Journal of
Marine Science and Engineering

MDPI

Article

Underwater Target Detection with High Accuracy and Speed Based on YOLOv10

Zhengliang Hu [1], Le Cheng [2], Shui Yu [3], Pan Xu [1,*], Peng Zhang [1], Rui Tian [1] and Jingqi Han [1]

1 College of Meteorology and Ocean, National University of Defense Technology, Changsha 410073, China
2 School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
3 School of Physical Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
* Correspondence: xupan09@nudt.edu.cn

Abstract: Underwater target detection exhibits extensive applications in marine target exploration and marine environmental monitoring. However, conventional images of underwater targets present challenges including blurred contour information, complex environmental conditions, and pronounced scattering effects. In this work, an underwater target detection method based on YOLOv10 is designed, and the detection performance is compared with the YOLOv5 model. Experimental results demonstrate that the YOLOv10 model has a mAP50 of 85.6% on the URPC 2020 dataset, improving the mAP50 by 1.2% than that of YOLOv5. This model exhibits high detection accuracy and high proceeding speed, which provides a promising support for precise and fast underwater target detection.

Keywords: deep learning; underwater target detection; YOLOv10

check for updates

Academic Editors: Woojae Seong, Youngmin Choo and Haesang Yang

Received: 15 December 2024
Revised: 13 January 2025
Accepted: 13 January 2025
Published: 14 January 2025

Citation: Hu, Z.; Cheng, L.; Yu, S.; Xu, P.; Zhang, P.; Tian, R.; Han, J. Underwater Target Detection with High Accuracy and Speed Based on YOLOv10. *J. Mar. Sci. Eng.* **2025**, *13*, 135. https://doi.org/10.3390/jmse13010135

1. Introduction

Underwater target detection exhibits substantial utility across a wide array of domains, including marine target exploration, oceanic environmental monitoring, and maritime security. These applications are pivotal for advancing fields such as marine biology, marine resource development, and environmental protection. As marine technology continues to evolve rapidly, the precise and swift detection of fish, biomimetic organisms, and other entities within the underwater environment has become increasingly critical for enhancing underwater defense frameworks. However, the intricate and dynamic nature of underwater environments poses significant challenges. Traditional detection methods, while effective in certain contexts, often fall short in achieving the required accuracy and processing speed necessary for modern applications [1–3]. Therefore, there is a pressing need for innovative approaches that can address these limitations and provide robust solutions for underwater target detection.

In recent years, deep learning (DL) methods have developed rapidly, which have exhibited great advantages in computer vision, natural language processing, speech recognition and target detection, etc. [4–8]. Meanwhile, DL techniques have been used for a fast and accurate identification of underwater targets, and they assume great importance in marine sciences. However, the traditional model for underwater target detection is usually influenced by the underwater complex environment, which leads to less feature information and low detection accuracy. In addition, the scale of underwater targets varies greatly and there are occlusions between targets, which make it easy to cause missed detection and wrong detection. To improve the detection efficiency of underwater targets, many studies have been carried out [9–11]. Majid et al. [12] presented a computationally efficient detection network for real-time vehicle detection in UAV imagery, leveraging channel shuffling and grouped convolutions for enhanced speed and incorporating inception and deformable modules to account for vehicle size and shape variations. Xiong et al. [13] revisited the role of regular convolutions in object detection models for mobile devices, demonstrating that their strategic placement via a neural architecture search can significantly improve the latency-accuracy trade-off, a family of models that achieve state-of-the-art results across various mobile accelerators. Yang et al. [14] proposed QueryDet, a novel query mechanism that accelerates the inference speed of feature pyramid-based object detectors by first predicting the coarse locations of small objects on low-resolution features and then refining the detections using high-resolution features guided by these coarse positions.

In the domain of object detection, the YOLOv5 series and its predecessors have achieved remarkable milestones, demonstrating substantial advancements in both accuracy and computational efficiency. However, despite their remarkable successes, these models exhibit certain limitations that warrant further exploration and refinement to elevate their performance and applicability. One of the primary challenges lies in the limited capacity of feature extraction. While the YOLOv5 architecture leverages convolutional neural networks (CNNs) to extract spatial features from input images, it may not fully exploit the rich contextual information present in higher-dimensional feature maps. This limitation can hinder the model's ability to discern subtle object characteristics, particularly in complex and cluttered scenes. Another critical issue pertains to the insufficiency of multi-scale feature fusion. Object detection often requires the integration of features from different spatial resolutions to accurately identify objects of varying sizes and scales. The current YOLOv5 framework, while incorporating strategies for multi-scale feature fusion, may not optimally balance the contribution of features across different layers. Moreover, the computational efficiency of YOLOv5, while superior to many other models, still presents a bottleneck, particularly when deployed on resource-constrained hardware environments. The trade-off between model complexity and computational requirements remains a significant challenge. Reducing the computational overhead while maintaining or improving detection accuracy is essential for expanding the applicability of these models to a broader range of devices and scenarios.

The You Only Look Once (YOLO) series, a pioneering framework in the realm of real-time object detection, has achieved a remarkable milestone with the unveiling of YOLOv10 in 2024. YOLOv10 incorporates cutting-edge advancements in deep learning architecture, optimization techniques, and computational efficiency, enabling it to outperform its predecessors and contemporary models in both accuracy and speed. Rigorous benchmarking and comparative studies have demonstrated that YOLOv10 not only achieves state-of-the-art performance in diverse detection tasks but also sets new standards for real-time processing in complex and dynamic environments [15]. The development of YOLOv10 underscores the critical role of innovation in addressing the growing demands of applications such as autonomous systems, surveillance, and environmental monitoring, where precision and rapid response are paramount. By leveraging novel neural network designs and enhanced training methodologies, YOLOv10 effectively mitigates challenges such as occlusion, scale variation, and background clutter, which have historically hindered the performance of object detection systems. Furthermore, its open-source nature fosters collaboration and accessibility, enabling researchers and practitioners worldwide to adapt and refine the model for specialized use cases. The success of YOLOv10 highlights the accelerating pace of technological progress in artificial intelligence and its transformative potential across industries. As the field continues to evolve, models like YOLOv10 exemplify the synergy between academic research and practical innovation, paving the way for future breakthroughs in real-time object detection and beyond. This achievement not only solidifies the YOLO series as a cornerstone of modern computer vision but also reaffirms the importance of interdisciplinary efforts in advancing the frontiers of machine learning and its applications. YOLOv10 demonstrates its prowess in real-time detection and multi-object recognition. However, the underwater environment often presents challenges in the form of noise, encompassing optical disturbances and blurring caused by the medium, which can compromise the accuracy of target detection.

