Simposio CEA de Robótica, Bioingeniería, Visión Artificial y Automática Marina 2026

RBVM
Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina

Autonomous 3D fish-farm net-defect localization for AUV intervention

Bonin-Font, F.J.[a,*], Colom-Pujol, A.[a], Perez-Gregori, V.[a], Oliver-Codina, G.[a], Lopez-Barajas, S.[b,*], Solís-Jimenez, A.[b], Marín-Prades, R.[b], Sanz-Valero, P.[b]

[a] *SRV-MARIS - Marine Robotics and Intelligent Systems, University of the Balearic Islands, ctra de Valldemossa, km 7.5. 07122 Palma de Mallorca*
[b] *IRS Lab - Interactive and Robotic Systems. University Jaume I. Avenida Vicent Sos Baynat, s/n 12071 Castellón de la Plana.*

Resumen

La inspección autónoma de infraestructuras acuícolas supone un reto debido a la complejidad de los entornos submarinos y la precisión requerida en las tareas de intervención. Este trabajo presenta un módulo de detección de defectos en redes de piscifactorías basado en aprendizaje profundo y su localización, todo integrado en una arquitectura de percepción-acción de un Vehículo Submarino Autónomo de Intervención (I-AUV). La detección de defectos se basa en un modelo YOLO-World ajustado para la identificación de agujeros o roturas. A partir de una cámara monocular calibrada y condiciones puramente geométricas, los centroides de los defectos devueltos por el modelo entrenado se proyectan en el sistema de referencia de la cámara, junto con una estimación del tamaño del defecto. Esta información se integra en el entorno Robot Operating System (ROS) y se publica en tiempo real para condicionar la navegación del i-AUV y alinear su pinza mecánica hacia el defecto. Los resultados experimentales confirman el buen desempeño del sistema para su uso en intervenciones en lazo cerrado.

Palabras clave: Neural networks, Autonomous robotic systems Perception and sensing

Autonomous 3D fish-farm net-defect localization for AUV intervention

Abstract

Autonomous inspection and maintenance of aquaculture infrastructures is a challenge due to the complexity of underwater environments and the precision required for physical intervention. This paper presents a module for fish-farm nets defect detection and localization, integrated within a perception-to-action pipeline in an Intervention Autonomous Underwater Vehicle (I-AUV) equipped with Deep-Learning-based perception modules and mechanical grippers. Defect detection is based on a fine-tuned YOLO-World model, optimized to detect holes and scratches in nets. Using a calibrated monocular camera and geometric constraints, defect centroids are projected to the camera world frame and delivered together with the size of the defect. This information is integrated into a Robot Operating System (ROS) framework and published in real time to support I-AUV navigation and robotic gripper alignment. Experimental results validate the robustness of the systems and demonstrate the suitability of the proposed approach for closed-loop intervention.

Keywords: Neural networks, Autonomous robotic systems Perception and sensing

1. Introduction

The rapid growth of aquaculture has led to large-scale fish-farm infrastructures, where the state of the net pens is vital for ensuring fish containment, animal welfare, and environmental protection. However, these nets are continuously exposed to harsh marine conditions, biofouling, predator interactions, and mechanical stress, which can result in defects such as holes, tears, or structural weakening. If left undetected, such damage may lead to fish escapes, economic losses, and negative ecological impacts, making regular inspection and timely maintenance essential. Traditionally, net inspection and repair are carried out by human divers and/or *Remotely Operated Vehicles* (ROV) (Caveen, 2007) (Wu et al., 2022a). The former present safety restrictions, are time-consuming, inherently limited by weather, visibility conditions and depth, and the later are usually costly, being the success of the missions conditioned to the ability of ROV operators during the suturation procedu-

