Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 53-2 (2020) 14589–14594

# Deep Learning for on-board AUV Automatic Target Recognition for Optical and Acoustic imagery

**Leonardo Zacchini** *,** **Alessandro Ridolfi** *,**
**Alberto Topini** *,** **Nicola Secciani** *,** **Alessandro Bucci** *,**
**Edoardo Topini** *,** **Benedetto Allotta** *,**

* *Department of Industrial Engineering, University of Florence, via di Santa Marta 3, 50139, Florence, Italy (e-mail: leonardo.zacchini@unifi.it).*

** *Interuniversity Center of Integrated Systems for the Marine Environment (ISME), www.isme.unige.it*

**Abstract:** In the widespread field of underwater robotics applications, the demand for increasingly intelligent vehicles is leading to the development of Autonomous Underwater Vehicles(AUVs) with the capability of understanding and engaging the surrounding environment. Consequently, to push the boundaries of cutting-edge smart AUVs, the automatic recognition of targets is becoming one of the most investigated topics and Deep Learning-based strategies have shown astonishing results. In the context of this work, two different neural network architectures, based on the Single Shot Multibox Detector (SSD) and on the Faster Region-based Convolutional Neural Network (Faster R-CNN), have been trained and validated, respectively, on optical and acoustic datasets. In particular, the models have been trained with the images acquired by FeelHippo AUV during the European Robotics League (ERL) competition, which took place in La Spezia, Italy, in July 2018. The proposed ATR strategy has then been validated with FeelHippo AUV in an on-board post-processing stage by exploiting the images provided by both a 2D Forward Looking Sonar (FLS) as well as an IP camera mounted on-board on the vehicle.

Copyright © 2020 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0)

*Keywords:* Marine Robotics, Artificial Intelligence, Automatic Target Recognition, Autonomous Underwater Vehicles, Neural Networks, Intelligent Robotics, Machine learning for environmental applications.

## 1. INTRODUCTION

Since the development of the first underwater robots, the demanded tasks for subsea operations have become more and more challenging. Pre-programmed missions, whose goal was to passively inspect and monitor areas of interest, have been being slowly but steadily supplanted by interactive tasks (e.g. recovering a target on the seafloor (Prats et al. (2012))); as a result, the need of intelligent vehicles, capable of actively and physically engaging the surrounding environment and selecting the optimal action to be performed (Vidal et al. (2019), Cashmore et al. (2014)), has been arisen and still plays a crucial role in the evolution of cutting-edge underwater system technologies. Nowadays, within the context of intelligent vehicle development, Automatic Target Recognition (ATR) is emerging as one of the most investigated topics by the scientific and industrial community. Indeed, understanding and gathering knowledge of the environment represents a preliminary and fundamental hierarchical stage for effectively accomplishing interactive tasks, creating fully autonomous robots that help human operators in challenging assignments.

Over the last few years, Deep Learning (DL) techniques have achieved significant success in digital image processing by fulfilling the object detection and classification tasks in increasingly challenging environments and scenarios. As a consequence, DL has recently resulted as the state-of-the-art approach in performing ATR by means of highly nonlinear feature extraction. Meanwhile, as far as marine robotics is concerned, there has been an exponential growth in the collection of underwater imagery for monitoring the subsea ecosystems. In light of the above-mentioned considerations, DL classifiers show the potential to address the automated object recognition task in the underwater environment and can be the key to outperform the previous feature-based approaches in terms of accuracy and robustness (Lowe (2004), Krizhevsky et al. (2012)).

Autonomous Underwater Vehicles (AUVs) are commonly equipped with several payload sensors, including cameras and sonars, with the aim of perceiving and inspecting the subsea environment. In particular, although modern cameras provide high-resolution images, optical data have the non-negligible drawback to significantly degrade in the presence of turbid water and low-light conditions. Conversely, acoustic sensors, such as Forward-Looking Sonar

2405-8963 Copyright © 2020 The Authors. This is an open access article under the CC BY-NC-ND license.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2020.12.1466

(FLS) or Side Scan Sonar (SSS), supply lower resolution, high-noise images with a wide range of coverage. As a result of the highlight patterns, several studies have been proposed in order to extend the traditional object recognition DL techniques to underwater scenarios by exploiting acoustic images (Kvasic et al. (2019), Valdenegro-Toro (2016)).

