Journal of Sustainable Development of Energy, Water and Environment Systems

http://www.sdewes.org/jsdewes

ISSN 1848-9257
Journal of Sustainable Development of Energy, Water and Environment Systems
https://www.sdewes.org/jsdewes

Year 2026, Volume 14, Issue 1, 1130635

Original Research Article

An Improved Deep Learning Method for Detecting Marine Plastic Litter

***Abdelaadim Khriss**[*1], **Aissa Kerkour Elmiad**[1,] **Mohammed Badaoui**[2],*
***Mimoun Yandouzi**[3], **Mounir Grari**[4], **Alae-Eddine Barkaoui**[5], **Yassine Zarhloule**[5]*
[1]Lab. LARI, FSO, Mohammed Premier University, Oujda, Morocco
[2]Lab. LaMSD, ESTO, Mohammed Premier University, Oujda, Morocco
[3]Lab. LSI, ENSAO, Mohammed Premier University, Oujda, Morocco
[4]Lab. MATSI, ESTO, Mohammed Premier University, Oujda, Morocco
[5]2GPMH Lab., FSO, Mohammed Premier University, Oujda, Morocco
e-mail: abdel.abdelkrs@gmail.com

Cite as: Khriss, A., Kerkour Elmiad, A., Badaoui, M., Yandouzi, M., Grari, M., Barkaoui, A. E., Zarhloule, Y., An Improved Deep Learning Method for Detecting Marine Plastic Litter, J.sustain. dev. energy water environ. syst., 14(1), 1130635, 2026, DOI: https://doi.org/10.13044/j.sdewes.d13.0635

ABSTRACT

Plastic pollution in the ocean is a widespread issue in the marine biosphere that requires large-scale monitoring systems. However, extending the use of deep-learning-based detection methodologies to real-world marine settings is difficult. Reflections from surfaces, small target objects, and partially blocked debris often hinder the effective functioning of such systems. To mitigate these challenges, this paper presents a detection system built on the You Only Look Once architecture that explicitly considers the constraints above. The architecture combines two mutually supportive modules: a directional coordinate attention module, which decodes spatial dependencies along horizontal and vertical axes, and a Sinkhorn Distance-based regularisation term, which stabilises feature focus across spatial dimensions. Experimental testing of image collections by aerial and underwater cameras shows significant performance improvements compared to the latest state-of-the-art assessments. The proposed system achieves a precision of 92% and a recall of 94% in aerial scenes and a precision of 90% and a recall of 92% in sub-aqueous scenes. An ablation study validates the hypothesis that the two modules work together to improve performance. Furthermore, visual inspection yields more reliable results for detecting typical marine debris, including reflective artefacts, small objects, and visually contaminated scenes.

KEYWORDS

Wastewater, Marine debris, Plastics detection, Environmental monitoring, Deep learning, Optimal transport.

INTRODUCTION

Plastic pollution has become a pressing issue of the current era, with significant implications for marine ecosystems. It is estimated that the amount of plastic in the world's oceans ranges between 100 and 150 million tons, with approximately 6.4 million tons of plastic deposited in the waters every year [1]. Plastics are difficult to decompose naturally and may take centuries to break down in the ocean. Over time, these polymers break down into microplastics and nanoplastics, which are difficult to see with the naked eye but deadly to marine life, the stability of the marine environment, and human health [2].

[*] Corresponding author

The growing crisis of plastic pollution is aggravated by the further increasing world production of plastics and the lack of recycling facilities. The extensive presence of plastic debris in the oceans and their deep layers prevents systematic surveillance and reduction efforts. The complexity of underwater ecology, including low visibility, unstable photic regimes, and refraction distortions, makes it difficult to identify and taxonomically assign sunken plastic debris successfully.

To overcome the above issues, more research has been conducted on the use of automated methods for detecting marine litter. Although manual shoreline monitoring and survey aerial technology have good datasets, they are labour-consuming, expensive and restricted in space and time [3]. In its turn, the recent advancements in robotics and sensor technologies, namely, remotely operated vehicle (ROV) and unmanned aerial vehicle (UAV), have made marine debris detection more effective and feasible. These systems provide safer, more economical, and scalable mechanisms for monitoring surface and underground conditions.

The fundamental component of such systems is computer vision algorithms. Convolutional neural networks (CNNs) are one such algorithm and have significantly advanced object detection capabilities by processing visual data efficiently and effectively [4]. Specifically, the You Only Look Once (YOLO) family of models has been widely adopted due to its excellent balance between inference speed and detection accuracy [5]. However, high visual similarity among unequal objects and high intra-class variability still limit detection in submerged data [6].

The current studies are still centred on reinforcing machine learning-based models to improve adaptability to the marine environment. Wang *et al.* proposed new network designs, such as Extended Efficient Layer Aggregation Networks, cascaded scaling planning, and specialised module layouts, to improve the efficiency of general detection [7]. Due to its small size and simple deployment, the YOLOv7-Tiny model has been highly successful in locating submerged marine debris. This model comes with adaptive anchor-box computation and massive data augmentation policies. Its feature extractors incorporate strong MaxPool blocks [8], thus enhancing detection performance [9]. However, the YOLOv7-tiny model performs poorly in identifying small or masked objects in low-contrast underwater data, primarily due to its limited attention ability.

To address these shortcomings, research studies have aimed to improve detection models by using lightweight neural network models and special attention mechanisms. For example, Chen *et al.* used the GhostNetv2 framework in WorldOv5s. They obtained a significant decrease in computational cost at the expense of improving the efficiency with which items are recognised in underwater conditions [10]. Qiang *et al.* improved the Single Shot MultiBox Detector (SSD) model by replacing the Visual Geometry Group (VGG) backbone with Residual Network (ResNet), achieving higher detection and processing speeds that can be used in oceanic surveys [11]. Huang *et al.* incorporated the Mobile Network (MobileNetv2) into the YOLOv3 framework, achieving a good balance of speed, feature extraction speed, and detection accuracy [12]. Furthermore, Wang *et al.* developed a lightweight, ocean exploration-specific attention mechanism that enhances environmental perception ability [13].

Attention mechanisms are now particularly relevant. Wen *et al.* simultaneously added a combination of Coordinates Attention (CA) and Squeeze-and-Excitation (SE) module pairs in YOLOv5s, effectively improving feature extraction for blurry ocean imagery [14]. Shen *et al.* added CA module units in YOLOv5 in a suitable position for advanced refined identification capability while inhibiting unnecessary data [15]. Chen *et al.* added the Convolutional Block Attention Module (CBAM) attention module in YOLOv7, considerably improving detection performance for smaller ocean debris fragments [16]. Liu *et al.* drove the expansion of attention strategies in YOLOv8n with the Simple Attention Module, improving efficient global feature understanding without a greater computational load [17]. As much as these innovations individually exhibited a form of performance, many are locally functional within networks, currently with a deficiency in global attention strategies for ensuring enduring, sustained comprehension within oceanic scenes.