*Autor para correspondencia: francisco.bonin@uib.es
Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

res. Manual inspections and interventions are difficult to scale as fish farms grow in size, and they often lack the consistency and frequency required for early defect detection. These limitations motivate the need for autonomous solutions that operate persistently and reliably in underwater environments. *Autonomous Underwater Vehicles* (AUVs), combined with recent advances in artificial intelligence and computer vision, offer a promising alternative for the inspection and maintenance of aquaculture nets. By enabling continuous and repeatable monitoring with cameras, AUV can significantly reduce operational costs while increasing mission times and improving inspection coverage. Learning-based perception methods allow AUV to detect and classify net defects directly from visual data, even under varying environmental conditions. When coupled with precise localization and robotic manipulation capabilities, such systems open a wide spectrum of innovation to fully autonomous inspection-and-repair pipelines. These technologies have the potential to improve both economic performance and environmental management in modern fish farming. In the context of the TANDEM (UdG and UJI and UIB, 2026) project, defects on complex submerged structures need to be identified and localized, in both rigid and non-rigid structures that are affected by marine growth or degraded by prolonged exposure to the submarine environment. One of the use cases addressed within TANDEM concerns the maintenance of submerged fish-farm nets, which are to be inspected and repaired using an Autonomous Underwater Vehicle (AUV) equipped with a mechanical gripper. This paper presents a key perception component required for autonomous intervention, with particular focus on the target acquisition. In this stage, fine-tuned deep-learning models are combined with geometric vision techniques to estimate the spatial coordinates of detected defects, providing the information necessary for the robotic system to generate navigation set-points towards the defect.

Initial approaches to autonomous net inspection relied on classical computer vision techniques for edge or salient point detection, including the Hough Transform (Betancourt et al., 2020) or Canny (Akram et al., 2022) to identify irregularities in net structures. Although such image processing methods can perform well in controlled or simplified scenarios, they often lack robustness when confronted with submarine real-world conditions with light absorption or blur, or there are net deformations, which can render classical feature-extraction methods unreliable.

Other approaches are based on acoustic sensors, such as Forward-Looking Sonar (FLS), which provide acoustic images that require additional processing to mitigate noise (Rosa et al., 2024). Acoustic sensors used for frontal or vertical inspection are effective for detecting large structural failures; however, they present an important limitation: they often lack the spatial resolution, descriptive textures or high-frequency detail required to identify fine structural defects, such as small holes in nets (Yang et al., 2026). In addition, acoustic measurements are inherently range-based, requiring further post-processing to extract localized features. Side-scan sonar (SSS) and multibeam systems can generate acoustic images or 3D point clouds with acceptable resolution; however, they are primarily designed for seabed mapping rather than frontal inspection of vertical or flexible structures. By contrast, optical cameras can provide milimeter-scale resolution and rich texture information at close range and in all directions, enabling finer identification of objects, despite the challenges related to submarine imaging.

The adoption of *Deep Neural Networks* (DNNs) in this field are an effective alternative to address the visual noise inherent in underwater environments, by incorporating into the training process images acquired under diverse conditions. Neural networks offer powerful non-linear feature extraction capabilities that increase robustness to underwater visual degradation. In recent years, several studies have explored the use of deep learning techniques for fish-farm nets defect detection. In (Paraskevas and Kavallieratou, 2023), two different approaches are compared: one based on traditional image-processing techniques and another based on early versions of Object Detection framework You Only Look Once (YOLO) (Ultralitics, 2026a). In this work, images of real fish-farm nets are acquired using a Remotely Operated Vehicle (ROV) to feed the dataset employed for training and validation. Akram *et al* (Akram et al., 2024) proposed an autonomous fish-farm net defect detection system based on a customized multi-scale semantic segmentation architecture implemented as a Convolutional Neural Network (CNN). Their approach targets the detection of biofouling, vegetation, and holes, with specifically designed layers and loss functions. The training dataset combines images obtained from public repositories with data recorded by an ROV in real submerged net environments. More recently, task-specific CNN architectures have been progressively replaced by general-purpose, large-scale models pre-trained on thousands or millions of images and subsequently fine-tuned using comparatively small, task-specific datasets. This enabled the transfer of knowledge acquired from general domains to improve learning efficiency and performance in new particular applications. Commonly referred to as *Deep Transfer Learning* (DTL) (Iman et al., 2023) (Zhu et al., 2023), this approach significantly simplifies DNN design and training procedures and is the one of the bases of the work presented in this paper.

