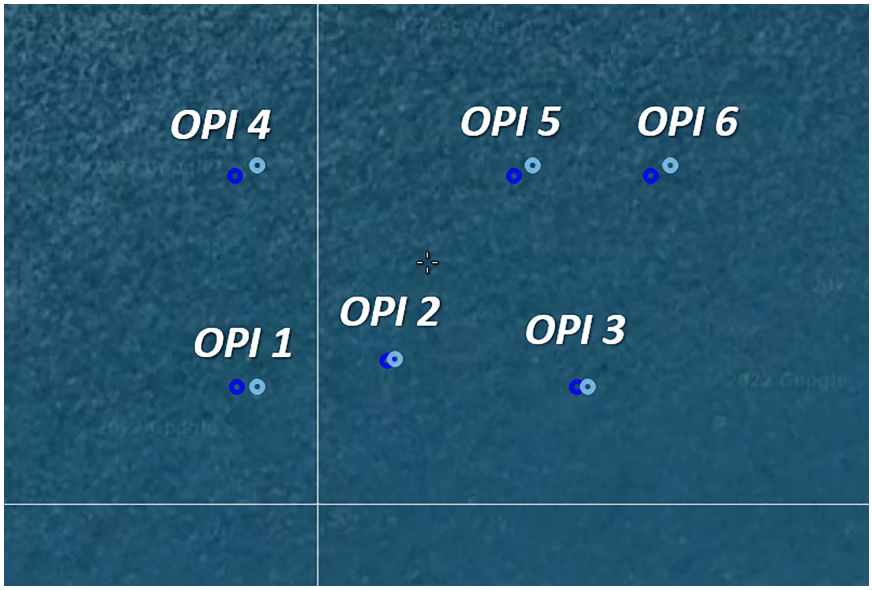


Fig. 11. Particular detail of the world modeling process achieved by employing the PMHA algorithm over the simulated UUV Simulator world geolocalized at the CSSN, La Spezia, Italy. The OPIs representing the world model objects are described by small *blue* circles, whereas the OPI ground-truth positions are reported in small light blue circles.

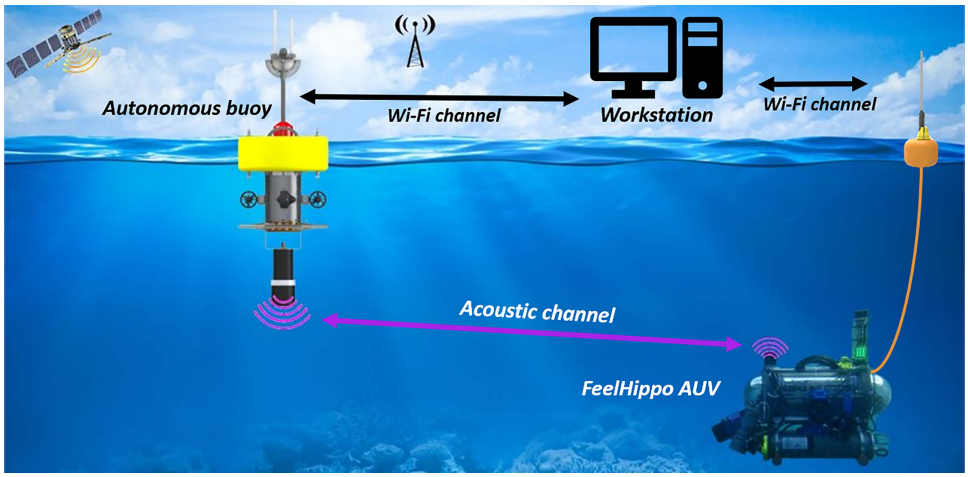


Fig. 12. FeelHippo AUV before an on-field underwater mission in Vulcano Island, Italy.

TABLE IV
FEELHIPPO AUV MAIN CHARACTERISTICS

FeelHippo AUV main characteristics	
Autonomy [h]	2-3
Controlled DOFs	4
Max longitudinal speed [m/s] (kn)	approx. 1 (2)
Max lateral speed [m/s] (kn)	approx. 0.2 (0.4)
Max depth [m]	30
Dimensions [mm]	approx. 600×640×500
Dry mass [kg]	35

Fig. 13. Autonomous moving buoy during an on-field mission at the CSSN basin, La Spezia, Italy.

Fig. 14. Overall ATR architecture used for the experimental campaign. FeelHippo AUV transmitted through the acoustical channel the ATR findings to the autonomous buoy, which was connected with the Wi-Fi to the workstation running the world modeling algorithm. Additionally, to allow an operator to supervise the recognition process during the validation stage, a towed buoy attached to FeelHippo AUV has been employed.

3) U-blox 7P precision global positioning system (GPS);
4) Orientus advanced navigation attitude and heading reference system (AHRS);
5) KVH DSP 1760 single-axis high precision FOG;
6) Nortek DVL1000 Doppler Velocity Log, measuring linear velocity and acting as depth sensor;
7) EvoLogics S2CR 18/34 acoustical modem;
8) Teledyne BlueView M900 2D FLS.