In this work, the YOLOv10-based model is utilized for underwater target detection. The performance is compared with the YOLOv5 model, where the YOLOv5 had demonstrated commendable performance in the field of target detection, as evidenced by various studies [16,17]. The rest of this paper is organized as follows: Section 2 introduces the YOLO method and the YOLOv10 model. Section 3 analyzes the datasets, results, and discussion of the experiment. Conclusions are given in Section 4.

2. Theory and Model

2.1. YOLO Algorithm

YOLO introduced the groundbreaking approach to image segmentation by employing regression-based methods. This innovative technique revolutionized the field of target detection, offering remarkable accuracy and high computational efficiency, making it particularly suitable for real-time systems.

In YOLO, this grid-based approach allows for parallel and independent detection processes across the image, which significantly enhances computational efficiency [18]. This fusion of class probability maps and bounding box predictions enables YOLO to simultaneously generate accurate target detection bounding boxes.

One of the key advantages of YOLO is its ability to process entire images in a single evaluation, unlike other methods that process images in regions or with sliding windows [19]. This holistic approach not only speeds up the detection process but also allows YOLO to capture contextual information across the entire image, leading to more accurate and reliable detections.

Moreover, the YOLO architecture is designed to be lightweight and efficient, making it deployable on a wide range of hardware, from high-performance GPUs to resource-constrained devices. This versatility has contributed to its widespread adoption in various real-time applications, including autonomous vehicles, surveillance systems, and mobile devices. The detection algorithm flow of YOLO is exhibited in Figure 1.

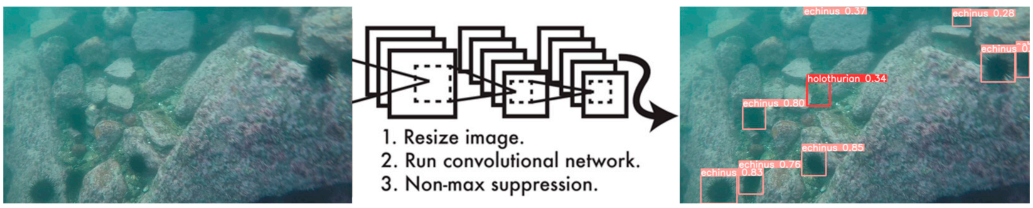


Figure 1. YOLO detection process.

The YOLO model performs predictions for each individual target class and generates a confidence score that quantifies the probability of the target belonging to a specific class [20]. This predictive capability facilitates rapid and precise target identification in images.

$$Confidence = P_r(Object) \times IOU_{truthpred}, \quad (1)$$

where $P_r(Object)$ is the probability of the bounding box containing the target and $IOU_{truthpred}$ is the accuracy of the bounding box.

2.2. YOLOv10-Based Underwater Target Detection Method

The YOLOv10 [12], released in 2024, represents the new and popular model in the YOLO family, as displayed in Figure 2. This advanced iteration of the YOLO series introduces a consistent dual assignment strategy, which significantly enhances its performance by addressing one of the most critical issues in target detection models, namely the need for non-maximum suppression (NMS) to eliminate duplicate prediction boxes during inference.

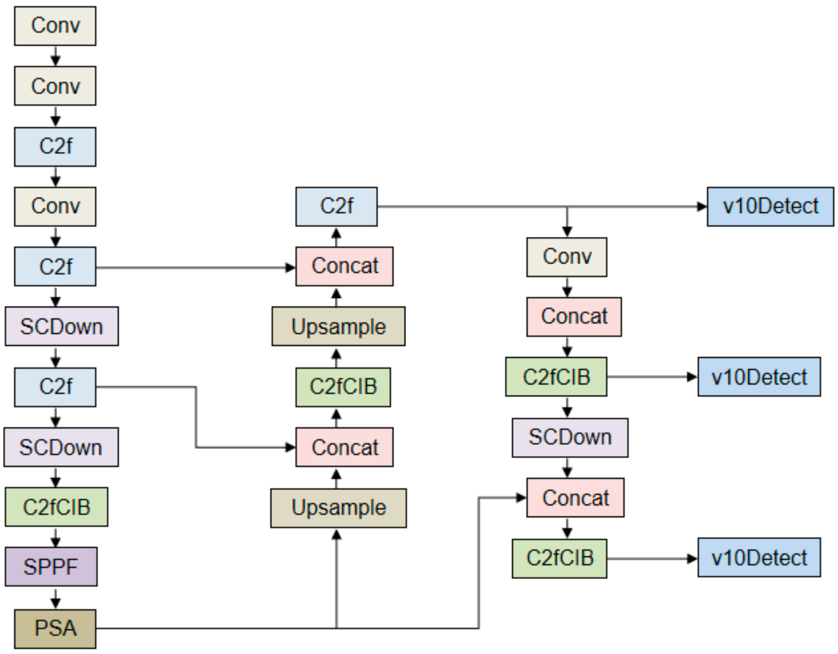


Figure 2. The network structure of YOLOv10.

Traditionally, NMS has been an indispensable post-processing step in target detection models, including earlier versions of YOLO. However, NMS is computationally expensive and can introduce latency, especially in real-time applications. YOLOv10's innovative dual assignment strategy eliminates the need for NMS by ensuring that each target is assigned to only one prediction box, thereby greatly reducing post-processing time. This approach not only accelerates inference speed but also maintains or even improves detection accuracy.

YOLOv10 has achieved remarkable results, setting new benchmarks in target detection performance. On the COCO dataset, YOLOv10 has achieved state-of-the-art results, demonstrating its ability to accurately detect a wide variety of targets in complex and diverse scenes. Similarly, on the VOC dataset, YOLOv10 has outperformed previous models, showcasing its robustness and reliability in handling a broad range of target classes.

