Automatic Extraction and Counting of Fish from Underwater Videos Using YOLO-Based Deep Learning Algorithms

Automatic Extraction and Counting of Fish from Underwater Videos Using YOLO-Based Deep Learning Algorithms

Francesca Gallitto [1], Andrea Maria Lingua [1], Francesca Matrone [1], Filiberto Chiabrando [2], Xinchen Li [2], Silvia Secco [3,4], Alessandro Acierno [3], Massimiliano Scalici [3]

[1] Department of Environment, Land and Infrastructure Engineering (DIATI), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino – (francesca.gallitto, andrea.lingua francesca.matrone)@polito.it
[2] Department of Architecture and Design (DAD), Politecnico di Torino, Viale Pier Andrea Mattioli, 39, 10125 Torino – (filiberto.chiabrando, xinchen.li)@polito.it
[3] Department of Sciences, University of Roma Tre, viale G. Marconi 446, 00146 Rome, Italy (silvia.secco, alessandro.acierno, massimiliano.scalici)@uniroma3.it
[4] Department of Integrative Marine Ecology (EMI), Genoa Marine Centre (GMC), Stazione Zoologica Anton Dohrn–National Institute of Marine Biology, Ecology and Biotechnology, Piazza del Principe 4, Genova, 16126, Italy

Keywords: fish monitoring, YOLO, underwater mapping, Posidonia Oceanica, deep learning

Abstract
In the context of increasing pressure on marine species, effective and automated methodologies for biodiversity monitoring are essential, particularly in sensitive ecosystems such as the Mediterranean Sea. This study presents an integrated deep learning approach for multi-object tracking and species recognition from underwater videos, aiming to automate and improve fish population censuses within Marine areas with a focus in Marine Protected Areas in Sardinia. Various detection models, including DeepFins, DeepEcomar, YOLOv8, and YOLOE, and tracking algorithms such as DeepSort, ByteTrack, and SAMURAI were evaluated. According to the achieved tests the YOLOE model, carefully trained on the Mediterranean-specific SardinIA dataset, demonstrated the best detection performance, while DeepSort proved most effective in maintaining individual identities across complex scenarios. The AI based achieved results compared with traditional visual census methods (underwater visual census, UVC and diver operated video census, DOVC), showing high accuracy in total abundance estimation and good agreement for dominant species. These findings suggest that deep learning techniques offer a promising, scalable solution for marine biodiversity monitoring, although challenges remain in species-level classification. In the present paper the following methodologies and the achieved results are reported.

1. Introduction

In a global context where the number of species threatened with extinction continues to rise, the adoption of effective conservation strategies is becoming increasingly essential. While risk assessment and the implementation of protective measures are already complex for terrestrial fauna, the challenge is even greater for marine organisms, particularly in delicate ecosystems such as the Mediterranean Sea.

Fish census specifically provides valuable insights into marine ecosystem conditions, informs fisheries management, and supports conservation efforts aimed at mitigating species decline and habitat degradation. Within this context, Marine Protected Areas (MPAs) play a pivotal role by preserving critical habitats, enhancing local biodiversity, and providing baseline ecological data necessary for comparative studies and long-term environmental assessments (Guidetti et al., 2014). Effective and accurate biodiversity monitoring in MPAs, particularly within sensitive ecosystems such as the Mediterranean Sea, becomes essential to assess the success of protection strategies, enabling adaptive management and informed policy-making (Sala et al., 2012).

Traditional manual census techniques for assessing fish biodiversity typically consist of two main methods: Underwater Visual Censuses (UVCs) and Diver-Operated Video Censuses (DOVCs), both based on surveying fish along predefined transects to record abundance and biometric data. While UVCs require real-time observation, DOVCs allow for post-dive data analysis (Brock, 1982; Harmelin-Vivien et al., 1985). While both UVCs methodologies are widely utilized due to their simplicity and cost-effectiveness, they inherently carry several limitations, including observer bias, reduced accuracy in detecting cryptic or small-sized species, variability in diver experience and skill, and extensive post-processing times for video analysis (Murphy & Jenkins, 2010; Thanopoulou et al., 2018). These factors underline the importance of the development of new methodologies and adoption of automated, more reliable, and reproducible alternatives.

1.1 Deep learning approaches for automatic fish census

In recent years, deep learning approaches have significantly advanced the automated recognition and tracking of marine biodiversity. Emerging automated techniques address these issues through efficient, accurate, and reproducible data analysis. Convolutional neural networks (CNNs), particularly those based on the YOLO (You Only Look Once) architecture, have become standard tools for real-time species identification tasks. Recent advancements in YOLO, such as YOLOv7 and YOLOv8, have improved both detection accuracy and computational speed (Redmon & Farhadi, 2018; Wang et al., 2023). Jalal et al. (2025) introduced the DeepFins model, a variant of YOLO specifically adapted for fish detection, able to achieve promising results in identifying fish species typical of Australian marine environments.

For tracking identified fish across video frames, DeepSort (Deep Simple Online and Real-time Tracking) is a widely adopted approach due to its effectiveness in handling occlusions and maintaining the identities of multiple moving objects (Wojke et al., 2017).

