symmetry

MDPI

Article

YOLOv7-SN: Underwater Target Detection Algorithm Based on Improved YOLOv7

Ming Zhao, Huibo Zhou * and Xue Li

School of Mathematical Sciences, Harbin Normal University, Harbin 150500, China; zm15765074886@163.com (M.Z.); lx15246066277@163.com (X.L.)
* Correspondence: zhouhuibo@hrbnu.edu.cn; Tel.: +86-150-0460-1579

Abstract: Exploring the ocean's resources requires finding underwater objects, which is a challenging task due to blurry images and small, densely packed targets. To improve the accuracy of underwater target detection, we propose an enhanced version of the YOLOv7 network called YOLOv7-SN. Our goal is to optimize the effectiveness and accuracy of underwater target detection by introducing a series of innovations. We incorporate the channel attention module SE into the network's key part to improve the extraction of relevant features for underwater targets. We also introduce the RFE module with dilated convolution behind the backbone network to capture multi-scale information. Additionally, we use the Wasserstein distance as a new metric to replace the traditional loss function and address the challenge of small target detection. Finally, we employ probe heads carrying implicit knowledge to further enhance the model's accuracy. These methods aim to optimize the efficacy of underwater target detection and improve its ability to deal with the complexity and challenges of underwater environments. We conducted experiments on the URPC2020, and RUIE datasets. The results show that the mean accuracy (mAP) is improved by 5.9% and 3.9%, respectively, compared to the baseline model.

Keywords: deep learning; underwater target detection; YOLOv7; symmetry

check for updates

Citation: Zhao, M.; Zhou, H.; Li, X. YOLOv7-SN: Underwater Target Detection Algorithm Based on Improved YOLOv7. *Symmetry* **2024**, *16*, 514. https://doi.org/10.3390/sym16050514

Academic Editor: Calogero Vetro

Received: 15 March 2024
Revised: 16 April 2024
Accepted: 20 April 2024
Published: 24 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The ocean occupies most of the Earth's surface solution and is the habitat of tens of thousands of marine organisms. Additionally, they hold immense potential as a valuable reservoir of resources, including minerals, oil, natural gas, and a variety of other aquatic resources. Underwater image target detection research is gaining momentum due to the growing need for intelligent underwater detection in academic, industrial, and military applications [1]. This includes various underwater tasks such as target localization, object search, aquatic life detection, seabed modeling, salvage and rescue, anti-mine, anti-submarine, etc. [2]. Complex underwater environments can affect detection results. Factors such as insufficient light due to weather conditions and variations in underwater brightness due to water depth can increase the difficulty of detection. In addition, the process of acquiring and transmitting underwater images is costly, further increasing the complexity of this research problem [3].

In the field of artificial intelligence and deep learning, symmetry plays an important role in helping to optimize model performance, simplify problems, and improve training efficiency. The concept of symmetry includes the property that the input data or model remains unchanged under certain transformations, such as spatial translation, rotation, scaling invariance, etc. Symmetry is utilized to reduce the model complexity, improve the generalization ability, and reduce the amount of data required.

In image processing, the symmetry of image translation, rotation, and other symmetries is used to design neural network structures with corresponding invariance, such as convolutional neural networks, to effectively capture image features and realize recognition and processing under symmetry transformation. In natural language processing,

the symmetry of sentence structure is utilized to design models with positional coding or attention mechanisms to improve semantic understanding and task performance. In deep learning, the symmetry of the environment is utilized to reduce the training complexity and improve the learning efficiency of the intelligences, e.g., learning symmetric strategies in symmetric environments to reduce state space search. In summary, symmetry is of great significance in AI and deep learning. By using symmetry to design efficient models, we can improve generalization ability and training efficiency and promote technology development and application.

Target detection can be classified into two main categories: traditional techniques and deep learning-based techniques. Traditional techniques use established algorithms and analytical methods to detect targets, while deep learning-based techniques employ complex neural networks to learn and recognize targets from data. The process of finding the object of interest in an image involves two subtasks: target localization and target classification. Along with classifying the object, it is also necessary to determine its position. The traditional method for detection involves using the sliding window method, which requires more accurate traversal for higher accuracy detection. However, this method also results in greater time overhead required for detection. Additionally, there is an issue with the binary classification samples not being balanced due to the large number of background images compared to foreground images. There is a significant difference between the frames corresponding to the background and foreground in a single image. The problems with traditional detection methods can be summarized into two points: high accuracy, demanding a lot of time and complex operations, and a large number of samples needing to be generated manually [4]. These advantages include the ability to effectively adapt to the complexity and variability of detection targets, reduced dependency on human intervention, and improved generalization capabilities. Currently, deep learning-based target detection algorithms are mainly classified into two categories: two-stage target recognition algorithms and single-stage target recognition algorithms. The R-CNN (area-CNN) [5] family of algorithms, which combines convolutional neural networks (CNNs) and area suggestions to achieve a notable performance boost, represents the former class of methods. However, this algorithm's poor computing performance makes it unsuitable for most real-world underwater target detection applications. Single-stage algorithms are currently the subject of extensive research in the field of target detection, and people have been working to improve their accuracy and performance. Numerous studies and experiments are being conducted to explore novel techniques and methodologies that can further improve the effectiveness of these algorithms. Researchers are actively looking for ways to optimize the existing single-stage algorithms by addressing various challenges and limitations associated with them. These endeavors aim to elevate the accuracy and overall performance of single-stage algorithms, making them a promising avenue for future advancements in target detection research. This is because single-stage algorithms may directly anticipate categorization and localization, which leads to faster detection speeds. The representative algorithms of this type are the SSD (Single Shot MultiBox Detector) [6] algorithm and the YOLO (You Only Look Once) [7] series algorithms (YOLO, YOLO9000 [8], YOLOv3 [9], YOLOv4 [10], YOLOv5 [11], YOLOv6 [12], and YOLOv7 [13]).

In recent years, the field of underwater target identification has seen remarkable advancements driven by numerous professionals, academics, and researchers. Li et al. (2015) utilized Fast R-CNN [14] to identify and detect fish species. To expedite the fish detection process, Faster R-CNN was subsequently employed. Zhou et al. [15] introduced a Faster R-CNN network with feature mapping in 2017. To address the challenge of underwater image blurring on extremely noisy backgrounds, Long Chen et al. proposed a novel sample-weighted super network called Fast Processing in 2020 [16]. In the same year, Qiao et al. [17] presented a combination of LWAP and MLP neural networks. Lei et al. [18] made significant improvements to the YOLO-v5s model through three key modifications: replacing the backbone network with a Swin Transformer.

However, the underwater environment poses numerous challenges, including severe color distortion and low visibility due to image capture in motion. These factors significantly impede the field of underwater target detection. As a more advanced algorithm within the YOLO family, YOLOv7 proves to be better suited for industrial applications. Based on the YOLOv7 model, this paper proposes an innovative approach for underwater imaging.