Nevertheless, defect detection in marine aquaculture is addressed to automatize the inspection and maintenance procedures. The perception–to-action loop is closed when the autonomous intervention capabilities of AUVs are combined with artificial intelligence for target recognition. Such systems are designed to autonomously observe, search for, detect, and localize fish-farm nets and their associated defects, while dynamically adapting the vehicle trajectory to the inferred information and the specific objectives (Karlsen et al., 2021) (Wu et al., 2022b) (Rosa et al., 2024). Barajas *et al.* (López-Barajas et al., 2024) integrated a YOLOv8 pre-trained net-defect detector, into the navigation and control architecture of a Girona 500 AUV (Ribas et al., 2012). The vehicle autonomously explores the net structure and, upon detecting a defect, navigates toward the target. Experimental validation was conducted in a controlled water-tank environment, where the AUV executed the *nano* version of the detection module. While the inferred bounding-box area was used to regulate the distance between the robot and the net, the position of the hole was estimated with ArUco OpenCv (2026) markers attached to the net.

Building upon (López-Barajas et al., 2024), this work introduces an enhanced deep-learning-based framework for net-defect detection and 3D localization. Our system offers two

primary contributions: (A) Hybrid detection and segmentation: we integrate a fine-tuned YOLO-World open-vocabulary detector (Ultralitics, 2026b) with the Fast Segment Anything Model (FAST SAM) (Zhao et al., 2023). While the object detector provides robust, coarse localization of defects, FAST SAM performs zero-shot segmentation within the resulting bounding boxes. This yields the precise geometric characterization, including shape and orientation, required for accurate gripper alignment during repair operations, (B) Marker-less 3D Localization: inferred bounding boxes are projected into metric world coordinates using the calibrated camera matrix (Hartley and Zisserman, 2003) via a revised blob median-area scale approach. By using the net geometry rather than contour lengths or external fiducial markers (e.g., ArUco), the AUV can localize and navigate toward defects using only onboard perception.

2. Fish farm net hole detection

YOLO-World Ultralitics (2026b) is a Open Vocabulary Object Detector, that it, has the ability to detect objects based on descriptive texts. It has demonstrated remarkable *zero-shot* transfer capabilities on benchmark datasets like COCO (Lin et al., 2014), which means that, being trained with thousands of images on the internet, it presents a high performance on classifying objects of the COCO dataset, without been learned from them. This indicates the model ability to generalize across diverse visual domains without task-specific training. However, the unique optical conditions of the geometry of aquaculture nets present a gap that exceeds its *zero-shot* limits, needing a previous fine-tuning step to learn the term *net hole* and incorporate it in its vocabulary.

The net hole detector was build upon the two native models of YOLO-World Ultralitics (2026b), the small one (yolov8s-worldv2) and the large one (yolov8l-worldv2). The aforementioned two new detectors were trained with 550 images of nets with damages and holes. A big group of these images were taken with the Girona 500 AUV in the water tank of the CIRTESU (UJI-Universitat Jaume I, 2026) moving in front of an old and deteriorated fish farm net hanged from one of the walls of the tank. The AUV was equipped with three cameras, one lateral, one frontal and another one on the gripper. The other group of images was taken from two public repositories: (Salvador López Barajas-UJI-CIRTESU, 2023) and (netinspection, 2023). Other 135 images of the same datasets were used for validation and 134 for test. As it is mandatory to avoid overfitting, images of the 3 groups were different. All images were labeled in YOLO format using the LABEL IMAGE (Label Studio, 2026) application and pre-proccessed with a CLAHE algorithm before training, validation and test, in order to improve textures and augment contrast. It is important to note that, the new YOLO-World model will be able to find all elements included in its native version plus the net holes, all from the preceptive terms included in its vocabulary.

Figure 1 shows some examples of the images used to train, validate and test the model, with the ground truth bounding box marked with LABEL IMAGE. Once the model is trained, it will be integrated in the AUV control architecture to detect and localize damages in real-time.

3. Hole net segmentation

In addition to detecting the image location of net holes, the small model of FAST SAM (Zhao et al., 2023) is applied within each inferred bounding box to segment the hole from the background. FAST SAM generates segmentation masks for all detected objects, which are here restricted to the bounding-box regions and highlight only the hole pixels. Compared to its predecessor SAM, FAST SAM significantly reduces segmentation time, enabling real-time operations. A key advantage of FAST SAM is that its pre-trained native model can accurately segment holes within the bounding boxes without requiring additional training or fine-tuning.