SAMURAI (Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Instance-Level Memory)

represents a significant leap in object tracking capabilities. Unlike traditional models, SAMURAI (Yang, 2023) employs a motion-aware memory mechanism that dynamically selects and refines past observations based on movement patterns. This approach allows the model to effectively handle crowded scenes, fast-moving objects, and occlusions without retraining or fine-tuning. Its zero-shot learning capability ensures adaptability across diverse scenarios, making it suitable for real-time applications in dynamic environments.

On the other hand, ByteTrack (Zhang Y. et al., 2022) offers a simplistic yet effective approach to multi-object tracking by associating every detection box, including those with low confidence scores. This strategy reduces the number of false negatives and enhances tracking performance in scenarios where objects might be partially occluded or detected with low confidence. However, Byte Track's reliance on basic metrics like Intersection over Union (IoU) for association can limit its accuracy in crowded scenes with multiple overlapping objects. Additionally, the absence of appearance modelling means it might struggle to maintain consistent object identities over extended periods, especially when objects appear similarly.
Integrating these advanced tracking algorithms into marine biodiversity monitoring systems could significantly enhance the accuracy and reliability of fish population assessments. By addressing the limitations of current models, such as handling occlusions and maintaining object identities, these algorithms offer promising approaches for improving conservation strategies in marine protected areas.

Alongside DeepFins and traditional YOLO variants, recent developments have highlighted the YOLOE (You Only Look at One Representation) model, an innovative object detection architecture designed to enhance both accuracy and inference speed by employing a more efficient representation learning framework (Xu et al., 2022). YOLOE integrates anchor-free techniques and efficient feature extraction strategies that could significantly improve the detection of small or partially occluded marine organisms, which frequently occur in underwater ecosystems such as Mediterranean seagrass meadows (i.e. Posidonia oceanica). Specifically, YOLOE's optimised architecture reduces computational overhead and improves accuracy in challenging scenarios characterised by complex backgrounds and varying illumination conditions, common traits of marine environments. Considering these advantages, YOLOE is a promising candidate for integration into fish biodiversity monitoring frameworks, potentially enhancing the identification accuracy and overall robustness of automated marine censuses compared to standard YOLO-based methods.

Integrating CNN-based detection with DeepSort (or other tracking algorithms) allows comprehensive monitoring of marine biodiversity, overcoming many limitations inherent to manual monitoring.

2. Methodology

To address these challenges, in the present paper an original method for deep learning-based fish recognition that integrates fish detection and multi-object tracking to automate fish counting and species identification from underwater video surveys is presented.

The workflow was structured into the following stages:

- *Data Acquisition:* traditional UVC and DOVC methods were collected at different depths across selected locations within Sardinia's Marine Protected Areas (MPAs);
- *Data Labelling:* the ground truth datasets were created through manual annotation of video frames, involved precisely labelling individual fish specimens across various scenarios;
- *Data processing for fish detection*: in the initial phase, established detection models were applied to underwater video data to detect fish presence as a single class. After this preliminary step, the research expanded to multi-class species recognition, testing various models, DeepFins (based on YOLOv11) and DeepEcomar. Model performance was evaluated using standard metrics including precision, recall, confidence scores, and mean Average Precision (mAP), enabling identification of the best-performing baseline detection approach;
- *Model Fine-tuning:* in this phase, selected detection models (YOLOv8 and YOLOE) were fine-tuned using an expanded and specifically curated Mediterranean marine dataset. Performance evaluations were systematically conducted to quantify improvements in detection accuracy, especially under challenging scenarios such as small or partially occluded fish. The performance of these newly trained models was systematically validated, and comparative analyses were carried out to quantify the improvements achieved through retraining, with particular emphasis on challenging detection conditions, such as smaller species or partially occluded individuals;
- *Data processing for multi-tracking*: following fish detection, several state-of-the-art tracking algorithms were explored, including DeepSort, SAMURAI, and ByteTrack. These methods were systematically tested to evaluate their effectiveness in handling multi-object scenarios typical of marine biodiversity monitoring. After careful performance analysis under conditions involving multiple overlapping fish and occlusions, DeepSort (Wojke et al., 2017) was identified as the optimal tracking algorithm, primarily due to its robust use of appearance embeddings, effectively reducing identity-switching issues encountered with other algorithms.

This structured methodological approach allowed rigorous validation of deep learning-based automated detection and tracking systems, effectively demonstrating their potential to outperform traditional visual census methods (UVC and DOVC) in terms of accuracy, efficiency, and scalability in biodiversity monitoring efforts. In the following scheme (Figure 1), the proposed workflow is reported.

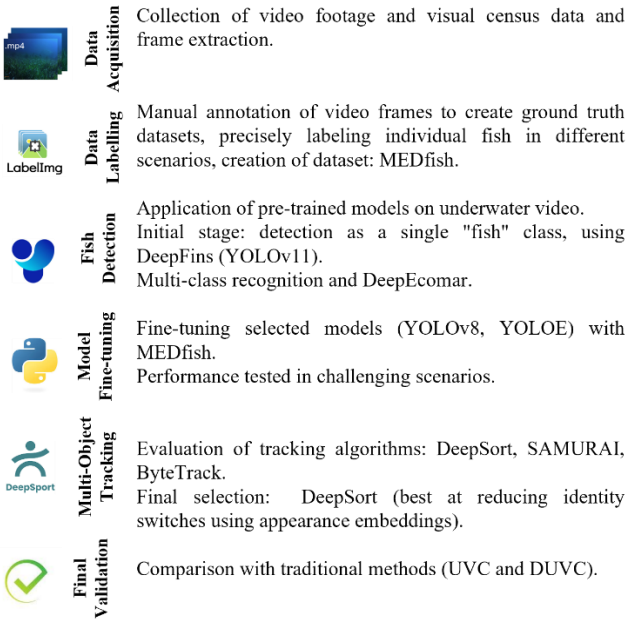

Figure 1. The proposed workflow.