The innovative contributions of this paper are as follows:

1. Some of the MP and ELAN modules have been modified, and the SE attention module has been embedded in some of the modules to increase the model's focus on regions of interest, and SE layers have been added at specific locations in the model to further enhance the model's focus on regions of interest. As a result, the model can focus its computational power on regions of interest;
2. To enhance detection accuracy, a decoupling head with implicit knowledge is employed, replacing the original detection head. This decoupling head captures additional implicit information without requiring additional processing steps;
3. To assess the similarity of BBox, a new metric called the Normalized Wasserstein Distance (NWD) is introduced, replacing the traditional IoU metric. The NWD is more appropriate for assessing similarity between small targets because it measures distributional resemblance independent of target overlap.

The following sections of this paper are organized as follows: Section 2 outlines the necessary knowledge for this paper. In Section 3, the theoretical foundations of the proposed YOLOv7-SN model are introduced. Section 4 details the experimental evaluation and performance analysis conducted on the underwater image dataset. Finally, Section 5 provides a summary of this work.

2. Related Work

2.1. YOLOv7

YOLOv7 is an algorithm for target detection that introduces model reparameterization [19], a label assignment strategy [20], an ELAN-efficient network architecture [21], and a training method with auxiliary heads in the network architecture. YOLOv7 and YOLOv5 are similar in general, but with replacements and improvements in the internal components of the network structure, the auxiliary training head, and the label assignment idea. According to Figure 1, the YOLOv7 network consists of three different parts.

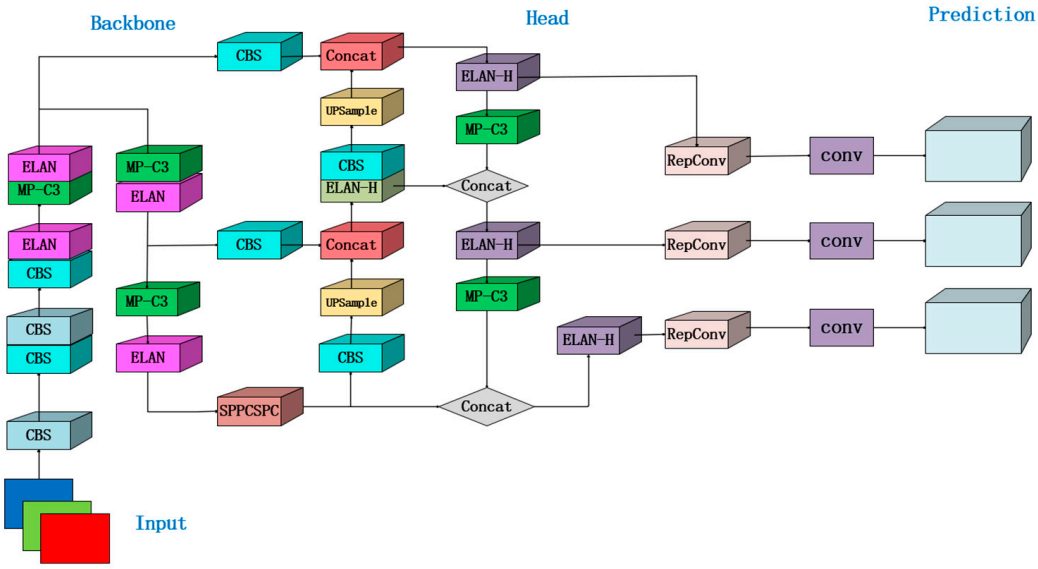


Figure 1. The network structure of YOLOv7.

The primary function of the input module in YOLOv7 is to perform essential processing tasks on the input image, such as data augmentation, adaptive image scaling, and anchor frame calculation. These tasks are crucial for pre-processing the input image and preparing it for further analysis and detection.

The main feature extraction module in the YOLOv7 network is called the backbone module. Its job is to acquire the input image and extract high-level semantic features that will be used in the target detection task. The backbone module in YOLOv7 employs several convolutional and other network layers to progressively decrease the resolution of the input image. Simultaneously, it extracts increasingly abstract and significant features.

The Head module in the YOLOv7 model is mainly responsible for the final regression prediction, which receives the multi-scale features from the Neck module and processes them through a series of operations to finally generate the output of target detection. The Head module usually includes convolutional layers, global average pooling layers, fully connected layers, etc., which convert the feature maps into the target's location and category information. As an integral component of the Head module, the prediction module typically encompasses convolutional and fully connected layers.

To enhance the YOLOv7 model, the proposed YOLOv7-SN network preserves the overall structure of YOLOv7 while integrating a channel attention mechanism. This mechanism introduces a dilated convolution module, which effectively replaces the conventional detection head. The inclusion of a channel focus mechanism is intended to improve the detection performance of the model by enhancing the network's ability to selectively focus on relevant features.

2.2. Attention Mechanisms

In the domain of computer vision, the human visual system tends to concentrate on salient objects, disregarding irrelevant parts of the visual field that do not contribute to object recognition [22]. Similar to this mechanism, the attention mechanism [23] in object detection models guides the model to focus on important objects and their respective locations within the acquired image [24]. Previous research has successfully integrated attention mechanisms into deep neural networks with promising results in various tasks such as object classification [25], image segmentation [26], and target detection [27].

Several attention modules have been proposed, including the SK module [28], the CBAM module [29], and the TA module [30]. The CBAM module is an attention module designed specifically for feedforward convolutional neural networks. By incorporating both spatial and channel attention, the CBAM module enhances the model's classification performance. Another commonly used channel attention module is the SK module, which was proposed in 2019. It leverages a convolutional kernel mechanism to assign varying levels of importance to different convolutional kernels concerning additional input images.

Momenta's Squeeze–Excite (SE) [31] network module, introduced in 2017, is another attention module that has achieved notable success in the ImageNet image recognition competition. The SE module models the interdependencies between feature channels and learns the significance of each channel by assigning weights to features. This process highlights key features while suppressing less-related ones. The SE module enhances the representativeness of the model and improves the detection of blurred images while maintaining a lightweight structure and minimal computational overhead.

2.3. Bounding Box Regression Loss Function

The Bounding Box Regression Loss Function [32] is a loss function used to measure the difference between the predicted box and the real box in a target detection task. Its goal is to optimize the accuracy of the target detection model by minimizing the distance between the predicted box and the real box.

In recent years of research, the Bounding Box Regression Loss Function has undergone an evolutionary process and mainly includes the following forms [33]:

SmoothL1 Loss: SmoothL1 Loss is a smooth L1 loss function that uses L2 loss when the difference between the predicted box and the real box is small and L1 loss when the difference is large to reduce the influence of outliers. IoU Loss: IoU Loss is a measure of the degree of overlap between the predicted and true frames by calculating the intersection and concurrency ratio (IoU) between them. GIoU Loss: GIoU Loss is an improved IoU Loss that takes into account information about the size and position of the bounding box between the predicted and true boxes. DIoU Loss: DIoU Loss is a further improved IoU Loss that takes into account the centroid distance between the predicted and real frames on top of GIoU Loss. CIoU Loss: CIoU Loss is a further improvement of DIoU Loss, which takes into account the difference in aspect ratio between the predicted frame and the real frame based on DIoU Loss.