The main computer (Intel i7-based LP-175-Commel motherboard) was used for onboard processing of guidance, navigation, and control algorithms, for supervising the state of the vehicle and for managing the communication channels. The NVIDIA Jetson Nano [21], used for running the developed ATR solution, was connected to the main computer through an Ethernet cable that guaranteed an elevated transmission speed (up to 1 Gb/s) and a satisfying bandwidth (250 MHz).

B. Autonomous Buoy

The autonomous buoy is a motorized buoy that can navigate on the sea surface and keep a position accurately, thanks to the use of GNSS signals. It has been designed to support FeelHippo AUV with the accomplishment of the localization and communication tasks. To deal with these functionalities, the buoy had been equipped with the following devices:

TABLE V
AUTONOMOUS MOVING BUOY MAIN FEATURES

Autonomous Moving Buoy main characteristics	
Autonomy [h]	10-12
Controlled DOFs	3
Cruise Speed [m/s] (kn)	approx. 0.3 (0.6)
Dimensions [mm]	approx. 1410×535×535
Dry mass [kg]	37

1) UDOO x86 ULTRA Processor;
2) RFD 868+ Radio Modem;
3) Ubiquiti Wi-Fi PicoStation M2-HP;
4) Xsens MTi-G IMU;
5) Ublox NEO 7P Evolution Kit GPS;
6) EvoLogics S2CR 18/34 USBL.

Furthermore, the main properties of the autonomous moving buoy are reported in Table V.

C. Results

This section reports the results obtained during the experimental campaign conducted in October 2020 at the CSSN, La Spezia, Italy. According to the performance comparison reported in Section VI, only the trained SSD network was tested during the experimental campaign. FeelHippo AUV conducted an autonomous lawnmower path at a constant altitude while achieving several OPI detections (see Fig. 15). To provide the whole experiment with an improved generalizing context, when the experimental campaign was conducted in October 2020, the OPIs, utilized for the the training data set acquired in May 2019, were replaced and relocated.

Since the testing site was an unknown environment and the OPIs ground-truth localization was not provided, the accuracy of the ATR methodology was assessed. To this end, an OPI (resembling a truncated cone) was deployed in a known position in the CSSN basin. FeelHippo AUV was used to enlighten the OPI with the Teledyne Blueview M900 2D FLS, and the ATR solution was run with the localization technique detailed in Section IV. In particular, the OPI position was estimated by using both the FLS standard 2-D approximation (4) and the 3-D version of (6), which makes use of additional assumptions that hold for the testing site (see Section IV). While with the former approach, the ATR solution achieved a localization error of about 2.5 m, the latter resulted more accurate with an error of less than 2 m (about 1.3 m). It is worth noting that considering the underwater navigation system accuracy (in position and orientation, in particular, the heading angle), localization errors below 2 m can be considered accurate.

Then, FeelHippo AUV was used to inspect the unknown region with the ATR strategy. As shown in Fig. 16, during the survey, the ATR solution with the SSD Mobilenet V2 network managed to detect and localize several OPIs of various forms and sizes online. Turning to quantitative analysis, during the survey, the ATR methodology provided 61 detections. Afterward, a human operator analyzed the ATR outputs in a postprocessing stage that allowed to classify the detections as true positives and

