Article

Not peer-reviewed version

Subaquatic Garbage Recognition Based on Improved YOLOv11 Network

Yinghao He , Wenjie Yin , Siyi Zhou * , Shimin Shan * , Xuezi Jiang

Posted Date: 11 August 2025

doi: 10.20944/preprints202508.0676.v1

Keywords: improved YOLOv11 network; subaquatic garbage; deep learning; fasterNet

Article

Subaquatic Garbage Recognition Based on Improved YOLOv11 Network

Yinghao He [1], Wenjie Yin [1], Siyi Zhou [1,*], Shimin Shan [2,*] and Xuezi Jiang [1]

[1] Electronics and Automation College, City Institute Dalian University of Technology, Dalian 116600, China

[2] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

* Correspondence: zhousiyi98@163.com (S.Z.); ssm@dlut.edu.cn (S.S.)

Abstract

In the field of marine ecosystem conservation and underwater garbage management, accurately identifying diverse debris is critical to improve cleanup efficiency and advance intelligent underwater operations. However, reflection, occlusion, color attenuation, object deformation in complex underwater environment render the traditional detection models impractical for dynamic scenarios. To address these issues, this paper proposes an underwater debris recognition model based on an enhanced YOLOv11 network. The improvements are as follows:a) FasterNet is adopted as the backbone network, which balances lightweight and high-fidelity feature extraction, effectively optimizing the retention of small-object features.b) SOAH, an occlusion-aware attention mechanism is introduced to reconstruct the detection head. This enables the module to strengthen the response to occluded targets during fusion process, thereby amplifying recognition performance in complex backgrounds.c) DAttention is utilized to replace the conventional spatial attention mechanism, which makes the DUAM module adaptable to partial texture deformation and increase the recognition accuracy of heterogeneous targets. The datasets comprises 9 major categories and 40 subcategories underwater debris images, collected from natural marine environment of varying depths and lighting conditions. It virtually contains typical challenges such as aquatic particle interference and uneven illumination. Experimental results show that the advanced network boosts from 66.9% to 77.3%, a 10.4% increase was observed. This proposed methodology demonstrates high precision and stability in complex conditions, providing robust support for underwater marine debris detection.

Keywords: improved YOLOv11 network; subaquatic garbage; deep learning ; fasterNet

1. Introduction

Amid the relentless tides of globalization and industrialization, marine debris has envolved into an intractable environmental challenge. Marine litter stems from diverse pathways, encompassing waste generated from land-based activities that enters the ocean via rivers and atmospheric deposition, as well as discarded items from maritime activities such as shipping and fishing [1]. With its progressive accumulation, a significant portion of debris eventually sinks to the seafloor, causing irreversible damage to marine ecosystem. Seafloor litter results in the death of marine organisms through ingestion, concurrently, disrupts the material cycling and energy transfer within the ocean [2]. Additionally, increasing debris constitutes a grave threat to marine economic activities, undermining the sustainability of fishing industry, increasing shipping costs, and adversely impacting the advancement of coastal tourism. These adverse effects ripple across the entire marine ecological chain and extend to the coastal socio-economic stability and sustainable development [3].

Considering the negative consequences on environment and economy caused by pollution, especially seafloor debris, a range of advanced technological approaches have been explored and applied in marine debris detection. Notably, deep learning has increasingly become the mainstream solution, demonstrating significant advantages in cluttered backgrounds, low-light conditions, and small-object identification. Ren et al. proposed Faster R-CNN to optimize the detection efficiency

 doi:10.20944/preprints202508.0676.v1

through Region Proposal Network (RPN), laying the technical foundation for underwater detection, but the real-time performance is inferior [4]. Based on this, Cai and Vasconcelos constructed Cascade R-CNN to enhance identification accuracy through a cascaded structure that shows greater robustness for low-quality marine debris images, however, this complex network demand higher experimental costs [5]. Shi and Wu studied SRP-UOD method to incorporate structural re-parameterization in a multi-branch architecture, which can improve the detection precision of small marine debris and demonstrate superior performance in feature diversity [6]. Liu and Li proposed PILLO algorithm to simulate plant behaviors, thus constructing a feature perception network. This structure refines recognition ability for low-contrast images, though its generalization needs to be validated on large-scale datasetss [7]. In conclusion, these CNN-based methodologies demonstrate both advantages and limitations in marine debris detection, providing invaluable insights for future technological evolution..

With the development of deep learning, Transformer architecture has also made notable strides in underwater detection. Cao et al. used Trf-net which integrated Transformer structure for multimodal feature modeling. supporting the fusion of multisource data. Nonetheless, it exhibits remarkably complexity and requires extensive training time [8]. Zhu et al. constructed Dformable DETR to address the traditional difficulty in tiny object detection. By using this, they achieved strong generalization in marine debris recognition, but struggled with training stability [9]. Carion et al. used DETR, a representative of end-to-end detection system, which introduced global modeling to improve accuracy. In contrast, its suboptimal inference speed adversely affects real-time underwater operations [10]. Zhao et al. conducted a systematic evaluation of DETR-like models in real-time scenarios, showing promising detection accuracy in underwater simulation environment, though the models heavily rely on hardware acceleration [11]. Zhu et al. proposed Vision Mamba, a network based on state space model, to reliaze the potential of identifying underwater video targets in dynamic backgrounds, nevertheless, practical implementations are scarce [12]. These Transformer-based approaches provide new perspectives and solutions for subaquatic garbage detection, and while challenges remain, their potential is substantial.