In the context of this work, a DL-based ATR architecture has been designed and implemented by using camera as well as sonar frames; firstly, the research activity has focused on evaluating the performance of Convolutional Neural Networks (CNNs) on visual and FLS recordings and, subsequently, the feasibility of the aforementioned system has been verified during real-time tests. More in detail, the first step of the proposed strategy consists in the training of CNN models by exploiting a custom gathered dataset of heterogeneous images and the open source machine learning library TensorFlow (2015): a large image dataset has been collected, pre-processed and labeled in order to train the SSD model (Liu et al. (2016)) and Faster R-CNN proposed by Girshick (2015). Afterward, the trained neural networks have been incorporated in a custom ATR software, developed in the Robot Operating System framework (ROS (2007)), the most commonly used set of software libraries and tools to build robot applications. All the presented results have been validated by sea trials conducted with FeelHippo AUV , one of the vehicles developed by Mechatronics and Dynamic Modeling Laboratory (MDM Lab) (Allotta et al. (2017), Allotta et al. (2015)) of the Department of Industrial Engineering of the University of Florence (UNIFI DIEF). According to the achieved experimental results, accurately presented and described in the paper, the proposed strategy arises as a noteworthy validation proof of the effectiveness of DL methodologies for accomplishing ATR tasks in subsea scenarios; indeed, the suggested approach succeeded in detecting and recognizing several targets on the seafloor.

This paper is organized as follows. Section 2 describes underwater ATR systems and reviews the most used CNN architectures in this field. Section 3 is dedicated to the description of the proposed methodology by accurately outlining the DL model training process. Section 4 presents the experimental results obtained by collecting data during a sea mission and processing them offline on the vehicle hardware. Finally, Section 5 summarizes the presented research by focusing on the major achieved results; furthermore, a brief description of the future trends is illustrated.

## 2. UNDERWATER ATR - STATE OF THE ART

A breakthrough object detection solution, based on modern CNNs, was proposed in Girshick (2015). Fast R-CNN processes the input image with convolutional and max-pooling layers to produce a set of Region of Interest (RoI). Each RoI is then fed into fully connected layers that branch into two sibling layers; one branch is in charge of classifying possible objects in the RoI, while the other one has to compute the corresponding bounding boxes. For the development of the You Only Look Once (YOLO) network Redmon et al. (2016), a different approach, based on optimized end-to-end networks, was chosen. Composed of 24 convolutional layers and 2 fully connected layers, YOLO achieves real-time image processing with an extremely high frame per second (fps) by predicting bounding boxes and class probabilities directly from full images in one evaluation.

The Single Short Multibox Detector (SSD) (Liu et al. (2016)) is a noteworthy real-time solution, constituted of convolutional layers, which predicts category scores and box offsets, making use of different predictors for different aspect ratio detections, for a fixed set of default bounding boxes using small convolutional filters applied to feature maps. This structure allowed the SSD to reach high-accuracy detections at high fps.

These ATR solutions have been tested and compared on several common datasets (Liu et al. (2016)). Region-based solutions, such as the Fast RCNN and the improved versions, are the more accurate networks but cannot reach extremely high inference speed. On the other hand, SSD and YOLO can work up to 45 fps but with reduced precisions. In particular, SSD is more accurate than the previous state-of-the-art for real-time single-shot detectors, such as YOLO.