Parallel developments in surface litter detection have also contributed valuable insights. Wang *et al.* proposed UAV-YOLOv8, integrating Wise – Intersection-over-Union v3 for improved localisation, BiFormer attention to enhance critical feature retention, and Focal FasterNet blocks for reducing small target miss rates in aerial imagery **[18]**. Qiao *et al.* introduced CA into a YOLO-based framework. They replaced the traditional Feature Pyramid Network (FPN) with a Bidirectional FPN module, achieving improved detection precision for very small floating debris **[19]**. Li *et al.* enhanced the feature extraction capacity of lightweight detectors by introducing a Feature Mapping Attention layer, supported by extensive data augmentation **[20]**. C. Hou *et al.* incorporated the HorBlock module and applied a genetic algorithm to optimise model generalisation in complex aquatic conditions **[21]**.

There are also a number of pre-processing techniques, which have brought about significant improvements. Zhang and Liu proposed an improved YOLOv5 algorithm for small target detection in aerial imagery, which inspired pre-processing strategies for underwater detection **[22]**. Ma *et al.* used the CBAM mechanism with a variant of YOLOv5s that used a Non-Weighted Distance Intersection-over-Union loss to enhance the accuracy of small floating objects detection **[23]**. Q. Hou *et al.* introduced an attention mechanism, which enables the model to capture long-range dependencies while maintaining fine-grained localisation capability, thereby improving feature extraction and detection performance in complex visual environments **[24]**.

Although the design of attention mechanisms, including CA, SE and CBAM, has shown clear improvements, existing methodologies continue to be inherently limited in terms of their ability to simulate global spatial structure **[25]**. These schemes mainly enhance the saliency of local features but often lead to the effect of the fragmentation or discontinuous distribution of attention throughout feature maps. These shortcomings are especially detrimental when data are collected in the ocean, where glare, low contrast, and complicated light scattering hinder the precise identification of little, concealed, or morphologically deformed plastic debris.

To address the limitations of current disjointed attention and the lack of spatial regularity in existing detection models, this paper proposes a YOLOv7-based model that incorporates CA and Sinkhorn Distance regularisation. The model's key innovation is its treatment of attention alignment as an optimal transport problem. In this context, directional distributions of attention are formulated as probability measures, which have to be compatible in spatial dimensions. CA combines contextual information in horizontal and vertical orientations through direction-conscious positional encoding. The design enables the network to obtain long-range dependencies without compromising on spatial resolution – which is a requirement in the process of recognition of irregular, partially obscured, or ambiguous marine waste **[26]**. However, local attention does not guarantee coherent attention across the entire feature map.

In order to obtain global structural alignment, the metric applied as a transport measure is the Sinkhorn Distance, which is regularised with entropy. It is a matching problem on distributions where attention has consistency minimised by redistributing attention in an entropic manner. The theory of this formalism offers a basis for the attainment of globally consistent and semantically compatible distributions of attention that outperform more general heuristic integration strategies, which dominate more modern models. The coherent components combine to form the unified attention system, which works on the optimal transport theory and balances the sensitivity of features and overall coherence of space. In this way, the accuracy and precision of detection under such harsh visual conditions, as in those observed in the air and subsurface environment conditions in the sea, are improved tremendously.

The remainder of the paper will be structured as follows: Section 2 will describe the proposed methodology. Section 3 covers such empirical findings as performance standards, ablation experiments, and visual detection examples. Section 4 addresses these findings alongside the rest of the problems of marine data localisation. Finally, Section 5 will provide the future research directions and implications.

PROPOSED METHOD

This section presents the approach to the process of refining the object-detection system to recognise marine plastic waste using deep learning and feature-based refinement. First, unmanned aerial vehicles equipped with specialised imaging sensors collect data, providing high-resolution images of plastic waste on beaches and underwater. After acquiring the image, the pre-processing protocol involves enlarging the image through rescaling and augmentation to meet the input format requirements of the detection model. The given pipeline is then detected using a YOLOv7-based backbone, which is lightweight and optimised for real-time use. To improve detection accuracy, CA mechanisms focus on crucial spatial features. Sinkhorn-based regularisation methods are used to ensure consistent, optimised detection results. **Figure 1** illustrates the overall methodology, which involves acquiring images with UAVs and ROVs, identifying objects, accurately classifying them, and analysing overall performance.

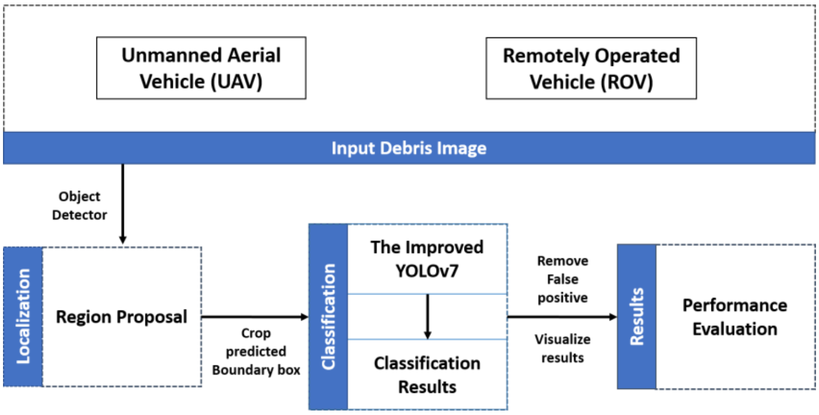


Figure 1. Overview of the proposed detection pipeline for marine plastic litter using UAV and ROV imagery

A direction-aware CA scheme is used to improve the limitations of attention modules using a conventional scheme. The scheme records long-range interactions by combining the spatial context in the horizontal and vertical dimensions separately. Directional encodings are then regularised by using Sinkhorn Distance, which helps enforce global consistency amongst the weights of spatial attention. Specifically, the attention maps are handled as normalised probability distributions and are trained through the entropy-regularised optimal transport and thus consistent and focused attention across feature maps [27]. The classification and localisation operations of the YOLOv7 detection head then use the modulated features. **Figure 2** illustrates the mechanisms of this attention module, emphasising the extraction of spatial descriptors, the creation of attention masks, and regularisation via Sinkhorn Distance.

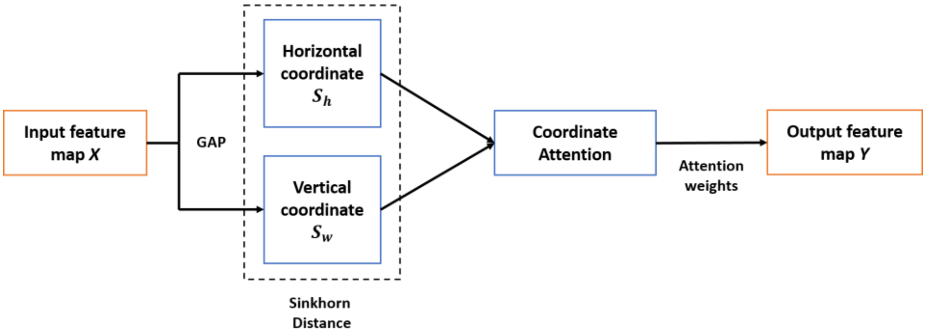


Figure 2. Directional Coordinate Attention with Sinkhorn regularisation: horizontal and vertical descriptors extracted via GAP and refined by 1×1 convolutions, and alignment using Sinkhorn Distance for coherent spatial attention

Coordinate Attention with Directional Context Encoding

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width, respectively, Global Average Pooling (GAP) is applied separately along the horizontal and vertical axes. This approach yields two aggregated feature descriptors that preserve spatial context in their respective directions:

$$s^h(c,j) = \frac{1}{W}\sum_{i=1}^{W} X(c,j,i), \quad s^w(c,i) = \frac{1}{H}\sum_{j=1}^{H} X(c,j,i) \tag{1}$$

Where $X(c,j,i)$ represents the activation value at channel c, row j, and column i of the feature map, where C, H and W denote the number of channels, height, and width, respectively. The terms $s^h \in \mathbb{R}^{C\times H}$ and $s^w \in \mathbb{R}^{C\times W}$ correspond to the vertically and horizontally aggregated descriptors for channel c at position j, and i, respectively. These descriptors capture directional spatial context through global average pooling along the horizontal and vertical axes.