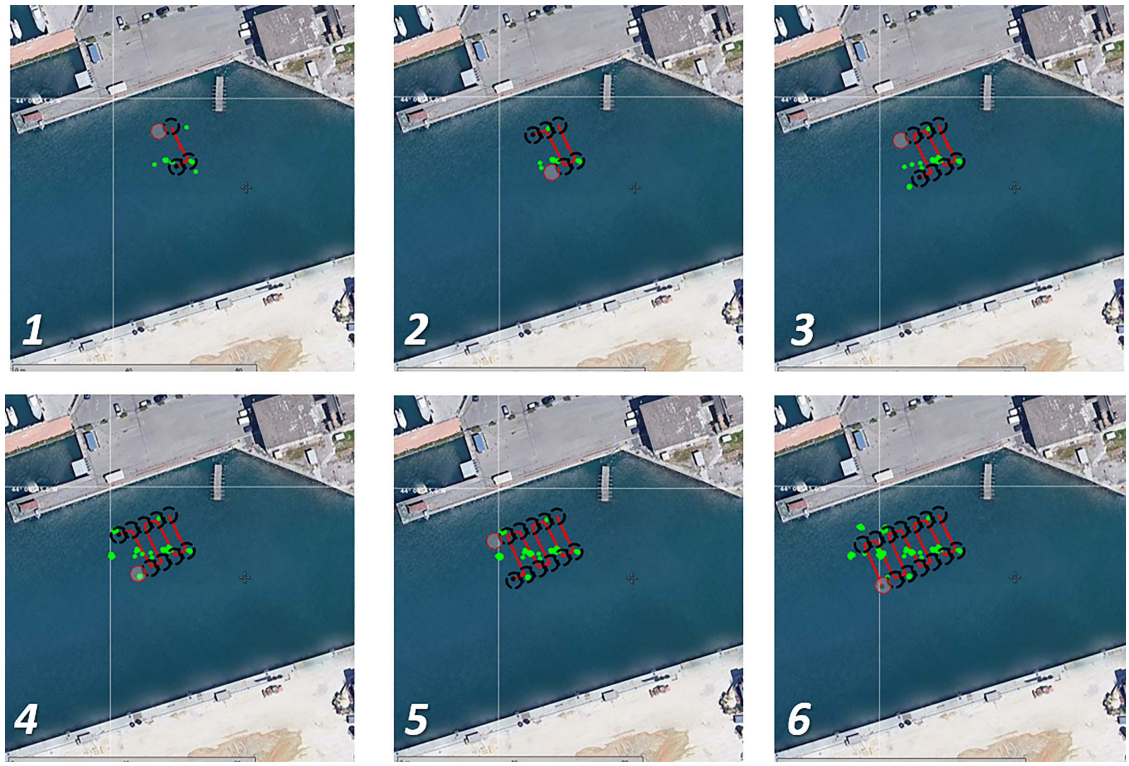


Fig. 15. Overview of the experimental campaign conducted in October 2020 at the CSSN, La Spezia, Italy. In a *red* line, the lawnmower path achieved by FeelHippo in a standard survey mission; every path waypoint is illustrated with a *black* dotted circle, where the current reference waypoint is highlighted in *red*; finally, the several OPI detections are shown in small *green* circles.

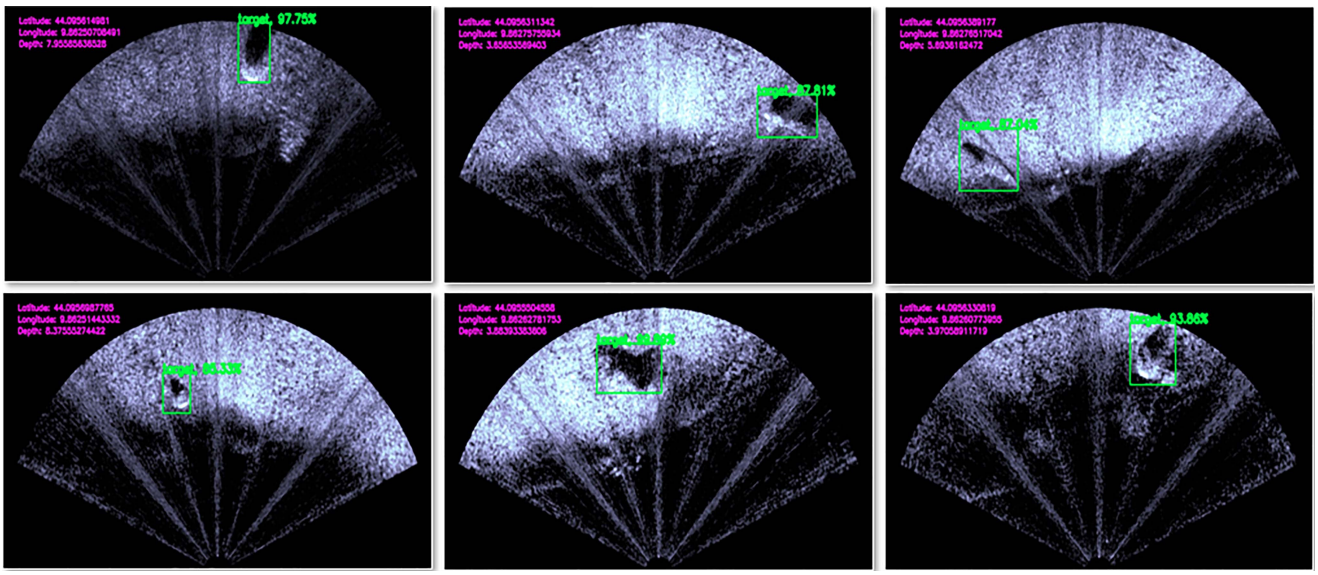


Fig. 16. Examples of the online detected and localized OPIs by means of the developed self-contained ATR with the SSD network during the experimental campaign conducted in October 2020 at the CSSN, La Spezia, Italy. In *purple* the localization outcomes in a latitude–longitude–depth representation whilst in *green* the bounding box traced around the OPIs as well as the class label and the detection probability.

false positives. In detail, 59 detections were true positives, while only two images were misidentified and were classified by the operator as false positives.

The resulting world model is shown in Fig. 17: the several OPI detections are handled by the described-above PMHA algorithm implementation over the FeelHippo AUV lawnmower motion. As summarized in Section V-A, the proposed world model represents the most probable hypothesis in the built multihypothesis tree. It is worth outlining the progressive update and change of the object positions (see Fig. 17) as long as new detections