2.1 Data acquisition

The study was carried out on the italian island of Sardinia, in the Mediterranean Sea, specifically in areas with pristine or minimally disturbed areas, such as the MPA of Capo Testa – Punta Falcone, the MPA of Capo Carbonara, the MPA of Tavolara –Punta Coda Cavallo and the Culuccia peninsula, and also areas heavily impacted by human activities such as the city of Porto Torres. In these sites (Figure 2), UVCs and DOVC were collected at two different depths, 5 and 15 meters, to support the analysis. These sites were selected due to their ecological relevance and different human disturbance, providing diverse conditions for robust method evaluation.

In particular, the results presented in the paper are based on the data collected in the MPA Tavolara – Punta Coda Cavallo, specifically nearby Isola Rossa (C zone of MPA), during August 2024 by scuba operators, at 15 m depth. The procedure was carried out along a 50-m linear transect at a constant velocity of 12 m per min. The acquisitions were replicated 3 times.

The videos were captured using a Becam 4k Eis Action Camera (Eis: Electronics Image Stabilizer; 170° Wide Angle; Glass Lens; Depth 50 M; Lcd 2.0‘’; Wi-Fi Connection; Memory Up To 32 Gb, Class 10, Micro Sd Card; Video Resolution: 4 K 30 Fps-4 K 25 Fps-2.7k 30 Fps-1080p 60/30fps-720p 120/60fps. Video Format: Mov H264; Photos 20-16-12-8-5-3 Mega Pixels; Photo Format: Jpg; Battery Life: 120 Min; 180° Auto Image Rotation).

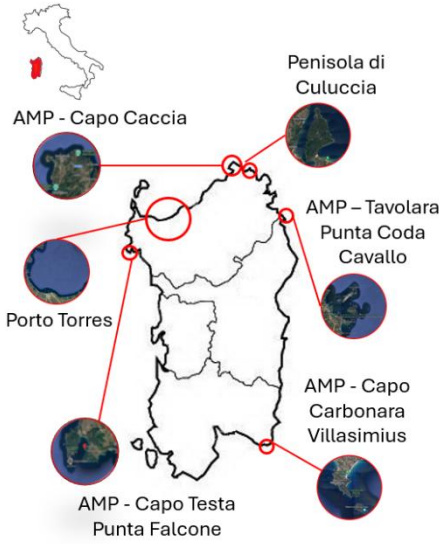

Figure 2. Location of case study area.

2.2 Data labelling

For the labelling phase, the *labelImg* software (Tzutalin, 2015) was used, as it allows annotations in a format compatible with the YOLO model. The resulting dataset includes 28 species of fish endemic to the Mediterranean Sea, comprising a total of 729 images. The software allows users to add labels, remove them, and assign labels to additional classes. In the YOLO model, a specific class is used to group unrecognized or ambiguous elements (Table 1). In this study, such elements were assigned to a dedicated class named *Afish poly*, which serves to reduce misclassifications by capturing instances that do not clearly belong to any of the predefined fish species.

This dataset, called SardinIA, was also used as ground truth for evaluating all the obtained results. This dataset is planned to be open source licensed, in the meantime, it is available upon request.

Table 1. List of classes of SardinIA dataset.

Afish poly	*Diplodus sp-*	*Pomatous salator*	*Symphodus sp-*
Chromis chromis	*Diplodus vulgaris*	*Sarpa salpa*	*Labrid unid-*
Coris julis	*Epinephelus marginatus*	*Seriola dumerilii*	*Small*
Dentex dentex	*Lithognathus mormyrus*	*Serranus cabrilla*	*Taca*
Diplodus annularis	*Mugilidae prob Chelon*	*Serranus scriba*	*Spicara smaris*
Diplodus puntazzo	*Mullus sp-*	*Sparus aurata*	*Symphodus tinca*
Diplodus sargus	*Oblada melanura*	*Sphyraena sp-*	*Thalassoma pavo*

Just to have an idea, the manual annotation time of the dataset, performed by two experts, took a total of about 6 hours.

2.3 Data processing for fish detection

As previously mentioned, the initial phase focused on defining an appropriate methodology by using a single class namely, detecting the presence of fish without species-level classification. During this phase, the processing pipeline was divided into two different steps: object detection and object tracking.

Once the dataset was fully established, the classification task was extended to include all target fish species. Multiple tests were conducted to identify the most suitable model architecture.

The models evaluated in this study include: DeepFins (Jalal et al., 2025), a YOLOv11-based model trained on fish species native to the Australian coasts; DeepEcomar (Catalan, 2023), developed using a dataset of Mediterranean species; YOLO-E, a recent version of the YOLO architecture; and finally, a custom YOLOv8-based model trained on the dataset explicitly developed for this study.