4. Monocular scale estimation and 3D localization

To overcome the scale ambiguity of the AUV monocular vision, a transformation pipeline that uses the known geometry of the aquaculture infrastructure was developed. The process is divided into three distinct phases: image-based scale recovery, geometric depth calculation, and 3D spatial projection.

4.1. Scale recovery from environmental priors

The system calculates a Meters-per-Pixel ratio (σ) by treating the fish farm net as a known reference grid. The length of each net square turns out to be known and equal to 1.5 cm in our particular case. Rather than relying on edge detection (Akram et al., 2022), which is susceptible to noise, we employ a method based on blob area analysis:

- Adaptive segmentation: The blue color channel is isolated to maximize contrast between the net (dark color) and the background (light color).
- Afterwards, an adaptive Gaussian threshold is applied to segment the net apertures in black and white. This step includes a morphological opening (erosion and dilation) with the same 5×5 kernel in order to remove noise and to ensure the integrity of the segmented apertures.
- To extract the contours and the geometric properties of the net apertures, we utilized the topological structural analysis algorithm proposed by Suzuki and Abe (Suzuki and Abe, 1985). Unlike traditional gradient-based edge detectors such as Canny, this approach extracts closed-loop boundaries on the segmented binary image. To optimize computational performance for the AUV onboard processor, we implemented a simple chain approximation, compressing each closed contour to a few essential vertices while maintaining the precision required to calculate the area of each contour. The precision of this approximation is as good as the net does not deform in excess losing the squared or rhomboidal appearance of the grid elements. Contours are stored as elements of a list, in such a way that their areas (so-called blobs) can be easily computed one by one.
- Then, a geometric filtering was applied. Contours are filtered by area: blobs with areas smaller than 10 pixels are assumed to be noise and blobs with areas larger than the 5 % of the image resolution are considered huge breakages and also filtered out. Additionally, blobs with aspect ratios (AR) of their