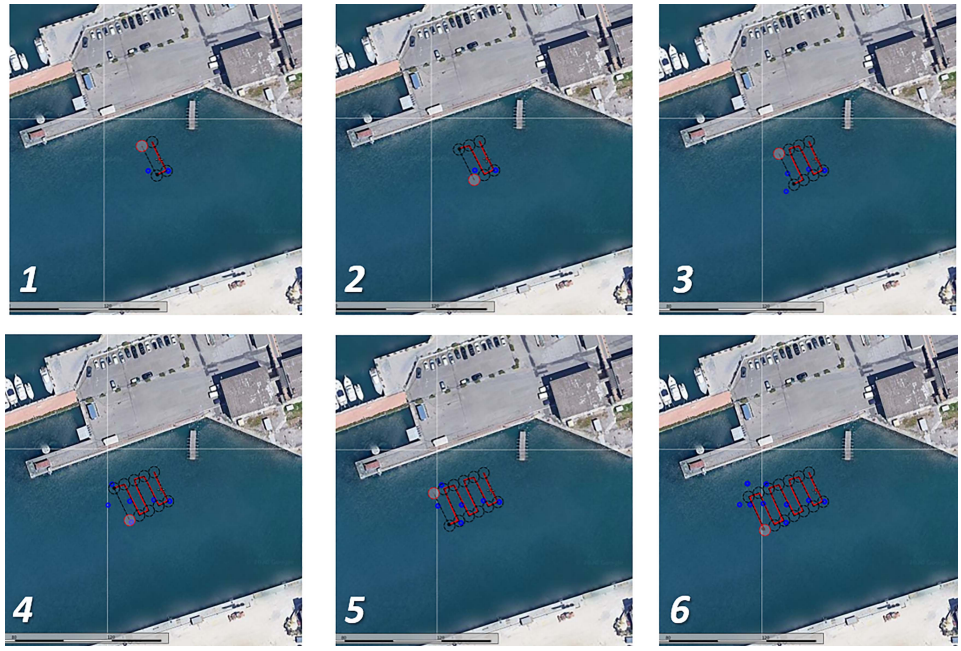


Fig. 17. Example of the world modeling process achieved by employing the PMHA algorithm during the experimental campaign conducted in October 2020 at the CSSN, La Spezia, Italy. Coherently with Fig. 15, in *red* line, the lawnmower path achieved by FeelHippo AUV; every path waypoint is illustrated with a *black* dotted circle, where the current reference waypoint is highlighted in *red*; finally, the OPIs representing the world model objects are described by small *blue* circles.

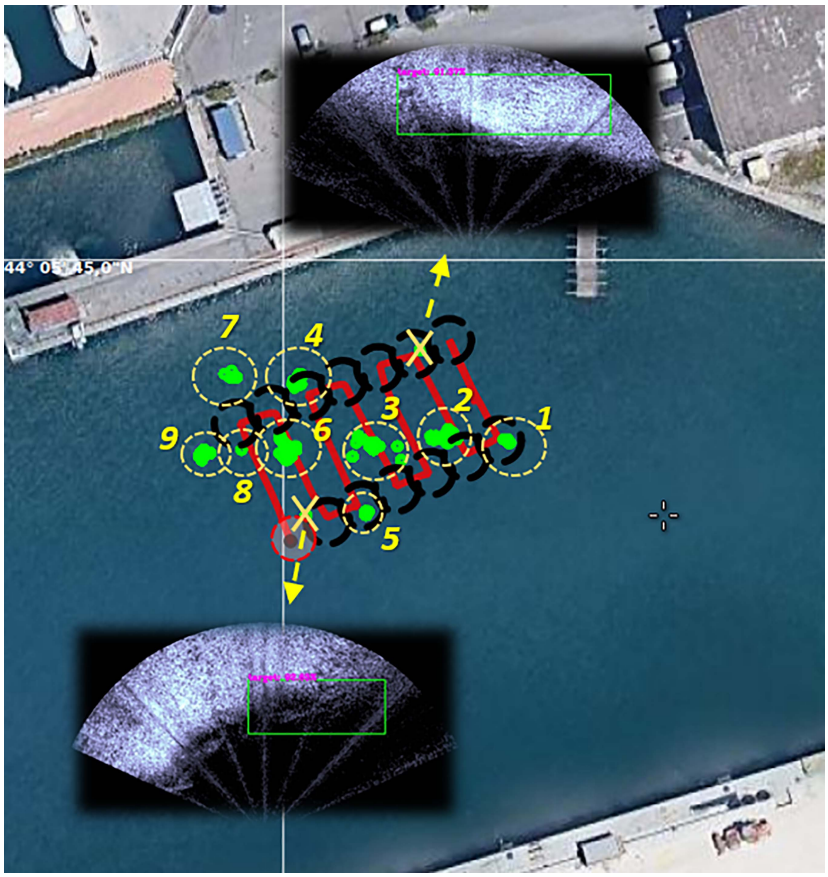


Fig. 18. Effect of the world modeling process achieved by employing the PMHA algorithm over the whole detection set during the experimental campaign. In dotted *yellow* circles, the data association outcomes are outlined with the dedicated OPI detections; finally, *yellow* crosses are employed to described rejected detections which, indeed, represent the two false OPI detections provided by the ATR architecture.

TABLE VI
PHMA OUTCOMES: NUMBER OF DETECTIONS FOR EACH OPI OBJECT AND RELATIVE COVARIANCE

OPIs	Number of detections	σ_x (m)	σ_y (m)
OPI1	3	0.17	0.13
OPI2	8	0.91	0.44
OPI3	10	1.46	0.88
OPI4	8	0.34	0.51
OPI5	5	0.13	0.14
OPI6	10	0.68	0.84
OPI7	6	0.44	0.38
OPI8	2	0.06	0.07
OPI9	7	0.52	0.62

are provided by the ATR architecture (see Fig. 15). More, in particular, as highlighted in Fig. 18, the whole set of detected OPIs has been clustered over a final world model comprised of nine geolocalized OPIs, which may be exploited for further more accurate inspections; Table VI reports the number of detections for each OPI object provided by the PMHA world modeling method along with the covariance with respect to a 2-D North–East reference system.