The YOLOv10 detection model has three components, which are the backbone, neck, and head. The backbone is responsible for feature extraction from the input image and includes the CBS module, SCDown module, C2f module, C2fCIB module, and the PSA self-attention mechanism. These modules work together to enhance feature extraction and capture spatial relationships within the feature maps. The neck employs an FPN-PAN structure to integrate and refine the feature information from the backbone, capturing information at multiple scales and improving feature fusion. The head features a lightweight decoupled design, implementing a consistent dual allocation strategy to address YOLO's reliance on NMS in post-processing [12]. This decoupling and strategy help in making more efficient and accurate detection predictions.

We have developed an efficient partial self-attention (PSA) module, as depicted in Figure 3a. Initially, the features, after processing through a 1×1 convolution, are divided into two segments along the channel dimension. One segment is directed to the NPSA block, which consists of a multi-head self-attention (MHSA) module and a feed-forward network (FFN) for further processing. Subsequently, these two segments are concatenated and fused through another 1×1 convolution. Notably, we have optimized the dimensions of the queries and keys within the MHSA, setting them to half the dimension of the values. This reduction in dimensionality significantly decreases computational complexity while preserving essential feature information. To further enhance computational efficiency, the PSA module is exclusively deployed after the fourth stage, which has the lowest resolution. This strategic placement effectively mitigates the quadratic computational complexity associated with self-attention, avoiding excessive computational overhead. Through this design, we have successfully integrated global representation learning capabilities into the YOLO framework with minimal computational cost, significantly enhancing the model's overall performance and efficiency.

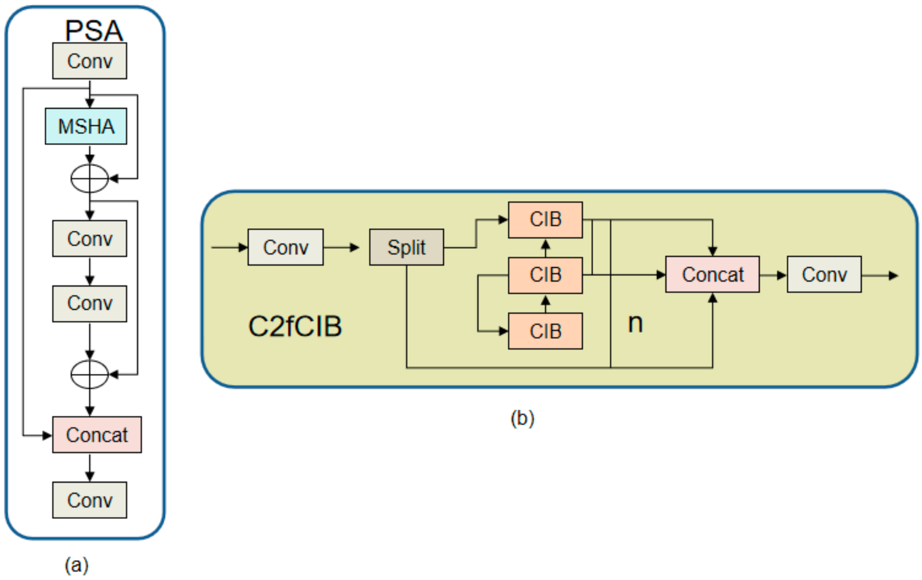


Figure 3. The YOLOv10 updated module. (**a**) PSA and (**b**) C2FCIB module.

The C2fCIB module represents a significant enhancement to the original C2f architecture by integrating a compact inverted block (CIB) in place of the traditional bottleneck module. In the original bottleneck design, standard convolutions are employed to facilitate feature transformation and dimensionality reduction. However, the C2fCIB module introduces a more efficient approach by replacing these standard convolutions with a combination of depthwise convolutions and pointwise convolutions.

The C2fCIB module, as depicted in Figure 3b, showcases a streamlined and optimized structure. Following this, the pointwise convolution projects the transformed features into a higher-dimensional space, ensuring that the model can effectively capture complex patterns and interactions among different channels. This dual approach allows the C2fCIB module to maintain a balance between computational efficiency and expressive power, making it particularly well-suited for resource-constrained environments while still delivering robust performance.

The integration of the CIB within the C2f framework signifies a strategic shift towards more efficient and scalable deep learning architectures. By leveraging the strengths of depthwise and pointwise convolutions, the C2fCIB module exemplifies a forward-thinking design that is poised to advance the state-of-the-art in various computer vision tasks.

3. Results and Discussion

3.1. Dataset and Evaluation Index

In this study, we employ the URPC2020 dataset [21], which serves as a pivotal resource for our experimental investigations. The dataset primarily encompasses research focused on stationary marine targets, including holothurian, echinus, scallop, and starfish. These organisms play crucial roles in their respective ecosystems, contributing to nutrient cycling and maintaining biodiversity. This dataset is meticulously crafted to mirror the intricate challenges inherent in detecting underwater biological targets in real-world scenarios. By incorporating a wide array of image conditions, including variations in brightness, background complexity, contrast, blur, and color deviation, the URPC2020 dataset aptly simulates the diverse and often unpredictable nature of underwater environments. This breadth of representation ensures that our model is rigorously tested and trained to perform effectively across different depths and lighting conditions, thereby enhancing its practical applicability. Among them, Ren et al. [22] introduced Faster R-CNN, integrating the region proposal network (RPN) with fast R-CNN for real-time object detection, achieving accuracy on datasets. Zhou et al. [23] proposed novel underwater object detection network AMSP-UOD, featuring AMSP-VConv, FAD-CSP modules, and enhanced NMS, demonstrating superior accuracy and noise immunity on URPC and RUOD datasets. Zhou et al. [24] proposed an underwater optical detection network (UODN), enhanced with CSMB and LKSP modules, which significantly improves feature extraction and object detection in underwater images, outperforming 12 models on the URPC 2020 dataset.