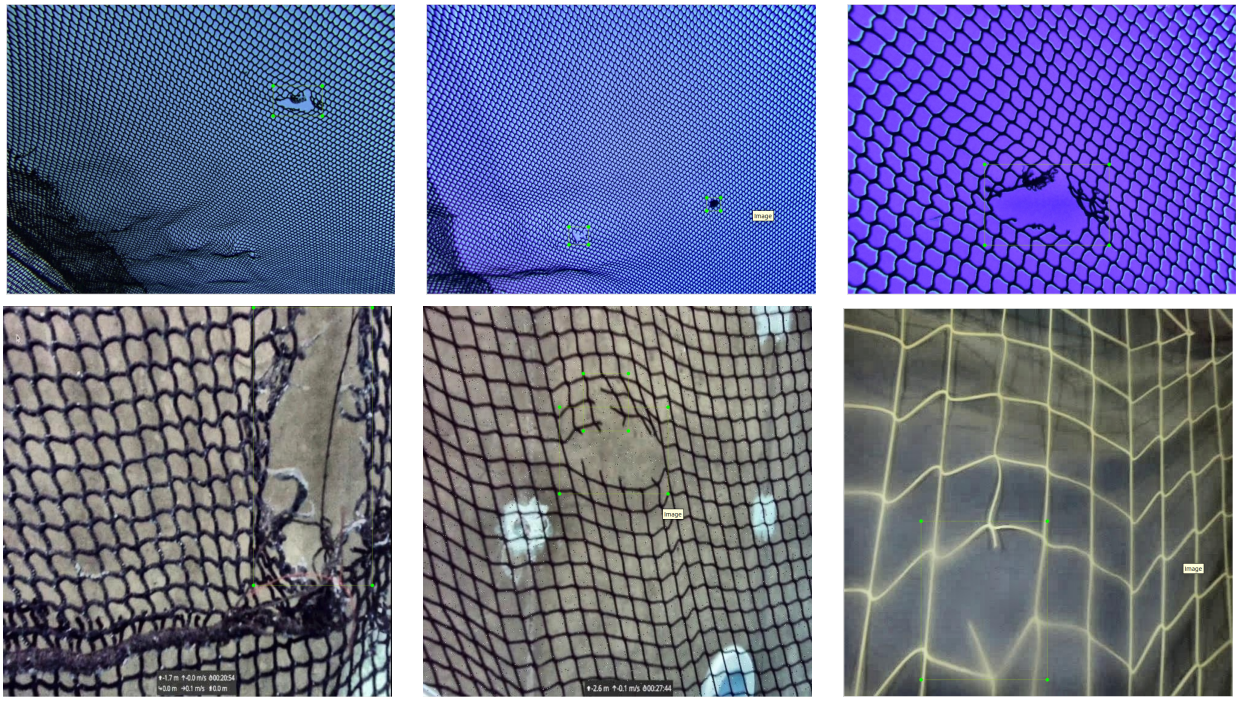


Figura 1: Some examples of images used to re-train the model with the ground truth bounding boxes.

bounding rectangles that accomplish $0{,}6 < AR < 1{,}6$ are kept in order to isolate valid, non-damaged net holes.

- Statistical Scale Inference: the final scale factor (σ) is calculated dividing the squared root of the real area of one element of the net grid ($\sqrt{A_{real}} = 0{,}000225m^2$) by the median squared area computed from all valid blobs $\sqrt{A_{px}}$, as: $\sigma = \frac{\sqrt{A_{real}}}{\text{median}(\sqrt{A_{px}})}$

4.2. *Geometric depth calculation*

Once the scale σ is established, the system determines the distance (Z) from the camera lens to the net plane. This is achieved using the known calibrated camera matrix which contains its intrinsic parameters:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

, being (f_x, f_y) and (c_x, c_y), respectively, the calibrated focal length and the optical center (principal point) of the lens. To take into account the possible sensor non-squareness, we utilized the Geometric Mean of the focal lengths (see equation 2). Finally, the depth Z was recovered using the relationship between the focal length ($f_{geometric}$) and the calculated scale, as shown in equation 3.

$$f_{geometric} = \sqrt{f_x \cdot f_y} \tag{2}$$

$$Z = f_{geometric} \cdot \sigma \tag{3}$$

4.3. *3D Spatial Projection*

In the last stage, the 2D detections given by the YOLO-World trained model are transformed into 3D coordinates relative to the camera. The algorithm identifies the normalized centroid of each YOLO bounding box (u_{norm}, v_{norm}) and denormalizes it to pixel coordinates (u, v) based on image resolution. Using the inverse pinhole camera model, the 3D position of each centroid (X,Y,Z) is calculated as follows (Hartley and Zisserman, 2003):

$$X = \frac{(u - c_x) \cdot Z}{f_x}, \qquad Y = \frac{(v - c_y) \cdot Z}{f_y} \tag{4}$$

5. Implementation and ROS integration

The proposed framework was developed in Python using the OpenCV library (Open Cv, 2026) and wrapped in a Robot Operating System (ROS) (Open Robotics, 2026) package. This integration ensures compatibility with the Girona 500 AUV, which utilizes ROS as its middleware for operation and control. For each detected defect, the node publishes to the ROS core the (X,Y,Z) world coordinates of the centroid along with its segmented area. This real-time data stream enables the AUV to run precise maneuvers towards the defect and plan more accurately the intervention. The source code is available via a GitHub repository upon request.

6. Experimental results

6.1. *Quantitative train evaluation*

The YOLO-World model was fine-tuned under four distinct configurations to integrate the *net hole* class into its vocabulary. All models utilized data augmentation, including random rotations, scaling, and perspective transformations, to ensure robustness against varying camera viewpoints. The specific hyperparameters for each configuration are as follows: (1) Model 1 (M1), small architecture, trained for 180 epochs with a batch size of 16. (2) Model 2 (M2), small architecture, trained for 100

epochs with a batch size of 8. (3) Model 3 (M3), large architecture, trained for 100 epochs with a batch size of 8. (4) Model 4 (M4), large architecture, trained for 180 epochs with a batch size of 8. Table 1 summarizes the main metrics obtained from the four trained models, including: (a) Precision (P.), which measures the reliability of the detections; (b) Recall or sensitivity (R.), which quantifies the proportion of relevant instances successfully detected; (c) the Mean Average Precision computed at an *Intersection over Union* (IoU) threshold of 0.50 (mAP50), which reflects the model accuracy under relatively permissive matching conditions; (d) the Mean Average Precision averaged over IoU thresholds from 0.50 to 0.95 (mAP50-95), providing a more stringent evaluation of performance with more challenging detection scenarios; and (e) the inference time (Inf. Speed), measured in milliseconds. Further details on the evaluation metrics can be found in (Everingham et al., 2015) and (Padilla et al., 2020).