2.4. Implicit Knowledge

Implicit knowledge refers to knowledge that is acquired and utilized subconsciously. However, there is currently no standard description or framework that comprehensively explains the mechanisms of implicit learning and the acquisition of implicit knowledge. In the context of neural networks, feature information acquired from shallow levels is often called explicit knowledge, while feature information learned from deeper levels is considered tacit knowledge. Implicit knowledge, on the other hand, is derived from deeper layers and encompasses the model's latent knowledge that is not directly related to the observed data or specific comments. Explicit knowledge corresponds directly to observable data and observations.

A network with implicit learning is a single, cohesive network that integrates the encoding of implicit and explicit knowledge [34], similar to how the human brain acquires knowledge through both conventional and subconscious learning. A unified network can produce a unified representation that fulfills multiple purposes at once. Kernel space alignment, multi-task learning convolutional neural networks, and prediction refinement are all possible.

3. Method

This section presents the methodology for underwater target identification utilizing the enhanced YOLOv7. The initial step involves dataset processing, encompassing both data labeling and augmentation. Following this, the enhanced YOLOv7 network is utilized to enhance the precision of the model's detection. To be more specific, we substituted the original detection header with the Efficient Decoupled Head with Implicit Learning, integrated a module featuring null convolution into the backbone network, amalgamated the SE attention mechanism with multiple modules, and ultimately bolstered the confidence loss function.

3.1. The Proposed YOLOv7-SN Model

The proposed YOLOv7-SN model incorporates the new ELAN-S and MP-S modules, replacing the original ELAN and MP-C3 modules. The SE attention mechanism is introduced to enhance the channel characteristics of the input feature maps and thus improve the accuracy and performance of the model. Following this, the RFE module [35], comprising three distinct dilated convolutions with varying dilated rates, is integrated after the SPPCSPC module. This addition serves to capture additional multi-scale information, expand the sensory area, and reduce the number of parameters, thus reducing the chance of overfitting. Furthermore, the accuracy of model detection is heightened through the replacement of the initial detection head with an efficient decoupling head utilizing implicit learning. The location of these attention modules concerning the RFE module and the structure of the modified YOLOv7-SN is illustrated in Figure 2.

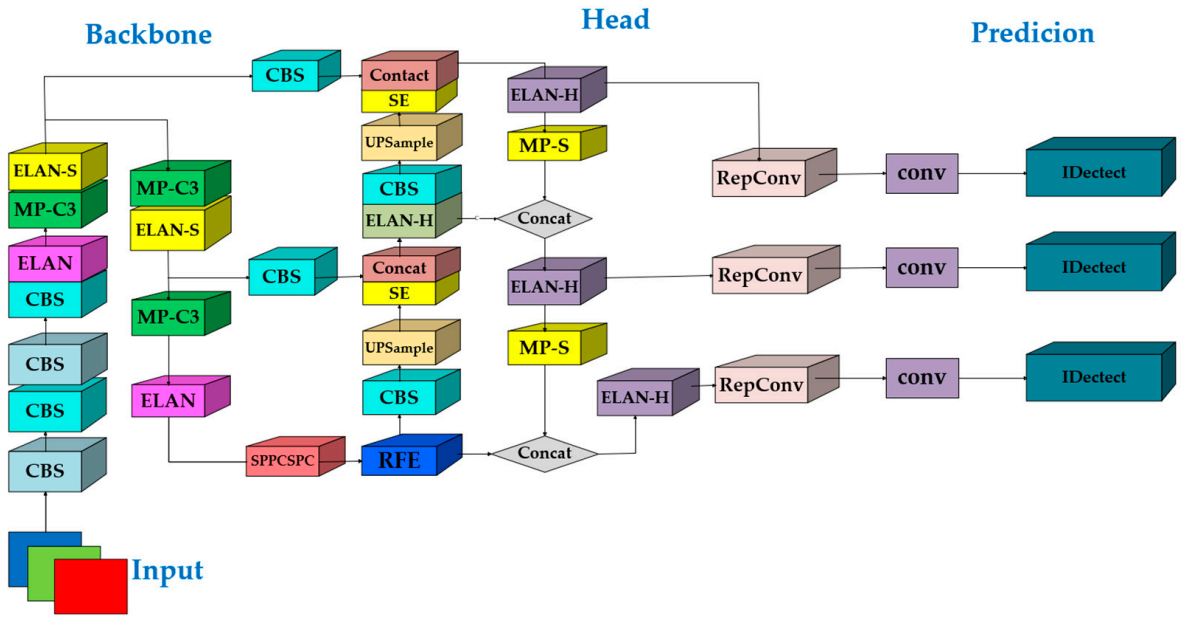


Figure 2. Structure of the proposed YOLO-SN network.

In terms of spectral feature extraction, the proposed YOLOv7-SN model is better suited for complex underwater environments. First, the image to be detected is input into the model. Through a series of improved convolutional and pooling layers, the model can extract the feature information of the image. These features can include edges, textures, etc. Then, after the convolutional layers, the model uses a fully connected layer to further process the feature information. The fully connected layer can map the features of the image to the target category. After the fully connected layer, the YOLO model uses a softmax function to make probabilistic predictions for each target category. Finally, in addition to the prediction of target categories, the model predicts bounding boxes for each target. These bounding boxes are used to locate the position of the target in the image.

3.1.1. Improvements Based on the Attention Mechanism

The model can extract higher-level characteristics and more information thanks to the enhanced backbone network's introduction of the ELAN-S structure, which substitutes the SE attention module for the CBS convolution module in the ELAN structure following the contact module. This will be more beneficial in navigating the intricate and crowded underwater landscape. The specific ELAN-S modules are shown in Figure 3.

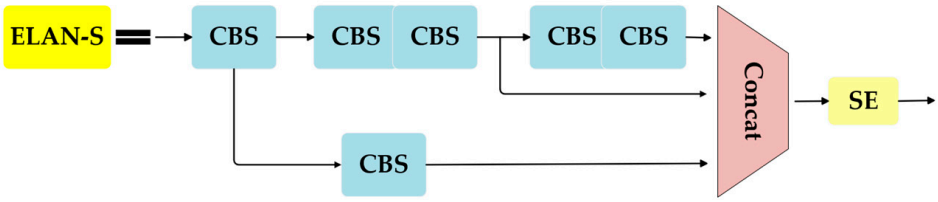


Figure 3. Structure of the ELAN-S module.

Furthermore, we use two enhanced MP-S modules to replace the original MP-C3 module. The MP-S module enables the network to adaptively adjust the channel weights, while the header network can learn and capture contextual information more effectively. The specific MP-S modules are shown in Figure 4.

The structural details of the SE Attention module are depicted in Figure 5. The module starts with a global pooling layer and passes through a series of fully connected layers that linearly transform the features using a weight matrix to achieve a nonlinear combination. The ReLU activation function is used to solve the vanishing gradient problem and increase

the sparsity of the model. Another activation function, Sigmoid, transforms the input into a probability distribution. The "Scale" operation then normalizes the information to ensure that the elements are appropriately weighted. By adaptively recalibrating the feature response, the attention module can effectively catch and emphasize important features.