The above-mentioned algorithms were designed to work with images and, since AUVs are commonly endowed with optical cameras either to acquire data or to aid the navigation algorithms (Salvi et al. (2008), Zacchini et al. (2019)), they can potentially be used to detect objects in the underwater environment. For instance, the Fast R-CNN network was successfully tested in the underwater domain in Xiu Li et al. (2015), where it was used to detect an recognize fish species in optical images. However, underwater object detection is an extremely challenging task due to water turbidity as well as illumination conditions. In fact, underwater images are affected by color distortion and low visibility, caused by the exponentially light attenuation while it penetrates through the water. Since the CNN-based algorithms make use of the input image color components, their results are biased by the color distortion. This aspect was investigated in Kolaman et al. (2019), where a novel light invariant video imaging system (LIVI) was proposed. The LIVI system managed to neutralize the effect of changing light condition, as could happen in the underwater applications, and increased the detection performance. However, the approach proposed in Kolaman et al. (2019) needed ad-hoc hardware to modulate the light source. Driven by these considerations, several software image enhancement techniques, such as Red Dark Channel Prior (Cheng et al. (2015)) and Contrast Limited Adaptive Histogram Equalization (CLAHE) (Ma et al. (2017)), could be used in order to increase the CNNs detection accuracy. Moreover, since the image information content can be extracted from the local contrast, the strategies, which work on it, should be preferred. For instance, CLAHE approach works on rectangular subregions instead of on the entire image, with the aim of realizing a local equalization. Furthermore, CLAHE algorithm is acceptable for real-time underwater applications thanks to its reduced computational cost and its adaptivity to different working conditions.

Although the image quality can be enhanced, the problem of the camera low range visibility caused by the water turbidity cannot be overcome. As a consequence, acoustic

sensors are commonly exploited in underwater applications as large scale mosaicing of unknown areas (Franchi et al. (2018)). Indeed, sonars can acquire acoustic images with various ranges, depending on the water and environment conditions, and they are not influenced by illumination conditions. However, acoustic images can be quite noisy and lack in details, making acoustic-based objects detection challenging. SSSs acquire cross-track slices of acoustic reflections that are combined along the direction of motion to create an image of the sea-bottom. Such sonars provide long-range high-resolution images that allow detecting objects in large survey areas. FLSs can provide good resolution images, more detailed than SSSs, but at shorter distances, at high frame rates. Additionally, the vehicle is not required to move to create an image.

The use of CNNs to classify FLS images was investigated in Valdenegro-Toro (2016), where the author proposed a performance comparison between CNNs and classical template-matching approaches. Finally, in Kvasic et al. (2019), modern CNN object detector architectures, modified versions of YOLO, were tested for diver detection in acoustic images, acquired by an FLS.

## 3. DEEP NEURAL NETWORK TRAINING

DL-based approaches for computer vision applications usually rely on CNN models trained with high-resolution images; as a matter of fact, such computationally expensive strategies aprioristically exclude the possibility for an ATR architecture to be performed in real-time. For this reason, since carrying out ATR while AUV navigating was the major purpose of the project, the focus has shifted to the SSD and Fast R-CNN architectures which guarantee the required trade-off between high-standard inference performance and the feasibility for real-time implementation. More in detail, the SSD network has been selected to fulfill a high-FPS recognition task with optical images. On the other hand, since the acoustic frames were captured with a lower frame-rate (3 Hz), the *mean Average Precision* (mAP) has been favored as model selection metric over the inference speed; as a consequence, Faster R-CNN has been preferred to faster but less accurate DL structures, as YOLO. Furthermore, since this research emphasizes on a practical application of state-of-the-art CNN techniques rather than a theoretical disquisition, using pre-trained model weights has resulted as the optimal solution in terms of learning and convergence timings. Moreover, from a practical perspective, since the process of gathering a large dataset in an underwater scenario is by no means straightforward, exploiting transfer learning, by fine-tuning higher-order feature representations, allows to remarkably speed up the training phase. As far as the specific selections are concerned, the SSDMobileNet v2 (Liu et al. (2016)) and the Faster R-CNN Inception v2 (Girshick (2015)) networks have been adopted to process, respectively, the optical and acoustic images.

Since the DL algorithm performance strongly depends on the training process, the training dataset is of utmost importance to achieve high-precision solutions; the larger and more heterogeneous the dataset is, the more accurate the network will be in detecting objects in new unseen images. Acquiring underwater data could be challenging and extremely expensive because of unexpected variations of the environmental conditions as well as unavoidably logistic constraints. Nonetheless, modern open-source machine learning platforms, such as Tensorflow, provide several data augmentation options able to create new images by modifying the existing ones and emerge as useful tools to increase the dataset dimension. Thus, plenty of images are not required and the underwater dataset gathering does result as a demanding but feasible task.

Fig. 1. The pipeline structure deployed in the basin in La Spezia, Italy, for the ERL 2018 challenge.