Comprising four primary classes of underwater biological targets—holothurians, scallops, starfish, and echinoids—the URPC2020 dataset offers a comprehensive visual repository of these species in various settings. With a total of 5543 images, each capturing these organisms under differing imaging conditions, the dataset provides an extensive and diverse training ground for our target detection model. This rich and varied dataset is essential for cultivating a robust and generalizable model capable of accurately identifying these underwater species despite the complexities and anomalies present in real-world underwater imagery.

Furthermore, the inclusion of such a broad spectrum of imaging conditions within the dataset is crucial for addressing the unique challenges posed by underwater photography. Water absorption and scattering of light lead to significant variations in image quality and color fidelity at different depths, which can profoundly impact the performance of computer vision models. By training our model on this diverse dataset, we aim to mitigate these challenges and enhance the model's ability to detect targets consistently and accurately across various underwater conditions.

To facilitate the training and evaluation of the model, the dataset is divided into a training set and a test set using an 8:2 ratio. Specifically, 4434 images are randomly selected to form the training set, which is used to train the target detection model. The remaining 1109 images are designated as the test set, which is used to evaluate the model's performance in terms of accuracy and robustness. This division ensures that the model is trained on a large and diverse set of images while allowing for a thorough and representative evaluation on a separate set of unseen data.

In this work, the confusion matrix, which serves as a prevalent method for evaluating the performance of classification models, is adopted for evaluation, as displayed in Table 1. The confusion matrix comprises predicted labels and actual labels, categorizing all samples into four distinct classes. These classes include true positive (TP), indicating the number of samples accurately predicted as positive and are indeed positive; false positive (FP), representing samples incorrectly predicted as positive while being negative; false negative

(FN), denoting samples wrongly predicted as negative despite being positive; and true negative (TN), signifying samples correctly predicted as negative and are indeed negative.

Table 1. Confusion matrix.

Confusion Matrix		**Reference**	
		True	**False**
Prediction	Positive	TP	FP
	Negative	FN	TN

In this experiment, the mean average precision (mAP) is employed as a critical metric to evaluate the performance of the model. The mAP is a composite metric that is closely related to both precision and recall rate, providing a comprehensive assessment of the model's ability to correctly identify and localize targets across different classes. This metric is particularly useful in target detection tasks, as it helps to quantify the overall effectiveness of the detection system in terms of both accuracy and completeness.

Precision measures the proportion of predicted positive instances (detected targets) that are correct, while recall rate quantifies the proportion of actual positive instances that are correctly identified by the model. The mAP integrates these two metrics by averaging the area under the precision–recall curve (PR curve) for each class and then taking the mean across all classes. This approach ensures that the model's performance is evaluated holistically, considering both false positives and false negatives, and the calculation formula is as follows:

$$AP = \int_0^1 p(r)dr, \tag{2}$$

$$mAP = \frac{\sum_{n=1}^{N} AP(n)}{N}, \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

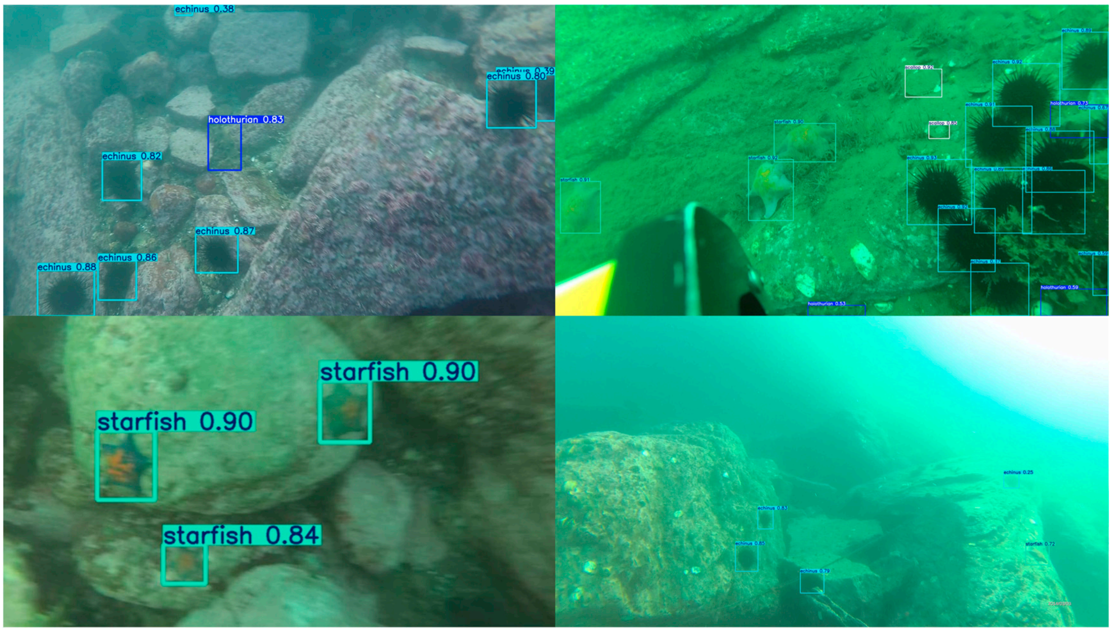
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{7}$$