To generate attention maps from these descriptors, independent 1×1 convolutional layers followed by sigmoid activation are applied:

$$M^h = \sigma\left(Conv^{1\times 1}(s^h)\right), \quad M^w = \sigma\left(Conv^{1\times 1}(s^w)\right) \tag{2}$$

Here, $Conv^{1\times 1}(\cdot)$ denotes a 1×1 convolutional operation that refines the direction-aware descriptors, and $\sigma(\cdot)$ is the sigmoid activation function that normalises the output values to the range $[0, 1]$. The resulting maps M^h and M^w are the vertical and horizontal attention masks, respectively, which emphasise the most informative spatial features in each direction.

Sinkhorn Regularisation for Global Attention Consistency

To ensure coherent and balanced attention distributions across spatial dimensions, Sinkhorn Distance is introduced as a regularisation mechanism. Given the unnormalised attention maps $M^h \in \mathbb{R}^{C\times H}$ and $M^w \in \mathbb{R}^{C\times w}$, they are flattened into probability vectors $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$, normalised so that $\sum p = \sum q = 1$.

The Sinkhorn Distance is defined as:

$$Sinkhorn(p,q) = \min_{M\in\mathbb{R}^{n\times m}} \langle M, C\rangle - \epsilon H(M) \tag{3}$$

Where p, q represent the normalised attention distributions derived from the horizontal and vertical attention maps. The matrix M denotes the transport plan that aligns the two distributions, while $C \in \mathbb{R}^{n\times m}$ is a cost matrix based on pairwise spatial distances. The notation $\langle M, C\rangle$ refers to the Frobenius inner product, computed as the element-wise product followed by summation. The coefficient ϵ is a small positive scalar controlling entropy regularisation, and $H(M) = -\sum_{i,j} M_{ij} \log M_{ij}$ is the entropy term that encourages smoother transport plans.

This formulation ensures that the learned attention weights for horizontal and vertical dimensions are mutually aligned, reducing redundancy and enforcing spatial structure. During training, the Sinkhorn loss is added to the overall objective:

$$L_{total} = L_{YOLO} + \lambda \times Sinkhorn(p,q) \quad (4)$$

Where L_{YOLO} represents the standard YOLOv7 detection loss, which combines classification, localisation, and confidence terms. The parameter λ is a balancing coefficient that adjusts the contribution of the Sinkhorn regularisation term, ensuring that attention alignment complements but does not dominate the overall learning objective.

Feature Modulation and Integration into you Only Look Once

After Sinkhorn-based regularisation, the attention weights are reshaped and broadcast across spatial dimensions to re-weight the original feature. This effect is achieved through element-wise multiplication:

$$Y = X \times M^h \times M^w \quad (5)$$

Where $M^h \in \mathbb{R}^{C\times H\times 1}$ and $M^w \in \mathbb{R}^{C\times 1\times W}$ are the vertically and horizontally broadcast attention masks, respectively, ensuring dimensional alignment with the original input feature map X. The resulting feature map $Y \in \mathbb{R}^{C\times H\times W}$ effectively combines directional context from both spatial dimensions, yielding enhanced feature representations with globally coherent and spatially attentive focus.

The obtained modulated feature map is sent to the YOLO detection head, which uses a multi-level aggregation scheme based on a path aggregation network and multi-layer, anchor-based predictions. This combination improves localisation accuracy, especially when detecting small or indistinct marine debris, by refining attention.

EXPERIMENTAL RESULTS

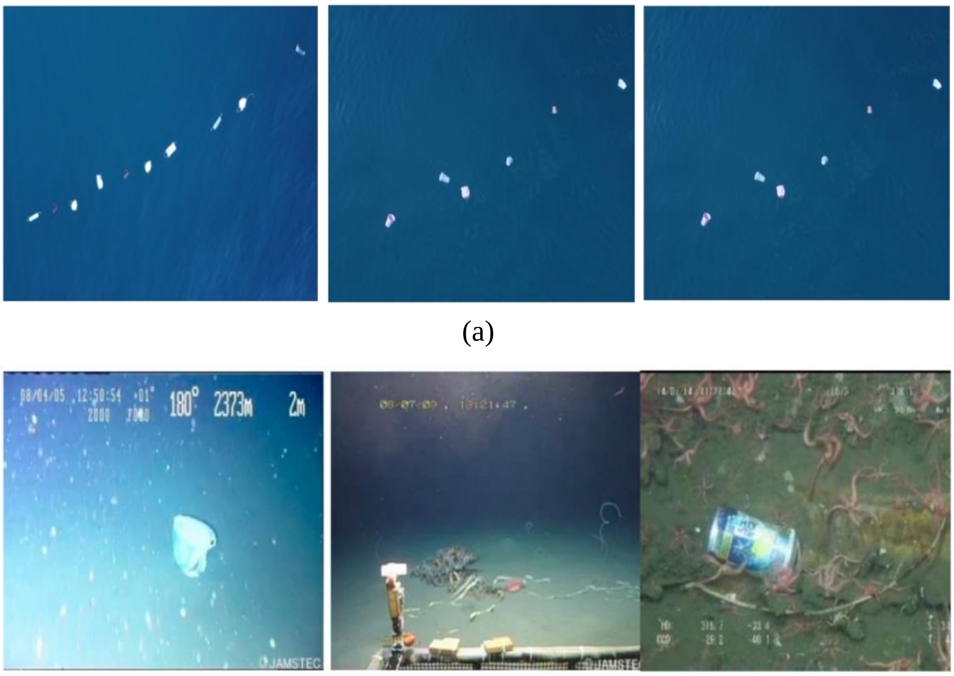
In this section, the proposed method is evaluated through quantitative benchmarks, ablation studies, and visual comparisons. Performance is assessed on both aerial and underwater datasets, focusing on detection accuracy, robustness, and the impact of the architectural enhancements in challenging marine environments.

Training Configuration

All the experimental processes were performed on an NVIDIA Tesla T4 card with 15360 MiB of memory and a NVIDIA driver version of 525.105.17 on Compute Unified Data Architecture (CUDA) 12.0, which ensured a smooth hardware acceleration. The training pipeline was set to have a batch size of 32, and the model was trained in 50 epochs.