Within the scope of this contribution, the training dataset was acquired with FeelHippo AUV, described in Section 4, during sea trials at the European Robotics League Emergency 2018 (Ferri et al. (2017)), which took place in La Spezia (Italy). Optical images were acquired with the downward-pointing IPCam ELP 720p, whilst acoustic images were provided by the 2D FLS Teledyne Blueview M900. The ERL Challenge was structured so as to address a simulated yacht accident in the basin of the NATO Science and Technology Organization Centre for Maritime Research and Experimentation (CMRE); as partial task of a more complex mission, the AUVs were required to automatically recognize a damaged pipeline (represented by a yellow pipeline with attached a red marker) alongside a whole pipeline structure (Fig. 1) assembled by the aforementioned damaged pipeline as well as other five pipelines. To accomplish such tasks, the camera was used to detect both the marker and the pipelines, whilst the FLS was exploited for structure detection. Hence, 500 optical images of the pipelines and the red marker, with a resolution of 704×576 pixels, constituted the SSD training dataset. In particular, before the training stage, the camera frames were processed by means of the CLAHE algorithm with the aim of enhancing the image sharpness and contrast. As reported above in this section, to build a robust training dataset, data augmentation options of the Tensoflow framework were used the training process. In particular, optical images were randomly horizontally and vertically flipped and randomly cropped, and image values, such as brightness, contrast, hue and saturation, were randomly modified. These augmentation options have been selected, considering that the AUV was supposed to detect interesting objects while it was performing different inspection surveys in different environmental conditions. The Faster

R-CNN was, instead, trained with a dataset made of 200 acoustic images, acquired by the acoustic sonar in a naive resolution of 894×477 pixels, depicting the structure. To further increase the dataset, data augmentation options of the Tensorflow framework were used: the dataset was augmented by randomly horizontally flipping the images and randomly varying their brightness.

The size configuration of the input images differs between SSD and Faster R-CNN; whereas the former relies on a fixed shape image resizing, the latter is trained by a shorter edge-based image scaling strategy. Concerning the camera frames, the images have been down-scaled (352×288) so as to trade inference accuracy for more efficient processing speed. On the other hand, due to the low-resolution and low-frame rate features of the sonar pictures, the Faster R-CNN training pipeline has been configured such that the image dimensions, as well as the aspect ratio, are maintained in order to prioritize the classification performance over the reduction of the computational cost. Turning to the optimizer selection and batch size design, the SSD model has been trained by using RMSProp (Tieleman and Hinton (2012)) with batch sizes of 24 whilst Faster R-CNN has exploited Stochastic Gradient Descent (SGD) with momentum with batch sizes of 1; for sake of completeness, the single-image batch size solution for Faster R-CNN has built upon the fact that this CNN architecture utilizes images with different sizes within the training stage. Finally, for each CNN architecture the learning rate schedules have been custom-defined so as to guarantee a fast convergence while achieving optimal inference results. The whole training process has been performed on a PC fitted with 32GB RAM, an Intel Core i7-8750H processor and an Nvidia GeForce GTX 1070 Ti card.

The loss function, which was defined as a weighted sum of the partial classification and localization losses, was minimized during the training stages of both the SSD and the Faster R-CNN networks. Turning to an analysis of the *mAP* and FPS metrics, evaluated on the validation dataset (provided by holding out 10% of the whole training dataset), the CNN architecture performances are highlighted in Table 1. The achieved outcomes reflect the aforestated considerations on the CNN model design selection: whereas the Faster R-CNN Inception v2 slightly outperforms SSDMobileNet v2 in terms of *mAP*, the acoustic image inference stage is considerably slower than the respective optical frame value.

Table 1. CNN Performance Indicator Score

| **Model** | **mAP** | **FPS** |
|---|---|---|
| SSDMobileNet v2 | 0.831 | 5.0 |
| Faster R-CNN Inception v2 | 0.851 | 1.0-1.1 |

## 4. EXPERIMENTAL TESTS AND RESULTS

The developed ATR strategy has been tested in the context of European Robotics League (ERL) (Ferri et al. (2017)), a robotic tournament for European university teams, in which the University of Florence has participated with FeelHippo AUV (Allotta et al. (2017)) (see Fig. 2). FeelHippo AUV, developed by the Department of Industrial Engineering of the University of Florence (UNIFI DIEF), is a light-weight and compact vehicle provided with the capability to actively perform complex tasks and missions. The main properties of FeelHippo AUV are reported in Table 2.