Table 1: Training and Validation metrics for the different configurations.

	P.	R.	mAP50	mAP50-95	Inf. Speed
$M1$	0.954	0.896	0.924	0.613	3.5 (ms)
$M2$	0.972	0.845	0.918	0.573	3.2 (ms)
$M3$	0.965	0.877	0.929	0.62	10.8 (ms)
$M4$	0.987	0.872	0.945	0.632	10.6 (ms)

6.2. *Quantitative test evaluation*

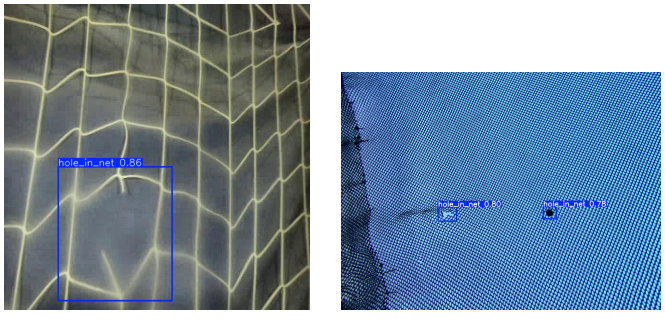
The four different models were tested with a set of images different from those used to train and validate. Table 2 shows the resulting metrics obtained from the predictions given by the re-trained model.

Table 2: Test metrics for the different configurations.

	P.	R.	mAP50	mAP50-95	Inf. Speed
$M1$	0.994	0.898	0.960	0.586	5 (ms)
$M2$	0.975	0.892	0.950	0.538	5.1 (ms)
$M3$	0.976	0.922	0.967	0.57	19 (ms)
$M4$	0.976	0.922	0.967	0.593	10.6 (ms)

6.3. *Qualitative results*

This section includes some examples of qualitative results. Figure 2 shows in (a) and (b) two test images with the blue bounding boxes indicating the holes inferred by M4 and their confidence detection, as a proof of success defect detection.



(a) (b)

Figura 2: Some examples of hole detection.

Figure 3 shows three processed test images with the bounding box marking the inferred hole (in blue), its centroid world coordinates (in red), and the calibrated and uncalibrated optical centers (in green and white, respectively). The examples illustrate how the estimated depth (Z) increases with the camera-to-net distance and how the (X, Y) coordinates are aligned with the camera motion. In particular, images (a) and (b) were taken from the AUV lateral camera, and the image 3-(c) was taken from its frontal camera. Image (c) shows two holes, which were both detected. The hole on the right of the image corresponds to the hole of image (b).