Subsequently, the YOLO family models gained widespread adoption. Bochkovskiy et al. proposed YOLOv4 that incorporated multi-scale path enhancement and residual structure to obtain a favorable balance between speed and precision. Conversely,.there remains room for improvement in recognizing extremely small targets [13]. Wang et al. used YOLOv7 to introduce E-ELAN and RepConv structure,, implementing cross-scale feature reorganization and enhancing adaptability to various marine debris [14]. Zheng et al. integrated sonar data to improve the modeling capability of YOLOv8, demonstrating strong scenario generalization [15]. Wang et al. utilized YOLOv9 to introduce programmable gradients and an improved path aggregation strategy, boosting detection accuracy of varying size objects while maintaining high speed [16]. Wang et al. presented YOLOv10 to further simplify the network and optimize multi-scale target matching, delivering stable performance on edge computing devices [17]. Zhou et al. proposed an advanced YOLOv5 model, enhancing detection efficiency by which introducing a multi-scale feature fusion module. It can significantly balance real-time and accuracy, though omission of small-objec still persists [18]. Zhu et al. combined TPH-YOLOv5 and Transformer prediction head to elevate the detection performance of distant and occluded targets, exhibiting solid transferability in underwater detection tasks, albeit with much more computational demands [19]. Wang et al. proposed YOLOv8-QSD, an evolved structure for small-object detection, demonstrating robustness in simulative underwater datasetss. However, its adaptability in genuine underwater conditions requires further validation [20]. Zheng et al. incorporate YOLOv8 and ScEMA module for sonar imagery, effectively improving detection efficacy of low-contrast targets though the direct-transfer proficiency of visual imagery requires improvement [21]. These YOLO-based methodologies have shown considerable potential in augmenting underwater detection performance, expanding the technological horizons of marine debris identification and highlighting unresolved research challenges simultaneously.

 doi:10.20944/preprints202508.0676.v1

To alleviate the above challenges, such as low contrast, color distortion, uneven lighting, and background interference in underwater recognition, this paper proposes an improved approach based on YOLOv11. This model improves the capability of feature extraction and multi-scale detection. By structurally optimizing backbone and detection head, it enhances both percision and robustness of key target identification and maintains lightweight at the same time.. The introduced improvements not only tailore to the distribution characteristics of visual information in complex environments but also take into account the needs of. high-recall expression of small-object.Chapter 2 details the datasets construction process and the implementation of the proposed network. Chapter 3 displays the experimental setup and evaluation metrics, compares the performance of different architectures in various underwater scenarios. Chapter 4 concludes the paper and outlines potential directions for future research.

2. Materials and Methods

2.1. Materials

2.1.1. Data Collection

The image data used in this study is sourced from the publicly shallow-water marine debris detection and segmentation datasets published in Scientific Data, a journal under Nature [22]. This datasets contains 8,610 annotated underwater images, with a total of 31,555 labeled object instances across 40 object subcategories. Moreover, it encompasses nine major categories: Animal, Plant, Robot, Metal, Plastic, Other debris, Rubber, Glass, and Fiber. All images were captured in natural ocean environment with deverse depth and brightness, which integrated typical underwater disturbances such as surface reflection, particulate masking, and color attenuation.During the preprocessing stage, the sizes of all images were uniformly cropped to 640×640 pixels. Based on the natural distribution of category labels, stratified random sampling was used to divide the datasets into training and validation sets in a ratio of 8:2. This improves robust feature extraction for target features without altering the inherent class distribution. Meanwhile, it reserves the generalization of validation set across different object sizes and environmental conditions. Figure 1 presents the categorical distribution of total annotations.

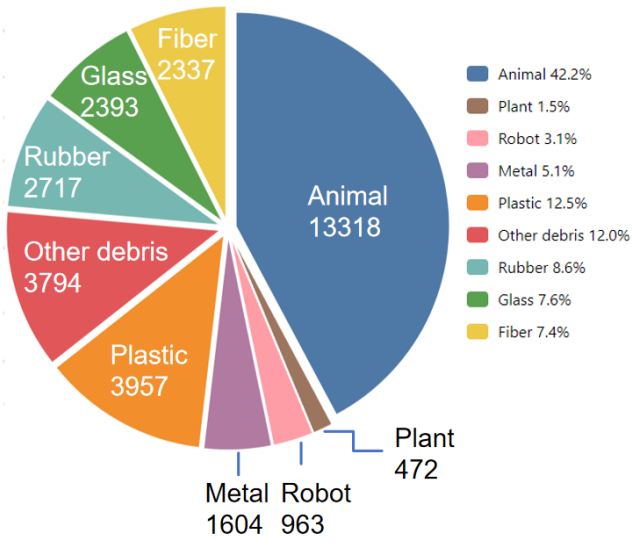


Figure 1. Proportion of each category in total annotations.

The datasets includes 40 subcategories collectively. Among these, the Animal pronouncedly outweighs others, including various subtypes such as fish, sea urchins, and shells. These targets often have vague boundaries and low salience, consequently, the camouflage-prone increases

 doi:10.20944/preprints202508.0676.v1

identification errors. The Plastic encompasses heterogeneous subclasses, for instance bottles, bags, nets, tubes. Some of them are fragile, deformable or impurity-covered, which causes unstable texture features and fails to extract unified representations. Metal objects generally have clearer contour but the various size and tiny objects generally decrease the precision of detection. Additionally, rust or damage may compromise the integrity of features. Rubber and Glass are highly susceptible to environmental conditions such as lighting and water clarity, generating reflection issue and confining the distinction of boundary and background. Fiber, Ceramic, and Wood have rare samples and irregular shape distribution, resulting in misclassification and instability during the training phase. Overall, the datasets presents significant imbalance and long-tail structure, with a high proportion of small objects. The complex underwater environment and image noise contribute to false and miss detection. Furthermore, other categories also compound robustness design challenges for model architecture.

2.1.2. Data Distribution

Small objects are widespread in real underwater conditions. Tiny fragments of various materials, cable ends, and parts of animal bodies are all within the sub-pixel scale range. Constrained by pixel resolution and receptive fields, missed detection and localization errors occur frequently. Moreover, due to the non-uniform illumination and suspended materials, partial images present obvious noise interference, escalating the feature confusion in edge, texture and color extraction. Throughout the amplifying, proper preservation and regulation of these interference characteristics directly impact training outcomes. Therefore, differentiated design strategies are essential to capture discriminative features instead of background noise.

These challenges stem from image acquisition conditions, object structure complexity and data distribution, which influences quality control,, class equilibrium and subsequent training strategies. In addition, separability disparity of different categories affects model selection and hyperparameter tuning as well, thereby promoting the demand of precise detection mechanisms in specific material and forms. Figure2 illustrates the size distribution of different categories of underwater garbage within the datasets.