Dataset

To train and test the proposed model, two datasets were utilised, namely TrashCAN [28] dataset on underwater detection and an aerial dataset, taken by the GreenTech Solution team [29]. TrashCAN contains 7212 annotated images of underwater debris and marine life, mainly obtained through the JAMSTEC E – Library of Deep Sea Images. The aerial dataset comprises 6,589 images obtained by drones, having 51,840 annotations, which are images of the floating litter of different types in diverse marine conditions. The aggregate set of 13,801 images was randomly divided into training (70%), validation (15%), and testing (15%) sets. A five-fold cross-validation scheme served to train the model in order to detect model robustness and reduce variance. Images were standardised in terms of pre-processing, which involved resizing to 640×640 pixels and preserving the aspect ratio by padding, and normalising with ImageNet statistics (mean: 0.485, 0.456, 0.225, standard deviation: 0.229, 0.224, 0.225). Bounding-box annotations were scaled in proportion to resizing, and corrupt images were filtered before partitioning. At training, the data augmentation was only used, which also entailed random rotations (90°, –25 to +25%), saturation modifications (–25 to +25%), and brightness modifications (0 to +20%). The probability of each augmentation was 0.5 to allow combinations of multiple augmentations on any image. These additions increased the generalisation of the model to underwater and aerial changes in imaging. The sample of both datasets is shown in **Figure 3**.



(a)

(b)

Figure 3. Sample images from marine debris detection datasets: UAV-based images from the MarineLitter dataset (a), ROV-based images from the TrashCAN dataset (b)

To quantitatively assess the performance of the proposed model, three standard metrics widely used in object detection were employed: *Precision*, *Recall*, and mean Average Precision *(mAP)*. *Precision* measures the proportion of correctly identified positive instances among all predicted positives, offering a sense of how reliable the detections are. *Recall* reflects the ability of the model to identify all relevant objects, indicating how many true positives were detected out of all actual positives. They said metrics are defined as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{6}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{7}$$

The mean Average Precision *(mAP)* serves as an aggregated measure, averaging the *Precision* across all classes and confidence thresholds. It is calculated as:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP(i) \tag{8}$$

Where $AP(i)$ denotes the average *Precision* for class *i*, and *n* is the total number of classes. In practice, *mAP* is reported at specific Intersection-over-Union (IoU) thresholds. *mAP@*0.5 indicates the *mAP* calculated at an IoU threshold of 0.5, while *mAP@*0.5:0.95 denotes the average *mAP* over IoU thresholds from 0.5 to 0.95 with a step of 0.05, providing a stricter and more comprehensive evaluation.

Quantitative Evaluation

The proposed model's performance was compared to the most popular baselines of detecting (Faster R-CNN) [30], SSD [31], and YOLOv7, on two marine datasets, and the results were summarised in **Table 1**. The model obtained significant gains in all the metrics tested, reaching a very high score of 94.00 per cent *mAP* at 0.5, which is a significant improvement over the standard YOLOv7 by more than 2.5 percentage points. Comparably, the proposed method performed better than all the baselines on the TrashCAN dataset with 92.80 at 0.5 and 72.00 at 0.5:0.95 *mAP*. Its sensitivity to difficult or partially occluded objects is further brought out by the fact that the *Recall* increased to 92.30%. These findings show that the technique can be used to generalise both in aerial and underwater environments with a high detection rate, even in different visual environments.

Table 1. Performance comparison of the proposed method and baseline models

Dataset	Model	*Precision* [%]	*Recall* [%]	*mAP*@0.5 [%]	*mAP*@0.5:0.95 [%]
Submarine (MarineLitter)	Faster R-CNN	87.51	88.30	89.62	70.92
	SSD	86.53	86.22	88.57	70.2
	YOLOv7	88.55	89.90	91.41	71.5
	Proposed method	92.01	94.17	94.00	72.2
Underwater (TrashCAN)	Faster R-CNN	85.20	84.85	87.10	69.8
	SSD	84.35	83.40	86.25	69.2
	YOLOv7	86.80	87.15	89.35	70.5
	Proposed method	90.15	92.30	92.80	72.0

In order to add more information about the performance of the classification, **Figure 4** shows the normalised confusion matrix of the given method on the TrashCAN dataset. The diagonal entries are an appropriate classification, where the majority of the classes have *Precision* values greater than 0.90, which means that they are very discriminant. It is important to note that items like "trash_bottle", "trash_pipe", and "trashcan" exhibit almost perfect classification as the values are 1.00, 1.00 and 0.96. The categories of marine life, such as "animal_starfish" (0.93), "animal_etc" (0.80), "animal_fish" (0.86) and ROV (0.94) also have strong detection accuracy, indicating that the model can differentiate between debris and the natural or working objects. The other categories of trash are all performing with a high degree of performance, with "trash_bag" (0.93), "trash_clothing" (0.93), "trash_container" (0.93), "trash_cup" (0.94), and "trash_rope" (0.94) all having a greater *Precision* of over 0.93. The *Precision* of the "trash_net" class is 0.83, and "trash_tarp" is 0.79, which both prove to be very useful in detecting flexible and deformable forms of garbage. A plant category reaches a value of 0.87 accuracy, as the model is able to find the organic marine elements precisely. The confusion matrix in general confirms the performance of the Sinkhorn-regularised attention scheme in improving feature differentiation, and a low amount of cross-category misclassification and high accuracy by classes is observed throughout the various range of underwater items.