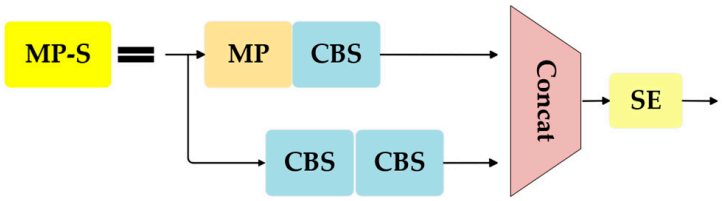


Figure 4. Structure of the MP-S module.

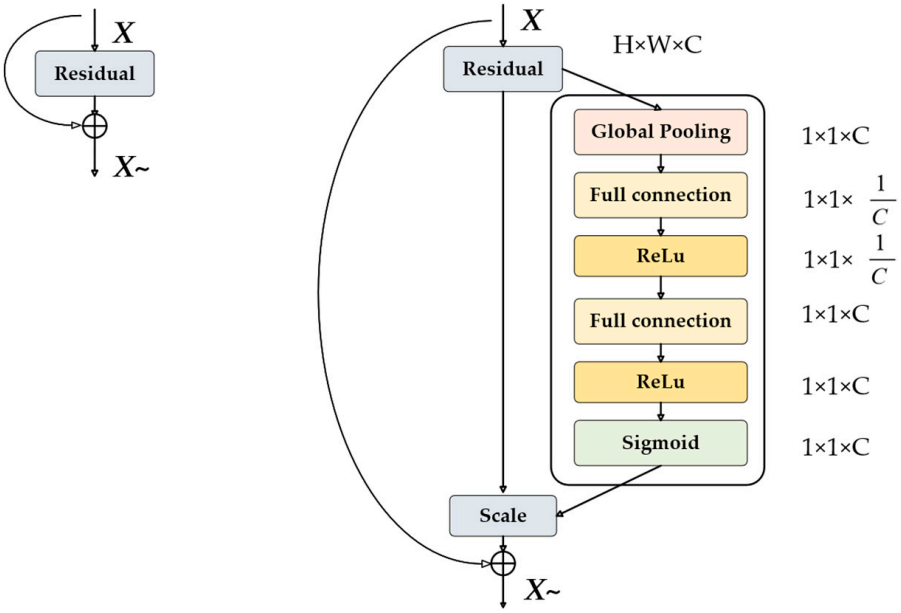


Figure 5. The structure diagram of the SE model.

3.1.2. RFE

After the SPPCSPC module, the upgraded head network adds an RFE module made up of a dilated convolution [36] module. The RFE module (Receptive Field Enhancement Module) makes full use of the different widths of the receptive fields of the feature maps to obtain multi-scale information through dilated convolution. As seen in Figure 6, the structure comprises two parts: an aggregated weighted layer based on dilated convolution and a multi-branch. In the first part of the structure, features are extracted using three different dilated convolutions with different dilated rates (1, 2, and 3). A fixed convolution kernel of 3×3 size is used to extract multi-scale information, and residual concatenation is used to avoid the gradient explosion problem. The output feature layer is finally obtained by adding the characteristics of the four branches.

Target detection tasks require a large practical receptive domain, which can be further increased using dilated convolution with several convolution kernels. The demand for shape information in the target identification task necessitates a greater number of convolutional kernels to extract more shape features. Therefore, models equipped with multiple kernel convolutions are more suitable for intense downstream detection tasks.

3.1.3. Implicit Leaning-Based Detection Head

In terms of detection head improvement, we replace the original detection head with an efficient decoupled head with implicit learning, applying implicit learning to multi-task deep learning. Implicit learning, as the name suggests, refers to learning outside of normal

learning (which we call epiphenomenal knowledge), just like the subconscious learning of the brain, where the subconsciously learned experiences will be encoded and stored in the brain. Utilizing this wealth of experience as a huge database, humans can efficiently process data, even if it is not seen beforehand. In this paper, we encode implicit and explicit knowledge together through a unified network to generate a unified representation that serves both underwater detection tasks.

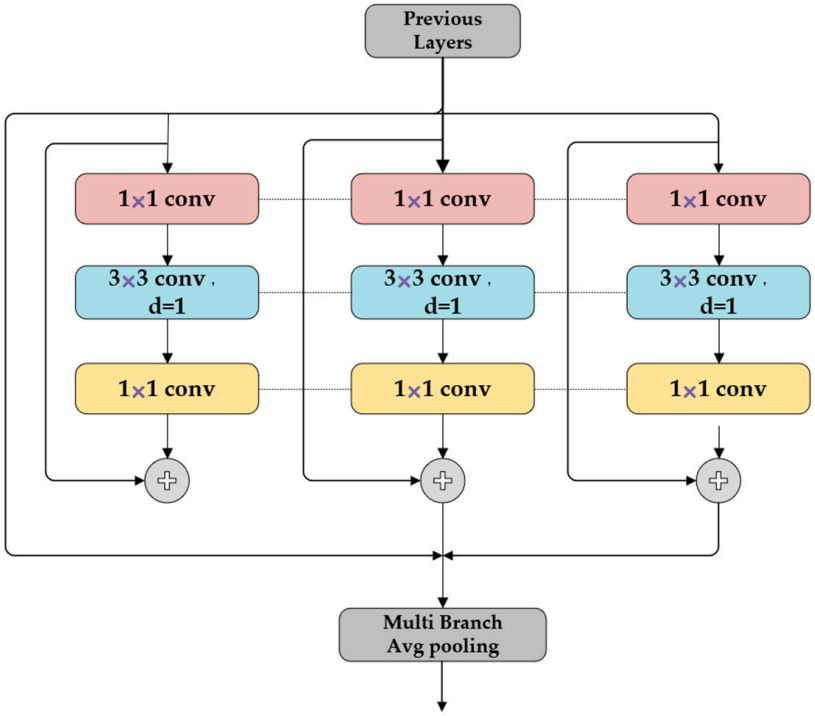


Figure 6. Structure of the RFE module.

The technique adds relatively little extra cost (less than 10,000 parameters and calculations) while improving the performance of the model. The combination of implicit and display learning is used to complete various tasks, as seen in Figure 7.

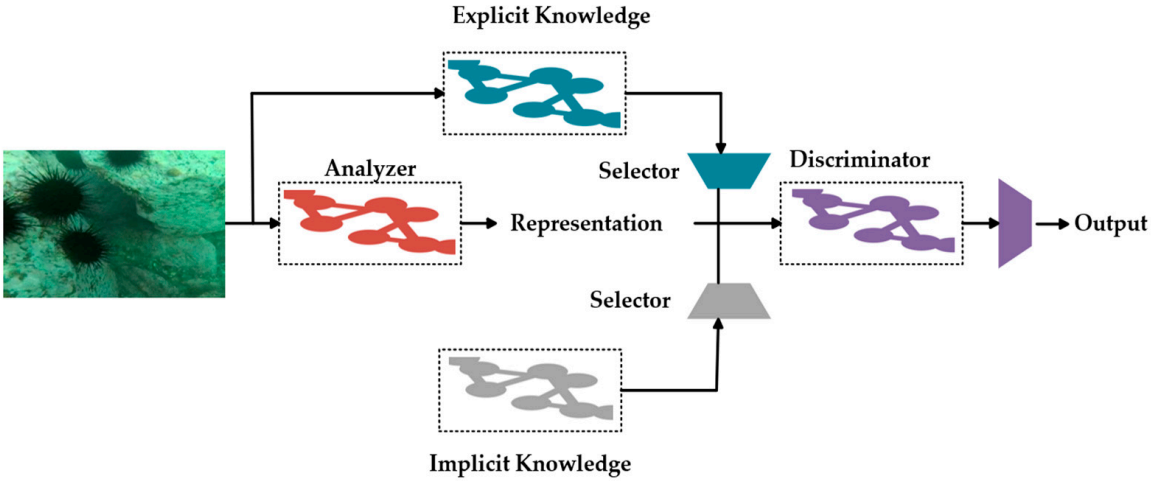


Figure 7. Diagram of network structure with implicit learning.