It is worth highlighting that the two OPI false detections (with a cross over them in Fig. 18) provided by the ATR strategy have been actually filtered by the world modeling procedure; indeed, standalone detections, due to the design guideline to establish the probability of a measurement representing a false detection higher than the probability of the measurement representing a new object, result as false detections in the PMHA algorithm outcome (i.e., most probable hypothesis). For the sake of completeness, since a precise OPI map of the underwater scenario has not been provided, a strict quantitative evaluation of the OPI localization metrics is not reported. However, the achieved world modeling procedure still represents a key hierarchical step in the multistage inspection process of unknown underwater scenarios. Indeed, providing a world representation (comprised of several OPI inspection points) is a preliminary but for sure fundamental task to achieve autonomous missions while actively engaging the surrounding environment.

VIII. CONCLUSION

This research work is an extension of [18] and presents an experimental campaign concerning a CNN-based ATR and world modeling architecture for AUVs with FLS acoustical frames.

Within the activities of the SEALab, the joint research laboratory between the CSSN of the Italian Navy and the Interuniversity Center of ISME, a data set constituted of 175 raw FLS images was, first, collected in May 2019 at the CSSN basin in La Spezia, Italy. This data set was used to train state-of-the-art CNN architectures that constituted the basis of the developed ATR solution. About the DNN-model picking, the SSD and Mask R-CNN have been selected as suitable solutions to supply accurate classification and localization outcomes.

Afterward, the proposed detection strategy was validated with a data set acquired on the same site in October 2019, where the targets were replaced and relocated. For both the CNN strategies, the detection results have been quantitatively evaluated in terms of classification precision and recall by exploiting an accurate analysis of the experimental environment provided by a human operator in a postprocessing stage. In particular, whereas the SSD has highlighted slightly better results regarding the detection of the underwater targets (approximately +5% recall performance), the Mask R-CNN has outlined an outperforming pattern in terms of precision (+24%) by avoiding false-positive misclassification. However, since this research work focuses on developing a self-contained ATR methodology capable of running onboard on compact AUVs, both the trained CNN models were also tested offline on an NVIDIA Jetson Nano, selected as dedicated payload hardware to achieve online detection performance. The SSD network managed to analyze up to 4 f/s, while the Mask R-CNN reached 0.5 f/s; as far as the FLS images were provided at 3 f/s, only the SSD met the online requirements and was considered during the sea trials.

Conversely, the world modeling architecture has been quantitatively validated by using a virtual replica of FeelHippo AUV within a custom-designed implementation of a UUV Simulator scenario. A realistic lawnmower path has been considered to survey over a predefined inspection area and search for an unknown number of targets (defined as semi-spheres) with a simulated FLS sensor. The achieved outcomes, with an average distance error of 0.73 meters, do provide a valid argument for the correctness of the world modeling procedure.

Finally, in October 2020, the proposed ATR architecture and environment modeling solution were validated online through suitable underwater missions. In this context, the testing framework incorporates three components: FeelHippo AUV, which real-time collected the FLS images and performing the OPI detection and localization, an autonomous moving buoy, exploited as an acoustical link, and a workstation, comprised of a laptop PC and a Wi-Fi bullet, where the data association algorithm took place. Thanks to the developed ATR methodology, FeelHippo AUV was able to identify and localize online onboard during the survey the present OPIs that were communicated to the workstation through the acoustical channel created with the autonomous buoy. The transmitted ATR results were employed to create by means of a PMHA data association and world modeling methodology a representation of the surveyed world alongside the position, a unique class label, and relative uncertainties of the detected OPIs.

As a summarizing statement, the achieved results have highlighted the capability for the proposed architecture to autonomously inspect an unknown underwater scenario by effectively detecting localize targets of interest and realizing a multiobject world model.

Future works will include the employment of the proposed ATR and data association strategy in an overall intelligent system leading the vehicle to actively exploit the achieved world model (i.e., performing motion planning to navigate towards them). Besides, the use of the ATR methodology for fully autonomous seabed inspections or manipulation tasks will be surely investigated. Additionally, the employment of the autonomous navigation buoy to enhance the FeelHippo AUV self-localization, utilizing the USBL acoustical device, and the target geolocalization accuracy will furtherly be taken into account and accurately deepened.

Acknowledgment

The authors would like to thank all the SEALab members, the joint applied research laboratory between the Naval Support and Experimentation Centre (CSSN) of the Italian Navy and the Italian Interuniversity Research Center of Integrated Systems for Marine Environment (ISME), who helped the research team during the missions at sea performed in La Spezia.

References

[1] M. Prats et al., "Reconfigurable AUV for intervention missions: A case study on underwater object recovery," *Intell. Serv. Robot.*, vol. 5, no. 1, pp. 19–31, 2012.

[2] G. Ferri, F. Ferreira, and V. Djapic, "Multi-domain robotics competitions: The CMRE experience from SAUC-E to the European robotics league emergency robots," in *Proc. IEEE OCEANS Conf.*, Aberdeen, U.K., 2017, pp. 1–7.