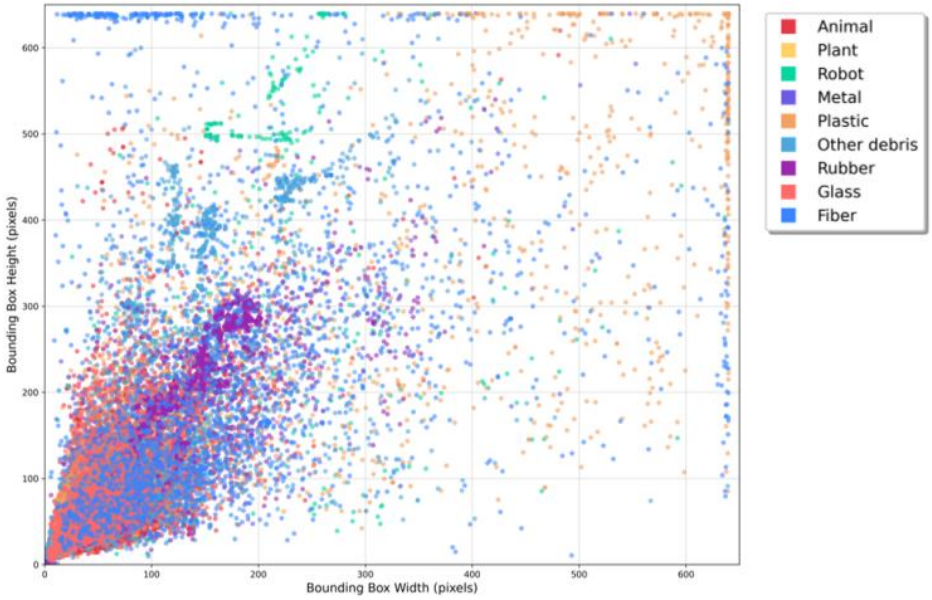


Figure 2. Datasets Size Distribution.

 doi:10.20944/preprints202508.0676.v1

2.2. Relevant Work

2.2.1. YOLOv11

The YOLO (You Only Look Once) network, evolved from YOLOv1 to YOLOv11 [23], has consistently displayed stable effectiveness in detecting medium-sized objects and possessed great potential for industrial applicability. However, when dealing with complex real-world scenarios involving diverse objects of varying sizes—particularly fragmented targets, partial occlusions, and edge blurring—the fixed receptive fields and multi-scale feature fusion mechanisms employed by YOLO often fail to provide sufficient discriminative power for small-object regions [24]. This readily leads to insufficient signal response, and feature representation corrupted by contextual noise., Ultimately, detection results commonly manifest as missed detection, misclassification, and positional offset, particularly in low-contrast or texture-deficient underwater conditions.

In the presence of dense occlusion, underwater turbidity, and complex overlapping backgrounds, edge information will be covered or entirely lost in general, negatively affecting both bounding box regression and classification. Moreover, it is vital to introduce pertinent mechanism for distinct material and morphological changes among various objects [25]. When dealing with the texture heterogeneity of underwater images, convolutional kernels with fixed size and structure constraint the dynamic adaptability to local deformation and rotation. This inflexibility hampers the detection of micro-feature, particularly, evident in ocean debris detection [26].In the original YOLOv11 architecture, a significant amount of small targets, for instance, cable ends, floating fragments, and animal remains generally occupy a small portion of the image. With blurred boundaries and color similarities, these targets are often ignored by compressed feature layers.To overcome these architectural shortcomings, this study introduces three improved key structures aimed at enhancing the detection of small-scale objects, occlusion, and complex material boundaries. Network architecture of the improved YOLOv11 as shown in Figure 3. Its effectiveness in underwater image recognition will be experimentally validated in subsequent sections.

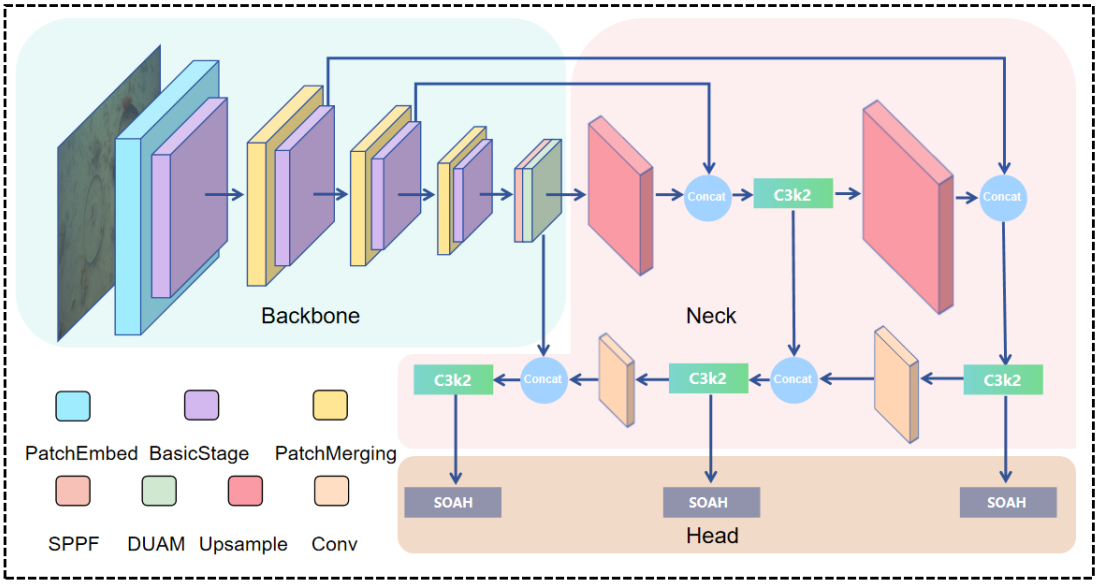


Figure 3. Network architecture of the improved YOLOv11.