In traditional network training, we can use the objective function of (1) to represent:

$$\begin{array}{l} y = f_\theta(x) + \epsilon \\ \min \epsilon \end{array} \quad (1)$$

where x is observation, θ is the set of parameters of a neural network, f_θ represents the operation of the neural network, ϵ is the error term, and y is the target of a given task.

We use the joint network, which is based on the traditional network described above. Together, we use explicit and implicit knowledge to model the error term and then use it to guide the training process of the multipurpose network. The corresponding training Formula (2) is as follows:

$$\begin{array}{l} y = f_\theta(x) + \epsilon + g_\phi(\epsilon_{ex}(x), \epsilon_{im}(z)) \\ \min \epsilon + g_\phi(\epsilon_{ex}(x), \epsilon_{im}(z)) \end{array} \tag{2}$$

where ϵ_{ex} and ϵ_{im} are the arithmetic operations that model the explicit and implicit errors of the observation x and latent code z, respectively. n is a mission-specific operation for selectively integrating information from explicit and tacit knowledge. There are now some ways to incorporate explicit knowledge, so we can further write (2) into (3).

$$y = f_\theta(x) \otimes g_\phi(z) \tag{3}$$

where $\otimes$ represents some possible operators that can combine f_θ and g_ϕ. When $\otimes$ represents a multiplication operator, if the subsequent layer is a convolutional layer, then use (4) for the integration. When $\otimes$ represents an addition operation, if the preceding layer is a convolutional layer and there is no activation function. In this case, use (5) for the integral operation.

$$\begin{array}{l} x_{(l+1)} = \sigma(W_l(g_\phi(z)x_l) + b_l) \\ = \sigma(W'_l(x_l) + b_l), where\ W'_l = W_l g_\phi(z) \end{array} \tag{4}$$

$$\begin{array}{l} x_{(l+1)} = W_l(x_l) + b_l + g_\phi(z) \\ = W_l(x_l) + b'_l, where\ b'_l = b_l + g_\phi(z) \end{array} \tag{5}$$

3.2. NWD-Based Loss Function Improvement

The original YOLOv7 model employs the CIoU (Compatible Intersection on Union) bounding box loss function [37], which aims to address the potential instability associated with typical IoU loss functions for bounding box regression by combining IoU with a distance factor. Despite its good performance in many target detection applications, CIoU loss still has certain shortcomings in tiny target recognition.

Due to the small size and few pixels of underwater targets, the distance and angle differences between their bounding boxes are relatively small, and a slight localization error can have a significant impact on the IoU, and the CIoU loss function may not be able to effectively distinguish the differences between their components. Therefore, the performance of CIoU in underwater target detection will be slightly insufficient.

To mitigate these issues, a Normalized Gaussian Wasserstein distance (NWD) [38] is used instead of a loss function. The NWD is defined by the Wasserstein distance between two normal distributions, as shown in Equation (6) below. This equation serves as the mathematical definition of the NWD loss function.

$$NWD(Na, Nb) = \exp\left(-\frac{\sqrt{W_2^2(Na, Nb)}}{C}\right) \tag{6}$$

where Na, Nb is the number of true frames in the neighborhood of each prediction frame, and the coefficient $W2^2(Na, Nb)$ is calculated as follows:

$$W_2^2(Na, Nb) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2 \tag{7}$$

Furthermore, the relative significance of actual and predicted frame loss is determined by representing the height, width, and centroid coordinates of the two frames.

In terms of small object recognition, the NWD loss function is superior to the CIoU in the following ways:

1. Stability: The model is more stable since the NWD loss function explicitly accounts for the distribution variations, making it more resistant to modest item localization mistakes;
2. Feature sensitivity: By better capturing the distinctive characteristics of small items, the NWD loss function raises the model's accuracy in identifying small objects;
3. Global Consideration: The model can better comprehend the relationship between small things and other objects in the image because the NWD loss function considers the global information of every object in the picture.

4. Experimental Verification and Analysis

This section describes the configuration of the experimental environment, hyperparameters, test datasets, and the optimization of the anchoring box. The experimental results show that the proposed YOLOv7-SN model improves the accuracy and speed of underwater target detection, thus verifying its effectiveness and superiority in the challenging underwater detection environment.

4.1. Experimental Platform

The experimental hardware comprised an Intel (R) Xeon (R) CPU E5-2630 v3 @2.40 GHz and an NVIDIA GeForce RTX 3090 GPU, and PyTorch was utilized for debugging. All experiments were trained for 200 episodes, and the performance of the YOLOv7-SN model was assessed by training and testing with the specified hyperparameters provided in Table 1.

Table 1. Experimental configuration.

Parameter	Configuration
learning rate	0.01
momentum	0.9
weight decay	0.0005
batch size	16
image size	640 × 640
training epochs	200

4.2. Evaluation Metrics

In this paper, two metrics, detection accuracy and speed, are used to evaluate model performance. Precision (P) and recall (R) are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where FP indicates the number of targets that were incorrectly detected, FN indicates the number of targets that were not detected, and TP indicates the number of targets that were correctly detected.

The performance of a model can be measured by observing the interaction between precision and recall. Ideally, a model with high precision and increasing recall signifies improved performance. The average precision (AP) is used to combine these metrics and quantify detection accuracy, as formulated in Equation (10).

$$AP = \int_0^1 Precision(Recall) dRecall \tag{10}$$

AP corresponds to the area under the precision–recall curve, with larger values representing higher network precision. In a multi-class target detection task, the average of all