Fig. 2. FeelHippo AUV during a sea trial.

Table 2. FeelHippo AUV Main Features

| | |
|---|---|
| Weight [kg] | 35 |
| Dimensions [mm] | 600×640×500 |
| Maximum Depth [m] | 30 |
| Maximum Longitudinal Speed [m/s] | 1 |
| Battery Life [h] | 3 |
| Controlled DOFs | 4 |

Additionally, the electronic devices and the sensor set mounted on board are listed as follows:

- U-blox 7P precision Global Positioning System (GPS);
- Orientus Advanced Navigation Attitude Heading Reference System (AHRS);
- KVH DSP 1760 single-axis high precision Fiber Optic Gyroscope (FOG);
- Nortek DVL1000 DVL, measuring linear velocity and acting as Depth Sensor (DS);
- EvoLogics S2CR 18/34 acoustic modem;
- Teledyne BlueView M900 2D FLS;
- Ubiquiti Bullet M2 WiFi access point;
- 868+ RFDesign radio modem;
- one bottom-looking ELP 720p MINI IP camera;
- one Microsoft Lifecam Cinema forward-looking camera;
- two lateral ELP 1080p MINI IP cameras;
- Intel i-7-based LP-175-Commel motherboard (used for onboard processing);
- two Intel Neural Compute Stick 2;
- one NVIDIA Jetson Nano.

In order to validate the training methodology, described in Section 3, a hierarchical-stage strategy has been employed. Firstly, several optical frames and acoustic images have been acquired by using, respectively, the bottom-looking ELP 720p MINI IP camera and the Teledyne BlueVIew M900 2D FLS during a FeelHippo AUV pre-programmed mission. As far as the optical image enhancement techniques are concerned, images have been processed with the CLAHE algorithm, running online on the Intel i-7-based

LP-175-Commel motherboard, to overcome the limitations introduced by the low-visibility conditions. For instance, a comparison between a camera raw image and a CLAHE enhanced one is shown in Fig. 3.

Fig. 3. Comparison between the original image (top) and the CLAHE enhanced image (bottom).

In a second post-processing stage, the SSD and Faster R-CNN trained models have been executed on different dedicated hardware platforms; indeed, this hardware-decoupling solution provides the ATR system with the capability to process the images with the requested FPS value and guarantees the trained CNN modes to be real-time on-board runnable. With regard to the optical ATR approach, the SSD trained architecture has been optimized as a compiled graph, which has been subsequently loaded onto the Intel Neural Compute Stick 2 (Intel (2018)). Turning to the Faster R-CNN trained network, the prediction task on the acoustic frames have been fulfilled by means of the NVIDIA Jetson Nano (NVIDIA (2018)). This hierarchical-stage strategy used to validate the proposed strategy, composed of a training stage and a post-processing inference stage, is summarized in the flowcharts depicted in Fig.4 and Fig. 5.

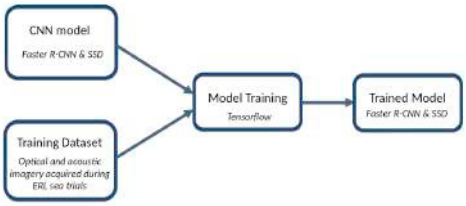


Fig. 4. The flowchart describing the training stage.

The developed strategy has guaranteed to reach the expected results; indeed, the classification task has been achieved with an overall accuracy which exceeds 90% in both the optical (Fig. 6) and acoustic (Fig. 7) image sets.