2.3.1 DeepFins: DeepFins is a hybrid deep learning model designed for fish detection in challenging underwater environments. It combines YOLOv11, a static object detector, with a lightweight motion-based segmentation module to capture fish dynamics and suppress background interference. This approach addresses challenges like poor visibility and environmental variability in unregulated underwater videos. DeepFins achieved a 90.0% F1 Score on the OzFish dataset, outperforming previous models by approximately 11%. It also obtained an 83.7% F1 Score on the Fish4Knowledge LifeCLEF 2015 dataset, marking a 4% improvement over the baseline YOLOv11. These results (Figure 3) allow to state that DeepFins is a practical solution for automated fish sampling and estimating relative fish abundance.
The DeepFins model was therefore applied to the frames previously extracted from the collected videos. Although the model produced promising results, several issues were observed. In particular, small fish in the background were often missed, and a high number of false positives were detected near *Posidonia oceanica meadows*, where many of its leaves were mistakenly identified as fish. These discrepancies are likely due, at least in part, to the fact that DeepFins was trained on a dataset composed of fish species from Australian waters, which differ significantly from the species found in the Mediterranean Sea.

The DeepFins model achieved the following metrics: an accuracy of 0.71, a precision of 0.78, a recall of 0.68, and an F1 score of 0.61. These results suggest that, while the model demonstrates relatively high precision, the lower recall and F1 score indicate room for improvement in consistently detecting all relevant instances. Furthermore, since the model is not capable of recognizing individual fish species, it became necessary to evaluate alternative models.

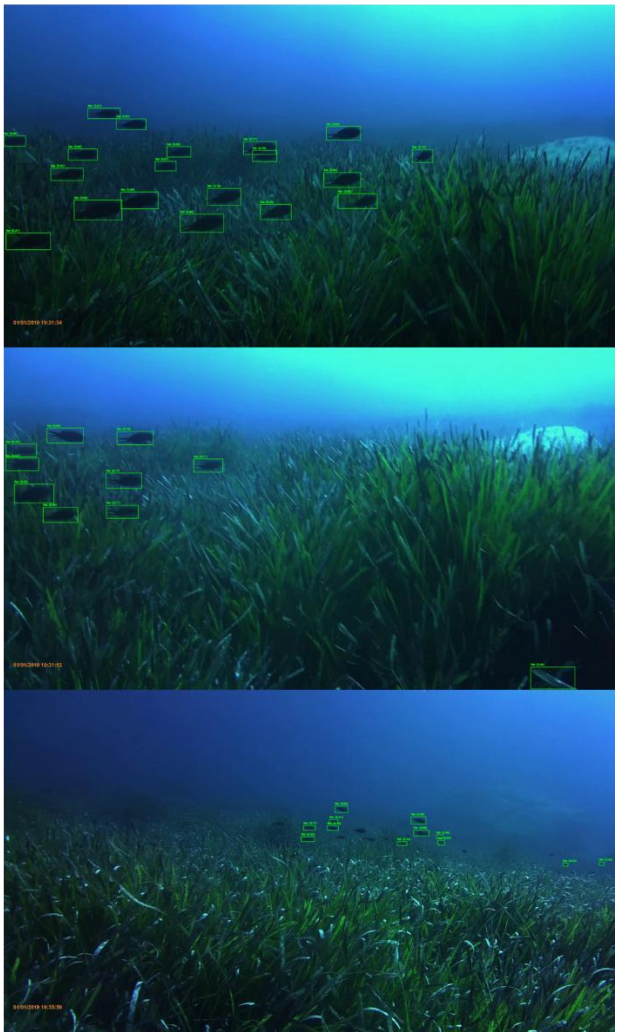

Figure 3. Results on selected frames using DeepFins. In top-down order: good results, false positive of PO, some fish were not recognized in the background.

2.3.2 DeepEcomar: To address the challenges proposed by endemic Mediterranean species, the DeepEcomar model (Catalan, 2023), was adopted. DeepEcomar was specifically developed to detect and classify fish species native to the Mediterranean Sea, offering a more context-appropriate alternative to models trained on non-Mediterranean datasets. The associated dataset comprises over 18,400 annotated instances of 20 fish species, extracted from more than 1,600 underwater images characterized by a wide variety of environmental conditions and backgrounds.

In addition to species-level annotations, the dataset includes broader categories such as "smudge" (e.g., visual obstructions covering more than 20% of the frame), "small" (fish smaller than 100 pixels squared), and "Afish poly" for unidentified specimens. Two state-of-the-art object detectors, YOLOv5m and Faster R-CNN, were compared during development. YOLOv5m demonstrated superior performance and was selected for further experimentation.

The issue with this dataset, in relation to the species present in our videos, is the absence of three species, *Spicara maena*, *Symphodus tinca*, and *Thalassoma pavo*, which are present, in some cases in high abundance.

In any case, the model was applied to the frames acquired in the present research to evaluate its performance on different species, the visual results are reported in Figure 4.