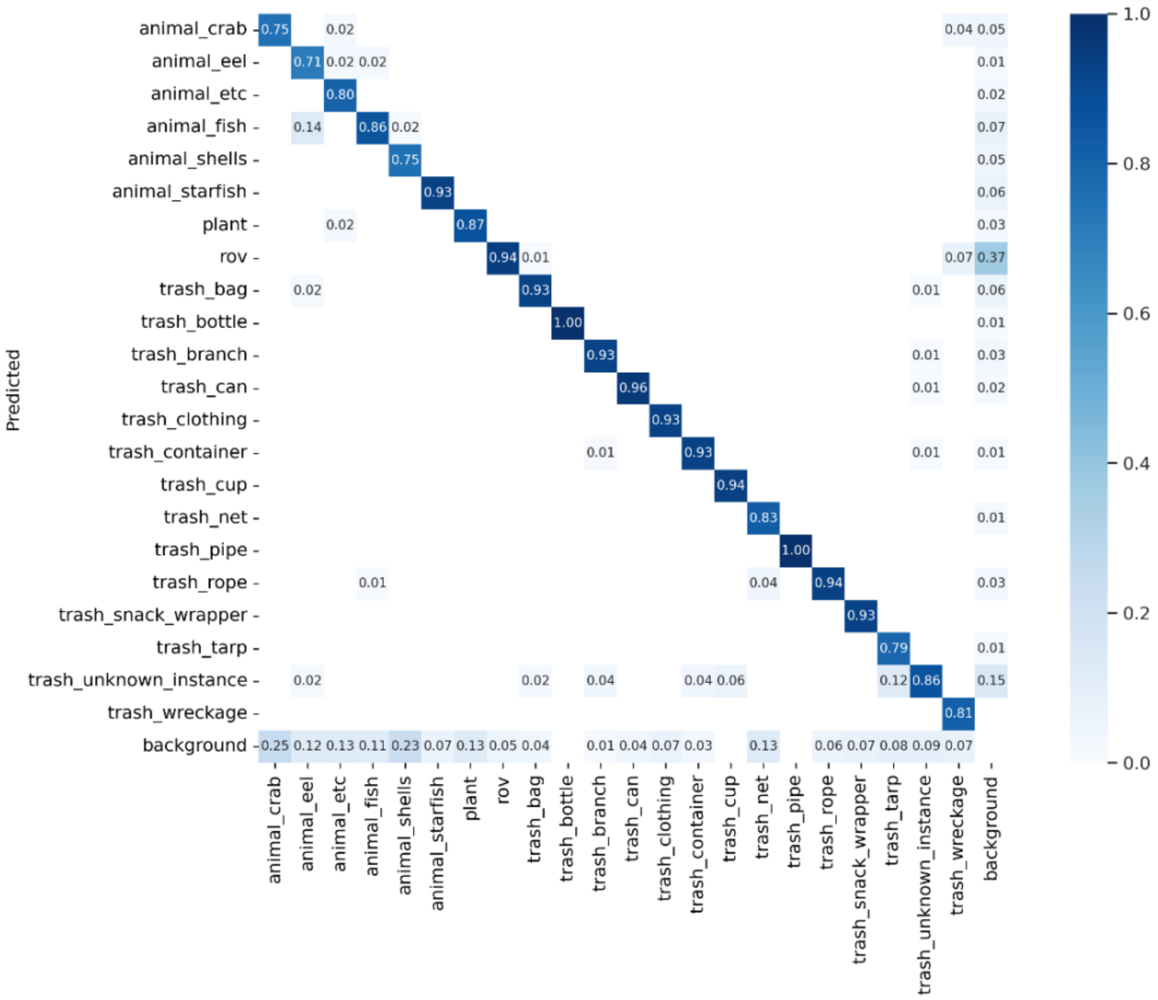


Figure 4. Normalised confusion matrix of the proposed method on the TrashCAN dataset, showing high classification accuracy across marine debris and biological classes

Ablation Study

Each of the proposed architectural improvements was quantitatively evaluated through an ablation study on the TrashCAN dataset. The results of four model configurations, including the baseline YOLOv7, the YOLOv7 with CA, the YOLOv7 with Sinkhorn Regularisation and the complete proposed method combining the two modules, are provided in Table 2. The *Precision* and *Recall*, with the addition of CA alone, showed the advantage of directional spatial information encoding in feature maps. The use of Sinkhorn Regularisation also helped to achieve better alignment and stability in attention maps, and performance improvement was found primarily in *mAP*. The model scored the best when both enhancements were realised, which is indicative of the complementary nature of the localised attention focus and global spatial coherence.

Table 2. Ablation results showing the impact of each module on detection performance

Model variant	*Precision* [%]	*Recall* [%]	*mAP*@0.5 [%]	*mAP*@0.5:0.95 [%]
Baseline YOLOv7	86.80	87.15	89.35	70.50
+ Coordinate Attention	88.05	89.50	90.60	71.10
+ Sinkhorn Regularisation	87.40	88.70	90.05	70.95
Proposed method	90.15	92.30	92.80	72.00

Figure 5 depicts the train loss curve of all the ablation variants. The original YOLOv7 model has a lower convergence speed and final loss than the improved ones. Either CA or Sinkhorn regularisation has a minor effect on accelerating convergence, but the largest acceleration is seen when the two modules are used together. The offered method not only yields the final lowest loss but also has a more stable and regular descent progression over the training epochs, which is a better sign of optimisation dynamics and better feature learning during the training process.

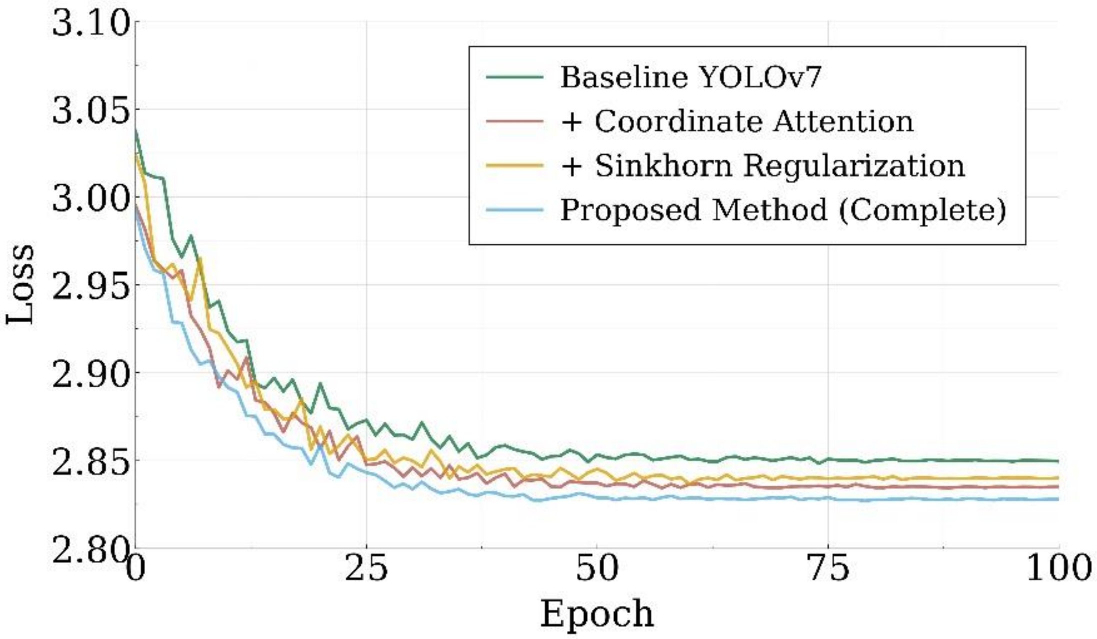
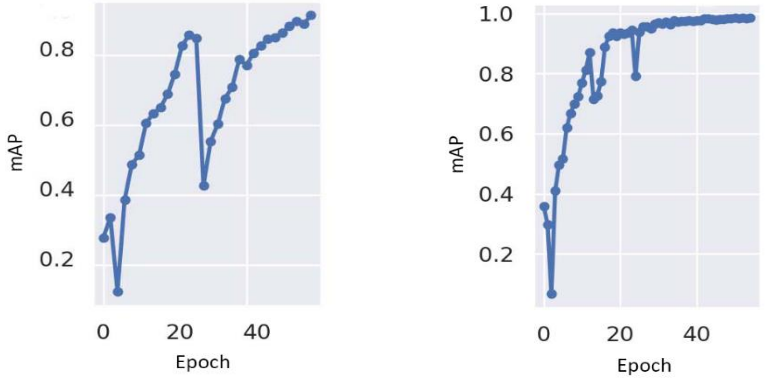


Figure 5. Loss curves for ablation variants on the TrashCAN dataset

Meanwhile, **Figure 6** demonstrates the curves of the *mAP* progression of both the baseline YOLOv7 (**Figure 6a**) and the proposed approach (**Figure 6b**) during the training period. The suggested model has more of an initial accuracy increase and stays on a higher *mAP* during the training. This result implies the model has better intermediate representations during the initial training phase, and thus it is more likely to converge faster and prevent overfitting or plateauing. The more continuous and higher *mAP* curve indicates better generalisation and efficient acquisition of discriminative features, even on difficult underwater imagery.