The selection of evaluation metrics directly determine the model performance. In this study, four standard evaluation indicators are utilized: Precision, Recall, AP, and $AP_{@0.5:0.95}$ [27] (Goutte & Gaussier, 2005). The definitions and calculation of these metrics are shown in Formulas (1), (2), (3), and (4).

Precision refers to the proportion of correctly positive samples among all predicted positives. It reflects the model's accuracy and reliability in predictions. A higher Precision indicates a more correct identification and a lower false detection rate. The formula is as follows:

 doi:10.20944/preprints202508.0676.v1

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

where, TP is true positive, which refers to the samples correctly identified as marine debris; FP is false positive, denotes the false discrimination of marine debris.

Recall measures the proportion of samples that are correctly predicted positive among all predictions. It reflects the model's copiability to retrieve positive instances, that is to say, how many true positives the model can find [28]. Higher recall indicates fewer missed detections and lower false negative rate. The formula is as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

FN means false negative, represents the incorrectly identification as background.

AP (Average Precision) represents the average accuracy at various recall levels which equals to the area under the Precision-Recall curve (PR curve). AP serves as a comprehensive indicator of model performance for each category, reveal the accuracy changes under different recall rates and the detectability of the model. The formula is as follows:

$$AP = \int_0^1 \text{Precision} * \text{Recall}\, dr \quad (3)$$

'r' indicates the integral variable and determine the integral of precision * recall, ranging between 0 and 1; AP is the area under the Precision-Recall (P-R) curve, which is the average precision at different recall values when the intersection over union (IOU) is 0.5; The average accuracy under different recall values when IOU is 0.95.

is the average value of AP calculated at multiple Intersection over Union (IoU) thresholds (ranging from 0.5 to 0.95, with a step size of 0.05). This metric evaluates model performance thoroughly across varying detection difficulties. Compared to AP, $AP_{@0.5:0.95}$ can reflects the actual behavior of the model ideally. It requires the model to maintain high precision across different IoU thresholds—crucial for real-world object detection tasks. and especially reflects than AP. Above all, it requires the model to maintain high accuracy under different IoU thresholds, which is particularly important for detection tasks in practical applications. The formula is as follows:

$$AP_{@0.5:0.95} = \frac{1}{10}\left(AP_{@0.50} + AP_{@0.55} + \ldots + AP_{@0.90} + AP_{@0.95}\right) \quad (4)$$

By integrate the improved network and a rational evaluation framework, this study seeks to optimize the model's capability to detect multi-scale, heterogeneous marine debris in complex underwater environments. Of greater importance, the adaptability and robustness of the typical challenges such as low visibility, small size and partial occlusion are specifically considered.

2.3. Improved Algorithm

2.3.1. FasterNet

FasterNet architecture is utilized to replace the backbone network [29], which can condense redundant convolutional computation while elevating the efficiency of feature retention .It rearranges massive stacked modules in the typical ResNet-like backbones. By introducing lightweight convolutional kernels and fusion channels, this network minimize overall computational load while preserving spatial awareness, finally, maintaining high inference efficiency for high-resolution images [30]. This innovation stands as particularly beneficial in fulfilling deeper feature extraction without high GPU memory and enhancing fine-grained modeling for spatial texture in shallow layers [31]. It contributes to preserve the integrity of features in small-object area [32]. Furthermore, FasterNet effectively separates essential target features from background interference, thereby restraining error accumulation during decoding.

In contrast with traditional backbones, FasterNet exhibits superior parameter efficiency and practical adaptability. It demonstrates stronger response intensity and classification activation during

 doi:10.20944/preprints202508.0676.v1

small-object feature extraction, ultimately providing more discriminative input features for detection head.

2.3.2. DUAM

The original C2PSA structure uses Parallel Attention Channel Enhancement to filter target activation regions through Spatial Attention Mechanism in the phase of multi-scale semantic fusion. This procedure evaluates the focusing ability of objects with clear edges and good contrast [33]. However, attributable to vague textures or deformed structures, its combination of fixed-scale convolution and unidirectional spatial perception mechanism decrease identification performance [34]. As established, there are numerous disturbances in underwater settings, such as shape deformation, broken edges or disrupted textures. These issues constraint traditional convolutional structure to capture full range of targets, which leads to edge response attenuation or detection region deviation. To solve this problem, DAttention [35] is introduced into C2PSA, forming the Deformable Undersea Attention Module (DUAM), as shown in Figure 4. DUAM maintains the original spatial attention extraction pathway and integrates a deformable receptive field at the same time. The advent of the enhancement enables the model to dynamically generate sampling positions for each pixel at different scales, hence, refining modelling ability of local deformation regions. This demonstrates prominent efficacy in soft materials and targets with irregular contour. Additionally, the improved module retains the parallel spatial attention branch to ensure an adaptive balance between long-range contextual dependency and short-range geometric perception. Experimental results in section 3 prove that DUAM dramatically gains a uplift in the aggregation of small-object region response during multi-scale fusion, which prevents shallow-layer features from misleading deeper classification branches.

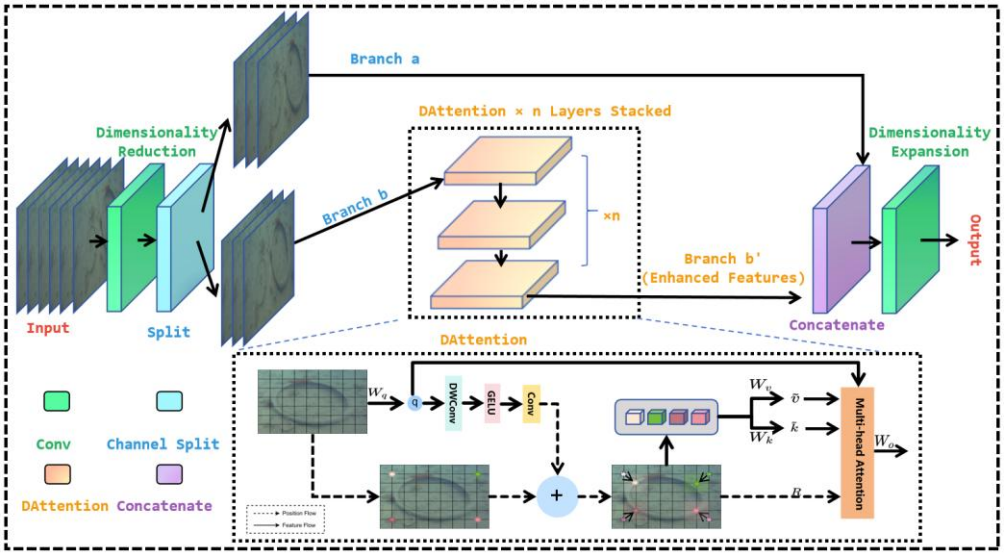


Figure 4. Structure of Deformable Undersea Attention Module (DUAM).