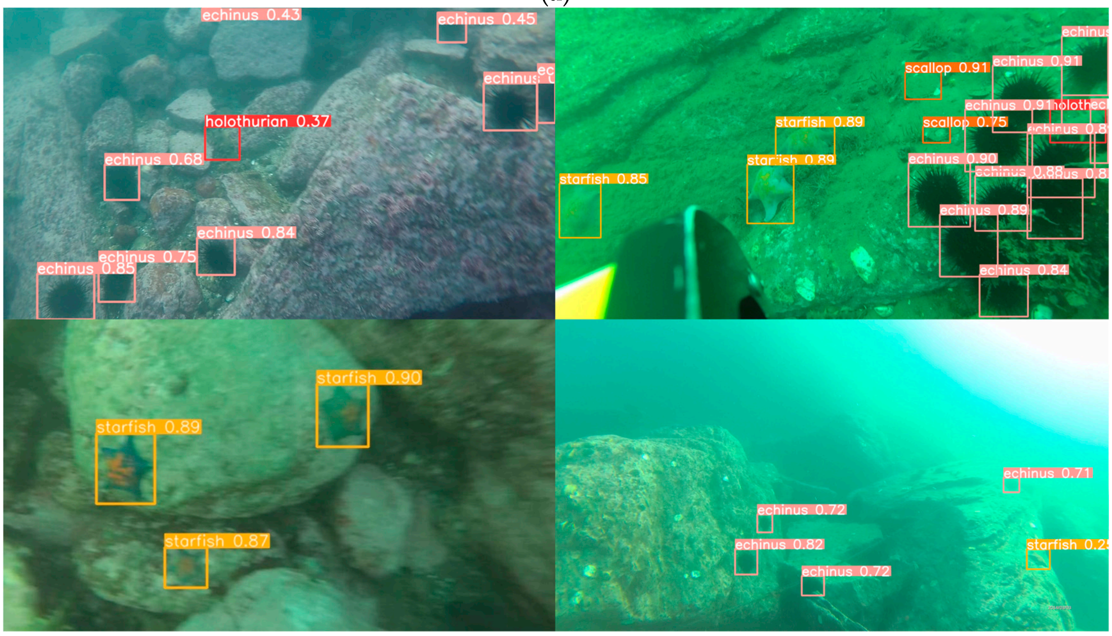
3.2. Comparative of Different Methods for Underwater Target Detection

In this research, all experiments are carried out with the operating system Windows 11. Our work is developed with the PyTorch 2.0.1 deep learning framework and with the GPU of NVIDIA GeForce RTX 3090 (Chengdu, China). The batch size is set to eight, and the epoch is set to 200.

In the case of a fixed IoU threshold of 0.5, the performance metrics of the URPC2020 dataset trained with the YOLOv10 and YOLOv5 algorithms are illustrated. The results of the YOLOv10 and YOLOv5 models' predictions on underwater images are exhibited in Figure 4. Based on the illustration in Figure 4, YOLOv10 exhibits a marked enhancement in object detection precision compared to its predecessor, YOLOv5. This advancement is reflected in higher intersection over union (IoU) values, indicating a closer alignment between the detected bounding boxes and the actual boundaries of the targets. Furthermore, the clarity of the IoU plots has improved, suggesting that the model may now handle complex scenarios more effectively by minimizing artifacts and confusion arising from overlapping objects, thereby offering more accurate and dependable detection outcomes.



(**a**)



(**b**)

Figure 4. Prediction results of underwater target detection: (**a**) YOLOv5; (**b**) YOLOv10.

In this study, Input X comprises the original image, the category of the detection target, and the location of the detection target. As for each image, the category and positional information of its detection targets should be stored in TXT files, which should be uniformly housed within another designated folder. Upon the reception of input images, both the YOLOv10 frameworks undergo an initial preprocessing stage that involves uniformly resizing the images to a standardized dimension of 640 × 640 pixels. After each training session, the model's optimal weight file is automatically saved in the 'weights' folder, with the filename 'best.pt'. When it is necessary to utilize these optimal weights, they can be loaded and invoked by running the 'val.py' script.

A confusion matrix is utilized to provide further insights into the model's performance across different classes of underwater biological targets. Figure 5 presents the confusion matrix, where the rows correspond to the real labels (actual classes of the targets in the test images) and the columns represent the predicted categories by the model. The diagonal elements of this matrix indicate the number of correct detections for each class, thereby reflecting the accuracy of the model in identifying those specific targets.

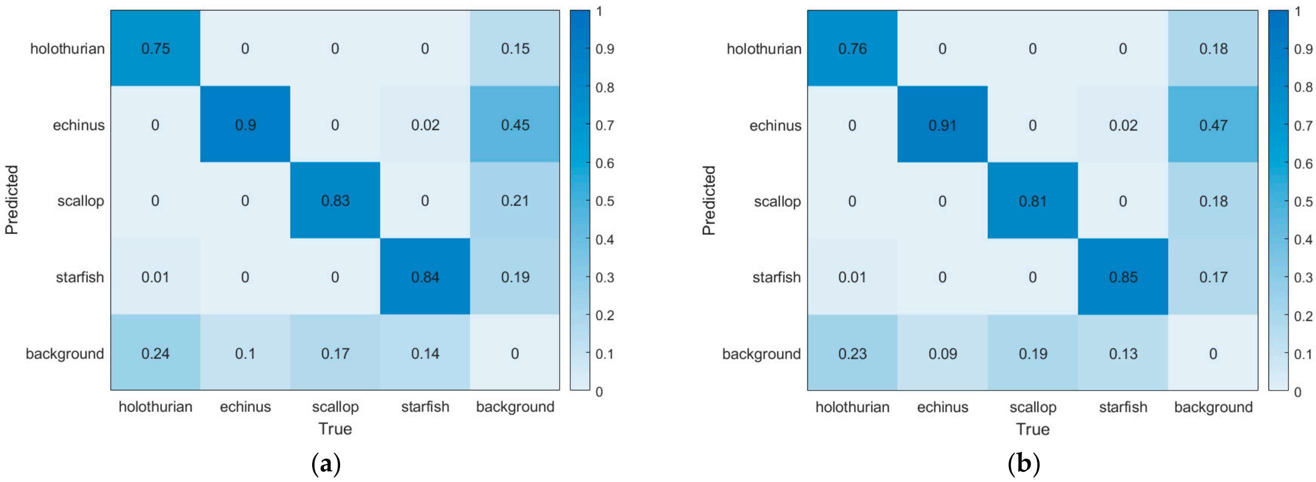


Figure 5. The confusion matrix of four targets: (**a**) YOLOv5; (**b**) YOLOv10.

From the confusion matrix, it can be observed that the detection accuracies for the four classes of underwater biological targets are as follows: holothurian at 76%, echinus at 91%, scallop at 81%, and starfish at 85%. These percentages denote the proportion of correctly identified instances for each class in the test set.

Upon analyzing the confusion matrix, it becomes evident that the primary source of classification errors stems from the model's confusion between targets and background elements. This phenomenon can be attributed to several factors inherent to underwater imaging. Firstly, underwater environments are characterized by low visibility due to factors such as water turbidity, which can make it difficult for the model to distinguish between targets and their surroundings. Additionally, the varying illumination conditions and color shifts caused by water absorption can further complicate the differentiation between targets and the background.