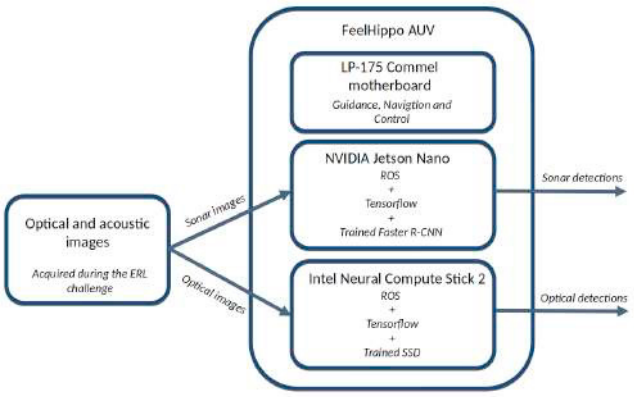


Fig. 5. The post-processing stage used to analyze collected images during a pre-planned mission.

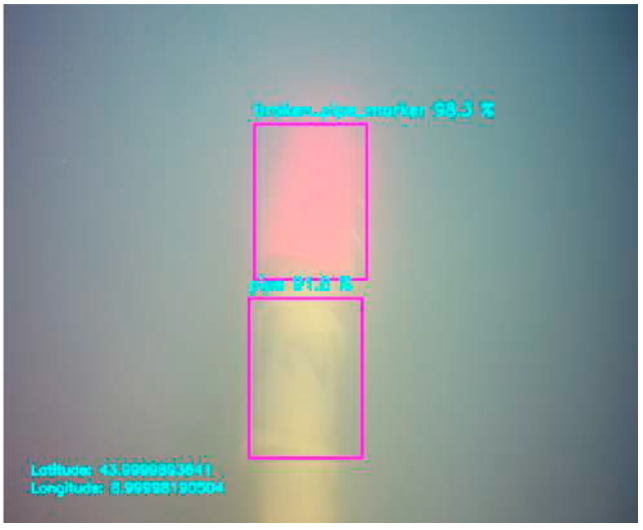

Fig. 6. Example of pipe and marker recognition in an optical image.

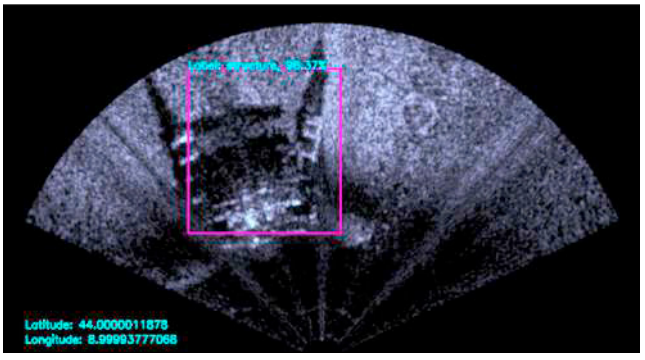

Fig. 7. Example of structure recognition in a 2D FLS acoustic image.

## 5. CONCLUSION AND FUTURE WORK

In light of all the above considerations, the proposed work arises as a validation proof of the feasibility of DL methodologies for ATR tasks in the underwater environment. In particular, the ATR strategy performances have been investigated on both the acoustic and optical subsea imagery. Within the regard of this research, an experimental dataset has been acquired by using the payload sensors

of FeelHippo AUV so as to optimally maximize the training stage effectiveness. Since the training dataset heavily affects networks performances, the dataset gathering task shall be tackled carefully. Concerning the underwater environment this process could be challenging and time expensive. Hence, the dataset augmentation options shall be investigated meticulously. As far as the CNNs, used in the training stages, are concerned, whilst a SSD network has been trained for the ATR task in optical images, a Faster R-CNN has been employed to develop an accurate trained model for the FLS acoustic imagery. Finally, experimental tests have been carried out in order to validate the abovementioned trained models loaded on dedicated hardware platforms. Different CNN architectures to fit the trade-off between inference speed and accuracy will be evaluated. Future works will also include the employment of the proposed ATR strategy in an overall intelligent system which led the vehicle to detect and recognize unknown targets as well as navigate towards them.

## REFERENCES


Allotta, B., Baines, S., Bartolini, F., Bellavia, F., Colombo, C., Conti, R., Costanzi, R., Dede, C., Fanfani, M., Gelli, J., Gundogdu, H.T., Monni, N., Moroni, D., Natalini, M., Pascali, M.A., Pazzaglia, F., Pugi, L., Ridolfi, A., Reggiannini, M., Roig, D., Salvetti, O., and Tekdemir, E.I. (2015). Design of a modular autonomous underwater vehicle for archaeological investigations. In *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*.