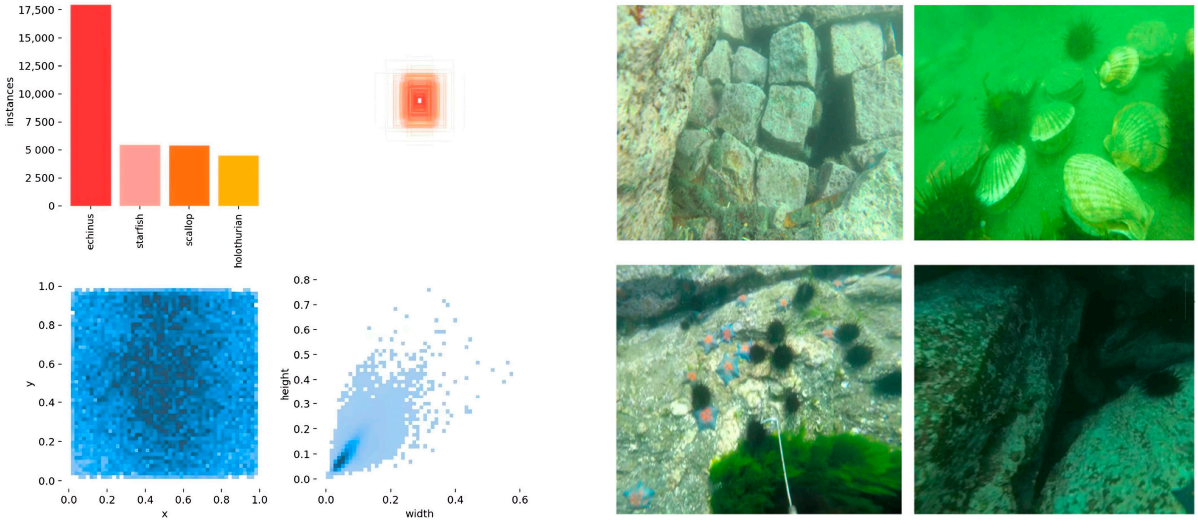
types of *AP* is the overall detection precision of the model, which is referred to as the mean accuracy (*mAP*) in Equation (11).

$$mAP = \frac{1}{class\_number} \sum_{1}^{class\_number} AP \tag{11}$$

4.3. Experimental Results and Analysis of the URPC Dataset

4.3.1. The URPC2020 Dataset

The dataset URPC2020 comprises 5543 pictures categorized into four groups: echinus, holothurian, scallop, and starfish. To facilitate the training and testing of the proposed algorithm, the dataset underwent partitioning into training and testing sets in an 8:2 proportion. This division resulted in 4434 images allocated for training and 1109 images designated for testing purposes. Within this dataset, various intricate scenarios are presented, including the clustering of underwater creatures leading to visual obstruction, disparities in illumination, and blurring due to motion shots. These complexities render the dataset a true-to-life depiction of the underwater setting, thereby enhancing the model's capacity for generalization. Nonetheless, the imbalanced distribution of samples across categories, coupled with variations in their resolutions, presents notable hurdles during model training. As shown in Figure 8a, the category statistics plot analyzes the number of targets, in which the largest number is echinus, followed by scallops and starfish, and the smallest number is holothurian. The boxplot demonstrates that the target box sizes are relatively concentrated, and the regularized target location plot shows that the targets are mainly concentrated in the horizontal direction, which is denser, while in the vertical direction, they are relatively dispersed. The normalized target size map shows that the target size is relatively concentrated, and most of them are small targets. Figure 8b shows some sample images from the URPC dataset.



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

Figure 8. The sample information for URPC2020 is (**a**) labels: Upper left is the distribution of categories; upper right is a visualization of all box sizes; lower left is the distribution of the position of box centroids; lower right is the distribution of box aspect ratios. (**b**) example images.

4.3.2. Ablation Experiments of the URPC Dataset

We tested each module in YOLOv7-SN on the URPC dataset and investigated how it affects the model in the ablation experiments. Table 2 displays the results of the ablation experiments, with "√" denoting the application of a specific improvement technique. According to the experimental results, the ELAN-S and MP-S modules, the RFE module, and the NWD, with the addition of the SE attention mechanism, improved the mAP

accuracy of the models by 1.3%, 2.4%, and 1.6%, respectively. In addition, the mAP@0.5:0.95 values increased by 0.7%, 0.5%, and 3.5%, respectively. By adding the RFE module and replacing the loss function NWD based on ELAN-S and MP-S, the accuracy of mAP@0.5 is improved by 5.4% and 4.3%, respectively. mAP@0.5:0.95 rose 3.2 percent and 1.5 percent, respectively. The three parts are added at the same time, based on replacing the detection head. The accuracy of mAP@0.5 is increased by 5.9%, and the value of mAP@0.5:0.95 is increased by 4.8%. Overall, the addition of several modules improved the detection accuracy of the model.

Table 2. Ablation comparison of model performance improvement on the URPC dataset.

Model	ELAN-S	MP-S	RFE	NWD	EDH	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv7						83.2	48.6
	✓					83.6	49.6
	✓	✓				84.5	49.3
			✓			85.6	49.1
				✓		84.8	52.1
	✓	✓	✓			88.6	51.8
	✓	✓		✓		87.5	50.1
	✓	✓	✓	✓		88.3	50.5
	✓	✓	✓	✓	✓	89.1	53.4

4.3.3. Comparison Experiments of the URPC Dataset

To further demonstrate the superiority of the proposed YOLOv7-SN model, we compare it with popular target detection models such as YOLOv7, YOLOv6, YOLOv5s, and Faster-RCNN. The URPC dataset is trained and tested, and their evaluation metrics, such as mean average precision (mAP), are compared. The comparison results are shown in Table 3. From the table, it can be seen that the YOLOv7-AC model outperforms other detection algorithms, with a mAP that is 5.9% higher than that of YOLOv7 and 6.5%, 9.3%, and 9.4% higher than that of YOLOv6, YOLOv5s, and Faster-RCNN, respectively. We also conducted experiments on YOLOv8, which is more novel than YOLOv7, and as shown by the results, our proposed network is 3.1% and 2.0% more accurate than YOLOv8n and YOLOv8s on the URPC2020 dataset, respectively. These experimental results demonstrate the superiority of the method for underwater target detection.

Table 3. Performance comparison of target detection model on the URPC2020 dataset.

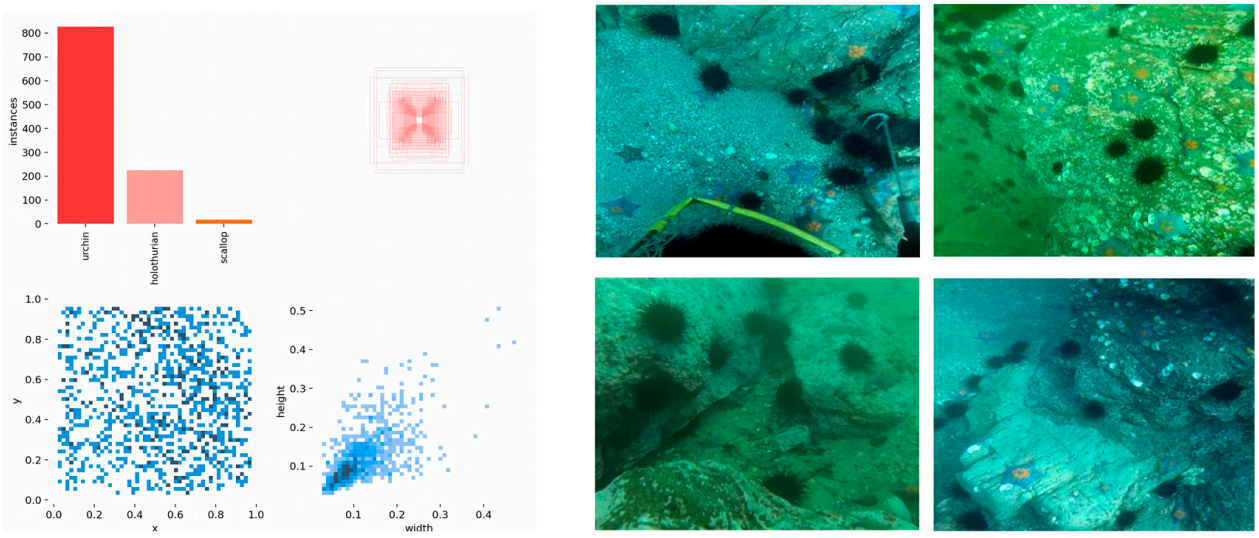
Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
SSD [39]	74.2	68.7	75.4	38.8
RetinaNet [40]	75.2	66.8	73.4	32.9
Faster-RCNN	78.8	73.1	76.1	35.4
YOLOX [41]	73.3	64.1	69.5	31.3
YOLOv5s	78.9	75.3	80.8	47.7
YOLOv6	81.7	79.1	80.4	45.1
YOLOv7	82.9	78.3	83.2	48.6
YOLOv8n	83.8	79.2	85.7	50.5
YOLOv8s	86.4	82.1	87.1	51.7
YOLOv7-SN	88.2	84.8	89.1	53.4