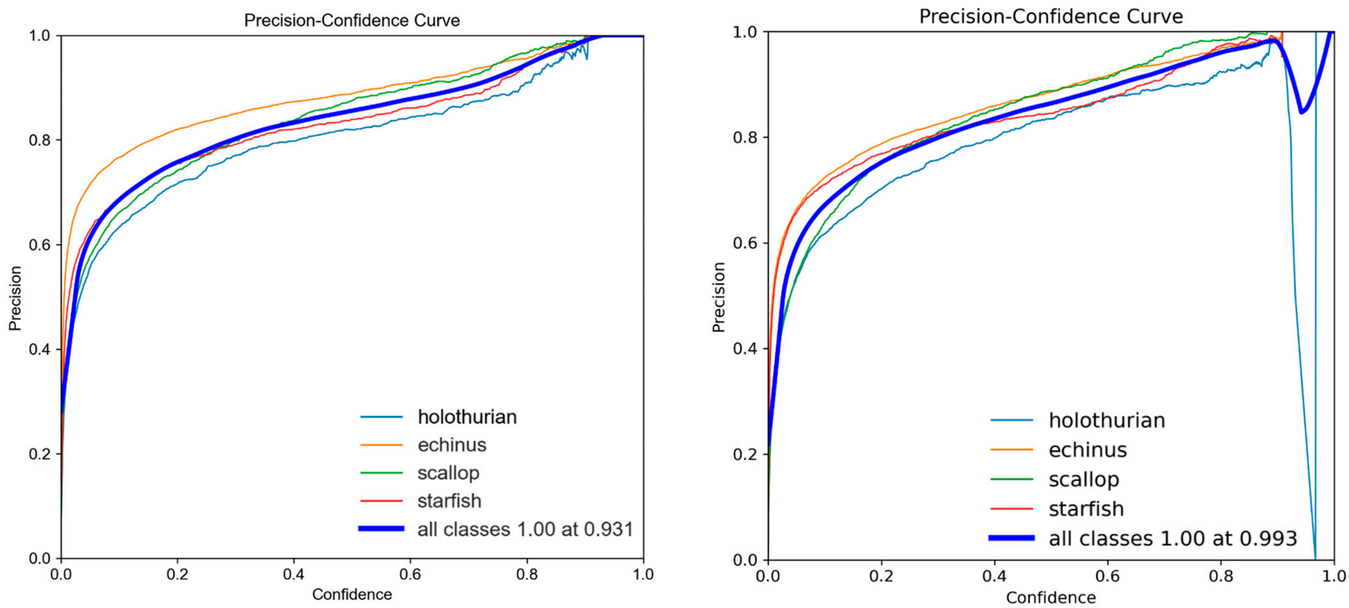
To provide a more comprehensive comparison, the performance metrics of YOLOv10 are evaluated against the widely used YOLOv5 and concurrently undertake a comparative analysis with the Faster R-CNN [22], AMSP-UOP [23], and UODN [24] object detection models. As depicted in Table 2, the performance metrics, specifically the mAP50, show significant improvements. Our model achieves an mAP50 score of 0.856, which represents a substantial enhancement compared to the baseline.

Table 2. Comparison of URPC2020 dataset detection accuracy with YOLOv10 and YOLOv5.

Model	Holothurian	Echinus	Scallop	Starfish	mAP50	FLOPs
Faster RCNN	\	\	\	\	0.802	90.9 G
AMSP-UOP	\	\	\	\	0.827	62.4 G
UODN	\	\	\	\	0.840	43.9 G
YOLOv5s	0.75	0.90	0.83	0.84	0.844	47.9 G
YOLOv10s	0.76	0.91	0.81	0.85	0.856	24.8 G

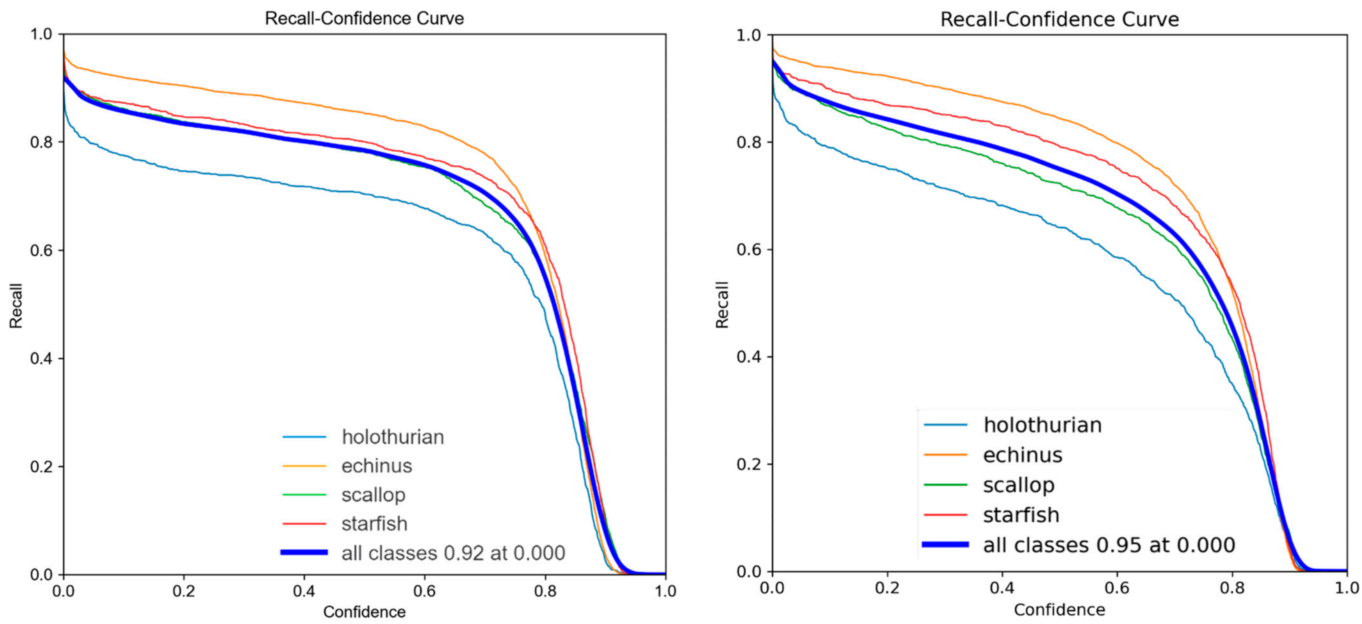
For the URPC2020 dataset, the overall improvement in mAP50 is noteworthy. Compared with the YOLOv5, Faster R-CNN, AMSP-UOP, and UODN models, the mAP50 score of YOLOv10 has increased from 0.802, 0.827, 0.840, and 0.844 to 0.856; the predicted accuracy increased by 1.2–6.3%. This improvement indicates that YOLOv10 is more effective in accurately detecting and localizing targets in the underwater environment. This significant improvement in mAP50 indicates that YOLOv10 is more effective in accurately detecting and localizing targets within the challenging underwater environment. The enhanced precision can be attributed to YOLOv10's advanced architecture, which likely includes refined feature extraction mechanisms, optimized loss functions, and improved object localization techniques. The improvement can be attributed to YOLOv10's refined architecture, which likely incorporates advanced feature extraction mechanisms and optimized loss functions, enabling it to better handle the complexities inherent in underwater imagery, such as low contrast, turbidity, and occlusions.