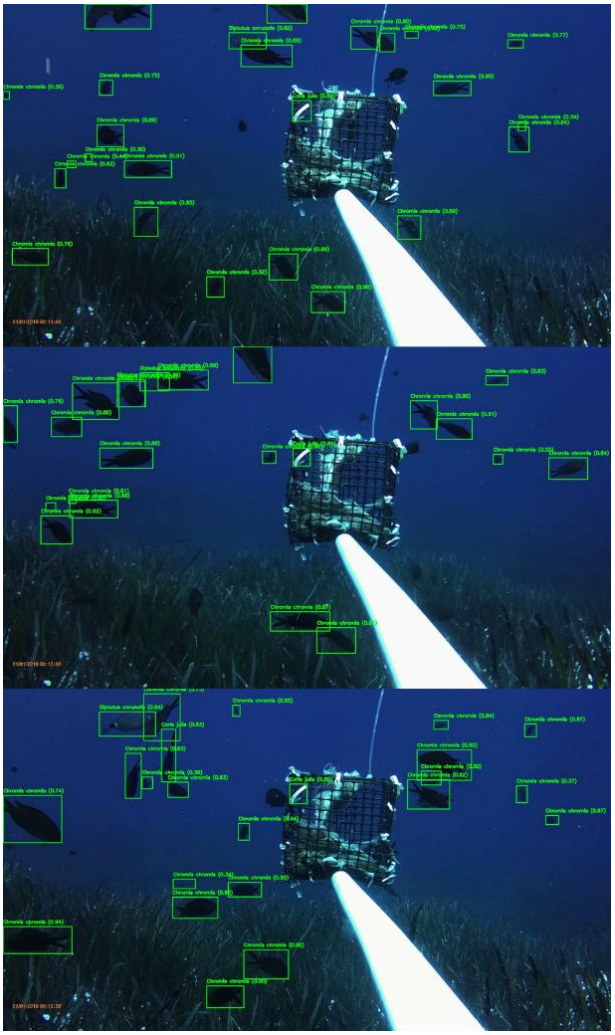

Figure 4. Results on selected frames using DeepEcomar, In the frames, a high number of recognized species can be observed; however, some small-sized fish are not detected.

Overall, it yielded good results, except for the previously mentioned species that were not included in the model's training data. The DeepEcomar model achieved relatively modest performance, with an accuracy of 0.52, precision of 0.31, recall of 0.23, and an F1 score of 0.20. These results suggest limitations in both correctly classifying fish and consistently identifying relevant instances, indicating room for improvement in detection and recognition capabilities

2.4 Model fine-tuning

2.4.1 YOLOv8 trained on SardinIA: Due to the absence of these species in the existing DeepEcomar dataset, it became necessary to develop a new model specifically optimized for the target species and objectives of this study. Using the dataset described in Section 2.2, a new model was trained based on the YOLOv8 architecture, which currently represents the most stable version available (Yaseen, 2024 and Terven, 2023).

The model was trained using the following parameters: epochs = 50 (Figure 5), image size (imgsz) = 640, and batch size = 16. The training was performed on Google Colab using a Tesla T4 GPU with 15,095 MiB in 0.592 hours.

Based on the selected quantitative evaluation metrics and the rationale behind their choice, the model achieved the following results: an accuracy of 0.84, precision of 0.78, recall of 0.69, and an F1 score of 0.70. These values reflect a balanced performance in terms of both correct predictions and the model's ability to identify relevant instances (Figure 6).

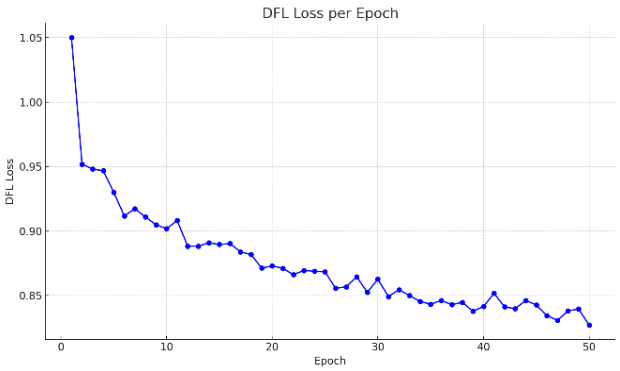

Figure 5. Loss as a function of training epochs. The curve illustrates the progressive decrease in loss over time, indicating an improvement in model performance during training.

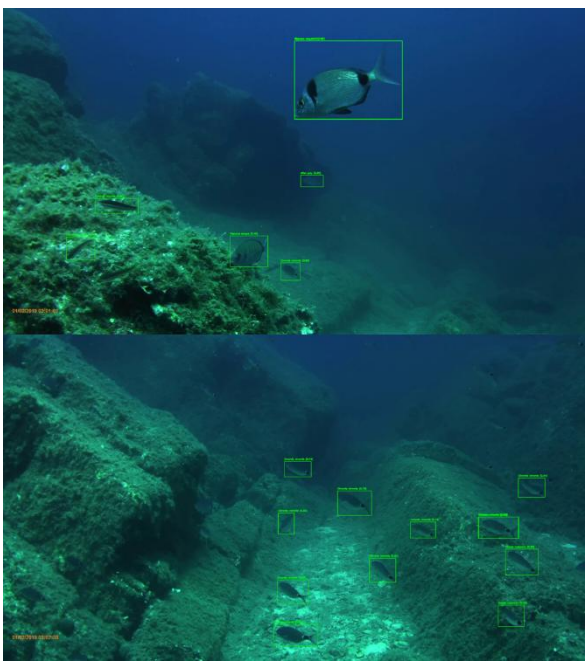

Figure 6. Example of fish detection for our trained model based on YOLOv8.

2.4.2 YOLOE trained on SardinIA dataset: Although the YOLOv8 version is the most stable, we also aimed to test the new YOLOE version by retraining the model on our dataset.