4.4. Experimental Results and Analysis of the RUIE Dataset

4.4.1. The RUIE Dataset

The RUIE dataset was proposed by Liu et al. [42]. The dataset is characterized by large data volumes, diverse scattering levels, rich color casts, and rich detection targets. It includes three subsets, UIQS (Underwater Image Quality Set), UCCS (Underwater Color Cast Set), and UHTS (Underwater Higher-Level Task-Driven Set), which can be used for underwater image enhancement, color correction, and target detection tasks, respectively.

The UIQS includes 3630 images, the UCCS includes 300 images, and the UHTS includes 300 images. Notably, 80% of the images were randomly sampled as the training set, and the remaining 20% were used as the testing set for validation. As shown in Figure 9a, the category statistic map analyzes the number of targets in three categories, in which the largest number is urchin, followed by holothurian, and the smallest number is scallop. The boxplot demonstrates that the target box sizes are relatively dispersed; the regularized target location map shows that the centroids of the target boxes show a relatively discrete distribution in the region, with no obvious aggregation or regularity. The normalized target size map shows that the target sizes are relatively concentrated, most of which are small targets. Figure 9b shows some sample images from the RUIE dataset.



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

Figure 9. The sample information for RUIE is (**a**) labels: Upper left is the distribution of categories; upper right is a visualization of all box sizes; lower left is the distribution of the position of box centroids; lower right is the distribution of box aspect ratios. (**b**) example images.

4.4.2. Ablation Experiments of the RUIE Dataset

In the second ablation experiment, we used the RUIE underwater dataset, focusing on the effect of the detection head with implicit learning (EHD) on the model. The results of the experiment are shown in Table 4, where the mAP@0.5 of the model is improved by 1.3% when only the detection head is modified and by 0.5%, 1.2%, and 0.7% when the modification network is added together with the other three modules. The addition of EHD was a 1.2% increase in the model's mAP@0.5 with the addition of the other three models. When several modules are added to the modified network at the same time, the mAP of the model is increased by 4.9%, and the value of mAP@0.5:0.95 is increased by 6.7%. In summary, our proposed YOLOv7-SN also works well on the RUIE underwater dataset.

4.4.3. Comparison Experiments of the RUIE Dataset

The model's validity was confirmed through comparison with many benchmark models. YOLOv5s, Faster-RCNN, YOLOv6, and YOLOv7 models used include Faster-RCNN, YOLOv6, and YOLOv7, among others. The same undersea dataset was used for testing, and each model was trained in the same context with identical parameter configurations. Table 5 displays the dataset and the comparison experiment findings. Our enhanced YOLOv7 detection framework outperforms all previous original YOLO models, achieving 83.5% for mAP@0.5. These results also show that it performs 7.0% and 8.9% better than RetinaNet and Faster-RCNN, respectively. We also conducted experiments on YOLOv8, which is more novel than YOLOv7, and as shown by the results, our proposed network is 0.8% and 1.1% more accurate than YOLOv8n and YOLOv8s on the RUIE dataset,

respectively. Based on these results, our proposed novel YOLOv7 network architecture exhibits excellent performance.

Table 4. Ablation comparison of model performance improvement on the RUIE dataset.

Model	ELAN-S	MP-S	RFE	NWD	EDH	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv7						79.6	43.1
					✓	79.9	41.2
	✓	✓			✓	79.1	41.0
			✓		✓	80.2	40.6
				✓	✓	79.3	41.1
	✓	✓	✓			80.7	44.4
	✓	✓		✓		81.8	45.7
	✓	✓	✓	✓		82.3	46.9
	✓	✓	✓	✓	✓	83.5	48.8

Table 5. Performance comparison of target detection model on the RUIE dataset.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
SSD	78.1	72.4	75.3	38.8
RetinaNet	76.9	63.6	76.5	40.7
Faster-RCNN	75.6	65.1	74.6	41.9
YOLOX	68.5	59.5	67.8	34.7
YOLOv5s	79.7	71.1	76.5	38.5
YOLOv6	74.2	68.7	75.4	40.8
YOLOv7	78.3	75.3	79.6	43.1
YOLOv8n	82.9	78.2	82.7	49.1
YOLOv8s	81.6	77.5	82.4	48.5
YOLOv7-SN	87.7	79.1	83.5	48.8

4.5. Comparison of the Model's Metrics with Other Classical Models

The proposed YOLOv7-SN model's performance in terms of computational complexity and speed was evaluated by comparing it with other popular target detection models applied to URPC and RUIE training and testing. The experimental results, presented in Table 6, show a slight increase in the number of parameters and computational complexity of the improved model. However, in terms of computational speed, it is only slightly inferior to the YOLOv5 model, which is significantly faster than the other models. This demonstrates that the model's detection accuracy and computing speed have been improved with only a 10% increase in computing cost. Furthermore, as represented by the F1 scores on both datasets, our proposed model has a clear superiority.

Table 6. Comparison of the model's metrics in the URPC dataset and the Brackish dataset.

Model	Parameters (M)	GFLOPS	FPS (URPC)	FPS (RUIE)	F1 (URPC)	F1 (RUIE)
YOLOX	9.0	26.8	32	28	68.4	63.7
YOLOv5	87.3	217.4	77	58	77.1	75.2
YOLOv6	59.6	150.7	61	36	80.4	66.8
YOLOv7	37.2	104.8	72	53	80.5	76.8
YOLOv7-SN	39.1	114.2	75	56	86.5	83.2

Figure 10 depicts the detection performance of the updated model in four distinct scenarios, showcasing its ability to provide accurate detection in challenging underwater conditions. The enhanced YOLOv7 model successfully identifies targets across different underwater scenarios, demonstrating its robust detection performance.