2.3.3. SOAH

Subaquatic Occlusion Aware Head(SOAH), integrated SEAM attention mechanism [36], presents occlusion-aware algorithm to solve precision reduction, as elaborated in Figure 5. By embedding localized perception channel into each output stage of the detection head [37], SOAH trains the model to recognize and distinguish occluded features and complete structures. It futher combines attention-guided mechanisms to amplify responses of occluded regions [38]. These adaptations enable the model to obtain enouogh class confidence and position accuracy while targets are partially covered. SOAH reveals exceptional boundary stability in multi-scale facial feature extraction and outstanding adaptability of small targets that are severely obscured. Besides, it

 doi:10.20944/preprints202508.0676.v1

integrates anchor orientation encoding and region mask perturbation control, introduce positional information and shape constraints as auxiliary loss feedback signal. This evolution allows model to infer the edge of uncertain targets with residual area. The above designs emphasize the response preservation of incomplete targets result from occlusion, improve the detection precision while striving to maintain inference speed, significantly reducing missed detection and localization drift in high-density occlusion conditions.

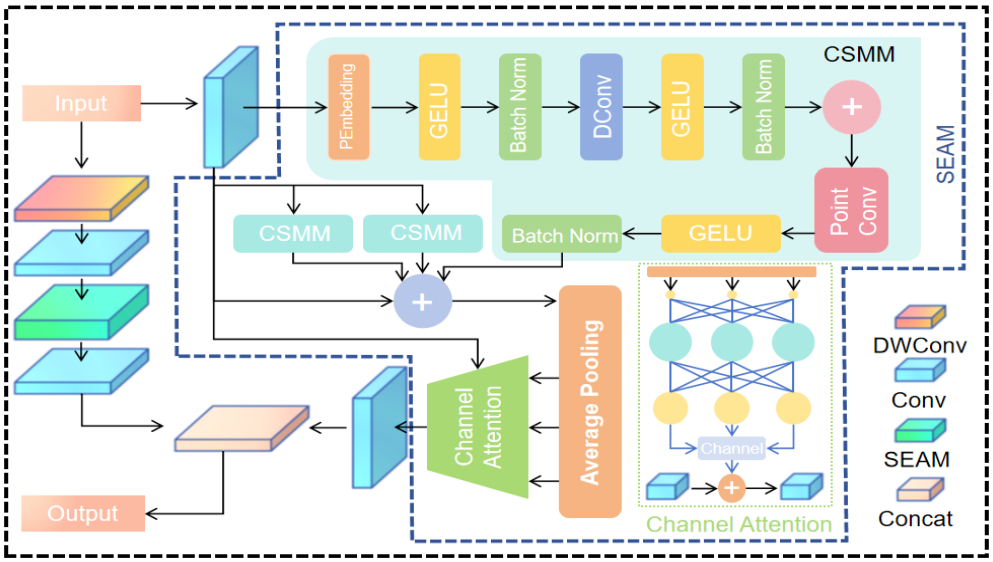


Figure 5. Structure of Subaquatic Occlusion-Aware Head (SOAH).

3. Results and Discussion

In this study, training procedure was conducted on a local desktop computer. The experimental environment is elaborated in Table 1. To achieve a preferable analysis, 200 iterations were executed. During the training process, relevant parameters were set to generate the pre-trained weight file, including a momentum of 0.949, an initial learning rate of 0.001, a weight decay coefficient of 0.0005, and a batch size of 16. To reduce performance fluctuations caused by varying sizes, all images were resized into a uniform resolution during training. Meanwhile, data augmentation strategy was applied to enhance the model's generalization ability in complex conditions. The loss function convergence and accuracy curve were recored in real-time and used to adaptively adjust the learning rate. Finally, model weights file was generated for the following experiment.

Table 1. Experimental Environment.

configuration	parameter
CPU	Intel Core i9-14900K
GPU	NVIDIA RTX 4090
Operating system	Windows 11
Accelerated environment	CUDA 11.8

NVIDIA Jetson Nano device was utilized to inference, as illustrated in Figure 6. To ensure efficiency under resource-constrained conditions, this paper appropriately pruned the network architecture and quantized the parameters. This reduced memory usage and calculation delay without compromising accuracy significantly. Through comparative analysis of pre- and post-intervention metrics such as inference frame rate, average detection accuracy and missed detection rate, the feasibility of applying the model on embedded platform was evaluated.

 doi:10.20944/preprints202508.0676.v1

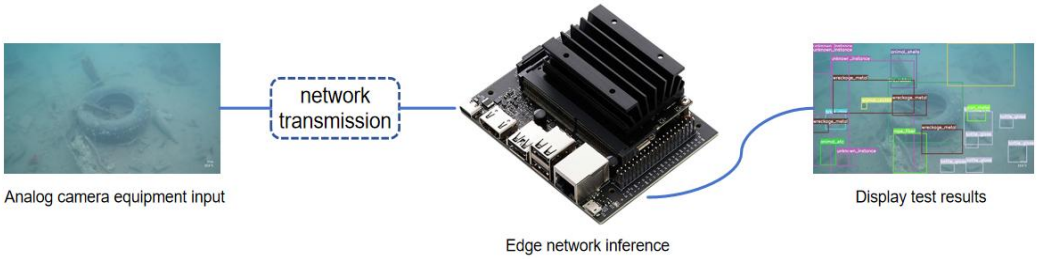


Figure 6. Edge deployment setup in a simulated underwater environment.