Allotta, B., Conti, R., Costanzi, R., Fanelli, F., Gelli, J., Meli, E., Monni, N., Ridolfi, A., and Rindi, A. (2017). A low cost autonomous underwater vehicle for patrolling and monitoring. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 231(3), 740–749.

Cashmore, M., Fox, M., Larkworthy, T., Long, D., and Magazzeni, D. (2014). AUV mission control via temporal planning. In *2014 IEEE international conference on robotics and automation (ICRA)*, 6535–6541.

Cheng, C.Y., Sung, C.C., and Chang, H.H. (2015). Underwater image restoration by red-dark channel prior and point spread function deconvolution. In *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 110–115.

Ferri, G., Ferreira, F., and Djapic, V. (2017). Multi-domain robotics competitions: The CMRE experience from SAUC-E to the European Robotics League Emergency Robots. In *OCEANS 2017-Aberdeen*, 1–7.

Franchi, M., Ridolfi, A., and Zacchini, L. (2018). A Forward-Looking Sonar-Based System for Underwater Mosaicing and Acoustic Odometry. In *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*, 1–6.

Girshick, R. (2015). Fast R-CNN. In *The IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

Intel (2018). Intel Neural Compute Stick 2. `https://software.intel.com/en-us/neural-compute-stick`.

Kolaman, A., Malowany, D., Hagege, R., and Guterman, G. (2019). Light Invariant Video Imaging for Improved Performance of Convolution Neural Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6), 1584–1594.

Krizhevsky, A., Sutskever, I., and Geoffrey, H. (2012). ImageNet classication with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Kvasic, I., Miskovic, N., and Vukić, Z. (2019). Convolutional Neural Network Architectures for Sonar-Based Diver Detection and Tracking. In *2019 MTS/IEEE Oceans 2019*, 1–6.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., and Berg, A.C. (2016). SSD: Single shot multibox detector. In *European conference on computer vision*, 21–37.

Lowe, D.G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Ma, J., Fan, X., Yang, S., Zhang, X., and Zhu, X. (2017). Contrast Limited Adaptive Histogram Equalization Based Fusion in YIQ and HSI Color Spaces for Underwater Image Enhancement. *International Journal of Pattern Recognition and Artificial Intelligence*, 32.

NVIDIA (2018). NVIDIA Jetson Nano . `https://developer.nvidia.com/embedded/jetson-nano-developer-kit`.

Prats, M., Ribas, D., Palomeras, N., García, J.C., Nannen, V., Wirth, S., Fernández, J.J., Beltrán, J.P., Campos, R., Ridao, P., et al. (2012). Reconfigurable AUV for intervention missions: a case study on underwater object recovery. *Intelligent Service Robotics*, 5(1), 19–31.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

ROS (2007). Official ROS website. `www.ros.org`. [Online; accessed November 2019].

Salvi, J., Petillo, Y., Thomas, S., and Aulinas, J. (2008). Visual SLAM for underwater vehicles using video velocity log and natural landmarks. In *OCEANS 2008*, 1–6.

TensorFlow (2015). Official TensorFlow website. `www.tensorflow.org`. [Online; accessed November 2019].

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2).

Valdenegro-Toro, M. (2016). Object recognition in forward-looking sonar images with Convolutional Neural Networks. In *OCEANS 2016 MTS/IEEE Monterey*, 1–6.

Vidal, E., Palomeras, N., Istenič, K., Hernández, J.D., and Carreras, M. (2019). Two-Dimensional Frontier-Based Viewpoint Generation for Exploring and Mapping Underwater Environments. *Sensors*, 19(6), 1460.

Xiu Li, Min Shang, Qin, H., and Liansheng Chen (2015). Fast accurate fish detection and recognition of underwater images with Fast R-CNN. In *OCEANS 2015 - MTS/IEEE Washington*, 1–5.

Zacchini, L., Bucci, A., Franchi, M., Costanzi, R., and Ridolfi, A. (2019). Mono visual odometry for Autonomous Underwater Vehicles navigation. In *2019 MTS/IEEE Oceans 2019*, 1–5.