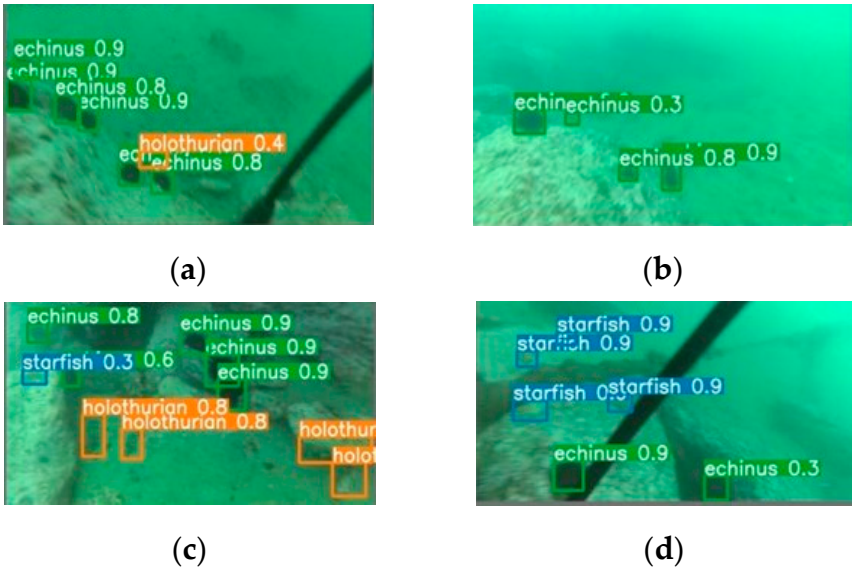


Figure 10. Detection results for multiple underwater scenarios: (**a**) close-clear scene; (**b**) close-fuzzy scene; (**c**) far-clear scene; (**d**) far-fuzzy scene.

5. Discussion

In underwater environments, target identification can be difficult due to the optical properties of water, which cause underwater light to attenuate. This results in under-illuminated, blurry images with low contrast. In this paper, we propose a method that uses the SE channel attention mechanism and an inflated convolutional ensemble module for feature extraction. This allows us to selectively emphasize informative features and suppress redundant ones while increasing the neural network's receptive field without adding too much computation. We also introduce NWD, a more suitable metric for small target detection that is insensitive to target scale. Underwater targets are typically small and dense, and the underwater environment is complex, so a lightweight network model is needed. Through experiments, we demonstrate that our proposed model outperforms other models in underwater environments, achieving improved accuracy.

6. Conclusions

This work presents an updated version of the YOLOv7 model that is applicable to underwater missions. The new model addresses the challenges of the underwater environment and the limited ability of underwater robots to recognize specific targets. This is achieved by changing the detecting head, using a novel metric called NWD instead of the conventional loss function, and integrating the SE attention and RFE modules into the YOLOv7 model. The enhanced model provides better accuracy in identifying fuzzy underwater objects. The experimental results demonstrate that our proposed YOLOv7-SN model performs well on both underwater datasets. The detection accuracy and computational speed of the model are improved with a slight increase in the computational cost. The proposed method has significant advantages for underwater target detection.

It is important to note that the lack of good-quality underwater datasets and image availability is a significant obstacle to target detection in underwater environments. Future research initiatives will focus on gathering sizable and varied underwater datasets and using image-enhancing methods to improve image quality.

Author Contributions: Conceptualization, M.Z.; methodology, M.Z. and X.L.; software, M.Z. and X.L.; validation, M.Z.; formal analysis, M.Z. and X.L.; investigation, M.Z. and X.L.; resources, M.Z.; data curation, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z. and H.Z.; visualization, M.Z.; supervision, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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

Abbreviations

The following abbreviations are used in this paper:

YOLO	You Only Look Once
CNN	Convolutional Neural Network
SE	Squeeze–Excite
SK	Selective Kernel Networks
CBAM	Convolutional Block Attention Module
IoU	Intersection over Union
URPC	Underwater Optical Target Detection Intelligent Algorithm Race
RUIE	Real-World Underwater Image Enhancement
mAP	Mean Average Precision

References

1. Ghafoor, H.; Noh, Y. An overview of next-generation underwater target detection and tracking: An integrated underwater architecture. *IEEE Access* **2019**, *7*, 98841–98853. [CrossRef]
2. Cai, S.; Li, G.; Shan, Y. Underwater object detection using collaborative weakly supervision. *Comput. Electr. Eng.* **2022**, *102*, 108159. [CrossRef]
3. Zhang, C.; Zhang, G.; Li, H.; Liu, H.; Tan, J.; Xue, X. Underwater target detection algorithm based on improved yolov4 with semidsconv and fiou loss function. *Front. Mar. Sci.* **2023**, *10*, 1153416. [CrossRef]
4. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
5. Pinto, F.; Torr, P.H.; Dokania, P.K. An impartial take to the cnn vs transformer robustness contest. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XIII; Springer Nature: Cham, Switzerland, 2022; pp. 466–480.
6. Yao, Y.; Qiu, Z.; Zhong, M. Application of improved MobileNet-SSD on underwater sea cucumber detection robot. In Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; pp. 402–407.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
8. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
9. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
10. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
11. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
12. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
13. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
14. Li, X.; Shang, M.; Qin, H.; Chen, L. Fast accurate fish detection and recognition of underwater images with fast r-cnn. In Proceedings of the OCEANS 2015-MTS/IEEE Washington, DC, USA, 19–22 October 2015; pp. 1–5.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Chen, L.; Liu, Z.; Tong, L.; Jiang, Z.; Wang, S.; Dong, J.; Zhou, H. Underwater object detection using Invert Multi-Class Adaboost with deep learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
17. Qiao, W.; Khishe, M.; Ravakhah, S. Underwater targets classification using local wavelet acoustic pattern and Multi-Layer Perceptron neural network optimized by modified Whale Optimization Algorithm. *Ocean Eng.* **2021**, *219*, 108415. [CrossRef]
18. Lei, F.; Tang, F.; Li, S. Underwater target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [CrossRef]

19. Anasosalu Vasu, P.K.; Gabriel, J.; Zhu, J.; Tuzel, O.; Ranjan, A. An Improved One Millisecond Mobile Backbone. *arXiv* **2022**, arXiv:2206.04040.
20. PGao, P.; Lu, J.; Li, H.; Mottaghi, R.; Kembhavi, A. Container: Context aggregation network. *arXiv* **2021**, arXiv:2106.01401.
21. Dollár, P.; Singh, M.; Girshick, R. Fast and accurate model scaling. *arXiv* **2021**, arXiv:2103.06877.
22. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2048–2057.
23. Tsotsos, J.K. *A Computational Perspective on Visual Attention*; MIT Press: Cambridge, MA, USA, 2021.
24. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; p. 27.
25. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 3286–3295.
26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
27. Shen, Z.; Nguyen, C. Temporal 3D RetinaNet for fish detection. In Proceedings of the 2020 Digital Image Computing: Techniques and Applications (DICTA), Melbourne, Australia, 29 November–2 December 2020; pp. 1–5.
28. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 3139–3148.
31. Wang, L.; Peng, J.; Sun, W. Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 884. [CrossRef]
32. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
33. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
34. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. You only learn one representation: Unified network for multiple tasks. *arXiv* **2021**, arXiv:2105.04206.
35. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *arXiv* **2021**, arXiv:2208.02019.
36. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
37. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI conference on artificial intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
38. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
39. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
41. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
42. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.