3.1. Algorithm Performance Evaluation

As shown in Figure 7, ablation experiment normalized the data to compare the model performance. Using our proposed method as control, each metric shows a consistent upward trend. In terms of , the precision of Yolov11 increased slightly after the insertion of SOAH or DUAM modules. However, neither of them surpassed the performance of FasterNet. This suggests that individually enhancing feature extraction or optimizing multi-scale adaptability provides limited benefit to boost overall detection. Conversely, the accuracy and recall of the mode obtain a remarkable leap under the combination of SOAH and DUAM or the module with FasterNet. What's more, the integration of FasterNet and DUAM, in particular, led to a substantial increase in recall, approaching the performance of the final model. Ultimately, the model, combining SOAH, DUAM, and FasterNet, presented the optimal efficiency among all metrics. The superiority of and Precision highlights the robustness and discrimination across diverse object scales .

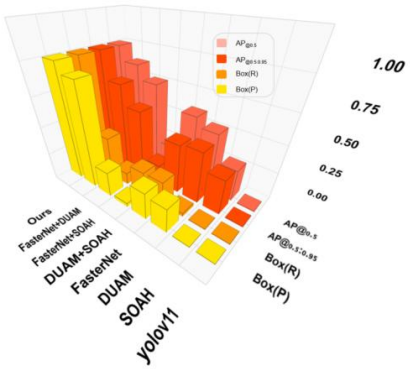


Figure 7. Comparison of Normalized algorithm performance.

Investigating on ablation of the models, the initial Yolov11 achieved only 0.669 , indicating limited detection boundary. The introduction of SOAH, resulted in a of 0.7, suggesting that the attention mechanism enhances the local semantic integrity of feature expression. At the same time, the baseline Yolov11 with DUAM led to a measurable in to 0.725, demonstrating a prime adaptability to large-scale-variation objects. Yolov11-FasterNet architecture brought to 0.733, revealing that the lightweight and efficient extraction of this backbone plays a vital role in detection. In terms of dual-module structures, SOAH+DUAM, FasterNet+SOAH and FasterNet+DUAM all marked a notable leap, but still fell short of Ours, the eventual model in this paper. Ours achieved the highest (0.773) and a remarkable increase in , confirming that multi-mechanism effectively enhances model generalization in occlusion, blury boundaries and complex backgrounds. The comparison of the experimental results after the enhancement is shown in Table 2.

 doi:10.20944/preprints202508.0676.v1

Table 2. Comparison of improved algorithm .

Method	yolov11	SOAH	DUAM	FasterNet	$AP_{@0.5}$(%)	$AP_{@0.5:0.95}$(%)
1	√				66.9%	47.2%
2	√	√			70.0%	49.9%
3	√		√		72.5%	51.1%
4	√			√	73.3%	50.9%
5	√	√	√		69.6%	68.6%
6	√	√		√	74.9%	52.1%
7	√		√	√	76.1%	53.5%
8(Ours)	√	√	√	√	**77.3%**	**55.3%**

Different categories have different detection performance through and . Yolov11 had no advantage in recognizing deformable soft targets like plastic tarp and clothing fiber, even failed to detect cardboard paper, exposing the defects in texture and structure modeling. FasterNet improved identification accuracy comprehensively, however, remained instability in complex backgrounds or low-visibility conditions. SOAH improved stability in detecting fine-grained targets such as plants and shells. With the addition of DUAM, the model significantly enhanced the identification of well-structured targets like cement tubes and glass bottles. It exhibits that the two modules are mutually complementary in detail extraction and interference resistance. Ultimately, the evolved model achieved a 5% enhancement in , strengthening the recognization capacity of long-tail categories including rope fiber and animal. That is to say, Ours maintains both constancy and consistency in low-contrast and complex scenarios thereby underscoring a more balanced performance. Comparison results of each category are shown in Figure 8.

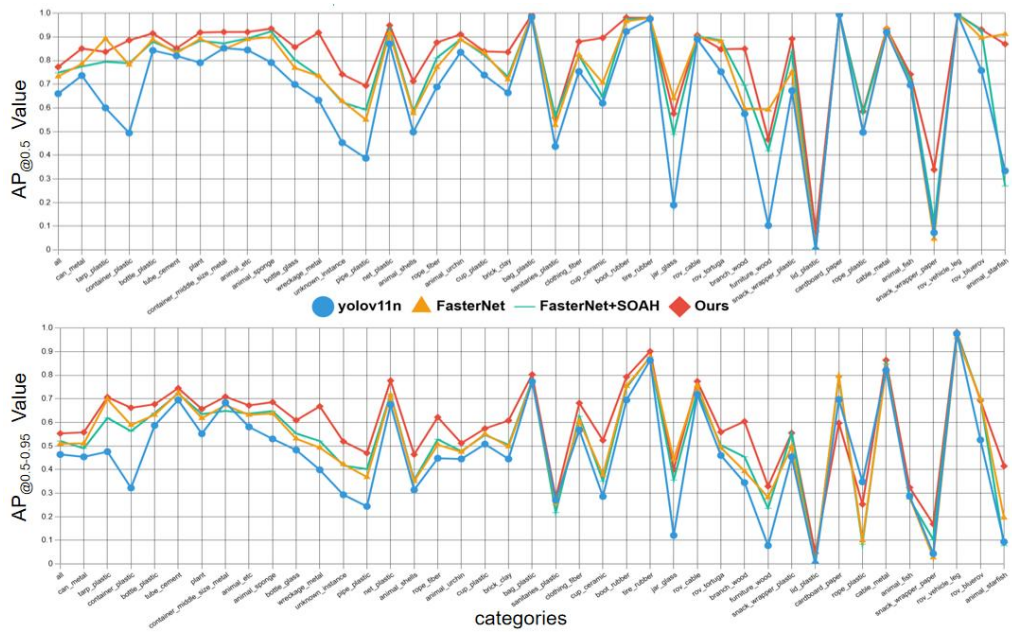


Figure 8. Comparison results of different categories and models.