The YOLOE model developed by the THU-MIG team integrates Text-prompt, Visual-prompt, and Prompt-free three open prompt mechanisms for detection and segmentation with zero-shot performance and transferability, and the model architecture composed of a backbone, PAN, regression head, segmentation head, and object embedding head (Wang, 2025), in this paper, we choose the text prompt with Re-parameterizable Region-Text Alignment (RRTA) strategy for subsequent transfer learning.

The labelled fish dataset used for YOLOE experiments is randomly divided into training, validation and testing sets in the ratio of 8:1:1. Notably, as the YOLOEPESegTrainer is selected as the trainer for full tuning transfer learning phase, it is necessary to preprocess the existing YOLO format dataset into a "pseudo-segmentation" format by converting bounding box annotations into polygon forms for subsequent training.

The YOLOE experiments are carried out using Google Colab with an A100 GPU. The hyperparameters are configured as follows: epochs = 80 (Figure 7), batch size = 16, image size = 640, and momentum = 0.9. The best-performing weight file obtained upon completion of training is then applied to evaluate model performance on the test set. Under evaluation conditions with IoU = 0.5 and confidence threshold (conf) = 0.4, the model achieves an Accuracy of 0.7631, a Precision of 0.8636, a Recall of 0.8683, and an F1-score of 0.8660; some results are shown in Figure 8.

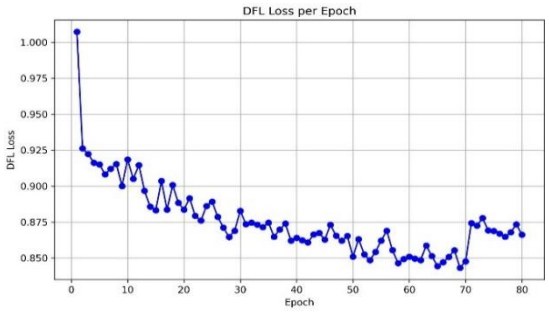

Figure 7 – Loss as a function of training epochs

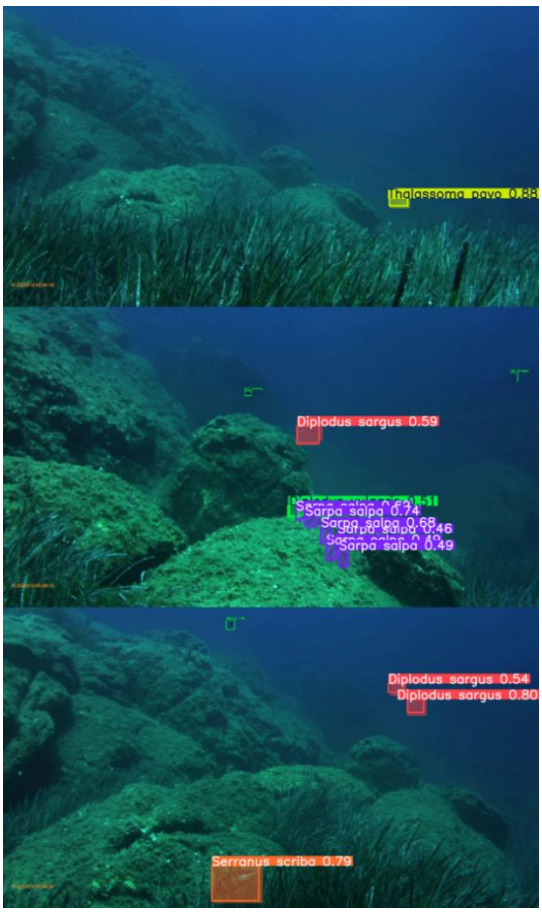

Figure 8. Some frames with the detection of YOLOE.

2.5 Data processing for Multi-Object Tracking

Multi-Object Tracking (MOT) is essential for completing fish species counting tasks, enabling an accurate assessment of aquatic fauna by maintaining consistent identities of individual fish across video frames. In this study, several trackers were considered, including, DeepSORT (Wojke et al., 2017) extends the original SORT algorithm by incorporating a deep appearance descriptor, which enhances its ability to re-identify objects after occlusions or long-term absence. This integration of a Re-Identification (ReID) model allows DeepSORT to maintain identity persistence even in crowded or complex scenes, making it widely adopted in applications such as pedestrian tracking and vehicle monitoring. However, DeepSORT relies heavily on the quality of the appearance model and can be computationally demanding. ByteTrack (Zhang et al., 2022) is a more recent tracker designed to improve tracking robustness by effectively utilizing low-confidence detections alongside high-confidence ones. By integrating both, ByteTrack reduces missed detections and identity switches, achieving state-of-the-art performance with high speed, especially when paired with YOLO detectors. Its simplicity and efficiency have led to its growing popularity in real-time tracking scenarios. More recently, SAMURAI Tracker (Yang, 2023) introduces a multi-cue fusion approach that integrates motion and appearance features to enhance tracking stability and reduce ID switches under challenging conditions such as abrupt motion or occlusion. Samurai's association mechanism leverages spatial-temporal constraints and deep feature embeddings to maintain object consistency over time. While DeepSORT emphasizes appearance-based re-identification and ByteTrack optimizes detection utilization, SAMURAI exemplifies the trend towards combining multiple cues for more reliable tracking. These trackers provide complementary strategies for MOT, enabling robust and efficient tracking across diverse application domains.