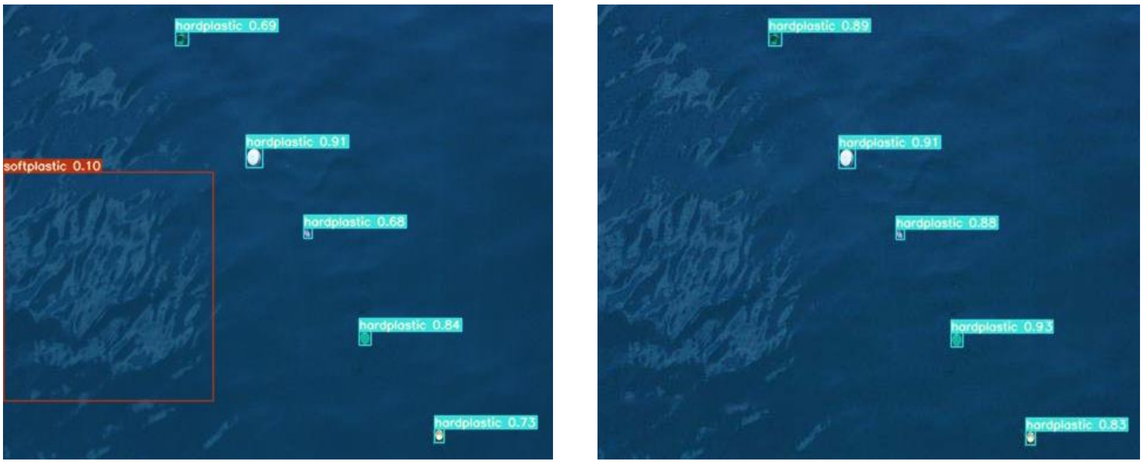
(a) (b)

Figure 6. mAP curves during training: standard YOLOv7 model (a); proposed method (b)

Detection Visualisations

Problematic situations requiring visual comparisons, such as reflection artefacts, small object detection, and underwater clutter, are offered to estimate the strength of the proposed approach in real-life settings. These examples are typical sources of error in the detection of plastic litter in the sea.

In aerial imagery, **Figure 7** illustrates the effects of surface reflections on the accuracy of detection. The original YOLOv7 baseline (**Figure 7a**) incorrectly identifies the reflective region as "soft plastic", assigning it a predetermined confidence score. In contrast, the proposed method (**Figure 7b**) completely prevents this error, enhancing the certainty of true positive findings. This effect makes the model less sensitive to noise, especially in images with high-frequency water textures, which can confuse conventional convolutional filters.

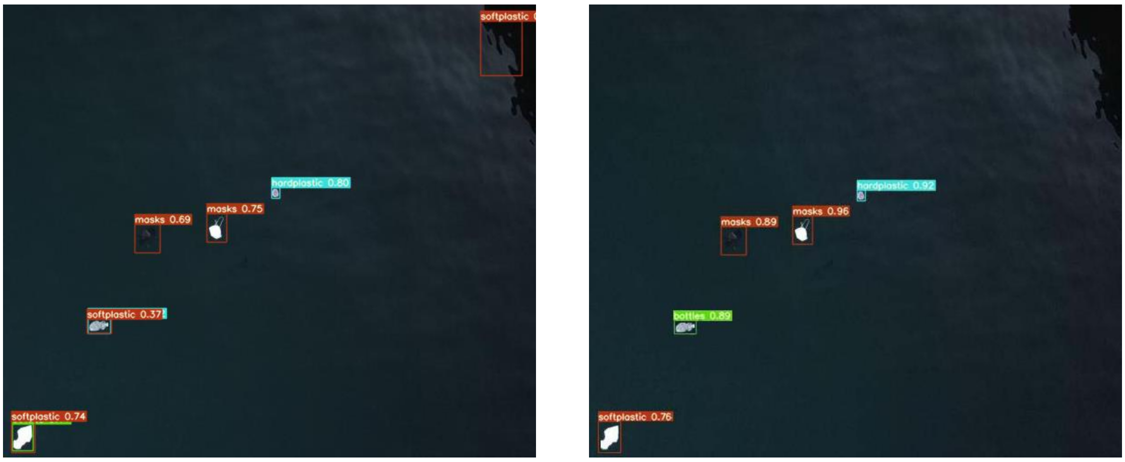


(a) (b)

Figure 7. Detection results in the presence of surface reflection artifacts:
standard YOLOv7 model (a); proposed method (b)

The direction-aware CA process ensures that the purely spatial functionality pays more attention to horizontal and vertical directions than to other appearance variations. This effect enables the network to distinguish structured debris from dynamic surface distortion in a dynamically evolving surface. Attention maps, further smoothed by Sinkhorn Distance, introduce spatial coherence and effectively reject non-salient patterns that generally resemble floating litter.

Figure 8 demonstrates the ability to find small-scale and low-contrast debris, which creates severe difficulties when using models that were trained on larger or brighter objects. The standard YOLOv7 model (**Figure 8a**) does not identify several targets and gives the others with low confidence scores; one “soft plastic” object is found (Figure 8b) identifies all of the visible objects of the scene, which is masks, “soft plastics”, and “bottles” but with much higher confidence approaching 0.90 and even higher in the vast majority of cases. The performance of this model is an indicator that it has performed better in the discrimination of features, especially the objects that consume fewer pixels or become partially covered by water. The lightweight CA module helps conserve fine spatial features in the extraction of features, and Sinkhorn-based regularisation maintains focus and consistency of attention.



(a) (b)

Figure 8. Detection performance on small and low-contrast floating debris: standard YOLOv7 model (a); proposed method (b)

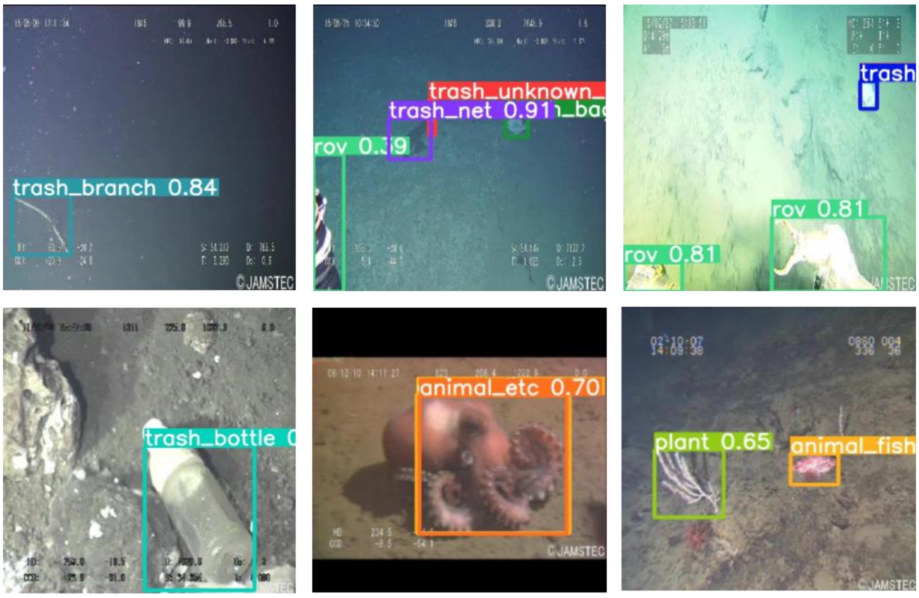


Figure 9. Detection of cluttered underwater scenes from the TrashCAN dataset

Figure 9 shows the output of detections on underwater images with cluttered backgrounds and overlapped objects like trash, aquatic life, and the structure of the seabed. The proposed approach effectively localises and categorises such objects as “trash nets”, “bottles”, and marine animals and generates narrow confined boxing with a confidence score that is always over 0.80. Such outcomes emphasise the fact that the model can be resilient to low-visibility and low-contrast conditions where the boundaries of objects are usually diffuse and the labels of the classes are semantically nearly similar. The spatially enriched attention to the entropy-regularised alignment enables the model to process the visual entropy in deep-sea imagery, preventing the model from being confused with what looks similar to a particular object.