To better illustrate the performance outcomes of YOLOv5 and YOLOv10 across various dimensions, we present evaluation metrics such as precision, recall, and the F1 score, as depicted in Figures 6–8. The precision, recall, and F1 score of YOLOv10 unequivocally demonstrate superior model performance across various metrics. Specifically, YOLOv10 achieves a markedly higher precision rate, signifying an enhanced accuracy in its positive predictions. This notable improvement is especially critical in contexts where the implications of false positives are profound, such as in real-time object detection systems and safety-critical applications where the reliability of detections is paramount. The elevation in precision exhibited by YOLOv10 not only drastically reduces the instances of incorrect detections but also significantly augments the overall reliability, efficiency, and robustness of the model. Furthermore, the superior recall and balanced F1 score highlight YOLOv10's comprehensive capability to maintain a high level of performance in identifying all relevant instances while minimizing false negatives. This multi-faceted enhancement underscores YOLOv10's suitability for demanding applications requiring both high accuracy and robust generalization.



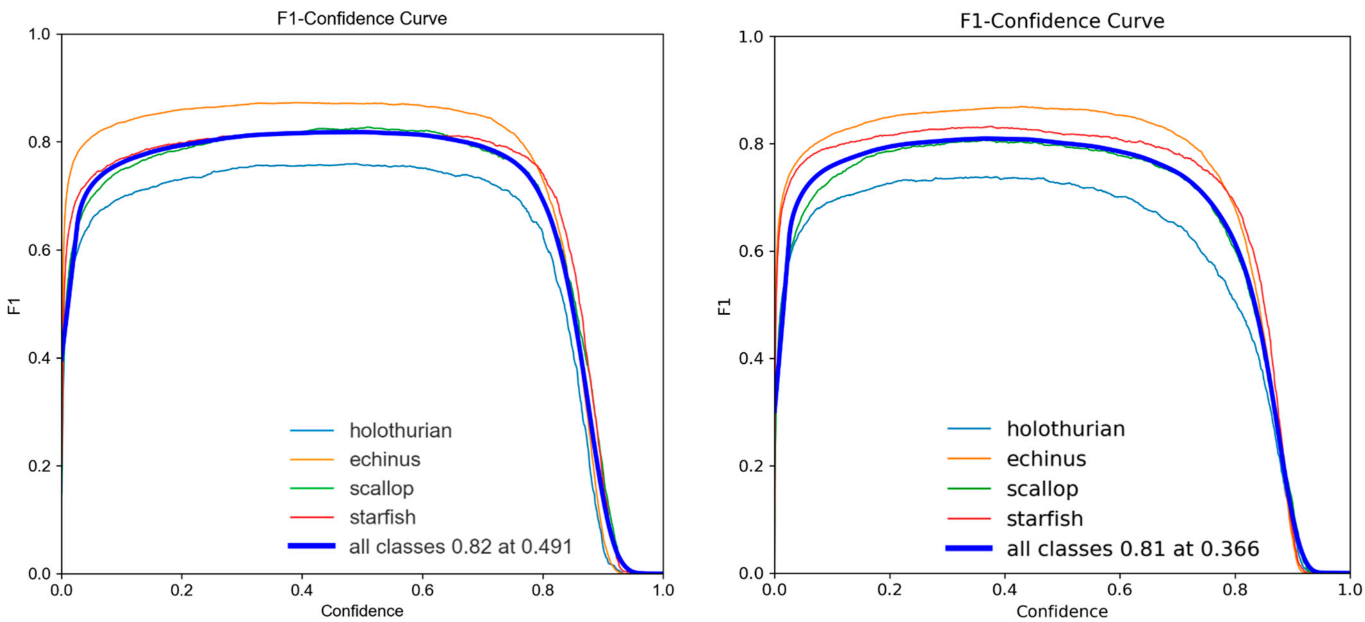
(**a**) (**b**)

Figure 6. The precision confidence curve: (**a**) YOLOv5; (**b**) YOLOv10.



(**a**) (**b**)

Figure 7. The recall confidence curve: (**a**) YOLOv5; (**b**) YOLOv10.



(**a**) (**b**)

Figure 8. The F1 score confidence curve: (**a**) YOLOv5; (**b**) YOLOv10.

In summary, the comparison with YOLOv5 clearly demonstrates the superior performance of YOLOv10, particularly in the context of the URPC2020 dataset. The 1.2% increase in mAP50 underscores the model's effectiveness in addressing the challenges of underwater target detection and its potential for widespread application in marine science and technology. In FLOPs, YOLOv10 demonstrates an enhancement over YOLOv5, reducing from 47.9 G to 24.8 G. Not only does YOLOv10 expedite the object localization process and facilitate the development of lightweight models, but it also accomplishes this task with greater precision.