The best result from the *Recognition* phase was used to test all the trackers. Since SAMURAI was primarily designed for single-object tracking in a video, it is not an ideal choice for monitoring multiple fish simultaneously. Therefore, both DeepSort and ByteTrack were tested, initialized as follows (in order):

DeepSort: max_iou_distance=0.8, max_age=100, n_init=8, nn_budget=120;
ByteTrack: track_thresh=0.3, match_thresh=0.5, track_buffer=30, mot20=False.

Figure 9 depicts the achieved results.

3. Results and discussion

The three models showed varying performance levels in fish detection and recognition. DeepEcomar achieved modest results, with low accuracy, precision, recall, and F1 score, indicating limited reliability in classification and detection. In contrast, SardinIA demonstrated a balanced and solid performance, achieving higher accuracy and a good balance between precision and recall, making it suitable for practical monitoring tasks. YOLOE outperformed both models, delivering high precision, recall, and F1 score, thanks in part to careful tuning and a powerful training setup. These results highlight that while simpler models like DeepEcomar need improvement, advanced and well-tuned models like YOLOE offer strong potential for accurate and scalable automated fish monitoring.

Regarding MOT, SAMURAI was initially tested but later discarded because it could not track multiple objects simultaneously. Between ByteTrack and DeepSort, DeepSort showed greater compatibility with YOLO in terms of coordinate handling, making it the more suitable choice for the tracking task. The results obtained (Table 2 and Figure 10) by applying the best fish detection method combined with DeepSort were then compared with those from DOVC and UVC techniques.

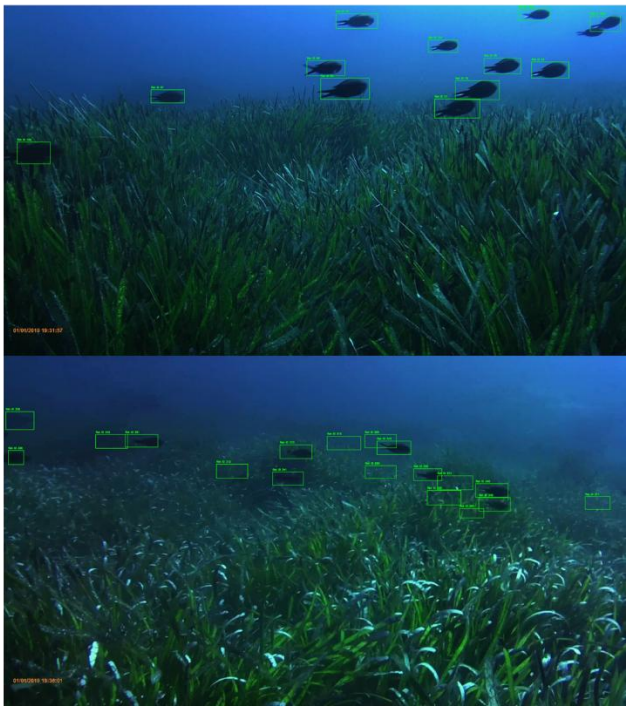

Figure 9. Top: result from DeepSort, Bottom: result from ByteTrack

Table 2. Comparison between traditional manual census techniques for fish biodiversity assessment (underwater visual census, UVC and diver operated video census, DOVC) typically involved and Deep Learning solution.

	DOVC	UVC	DeepSort
Afish poly	0	0	1
Chromis chromis	47	11	7
Coris julis	9	13	13
Diplodus sargus	8	3	15
Diplodus vulgaris	9	1	15
Ephinephelus marginatus	1	0	0
Mullus sp-	0	0	1
Oblada melanura	94	100	180
Sarpa salpa	12	8	9
Serranus scriba	2	5	2
Sparus aurata	2	0	2
Spicara maena	0	45	0
Spicara smaris	71	0	6
Thalassoma pavo	1	1	10
TOT	260	190	261

The DeepSort-based automated method achieved a total fish count (261) almost identical to DOVC (260), indicating strong reliability in overall abundance estimation. However, differences emerged at the species level. Some species, like *Spicara smaris*, were underestimated, while others, like *Oblada melanura*, were overestimated.

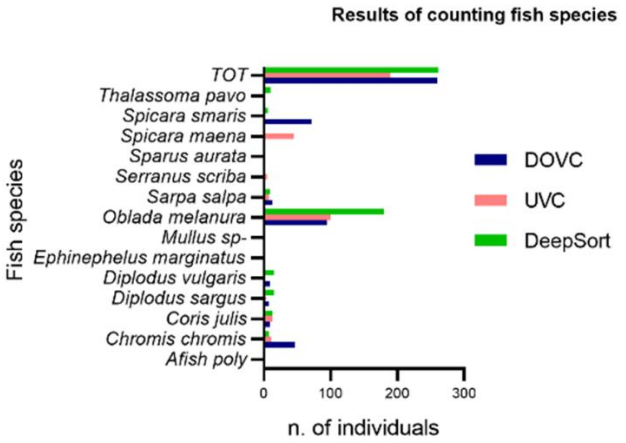

Figure 10. Number of fish individuals per species detected using the three analysis techniques: underwater visual census (UVC), diver operated visual census (DOVC) and Deep Sort.