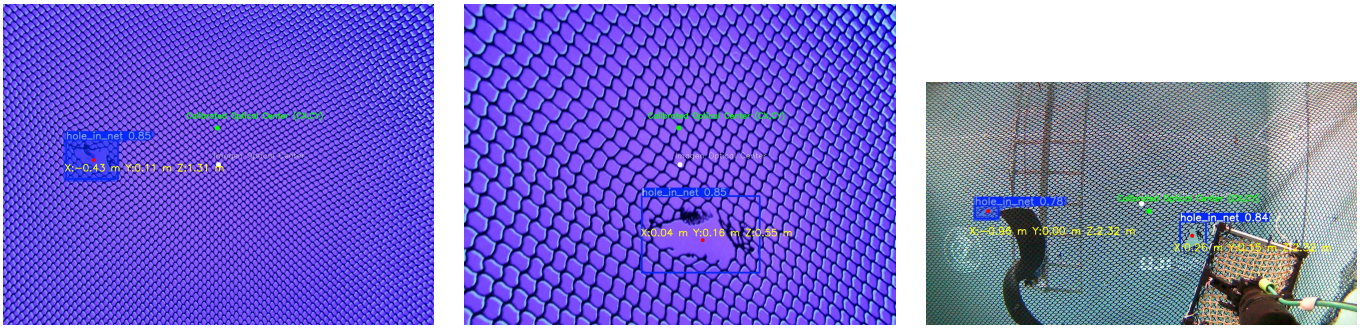
7. Discussion

The experimental results across four model configurations (M1 through M4) show the effectiveness of the fine-tuning approach for net-hole detection. Table 1 and Table 2 summarize the performance on the validation and test sets, respectively. All models exhibited high precision, with M4 achieving a peak of 0.987 during training and 0.976 during testing. This high precision is critical for subsea intervention, as it minimizes the risk of false-positive detections that could lead to unnecessary or damaging autonomous maneuvers. A comparison between validation and test metrics indicates strong generalization. For instance, M4 maintained a consistent $mAP50$ of 0.945 in validation and improved to 0.967 in testing. This issue points out that *easy* detections are all found. The $mAP50-95$ scores, which represent localization stringency, remained stable across all models (ranging from 0.538 to 0.632), suggesting that the model is highly capable of identifying holes, but the predicted bounding boxes are displaced from their true center. Recalls are between 0.84 and 0.92 in all cases, showing high levels of sensitivity.

The models present a clear distinction in computational demand. Models M1 and M2 represent the *lightweight* configurations, with inference speeds between 3.2 and 5.1 ms. Conversely, M3 and M4 show higher complexity, with M3 reaching an inference time of 19 ms in test conditions. While M4 achieves the highest overall accuracy ($mAP50$ and $mAP50-95$) on the test set, it requires a longer inference time than M1. Nevertheless, with an inference time of 10.6 ms per frame, M4 still supports a frame rate of approximately 94 *Frames per Second* (FPS) (FPS=$1 \div 10.6$ ms) , which comfortably exceeds the real-time requirements of the AUV navigation and control system, typically ranging from 5 to 10 FPS. M4 therefore is the optimal configuration, offering the best trade-off between performance and efficiency. It combines a robust recall of 0.922, ensuring that very few defects are missed, with the highest localization accuracy. Although M1 achieves faster inference, M4 delivers superior performance without compromising real-time operation.

8. Conclusions

This paper presents a real-time 3D localization pipeline for AUV-based fish-farm net inspection, combining a fine-tuned YOLO-World model with a monocular scale-recovery method. The system achieves high detection accuracy (mAP50 of 0.967) while operating in real time, with low inference latency (10.6



(a) (b) (c)

Figura 3: Test samples: (X,Y,Z) for Images (a): (0.43, 0.11, 1.31) m , (b): (0.04 , 0.16, 0.55) m, and (c): Right hole coord. (0.26, 0.15, 2.32) m

ms). Results demonstrate reliable defect detection and localization, along with robust scale estimation, enabling accurate 3D positioning without the need for stereo vision or multiple views.

Future work focuses on completing the perception-to-action loop through several key objectives: improving defect orientation estimation using geometric fitting of segmented regions, enhancing robustness by expanding the training dataset with more challenging underwater conditions, benchmarking against other object-detection models, validating the system in real aquaculture environments, and quantitatively assessing the 3D localization error of detected defects.

Acknowledgements

This work is partially supported by Grants PID2023-149910OB-C33 and PID2023-149910OB-C32 funded by MICIU/AEI/10Z.13039/501100011033 and FEDER-UE funding, and by Project AIA2025-163346-C42 funded by MICIU/AEI/10.13039/501100011033.

Referencias

Akram, W., Casavola, A., Kapetanović, N., Miškovic, N., 2022. A Visual Servoing Scheme for Autonomous Aquaculture Net Pens Inspection Using ROV. Sensors 22 (9).

Akram, W., Hassan, T., Toubar, H., Ahmed, M., Miškovic, N., Seneviratne, L., Hussain, I., 2024. Aquaculture Defects Recognition via Multi-scale Semantic Segmentation. Expert Systems with Applications 237, 121197.

Betancourt, J., Coral, W., Colorado, J., 2020. An Integrated ROV Solution for Underwater Net-cage Inspection in Fish Farms Using Computer Vision. SN Applied Sciences 2 (12).

Caveen, A., 12 2007. Occupational Health and Safety in the Aquaculture Industry -a Global Review. Lloyd's Register Foundation.

Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J., Zisserman, A., 2015. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision 111, 98–136.

Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision, 2nd Edition. Cambridge University Press, New York, NY, USA.

Iman, M., Arabnia, H. R., Rasheed, K., 2023. A Review of Deep Transfer Learning and Recent Advancements. Technologies 11 (2).

Karlsen, H. Ø., Amundsen, H. B., Caharija, W., Ludvigsen, M., 2021. Autonomous Aquaculture: Implementation of an Autonomous Mission Control System for Unmanned Underwater Vehicle Operations. In: IEEE OCEANS. pp. 1–10.

Label Studio, 2026. label Studio: Open Source Data Labeling Platform. `https://labelstud.io/`.

Lin, T.-Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Lecture Notes in Computer Science. Vol. 8693. Springer International Publishing, pp. 740–755. URL: `https://cocodataset.org/#home`

López-Barajas, S., Sanz, P. J., Marín-Prades, R., Gómez-Espinosa, A., González-García, J., Echagüe, J., 2024. Inspection Operations and Hole Detection in Fish Net Cages through a Hybrid Underwater Intervention System Using Deep Learning Techniques. Journal of Marine Science and Engineering 12 (1).

netinspection, 2023. Net Holes Dataset. `https://universe.roboflow.com/netinspection/net_holes`.

Open Cv, 2026. Open Computer Vision Library. `https://opencv.org/`.

Open Robotics, 2026. ROS-Robot Operating System. `https://www.ros.org/`.

OpenCv, 2026. ArUco Detections. `https://docs.opencv.org/4.x/d5/dae/tutorial_aruco_detection.html`.

Padilla, R., Netto, S. L., da Silva, E. A. B., 2020. A Survey on Performance Metrics for Object-Detection Algorithms. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). pp. 237–242.

Paraskevas, K., Kavallieratou, E., 2023. Detecting holes in fish farming nets: A two–method approach. In: 2023 International Conference on Control, Automation and Diagnosis (ICCAD). pp. 1–7.

Ribas, D., Palomeras, N., Ridao, P., Carreras, M., Mallios, A., 2012. Girona 500 AUV: From Survey to Intervention. IEEE/ASME Transactions on Mechatronics 17 (1), 46–53.

Rosa, D., Cabecinhas, D., Ferreira, F., 2024. Forward-looking Sonar Based Autonomous Aquaculture Inspection. In: OCEANS 2024 - Singapore. pp. 1–8.

Salvador López Barajas-UJI-CIRTESU, 2023. Realholes Dataset. `https://universe.roboflow.com/salvador-lpez-barajas/realholes`.

Suzuki, S., Abe, K., 1985. Topological Structural Analysis of Digitized Binary Images by Border Following. Computer Vision, Graphics and Image Processing 30 (1), 32–46.

UdG and UJI and UIB, 2026. `https://tandemproject.es/`.

UJI-Universitat Jaume I, 2026. Research Center for Robotics and Underwater Technologies. `https://blogs.uji.es/cirtesu/`.

Ultralitics, 2026a. `https://docs.ultralytics.com/models/`.

Ultralitics, 2026b. Yolo World: Real Time Open-Vocabulary Object Detection. `https://docs.ultralytics.com/models/yolo-world/`.

Wu, Y., Duan, Y., Wei, Y., An, D., Liu, J., 2022a. Application of Intelligent and Unmanned Equipment in Aquaculture: A review. Computers and Electronics in Agriculture 199, 107201.

Wu, Y., Liu, J., Wei, Y., An, D., Duan, Y., Li, W., Li, B., Chen, Y., Wei, Q., 2022b. Intelligent Control Method of Underwater Inspection Robot in Net-cage. Aquaculture Research 53 (5), 1928–1938.

Yang, Z., Zhao, J., Zhao, X., Huang, C., 2026. Super-resolution Reconstruction of Side-Scan Sonar Images Based on Texture Consistency. Expert Systems with Applications 296, 129033.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J., 2023. Fast Segment Anything. `https://docs.ultralytics.com/models/fast-sam/`.

Zhu, Z., Lin, K., Jain, A. K., Zhou, J., 2023. Transfer Learning in Deep Reinforcement Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (11), 13344–13362.