[3] M. Cashmore, M. Fox, T. Larkworthy, D. Long, and D. Magazzeni, "AUV mission control via temporal planning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 6535–6541.

[4] E. Vidal, N. Palomeras, K. Istenič, J. D. Hernández, and M. Carreras, "Two-dimensional frontier-based viewpoint generation for exploring and mapping underwater environments," *Sensors*, vol. 19, no. 6, 2019, Art. no.1460.

[5] L. Paull, S. Saeedi, M. Seto, and H. Li, "Sensor-driven online coverage planning for autonomous underwater vehicles," *IEEE/ASME Trans. Mechatronics*, vol. 18, no. 6, pp. 1827–1838, Dec. 2013.

[6] L. Zacchini, A. Ridolfi, and B. Allotta, "Receding-horizon sampling-based sensor-driven coverage planning strategy for AUV seabed inspections," in *Proc. IEEE/OES Auton. Underwater Veh. Symp.*, 2020, pp. 1–6.

[7] É. Pairet, J. D. Hernández, Y. Petillot, and M. Lahijanian, "Online mapping and motion planning under uncertainty for safe navigation in unknown environments," 2020, *arXiv:2004.12317*.

[8] M. Franchi, A. Bucci, L. Zacchini, E. Topini, A. Ridolfi, and B. Allotta, "A probabilistic 3D map representation for forward-looking SONAR reconstructions," in *Proc. IEEE/OES Auton. Underwater Veh. Symp.*, 2020, pp. 1–6.

[9] D. Youakim, P. Cieslak, A. Dornbush, A. Palomer, P. Ridao, and M. Likhachev, "Multirepresentation, multiheuristic A* search-based motion planning for a free-floating underwater vehicle-manipulator system in unknown environment," *J. Field Robot.*, vol. 37, no. 6, pp. 925–950, 2020.

[10] E. Simetti et al., "Sea mining exploration with an UVMS: Experimental validation of the control and perception framework," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 3, pp. 1635–1645, Jun. 2021.

[11] R. B. Wynn et al., "Autonomous underwater vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience," *Mar. Geol.*, vol. 352, pp. 451–468, 2014.

[12] M. A. Moline, D. L. Woodruff, and N. R. Evans, "Optical delineation of benthic habitat using an autonomous underwater vehicle," *J. Field Robot.*, vol. 24, no. 6, pp. 461–471, 2007.

[13] B. Allotta, R. Costanzi, A. Ridolfi, M. Reggiannini, M. Tampucci, and D. Scaradozzi, "Archaeology oriented optical acquisitions through MARTA AUV during ARROWS European project demonstration," in *Proc. IEEE/MTS OCEANS Conf.*, Monterey, CA, USA, 2016, pp. 1–4.

[14] D. Terracciano, L. Bazzarello, A. Caiti, R. Costanzi, and V. Manzari, "Marine robots for underwater surveillance," *Curr. Robot. Rep.*, vol. 1, pp. 159–167, 2020.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[16] M. Valdenegro-Toro, "Object recognition in forward-looking sonar images with convolutional neural networks," in *Proc. IEEE/MTS OCEANS Conf.*, Monterey, CA, USA, 2016, pp. 1–6.

[17] L. Zacchini et al., "Deep learning for on-board AUV automatic target recognition for optical and acoustic imagery," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 14589–14594, 2020.

[18] L. Zacchini et al., "Forward-looking sonar CNN-based automatic target recognition: An experimental campaign with FeelHippo AUV," in *Proc. IEEE/OES Auton. Underwater Veh. Symp.*, 2020, pp. 1–6.

[19] M. Abadi et al., "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation*, 2016, pp. 265–283.

[20] M. Quigley et al., "ROS: An open-source robot operating system," *ICRA Workshop Open Source Softw.*, vol. 3, no. 3.2, 2009.

[21] NVIDIA, "NVIDIA Jetson Nano," 2018, [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit

[22] B. Allotta et al., "A low cost autonomous underwater vehicle for patrolling and monitoring," *Proc. Inst. Mech. Eng. M, J. Eng. Maritime Environ.*, vol. 231, no. 3, pp. 740–749, 2017.

[23] M. Franchi, A. Ridolfi, and B. Allotta, "Underwater navigation with 2D forward looking SONAR: An adaptive unscented Kalman filter-based strategy for AUVs," *J. Field Robot.*, vol. 38, no. 3, pp. 355–385, 2021.

[24] A. Meschini et al., "Design of a self-moving autonomous buoy for the localization of underwater targets," in *Proc. IEEE Marseille*., 2019, pp. 1–6.

[25] V. Myers and J. Fawcett, "A template matching procedure for automatic target recognition in synthetic aperture sonar imagery," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 683–686, Jul. 2010.

[26] Y. Petillot, Y. Pailhas, J. Sawas, N. Valeyrie, and J. Bell, "Target recognition in synthetic aperture and high resolution side-scan sonar," in *Proc. Eur. Conf. Underwater Acoust.*, 2010, pp. 99–106.