4. Conclusions

In summary, YOLOv10 has made significant strides in the field of object detection, showcasing exceptional speed, high precision, robust multiscale detection capabilities, and a lightweight design, making it an ideal choice for applications. Especially in the detection of underwater targets, the YOLOv10-based method illustrates better performance for underwater target detection. For the URPC2020 dataset, there is an overall of 85.6% in mAP50, accompanied by a substantial increase of 1.2% than that of the YOLOv5 model. In terms of mAP50 performance, the detection accuracy for the holothurian category demonstrates an enhancement of 1.0, while the scallop category shows a 1.0% improvement by using YOLOv10, where it gives a promising method for applications of underwater target detection. Future research on YOLOv10 for underwater target detection will focus on enhancing the model's adaptability to complex underwater environments, encompassing the challenges of low visibility, the refractive and scattering effects of light, and the interference caused by diverse underwater backgrounds.

Author Contributions: Conceptualization, Z.H., L.C., S.Y. and P.X.; funding acquisition, Z.H., P.X., P.Z. and J.H.; experiments, Z.H., L.C., S.Y., P.X., P.Z., R.T. and J.H.; method, Z.H., L.C., S.Y., P.X., P.Z. and J.H.; data analysis, Z.H., L.C., S.Y., P.X., P.Z. and J.H.; writing, Z.H., L.C., S.Y., P.X., P.Z. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (no. 62401601) and the Fundamental Research Funds for the Central Universities (no. 2682024GF009).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Lin, S.; Zhao, Y. Review on key technologies of object exploration in underwater optical images. *Laser Optoelectron. Prog.* **2020**, *57*, 060002. [CrossRef]
2. Lei, F.; Tang, F.; Li, S. Underwater object detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [CrossRef]
3. Teng, B.; Zhao, H. Underwater object recognition methods based on the framework of deep learning: A survey. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1–12. [CrossRef]
4. Xu, S.; Zhang, M.; Song, W.; Mei, M.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [CrossRef]
5. Shao, Y.; Zhang, D.; Chu, H.; Zhang, X.; Rao, Y. A review of yolo object detection based on deep learning. *J. Electron. Inf. Technol.* **2022**, *44*, 3697–3708.
6. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
7. Barhoumi, C.; BenAyed, Y. Real-time speech emotion recognition using deep learning and data augmentation. *Artif. Intell. Rev.* **2024**, *58*, 49. [CrossRef]
8. Cao, C.; Zheng, J.; Huang, Y.; Liu, J.; Yang, C. Investigation of a Promoted You Only Look Once Algorithm and Its Application in Traffic Flow Monitoring. *Appl. Sci.* **2019**, *9*, 3619. [CrossRef]
9. Shi, K.; Wang, J.; Zhu, H.; Li, Q.; Zhang, M.; Sui, H.; Xia, K.; Ni, Q.; Mao, R. Intelligent target detection of underwater optical images based on image enhancement. In Proceedings of the Fourteenth International Conference on Information Optics and Photonics, Xi'an, China, 7–10 August 2023; Volume 12935, pp. 198–204.
10. Chen, X.; Yuan, M.; Yang, Q.; Yao, H.; Wang, H. Underwater-YCC: Underwater Target Detection Optimization Algorithm Based on YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 995. [CrossRef]
11. Yi, W.; Yang, J.; Yan, L. Research on underwater small object detection technology based on single-stage USSTD-YOLOv8n. *IEEE Access* **2024**, *12*, 69633–69641. [CrossRef]

12. Majid Azimi, S. ShuffleDet: Real-time vehicle detection network in on-board embedded UAV imagery. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
13. Xiong, Y.; Liu, H.; Gupta, S.; Akin, B.; Bender, G.; Wang, Y.; Kindermans, P.; Tan, M.; Singh, V.; Bo, C. Mobiledets: Searching for object detection architectures for mobile accelerators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3825–3834.
14. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13668–13677.
15. Tan, L.; Liu, S.; Gao, J.; Liu, X.; Chu, L.; Jiang, H. Enhanced Self-Checkout System for Retail Based on Improved YOLOv10. *J. Imaging* **2024**, *10*, 248. [CrossRef] [PubMed]
16. Jiang, W.; Jiang, Z. Research on early fire detection of Yolo V5 based on multiple transfer learning. *Fire Sci. Technol.* **2021**, *40*, 109–112.
17. Wu, Z.; Xue, R.; Li, H. Real-time video fire detection via modified YOLOv5 network model. *Fire Technol.* **2022**, *58*, 2377–2403. [CrossRef]
18. Liu, K.; Sun, Q.; Sun, D.; Peng, L.; Yang, M.; Wang, N. Underwater Target Detection Based on Improved YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 677. [CrossRef]
19. Ou, J.; Shen, Y. Underwater target detection based on improved YOLOv7 algorithm with BiFusion Neck structure and MPDIoU loss function. *IEEE Access* **2024**, *12*, 105165–105177. [CrossRef]
20. Ru, J.; Jiang, H.; Xu, H. Underwater Small Target Detection Based on Improved YOLOv7. In Proceedings of 3rd 2023 International Conference on Autonomous Unmanned Systems (3rd ICAUS 2023), Nanjing, China, 9–11 September 2023; Springer: Singapore, 2024; pp. 102–111.
21. Chen, X.; Yuan, M.; Fan, C.; Chen, X.; Li, Y.; Wang, H. Research on an Underwater Object Detection Network Based on Dual-Branch Feature Extraction. *Electronics* **2023**, *12*, 3413. [CrossRef]
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
23. Zhou, J.; He, Z.; Lam, K.M.; Wang, Y.; Zhang, W.; Guo, C.; Li, C. AMSP-UOD: When vortex convolution and stochastic perturbation meet underwater object detection. *AAAI Conf. Artif. Intell.* **2024**, *38*, 7659–7667. [CrossRef]
24. Zhou, H.; Kong, M.; Yuan, H.; Pan, Y.; Wang, X.; Chen, R.; Lu, W.; Wang, R.; Yang, Q. Real-time underwater object detection technology for complex underwater environments based on deep learning. *Ecol. Inform.* **2024**, *82*, 102680. [CrossRef]