DISCUSSION

The proposed method shows an enhancement in the detection of marine plastic litter both in aerial and underwater settings as compared to baseline models. According to the per-class analysis, there are high recognition results for the categories with a clear visual appearance.

These performance improvements are achieved through two complementary architectural elements. CA breaks down context extraction in horizontal and vertical directions, and long-range structure is preserved, with positional accuracy, which is more effective in identifying elongated objects. Sinkhorn regularisation follows a global coherence principle on the attention distributions by applying entropy-regularised optimal transport, which encourages smooth attention assignment and discourages disjointed focus patterns, which are prevalent in underwater scenes with cluttered backgrounds. Ablation experiments demonstrate that the use of both modules is more effective in detection accuracy than either of the two, where visual analysis has proven that the model reduces false positives due to surface reflection, small debris occupying a few pixels in the image and the ability to see visual similarities to categories in difficult underwater conditions.

To identify the statistical significance of reported performance gains between the proposed approach and baseline YOLOv7, paired Wilcoxon signed-rank tests were performed in both datasets. **Table 3** provides the average performance values, standard deviations, 95% confidence intervals, and p-values of all evaluation measures. The analysis shows that all measures indicate statistically significant improvements ($p < 0.05$) in both datasets, and most results are highly significant ($p < 0.01$). The small confidence interval and low standard deviations support the execution of patterns across cross-validation folds, which justifies the stability of the Sinkhorn-regularised attention mechanism in working marine monitoring systems.

Table 3. Statistical analysis using Wilcoxon signed-rank tests across 5-fold cross-validation; Mean ± Std (standard deviation), 95% CI (Confidence Interval), and p-values are reported

Dataset	Metric [%]	YOLOv7 Mean ± Std	Proposed Mean ± Std	95% CI	p-value
Submarine (MarineLitter)	*Precision*	88.55 ± 1.23	92.01 ± 0.87	[2.89, 4.03]	0.0043
	Recall	89.90 ± 1.45	94.17 ± 0.92	[3.51, 5.03]	0.0021
	mAP@0.5	91.41 ± 1.12	94.00 ± 0.76	[1.98, 3.20]	0.0035
	mAP@0.5:0.95	71.50 ± 1.67	72.20 ± 1.21	[0.12, 1.28]	0.0412
Underwater (TrashCAN)	*Precision*	86.80 ± 1.56	90.15 ± 1.08	[2.54, 4.16]	0.0038
	Recall	87.15 ± 1.72	92.30 ± 0.95	[4.23, 6.07]	0.0015
	mAP@0.5	89.35 ± 1.34	92.80 ± 0.89	[2.67, 4.23]	0.0028
	mAP@0.5:0.95	70.50 ± 1.89	72.00 ± 1.34	[0.71, 2.29]	0.0187

Although there are a number of considerations that will guide future research directions, the combined size of these data, about 14,000 images, is enough to justify the proposed method, but it is small with regard to large-scale benchmarks in computer vision. Future research will consider semi-supervised training or synthetic data generation to increase training diversity, especially for low-represented debris groups. The existing structure uses only the RGB images, which reduces the possibilities of detection in the case of high turbidity or in the dark. The combination of multimodal sensing capabilities, like sonar to map volumetric debris, hyperspectral imaging to analyse material composition or thermal sensors to distinguish litter on the surface, is a potential solution in strengthening the detection capability in a variety of ocean environments.

CONCLUSION

This study presented an enhanced YOLOv7-based framework for automated marine plastic litter detection that integrates direction-aware CA with Sinkhorn Distance regularisation. By formulating attention alignment as an entropy-regularised optimal transport problem, the proposed method addresses critical limitations in existing detection systems, particularly the fragmented attention distributions and lack of global spatial coherence that impair performance in challenging marine environments.

Validation experiments show a significant increase in value compared to baseline methods for both aerial and underwater approaches. The proposed approach performed better in terms of surface and submerged imagery. *Recall* improvement was significantly larger, and sensitivity to small, occluded, or low-contrast debris improved as well. Cross-validation statistically validated that the improvements found are significant and consistent across all evaluation metrics, proving that the proposed architectural improvements are robust and reliable.

In terms of architecture, this work has two contributions. First, CA enables the model to extract long-range spatial dependencies through independent horizontal and vertical context aggregation, which is critical for identifying irregularly shaped marine debris due to its positional accuracy. Second, Sinkhorn regularisation applies global structure consistency, reducing transportation costs between directional attention distributions with entropic hedges. This approach essentially handles frigid activations due to reflections on the surface, turbidity of the water, and complicated seabed structures. Ablation experiments and visual analysis show that the interactive combination of both elements provides better performance than either module alone. The hybrid structure of the two modules generalises well in varying visual conditions.

Despite these developments, there are still areas in which future research can be meaningful. Using semi-supervised learning or physics-based synthetic data to increase the variety of data might enhance the robustness of the models to underrepresented debris types and severe environmental impacts. Using a combination of multimodal sensing, such as sonar for low-visibility detection, hyperspectral imaging for material identification, and thermal sensors for surface detection, would be superior to using RGB imagery for object detection.

ACKNOWLEDGMENT

The authors are grateful to the assistance of the REMEDIES project under the HORIZON Innovation Actions (HORIZON-MISS-2021-OCEAN-03) for their significant contribution to the development of solutions for marine plastic pollution. The aid of the PhD Association Scholarship-PASS offered by the National Centre of Scientific and Technical Research (CNRST), Morocco, with grant 77 UMP2023, is also recognised by A. Khriss.

NOMENCLATURE

Symbols

H	height of the feature map
L_{YOLO}	YOLOv7 loss
L_{total}	total loss (detection + regularization)
W	width of the feature map
X	input feature map
Y	modulated feature map

Greek letters

ϵ	entropy regularisation factor
λ	regularisation weight for Sinkhorn loss
ρ	density

Subscripts and superscripts

h	horizontal
w	vertical

Abbreviations

CA	Coordinate Attention
GAP	Global Average Pooling
ROV	Remotely Operated Vehicle
SGD	Stochastic Gradient Descent
SSD	Single Shot MultiBox Detector
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once