[27] E. Galceran, V. Djapic, M. Carreras, and D. P. Williams, "A real-time underwater object detection algorithm for multi-beam forward looking sonar," *IFAC Proc. Vol.*, vol. 45, no. 5, pp. 306–311, 2012.

[28] F. Ferreira, V. Djapic, M. Micheli, and M. Caccia, "Improving automatic target recognition with forward looking sonar mosaics," *IFAC Proc. Vol.*, vol. 47, no. 3, pp. 3382–3387, 2014.

[29] M. Dos Santos, P. O. Ribeiro, P. Núñez, P. Drews, Jr., and S. Botelho, "Object classification in semi structured environment using forward-looking sonar," *Sensors*, vol. 17, no. 10, 2017, Art. no. 2235.

[30] D. M. Montserrat, Q. Lin, J. Allebach, and E. J. Delp, "Training object detection and recognition CNN models using data augmentation," *Electron. Imag.*, vol. 2017, no. 10, pp. 27–36, 2017.

[31] J. Talukdar, S. Gupta, P. Rajpura, and R. S. Hegde, "Transfer learning for object detection using state-of-the-art deep neural networks," in *Proc. IEEE 5th Int. Conf. Signal Process. Integr. Netw.*, 2018, pp. 78–83.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[33] I. Kvasić, N. Mišković, and Z. Vukić, "Convolutional neural network architectures for sonar-based diver detection and tracking," in *Proc. IEEE OCEANS Conf.,-Marseille*, 2019, pp. 1–6.

[34] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

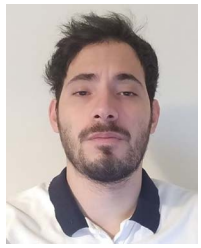
[36] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[37] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[39] J. Huang et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3296–3297.
[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[41] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with Fast R-CNN," in *Proc. IEEE/MTS OCEANS Conf.*, Washington, DC, USA, 2015, pp. 1–5.
[42] J. Elfring, S. van den Dries, M. Van De Molengraft, and M. Steinbuch, "Semantic world modeling using probabilistic multiple hypothesis anchoring," *Robot. Auton. Syst.*, vol. 61, no. 2, pp. 95–105, 2013.
[43] N. Blodow, D. Jain, Z.-C. Marton, and M. Beetz, "Perception and probabilistic anchoring for dynamic world state logging," in *Proc. IEEE-RAS 10th Int. Conf. Humanoid Robots*, 2010, pp. 160–166.
[44] M. Baum, I. Gheţa, A. Belkin, J. Beyerer, and U. D. Hanebeck, "Data association in a world model for autonomous systems," in *Proc. IEEE Conf. Multisensor Fusion Integration*, 2010, pp. 187–192.
[45] T. De Laet, "Rigorously Bayesian multitarget tracking and localization (rigoureus Bayesiaans detecteren en volgen van meerdere objecten)," 2010.
[46] R. P. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.
[47] D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979.
[48] Y. Bar-Shalom, P. K. Willett, and X. Tian, *Tracking and Data Fusion*. Storrs, CT, USA: YBS, 2011.
[49] X. R. Li and Y. Bar-Shalom, "Tracking in clutter with nearest neighbor filters: Analysis and performance," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 3, pp. 995–1010, Jul. 1996.
[50] A. Mallios, P. Ridao, D. Ribas, and E. Hernández, "Scan matching SLAM in underwater environments," *Auton. Robots*, vol. 36, no. 3, pp. 181–198, 2014.
[51] D. Ribas, P. Ridao, J. D. Tardós, and J. Neira, "Underwater SLAM in man-made structured environments," *J. Field Robot.*, vol. 25, no. 11/12, pp. 898–921, 2008.
[52] A. Mallios, P. Ridao, D. Ribas, M. Carreras, and R. Camilli, "Toward autonomous exploration in confined underwater environments," *J. Field Robot.*, vol. 33, no. 7, pp. 994–1012, 2016.
[53] J. Melo and S. Dugelay, "AUV mapping of underwater targets," in *Proc. IEEE/MTS OCEANS Conf.*, Seattle, WA, USA, 2019, pp. 1–6.
[54] R. M. Rogers, *Applied Mathematics in Integrated Navigation Systems*. Reston, VA, USA: Amer. Inst. Aeronaut. Astronaut., 2007.
[55] L. Zacchini, V. Calabrò, M. Candeloro, F. Fanelli, A. Ridolfi, and F. Dukan, "Novel noncontinuous carouseling approaches for MEMS-based north seeking using Kalman filter: Theory, simulations, and preliminary experimental evaluation," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 5, pp. 2437–2448, Oct. 2020.
[56] T. I. Fossen, *Guidance and Control of Ocean Vehicles*. Chichester, U.K.: Wiley, 1994.
[57] R. Costanzi, F. Fanelli, N. Monni, A. Ridolfi, and B. Allotta, "An attitude estimation algorithm for mobile robots under unknown magnetic disturbances," *IEEE/ASME Trans. Mechatronics*, vol. 21, no. 4, pp. 1900–1911, Aug. 2016.
[58] B. Allotta et al., "A new AUV navigation system exploiting unscented Kalman filter," *Ocean Eng.*, vol. 113, pp. 121–132, 2016.
[59] S. Negahdaripour, "On 3D motion estimation from feature tracks in 2D FS sonar video," *IEEE Trans. Robot.*, vol. 29, no. 4, pp. 1016–1030, Aug. 2013.
[60] N. Hurtós, D. Ribas, X. Cufí, Y. Petillot, and J. Salvi, "Fourier-based registration for robust forward-looking SONAR mosaicing in low-visibility underwater environments," *J. Field Robot.*, vol. 32, no. 1, pp. 123–151, 2015.
[61] N. Hurtós Vilarnau et al., "Forward-looking sonar mosaicing for underwater environments," 2014.
[62] F. Ferreira, V. Djapic, M. Micheli, and M. Caccia, "Forward looking sonar mosaicing for mine countermeasures," *Annu. Rev. Control*, vol. 40, pp. 212–226, 2015.
[63] 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 *Proc. IEEE/MTS OCEANS Conf.*, Monterey, CA, USA, 2016, pp. 1–8.