3.2. Detection Results

As illustrated in Table 3, Ours achieved 77.3% in , exceeded the baseline Yolov11 (66.9%) and Yolov12 (66.0%) by 10.4% and 11.3%, and considerably surpassed SSD (40.1%) and Faster-RCNN (67.2%). In terms of , Ours reached 55.3%, outperforming yolov11 (47.2%) and yolov12 (46.4%) by 8.1% and 8.9%.

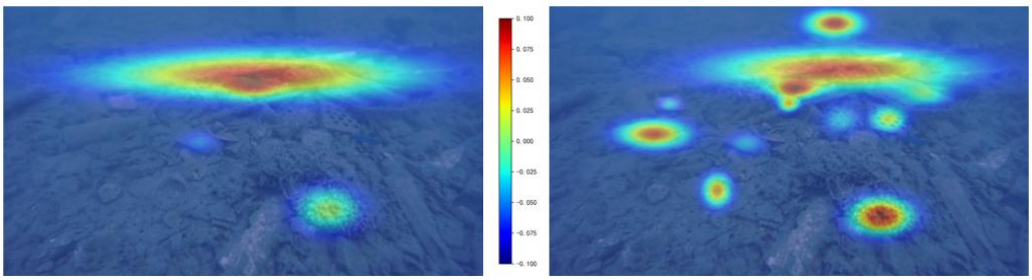
 doi:10.20944/preprints202508.0676.v1

Accrodding to the detection results of different FasterNet models, FasterNet_L attained the highest (79.1%) and (56.6%). Futhermore, when it comes to the balance of complexity and actual demand, FasterNet_T0 not only ensures the detection accuracy but also retains lightweight.

Table 3. Performance Comparison of mainstream detection algorithms.

Method	$AP_{@0.5}$ （%）	$AP_{@0.5:0.95}$ （%）	Box(P)	Box(R)
SSD	40.1%	-	86.4%	22.8%
Faster-RCNN	67.2%	-	53.7%	69.1%
yolov11	66.9%	47.2%	81.8%	62.8%
yolov12	66.0%	46.4%	73.9%	60.3%
Ours(FasterNet_T0+DUAM+SOAH)	**77.3%**	**55.3%**	**87.8%**	**75.3%**
FasterNet_T1+DUAM+SOAH	77.7%	54.9%	84.5%	69.1%
FasterNet_T2+DUAM+SOAH	77.5%	54.6%	88.1%	66.2%
FasterNet_S +DUAM+SOAH	77.7%	54.9%	85.8%	69.0%
FasterNet_L +DUAM+SOAH	79.1%	56.6%	87.9%	69.1%
FasterNet_M +DUAM+SOAH	79.0%	56.3%	87.6%	71.0%

As the heat maps delineated in Figure 9, the improved version of YOLOv11, incorporating FasterNet, SOAH, and DUAM modules, precisely detected various underwater debris in challenging conditions, for instance dense clutter, specular reflection, multi-occlusion, and edge blurring. Its attention regions further focused on key features about objects, substantially outperforming the original Yolov11 in feature extraction and localization.



(**a**) Heat map of Yolov11 (**b**) Heat map of Ours

Figure 9. Comparison of Heat maps.

Test environment contained extremely small objects, densely packed debris, specular glare interference, multi-layer occlusion, and blurry edges, which realistically simulated various obstacles in real marine environment. These conditions pose significant detection challenges and better align with practical requirements. Detection results of different model are displayed in Figure 10.

 doi:10.20944/preprints202508.0676.v1

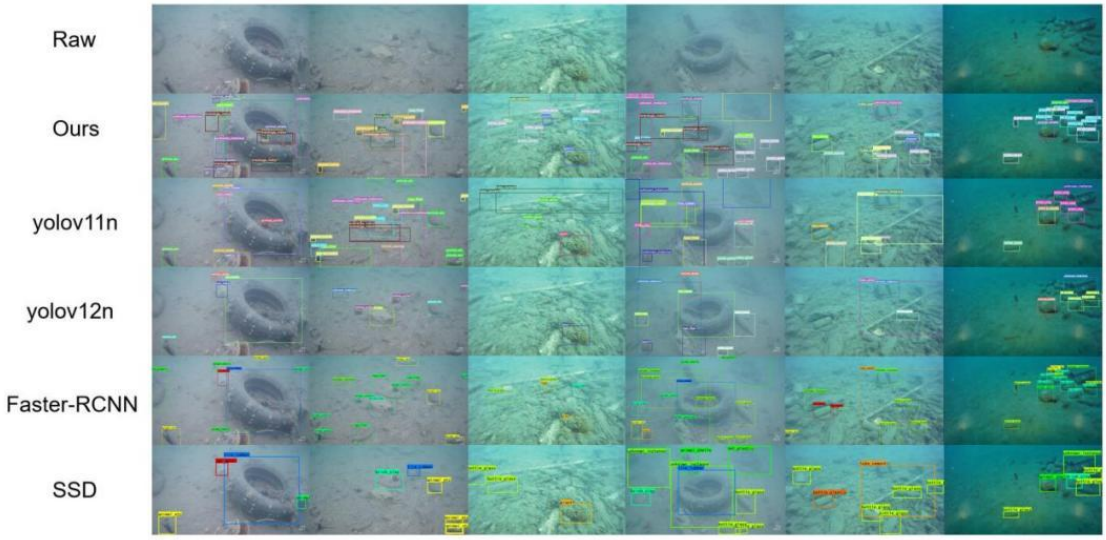


Figure 10. Comparison of detection results under complex scenarios.

4. Conclusions

This paper proposes an enhanced version of YOLOv11 network, specifically tailored for underwater debris recognition under complex ocean environment. In the advanced approach, FasterNet backbone is introduced to reduce the model size while optimizing the representation of spatial detail. When integrating SOAH and DUAM, a strategic improvement is evident in robustness and precision while handling partially occlusion, texture distortion and background interference. Experimental results reveal that the enhanced model achieves exceptional detection efficiency across a multi-source marine debris datasets comprising 9 major categories with 40 subcategories. Especially, it demonstrates superior robustness in detecting small objects, heterogeneous targets and heavy occluded conditions. In summary, this research provides a emphatic foundation for intelligent marine debris detection and a basis for further underwater monitoring and environmental assessment.