REFERENCES

1. X. Zhu, C. M. Rochman, B. D. Hardesty, and C. Wilcox, "Plastics in the deep sea – A global estimate of the ocean floor reservoir," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 206, p. 104266, Apr. 2024, https://doi.org/10.1016/j.dsr.2024.104266.
2. I. O. Musa *et al.*, "Micro- and Nanoplastics in Environment: Degradation, Detection, and Ecological Impact," *Int J Environ Res*, vol. 18, no. 1, p. 1, Feb. 2024, https://doi.org/10.1007/s41742-023-00551-9.
3. T. L. C. Tran, Z.-C. Huang, K.-H. Tseng, and P.-H. Chou, "Detection of Bottle Marine Debris Using Unmanned Aerial Vehicles and Machine Learning Techniques," *Drones*, vol. 6, no. 12, p. 401, Dec. 2022, https://doi.org/10.3390/drones6120401.
4. A. Khriss, A. Kerkour Elmiad, M. Badaoui, A.-E. Barkaoui, and Y. Zarhloule, "Exploring Deep Learning for Underwater Plastic Debris Detection and Monitoring," *J. Ecol. Eng.*, vol. 25, no. 7, pp. 58–69, July 2024, https://doi.org/10.12911/22998993/187970.
5. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, June 2016, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.
6. B. Xue *et al.*, "An Efficient Deep-Sea Debris Detection Method Using Deep Neural Networks," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 14, pp. 12348–12360, 2021, https://doi.org/10.1109/JSTARS.2021.3130238.
7. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, June 2023, pp. 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721.
8. Y. Mei *et al.*, "Pyramid Attention Network for Image Restoration," *Int J Comput Vis*, vol. 131, no. 12, pp. 3207–3225, Dec. 2023, https://doi.org/10.1007/s11263-023-01843-5.
9. Q. Wu, L. Cen, S. Kan, Y. Zhai, X. Chen, and H. Zhang, "Real-time underwater target detection based on improved YOLOv7," *J Real-Time Image Proc*, vol. 22, no. 1, p. 43, Feb. 2025, https://doi.org/10.1007/s11554-025-01621-1.

10. X. Chen, C. Fan, J. Shi, H. Wang, and H. Yao, “Underwater target detection and embedded deployment based on lightweight YOLO_GN,” *J Supercomput*, vol. 80, no. 10, pp. 14057–14084, July 2024, https://doi.org/10.1007/s11227-024-06020-0.
11. W. Qiang, Y. He, Y. Guo, B. Li, and L. He, “Exploring Underwater Target Detection Algorithm Based on Improved SSD,” *JNWPU*, vol. 38, no. 4, pp. 747–754, Aug. 2020, https://doi.org/10.1051/jnwpu/20203840747.
12. S. Huang, Y. He, and X. Chen, “M-YOLO: A Nighttime Vehicle Detection Method Combining Mobilenet v2 and YOLO v3,” *J. Phys.: Conf. Ser.*, vol. 1883, no. 1, p. 012094, Apr. 2021, https://doi.org/10.1088/1742-6596/1883/1/012094.
13. Q. Wang, Y. Zhang, and B. He, “Intelligent Marine Survey: Lightweight Multi-Scale Attention Adaptive Segmentation Framework for Underwater Target Detection of AUV,” *IEEE Trans. Automat. Sci. Eng.*, vol. 22, pp. 1913–1927, 2025, https://doi.org/10.1109/TASE.2024.3371963.
14. G. Wen *et al.*, “YOLOv5s-CA: A Modified YOLOv5s Network with Coordinate Attention for Underwater Target Detection,” *Sensors*, vol. 23, no. 7, p. 3367, Mar. 2023, https://doi.org/10.3390/s23073367.
15. L. Shen, B. Lang, and Z. Song, “CA-YOLO: Model Optimization for Remote Sensing Image Object Detection,” *IEEE Access*, vol. 11, pp. 64769–64781, 2023, https://doi.org/10.1109/ACCESS.2023.3290480.
16. X. Chen, M. Yuan, Q. Yang, H. Yao, and H. Wang, “Underwater-YCC: Underwater Target Detection Optimization Algorithm Based on YOLOv7,” *JMSE*, vol. 11, no. 5, p. 995, May 2023, https://doi.org/10.3390/jmse11050995.
17. Q. Liu *et al.*, “DSW-YOLOv8n: A New Underwater Target Detection Algorithm Based on Improved YOLOv8n,” *Electronics*, vol. 12, no. 18, p. 3892, Sept. 2023, https://doi.org/10.3390/electronics12183892.
18. G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, “UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios,” *Sensors*, vol. 23, no. 16, p. 7190, Aug. 2023, https://doi.org/10.3390/s23167190.
19. G. Qiao, M. Yang, and H. Wang, “A Detection Approach for Floating Debris Using Ground Images Based on Deep Learning,” *Remote Sensing*, vol. 14, no. 17, p. 4161, Aug. 2022, https://doi.org/10.3390/rs14174161.
20. Y. Li, X. Bai, and C. Xia, “An Improved YOLOV5 Based on Triplet Attention and Prediction Head Optimization for Marine Organism Detection on Underwater Mobile Platforms,” *JMSE*, vol. 10, no. 9, p. 1230, Sept. 2022, https://doi.org/10.3390/jmse10091230.
21. C. Hou, Z. Guan, Z. Guo, S. Zhou, and M. Lin, “An Improved YOLOv5s-Based Scheme for Target Detection in a Complex Underwater Environment,” *JMSE*, vol. 11, no. 5, p. 1041, May 2023, https://doi.org/10.3390/jmse11051041.
22. P. Zhang and Y. Liu, “A small target detection algorithm based on improved YOLOv5 in aerial image,” *PeerJ Computer Science*, vol. 10, p. e2007, Apr. 2024, https://doi.org/10.7717/peerj-cs.2007.
23. J. Ma *et al.*, “CBS-YOLO for ship target detection in complex marine environments,” in *International Conference on Computer Application and Information Security (ICCAIS 2024)*, S. Ali Safaa, P. Hari Mohan, and B. Farid, Eds., Wuhan, China: SPIE, Apr. 2025, p. 68. https://doi.org/10.1117/12.3061586.
24. Q. Hou, D. Zhou, and J. Feng, “Coordinate Attention for Efficient Mobile Network Design,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, June 2021, pp. 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350.

25. X. Jin, Y. Xie, X.-S. Wei, B.-R. Zhao, Z.-M. Chen, and X. Tan, "Delving deep into spatial pooling for squeeze-and-excitation networks," *Pattern Recognition*, vol. 121, p. 108159, Jan. 2022, https://doi.org/10.1016/j.patcog.2021.108159.
26. L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré, "Faster Wasserstein Distance Estimation with the Sinkhorn Divergence," Oct. 29, 2020, *arXiv*: arXiv:2006.08172. https://doi.org/10.48550/arXiv.2006.08172.
27. A. Khriss, A. K. Elmiad, and M. Badaoui, "OTM-UNet: Optimized Semantic Segmentation of Remote Sensing Imagery with Learned Optimal Transport Maps," *International Journal of Artificial Intelligence*^{TM}, vol. 23, no. 1, pp. 71–91, 2025.
28. J. Hong, M. Fulton, and J. Sattar, "TrashCan: A Semantically-Segmented Dataset towards Visual Detection of Marine Debris," July 16, 2020, *arXiv*: arXiv:2007.08097. https://doi.org/10.48550/arXiv.2007.08097.
29. "Greentech Solution | automazione industriale e sviluppo tecnologico | Via Elvira Notari, 38, 80147 Napoli, Napoli Metropolitan City of Naples, Italy," greentechsolution, https://www.greentechsolution.it/en, [Accessed: Oct. 15, 2025].
30. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 06, 2016, *arXiv*: arXiv:1506.01497. https://doi.org/10.48550/arXiv.1506.01497.
31. W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," vol. 9905, 2016, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

Paper submitted: 15.08.2025
Paper revised: 17.10.2025
Paper accepted: 19.10.2025