These discrepancies are likely due to visual similarities between species, which can confuse even advanced models. Despite this, DeepSort showed good agreement with DUVC for several common species. Overall, the results suggest that DeepSort is a promising tool for scalable fish monitoring, with performance comparable to traditional digital census methods, though species-level classification can still be improved. Importantly, the most abundant species identified by DeepSort was the same as in DOVC and UVC (*Oblada melanura*), confirming consistency in identifying dominant species. In general, orders of magnitude across species remained coherent, reinforcing the method's reliability for ecological assessments.

The next steps will involve improving the recognition phase, aiming to obtain values that more closely match manual counts, particularly for species that are very similar to each other. This will be achieved by refining both the dataset and the training process.

4. Conclusions

This study demonstrated the potential of deep learning techniques for automated monitoring of fish biodiversity in complex marine environments like the Mediterranean. Through a comparative evaluation of various detection models and tracking algorithms, the YOLOE model trained on a localized dataset (SardinIA and the DeepSort multi-object tracker emerged as the most effective combination. The developed system delivered performance comparable to traditional visual census methods, ensuring reliability in overall abundance estimation and accurate identification of dominant species. However, some difficulties in classifying visually similar species highlight the need for further improvements, such as incorporating additional data modalities or adopting multimodal models. Overall, this work supports the adoption of automated technologies to enhance the efficiency, scalability, and reproducibility of marine monitoring efforts, providing valuable tools for the management and conservation of protected ecosystems.

Acknowledgements

The acquisition part of this study is a working package of a research project of Italian national relevance initiative "Italia Domani - Piano Nazionale di Ripresa e Resilienza" (PNRR,

European Commission, Next Generation EU) with the name "multitemPOral SEagrass mapping and monitoring of posIDONia meadows and banquettes for blue carbon conservation (POSEIDON)".

The application of artificial intelligence for fish recognition and tracking and the definition of dataset of Mediterranean Sea was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

The authors would like to express our sincere gratitude to Ahsan Jalal, Ahmad Salman, Ajmal Mian, Salman Ghafoor, and Faisal Shafait for their kind availability and support in allowing us to use DeepFins and to Dott. Fabio Menna for the suggestion related to the Labelled Mediterranean Fish Dataset.

References

Brock, R. E., 1982. A critique of the visual census method for assessing coral reef fish populations. *Bulletin of Marine Science*, 32, 269–276.

Catalán, I., & Alvarez-Ellacuria, A., 2023. Labelled Mediterranean fish dataset. doi.org/10.5281/zenodo.7534425

Guidetti, P., Baiata, P., Ballesteros, E., Di Franco, A., Hereu, B., Macpherson, E., ... & Sala, E. 2014. Large-scale assessment of Mediterranean marine protected areas effects on fish assemblages. PLoS ONE, 9(4), e91841.

Harmelin-Vivien, M. L., Harmelin, J. G., Chauvet, C., Duval, C., Galzin, R., Lejeune, P. Lasserre, G. 1985. Evaluation visuelle des peuplements et populations de poissons: méthodes et problèmes. Revue d'Écologie,, 467–539. doi.org/10.1371/journal.pone.0091841.

Jalal A, Salman A, Mian A., Ghafoor S., Shafait F., 2025 DeepFins: Capturing dynamics in underwater videos for fish detection, *Ecological Informatics, Volume 86*. doi.org/10.1016/j.ecoinf.2025.103013.

Murphy, H. M., Jenkins, G. P., 2010. Observational methods used in marine fish monitoring: a review. Marine and Freshwater Research, 61(2), 236–252.

Sala, E., Ballesteros, E., & Starr, R.M. ,2012. Rapid decline of a Mediterranean ecosystem: the case of Fucus forests. *Marine Ecology Progress Series*, 452, 63-72.

Redmon, J., & Farhadi, A., 2018. YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.

Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A. .2023. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction, 5(4), 1680–1716. doi.org/10.3390/make5040083.

Thanopoulou Z, Sini M, Vatikiotis K, Katsoupis C, Dimitrakopoulos PG, Katsanevakis S. 2018. How many fish? Comparison of two underwater visual sampling methods for monitoring fish communities. PeerJ 6:e5066 doi.org/10.7717/peerj.5066 .

LabelImg, Tzutalin, 2015. Sofware. Open source Git code. github.com/tzutalin/labelImg.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M., 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang A, Liu L, Chen H, et al. Yoloe: Real-time seeing anything[J]. arXiv preprint arXiv:2503.07465, 2025.

Wojke N., Bewley A. and Paulus D., 2017. Simple online and realtime tracking with a deep association metric, IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 3645-3649, doi: 10.1109/ICIP.2017.8296962.

Xu, Y., Wang, X., Zhang, Z., Feng, X., Chen, J., & Song, M. 2022. YOLOE: You Only Look at One Representation for Efficient Object Detection. *arXiv preprint arXiv:2203.16250*.

Yang, C. 2023. Samurai Tracker: A Multi-Cue Fusion Approach for Robust Multi-Object Tracking. Sofware. Open source Git code. GitHub repository: github.com/yangchris11/samurai.

Yaseen, M. , 2024. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. arXiv:2408.15857. arxiv.org/abs/2408.15857.

Zhang, Y., Wang, C., Wang, X., & Yuan, Y., 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. AAAI Conference on Artificial Intelligence. ojs.aaai.org/index.php/AAAI/article/view/21037.