Leonardo Zacchini (Member, IEEE) received the M.S. degree in automation and control engineering and the Ph.D. degree with a thesis focused on autonomous inspection strategies for underwater robots from the University of Florence, Florence, Italy, in 2018 and 2021, respectively.

He is currently a Postdoctoral Researcher in Robotics with the University of Florence, Italy. His research interests include guidance, navigation, and control systems for mobile robots, underwater robotics, robotics exploration, motion planning, and AI for robotics.



Alberto Topini (Student Member, IEEE) received the M.Sc. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K., in 2019. He is currently working toward the Ph.D. degree with the Department of Industrial Engineering, University of Florence, Florence, Italy.

His current research interests include decision-making, AI planning, probabilistic world modeling, and deep learning-based automatic target recognition methods for underwater robotics.

Matteo Franchi received the bachelor's degree in mechanical engineering, the master's degree in electrical and automation engineering, and the doctoral degree in industrial and reliability engineering from the School of Engineering, University of Florence, Florence, Italy.

He was a Postdoctoral Researcher within the Mechatronics and Dynamic Modeling Laboratory of the Department of Industrial Engineering, University of Florence. His research interests focus on localization, mapping, and control systems for autonomous underwater vehicles. Moreover, his work included forward-looking SONARs.

Nicola Secciani received the bachelor's degree in mechanical engineering, the master's degree in electrical and automation engineering, and the doctoral degree in industrial and reliability engineering from the School of Engineering, University of Florence, Florence, Italy.

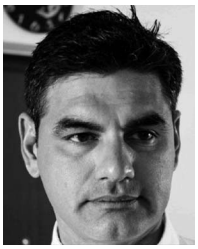
He is currently a Postdoctoral Researcher within the Mechatronics and Dynamic Modeling Laboratory of the Department of Industrial Engineering, University of Florence. His research interests focus on control strategies based on electromyographic signal classification for wearable robotic devices.

Vincenzo Manzari received the M.Sc. degree in telecommunications engineering in 2013 and the Ph.D. degree in underwater robotics both from the University of Pisa, Pisa, Italy.

He joined the Italian Navy in 2006 and he completed the Naval Academy of Livorno in 2010. Since 2015, he has been an Engineer Officer with the Naval Support and Experimentation Centre of the Italian Navy in the Autonomous Systems branch. His main research interests lie in increasing the autonomy of underwater operations.

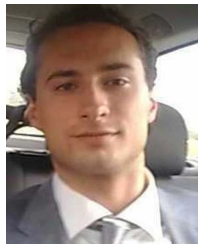
Lorenzo Bazzarello received the M.Sc. degree in telecommunications engineering in 2013 and the postgraduate degree in underwater electroacoustics and application from the University of Pisa, Pisa, Italy, where he is currently working toward the Ph.D. degree in information engineering under the program in Underwater Electroacoustics and Robotics.

He joined the Italian Navy in 2006 and completed the Naval Academy of Livorno in 2010. From 2013 to 2019, he was a part of the research branch in the Mine Counter Measure Headquarter of the Italian Navy. In 2015, he attended the course in Mine Counter Measure with the Naval Academy of Livorno. He is an Engineer Officer with the Naval Support and Experimentation Centre of the Italian Navy, La Spezia, Italy. His research interests include high-frequency imaging sonar, acoustic measurement techniques, electroacoustics, and artificial intelligence applied to mine countermeasure and underwater robotics.



Mirko Stifani received the postgraduate degree in underwater electroacoustics and its applications from the University of Pisa, Pisa, Italy, in 2005.

He joined Italian Navy in 1993 and he completed the Naval Academy of Livorno in 2000. He is currently the Chief of the Underwater Warfare Office with the Naval Support and Experimentation Centre of the Italian Navy, La Spezia, Italy, and the Director of the SEALab, La Spezia.



Alessandro Ridolfi (Senior Member, IEEE) received the Ph.D. degree in industrial engineering from the University of Florence, Florence, Italy, in 2014.

He is currently a Ph.D. Researcher and an Assistant Professor of machine theory and robotics with the School of Engineering, Department of Industrial Engineering, University of Florence, Italy. His current research interests include biorobotics, vehicle dynamics, mechanical systems modeling, robotics, and underwater robotics.