Author Contributions: Conceptualization, S.Z.and Y.H.; methodology, S.Z. and W.Y.; software, Y.H. and W.Y.; validation, S.Z. and W.Y.; formal analysis, S.S and Y.H.; writing—original draft preparation, S.Z.and W.Y.; writing—review and editing, S.S.; visualization,J.X and W.Y.; supervision, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Issifu, I. & Sumaila, U. R. A Review of the Production, Recycling and Management of Marine Plastic Pollution. J. Mar. Sci. Eng. 8, 945 (2020).
2. Pillai, R. R., Lakshmanan, S., Mayakrishnan, M., George, G. & Menon, N. N. Impact of Marine Debris on Coral Reef Ecosystem and Effectiveness of Removal of Debris on Ecosystem Health – Baseline Data From Palk Bay, Indian Ocean. Res. Square (2023).
3. McIlgorm, A., Campbell, H. F. & Rule, M. J. The cost of marine litter damage to the global marine economy: Insights from the Asia-Pacific into prevention and the cost of inaction. Mar. Pollut. Bull. 174, 113167 (2022).
4. Ren, S. et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149 (2017).

 doi:10.20944/preprints202508.0676.v1

5. Cai, Z. & Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the CVPR, pp 6154–6162 (2018).
6. Shi, J. & Wu, W. SRP-UOD: multi-branch hybrid network framework based on structural re-parameterization for underwater small object detection. In ICASSP 2024, pp 2715–2719.
7. Liu, L. & Li, P. Plant intelligence-based PILLO underwater target detection algorithm. Eng. Appl. Artif. Intell. 126, 106818 (2023).
8. Cao, H. et al. Trf-net: a transformer-based RGB-D fusion network for desktop object instance segmentation. Neural Comput. Appl. 35, 21309–21330 (2023).
9. Zhu, X. et al. Deformable DETR: Deformable transformers for end-to-end object detection. in Proceedings of the International Conference on Learning Representations, pp 1–16 (ICLR, 2021).
10. Carion, N. et al. End-to-end object detection with transformers. In European Conference on Computer Vision, pp 213–229 (ECCV, 2020).
11. Zhao, Y. et al. DETRs beat YOLOs on real-time object detection. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16965–16974 (CVPR, 2024).
12. Zhu, L. et al. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In ICML, pp 62429–62442 (2024).
13. Bochkovskiy, A. YOLOv4: Optimal Speed and Accuracy of Object Detection. Preprint at arXiv (2020).
14. Wang, C. Y. et al. YOLOv7: Trainable bag-of-freebies sets new SOTA for real-time object detectors. In CVPR, pp 7464–7475 (2023).
15. Zheng, L., Hu, T. & Zhu, J. Underwater sonar target detection based on improved ScEMA-YOLOv8. IEEE Geosci. Remote Sens. Lett. 21, 1–5 (2024).
16. Wang, C. Y. et al. YOLOv9: Learning what you want to learn using programmable gradient information. In ECCV, pp 1–21 (2024).
17. Wang, A. et al. YOLOv10: Real-time End-to-End Object Detection. Preprint at arXiv (2024).
18. Zhou, H. et al. Real-time underwater object detection technology for complex underwater environments based on deep learning. Ecol. Inform. 82, 102680 (2024).
19. Zhu, X. et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2778–2788 (ICCV, 2021).
20. Wang, H. et al. YOLOv8-QSD: an improved small object detection algorithm for autonomous vehicles based on YOLOv8. IEEE Trans. Instrum. Meas. 73, 1–16 (2024).
21. Zheng, L., Hu, T. & Zhu, J. Underwater sonar target detection based on improved ScEMA-YOLOv8. IEEE Geosci. Remote Sens. Lett. 21, 1–5 (2024).
22. Đuraš, A., Wolf, B.J., Ilioudi, A. et al. A Dataset for Detection and Segmentation of Underwater Marine Debris in Shallow Waters. Sci Data 11, 921 (2024).
23. Redmon, Joseph, et al. You only look once: Unified, real-time object detection.Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
24. Liu, P. et al. YWnet: A convolutional block attention-based fusion deep learning method for complex underwater small target detection. Ecol. Inform. 79, 102401 (2024).
25. Woo, S. et al. CBAM: convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision, pp 3–19 (ECCV, 2018).
26. Hu, J. et al. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141 (CVPR, 2018).
27. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall, and F-score, with implication for evaluation. In Advances in Information Retrieval; Lecture Notes in Computer Science; Losada, D.E., Fernández-Luna, J.M., Eds.; Springer:Berlin/Heidelberg, Germany, 2005; pp. 345–359.
28. Cao, D.; Chen, Z.; Gao, L. An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks. Hum.-Cent. Comput. Inf. Sci. 2020, 10, 14.
29. John Olafenwa et al. FastNet: An Efficient Convolutional Neural Network Architecture for Smart Devices. arXiv, 2018.

 doi:10.20944/preprints202508.0676.v1

30. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. PMLR, 2015.
31. Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.
32. Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In Icml, 2010.
33. Wang, Q. et al. ECA-Net: efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11531–11539 (CVPR, 2020).
34. Hu, J. et al. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141 (CVPR, 2018).
35. Xia Z , Pan X , Song S ,et al.Vision Transformer with Deformable Attention[J]. 2022.
36. Yu Z , Huang H , Chen W ,et al.YOLO-FaceV2: A scale and occlusion aware face detector[J].Pattern Recognition, 2024, 155(000):10.
37. Hou, Q. et al. Coordinate attention for efficient mobile network design. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13713–13722 (CVPR, 2021).
38. Ouyang, D. et al. Efficient Multi-scale Attention Module with Cross-Spatial Learning. In ICASSP, 2023.